



Model under test: Meta-Llama-3.1-70B-Instruct

Audit executed: 2026-06-01 17:38:37 – 2026-06-01 19:57:24

Report generated: 2026-07-13 23:17:20 (engine v9)

Source: C:\VTRunRecords\OSHA_llama-3.1-70b

Contents (click a section to jump)

Headline Accuracy (bad-test cases excluded from denominator)

Release disposition

Human sign-off

Paired outcome (same questions, both arms)

Per-Category Accuracy & Failure Recovery

Adjudication reconciliation

Test-set coverage

Estimated audit resource usage

Without veritrooper — Baseline LLM Performance

With veritrooper — Pipeline Performance

Independent Verification

Bottom Line

Deterministic Guardrails

Complete Failure Listing — With VERITROOPER (Pipeline)

Complete Failure Listing — Without VERITROOPER (Baseline)

EU AI Act — Conformity Evidence Statement

Audit scope — point-in-time snapshot. This audit is a single point-in-time measurement of **Meta-Llama-3.1-70B-Instruct** answering **1000 purpose-generated questions** drawn from **Generated — OSHA_29CFR_full_2024 — veritrooper v9** (OSHA_29CFR_full_2024.txt), under the recorded retrieval, prompt, and runtime configuration (temperature 0.1). The baseline arm is a standardized BM25 vanilla-RAG reference retrieval.

It measures **only this endpoint × this corpus × this configuration at this point in time**. It does **not** measure other deployments, later model or prompt versions, a different corpus, or live production traffic — re-audit when any of those change. It is not a general rating of the underlying model.

Headline Accuracy (bad-test cases excluded from denominator)

Scoring set: **1000 questions** (of 1000 generated; 0 identified as bad tests — questions where the ground truth itself was malformed — excluded symmetrically from both arms' denominators per audit-grade methodology.).

Metric	Without VERITROOPER (vanilla-RAG baseline)	With VERITROOPER (pipeline)
Final verified accuracy	79.90% (799 / 1000)	95.80% (958 / 1000)
Confirmed errors	201	42
Δ vs baseline		+15.90pp
Failure-Recovery Rate (pipeline recovers what % of baseline failures)		85.1%
95% confidence interval (Wilson)	77.3–82.3%	94.4–96.9%

The 95% confidence interval reflects sampling uncertainty on 1000 scored items: a point estimate of 100% does not prove a 0% error rate in the field — it means no error was observed in this sample; a tighter bound needs more questions.

Release disposition

Acceptance profile: General enterprise · selected by run configuration (default profile). The audited configuration is measured against this profile's thresholds (below). This is a disposition on the audit **evidence** ; final release additionally requires the responsible person's recorded sign-off and remains the provider's decision.

Configuration	Disposition
Without VERITROOPER (baseline)	FAIL
With VERITROOPER (pipeline)	CONDITIONAL PASS

Acceptance criteria & results:

Criterion	Threshold	Baseline	Pipeline
Final verified accuracy	$\geq 95\%$	79.90%	95.80%
Confirmed model errors	$= 0$	201	42
Unsupported-question accuracy	$\geq 95\%$	1.50%	86.00%
Scored sample size	≥ 30	1000	1000
Connection errors (execution failures)	≤ 2	0	0

PASS = all criteria met · CONDITIONAL PASS = accuracy target met but a secondary criterion (confirmed errors, unsupported-question handling, or connection errors) missed · FAIL = accuracy target not met · INSUFFICIENT EVIDENCE = sample too small for a reliable call. Connection errors are execution failures (no answer returned), not model verdicts; under this profile a run exceeding the tolerance should re-run the affected item(s) before release. Other selectable profiles (High assurance, Regulated, Safety-critical) raise these bars; all are configurable in the run configuration.

Human sign-off

No human sign-off is on record for these exact results yet. The responsible person records it — name, title, decision, and date (system clock) — at the end of the run in the results screen, or via vt.py signoff; it then binds to this result set's fingerprint.

Paired outcome (same questions, both arms)

Both arms answered the identical scored questions, so the improvement reads pair-by-pair rather than as two separate accuracy figures. **Recovered** = baseline wrong and pipeline right; **regressed** = the reverse.

Paired outcome	Questions
Both correct	787
Recovered (baseline wrong → pipeline correct)	171
Regressed (baseline correct → pipeline wrong)	12
Both wrong	30
Total scored	1000

Of the 183 questions where the arms disagreed, 171 favoured the pipeline and 12 favoured the baseline. Exact McNemar two-sided $p = 0.0000$.

Per-Category Accuracy & Failure Recovery

Per-category accuracy with and without VERITROOPER. **Failure-Recovery Rate** measures the fraction of vanilla-RAG baseline failures the pipeline recovers — the most direct signal of where the architecture adds value on this material. Every category is shown, including any where the pipeline matches or underperforms baseline.

Question Type	Questions	Baseline Accuracy	Pipeline Accuracy	Improvement	Failure-Recovery Rate
Negative (hallucination traps)	200	1.50%	86.00%	+84.50pp	85.8%
Calculation	12	100.00%	100.00%	—	—
Exception	93	100.00%	100.00%	—	—
Precision	218	99.54%	99.08%	-0.46pp	0.0%
Cross-Reference	251	99.20%	97.61%	-1.59pp	100.0%
Conditional	211	99.53%	97.63%	-1.90pp	0.0%
Cause & Effect	15	100.00%	93.33%	-6.67pp	—
All Categories	1,000	79.90%	95.80%	+15.90pp	85.1%

Adjudication reconciliation

How each arm's raw automated sweep becomes the final verified accuracy, stage by stage. Every count reconciles to the total generated; "cleared by verification" are answers the validator flagged but the Doctor + independent verifier confirmed were actually correct. Bad tests (malformed questions) and connection errors (no answer returned — an execution fault) are excluded symmetrically from both arms.

Stage	Baseline	Pipeline
Passed automated review	796	940
Flagged for review	204	60
— cleared by verification (false flags)	3	18
— confirmed model errors	201	42
Final correct / scored	799 / 1000	958 / 1000
Bad tests excluded (symmetric)	0	0
Connection errors excluded (no answer returned)	0	0
Total generated	1000	1000

Test-set coverage

What this question panel actually exercised. A point-in-time audit speaks only to the material it covered, so coverage is stated explicitly. "Source passages" are the indexed evidence chunks the questions were generated from — not formal document headings.

Coverage measure	Value
Distinct source passages probed	226
Questions per passage (mean / median)	4.4 / 4
Most / least on a single passage	15 / 1
Passages with only one question	24
Share of questions from the top 5 passages	7%
Adversarial (hallucination-trap) questions	200 of 1000 (20%)

- **Category coverage:** Cross-Reference (251), Precision (218), Conditional (211), Negative (hallucination traps) (200), Exception (93), Cause & Effect (15), Calculation (12).

Passage-level counts above are exact. Whole-corpus coverage (% of ALL source passages represented) is not computed here — the corpus's total passage count is not recorded in this run; it requires the corpus manifest (a planned addition).

Estimated audit resource usage

What running this audit consumed. **Measured** values are exact; **estimated** values (tokens, cost) are approximations — the audit is not per-call token-metered — for planning rather than billing.

Measured

- Wall-clock duration: 214 min 3 s.
- Answer calls: 1000 baseline + 1000 pipeline = 2000.
- Compute locality: MUT **Meta-Llama-3.1-70B-Instruct** (local) · Doctor **Meta-Llama-3.1-70B-Instruct** (local)
- Verifier **gpt** (hosted API).

Estimated

- Adjudication calls: ~857 Doctor + ~857 independent-verifier review(s) (inferred from the result records, not directly metered).
- Answer-phase tokens: ~215,957 (chars÷4 over system prompt + question + answer; retrieved-evidence context not stored, so this UNDERSTATES real usage).
- Answer-phase API cost: n/a — no list price on file for the model under test; multiply the estimated tokens above by your provider's rate.

Estimated cost is a lower bound — retrieval at answer time may pull more context than the gold evidence chunk, and Doctor / verifier token usage is not included. Local / self-hosted models incur compute, not API cost. Per-call token metering is a planned addition for exact figures.

Without veritrooper — Baseline LLM Performance

What the LLM Did On Its Own

Without veritrooper, Meta-Llama-3.1-70B-Instruct answered a meaningful minority of questions incorrectly. A notable pattern in these failures was the model's tendency to answer questions that it should have refused because the provided documents did not contain the answer. For instance, when asked for the filing fee for a discrimination complaint, the model attempted to provide an answer instead of stating that the information was not available in the source material. This behavior of inventing or guessing answers for unanswerable questions was a primary driver of errors in the baseline configuration.

Failure Patterns (Baseline)

The diagnostic review identified several patterns in the baseline model's errors:

- * **Knowledge Gap:** The most significant pattern was the model answering questions for which the provided documents contained no information. The model frequently failed to recognize when an answer was not present and attempted to respond anyway, leading to incorrect or fabricated information.
- * **Other:** This category includes a variety of miscellaneous errors that did not fit into a more specific pattern. These were isolated, one-off mistakes in logic or interpretation.
- * **Computed vs. Quoted:** In a small number of cases, the model incorrectly performed a calculation or synthesis of information instead of quoting a direct value or statement from the source text as required.

Baseline Remediation Recommendations

To address the failures observed in the baseline configuration, the following system-level remediations could be considered:

1. **Strengthen "Refusal" Instructions:** The system prompt for Meta-Llama-3.1-70B-Instruct should be enhanced with more explicit and forceful instructions to refuse to answer any question if the answer is not explicitly present in the provided source documents. Emphasize that stating "unanswerable" or a similar refusal is the correct behavior when information is missing.
 2. **Implement Answer Validation:** Before presenting an answer to a user, an application-level check could be implemented. This could involve a secondary, simpler query to verify that the key entities or facts in the model's answer are actually present in the source text that was used to generate it.
 3. **Route Unanswerable Questions for Review:** To mitigate the risk of fabricated answers, develop a process to flag and route questions that the model should have identified as unanswerable. This allows a human expert to confirm that the information is truly absent from the knowledge base and handle the user's query appropriately.
-

With veritrooper — Pipeline Performance

What the LLM Did With veritrooper

With the veritrooper engine, Meta-Llama-3.1-70B-Instruct still produced a meaningful minority of incorrect answers. While the pipeline recovered a portion of the errors seen in the baseline, a distinct set of failures remained. The most common residual error pattern involved the model failing to find an answer that was present in the source documents. For example, when asked for the street address of the American National Standards Institute, the model did not extract the correct address from the provided text. Another observed pattern was the misreading of tabular data.

Pipeline Remediation Recommendations

To address the residual failures observed in the pipeline configuration, the following system-level remediations could be considered:

1. **Improve Retrieval for Specific Entities:** The remaining knowledge gap errors suggest that for certain types of specific data points (like addresses or technical specifications), the retrieval mechanism may not be surfacing the most relevant text snippet. Investigate whether pre-processing documents to better tag or structure such entities could improve retrieval performance for these query types.
 2. **Refine Table Extraction Logic:** For errors involving misread tables, consider implementing a specialized pre-processing step that extracts tables from documents and converts them into a more structured format (like JSON or Markdown) before they are passed to the model. This can make it easier for the model to parse and correctly query tabular data.
 3. **Targeted Human Review:** Implement a review queue specifically for answers generated from complex documents containing multiple tables or dense regulatory text. This allows human experts to focus their oversight on the types of content where the model is most likely to make a mistake.
-

Independent Verification

An independent verifier audited the diagnostic verdicts for both the baseline and pipeline results. This audit applied the same standard to both sets of outputs, ensuring a fair and balanced comparison of the model's performance in each configuration.

Bottom Line

In this audit, the baseline model's primary weakness was inventing answers to unanswerable questions. The veritrooper pipeline addressed some of these failures but introduced different error patterns, primarily related to failing to find existing information or misreading tables. While the pipeline configuration reduced certain risks, both configurations produced errors and did not meet a zero-confirmed-error standard on this audit panel, underscoring the continued need for human oversight before using answers in a business context.

Deterministic Guardrails

Code-level checks you can deploy in your own stack to catch the failure patterns found in this run — no model retraining required. Because they are deterministic, they behave identically on every run. **Block** = the check is authoritative and can reject or correct the answer automatically; **Flag** = the check detects the error class and routes it to human review or abstention. This complements the remediation recommendations above — it does not replace them.

Block — reject or correct automatically:

- **Grounding check** — Require every figure and factual claim in the answer to appear in — or be directly entailed by — the retrieved evidence; reject answers with unsupported spans (the model leaning on its own knowledge over the source). (Addresses: Out of scope · 146 cases this run.)
- **Quote-vs-compute guard** — Decide whether the answer should be a verbatim quote or a computed value; validate a quote as a substring of the source and a computed value by recomputation. (Addresses: Wrong calc/quote · 18 cases this run.)
- **Numeric tolerance** — Normalize both values to a common precision and accept only within an explicit rounding tolerance (e.g. ± 0.5 of the least-significant digit). (Addresses: Out of scope, Unit/format mismatch, Wrong calc/quote · 5 cases this run.)
- **Unit/format normalization** — Parse the value AND its unit, convert to a canonical unit/scale (% , \$, thousands vs millions) before comparison, and reject unit or scale mismatches. (Addresses: Unit/format mismatch · 4 cases this run.)
- **Recompute check** — Recompute the arithmetic from the source figures and reject any answer whose stated result differs beyond a rounding tolerance. (Addresses: Out of scope, Wrong calc/quote · 2 cases this run.)

Flag — detect and route to review:

- **Cell-provenance check** — Confirm the answered value comes from the exact row + column the question names (row/column-header match); flag row/column drift. (Addresses: Table misread · 4 cases this run.)
- **Rule/citation check** — Verify the cited rule/section actually governs the question's scenario; flag mis-applied rules. (Addresses: Misinterpreted, Out of scope · 2 cases this run.)
- **Period-match check** — Extract the period/date from the question and require the answer's figure to come from that period; flag period mismatches. (Addresses: Out of scope · 1 case this run.)
- **Source-span verification** — Require the answer's key value to be extractable verbatim from the retrieved passage; flag when it isn't. (Addresses: Misinterpreted · 1 case this run.)

Case counts are per flagged answer and can overlap — a single answer may exhibit more than one failure

pattern — so they need not sum to the number of failed questions. Patterns without a listed guard are semantic reasoning slips with no clean deterministic check; the remediation recommendations above address those.

Complete Failure Listing — With VERITROOPER (Pipeline)

220 confirmed model errors + 4 excluded bad tests (malformed question, not counted against the model). Each entry shows the full question, the model's full answer, the ground truth, the Doctor's diagnosis, and (where available) the Verifier's independent audit verdict.

Failure shape at a glance — With VERITROOPER (Pipeline)

Category	Failures	Model error	Bad test	Review recommended
Cross-Reference	86	86	0	0
Conditional	68	68	0	0
Precision	33	33	0	0
Negative (hallucination traps)	15	11	4	0
Exception	12	12	0	0
Cause & Effect	7	7	0	0
Calculation	3	3	0	0
TOTAL	224	220	4	0

Severity mix (confirmed errors): 18 HIGH, 189 MEDIUM, 13 LOW. Severity is a domain-agnostic heuristic from the error type — a confidently-wrong answer outranks a missing one; re-weight for your risk context.

Failure 1 — `OSHA\_29CFR\_full\_2024\_2250\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** Equipment may have become contaminated with blood, but the employer can demonstrate that decontamination of that equipment is not feasible. Must it still be decontaminated before servicing?
- ✓ **Ground truth:** No, decontamination is required unless the employer can demonstrate that decontamination of such equipment is not feasible.
- ✗ **Model answered:** {"name": "decontaminate_equipment", "parameters": {"contaminated": "True", "feasible": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 2 — `OSHA\_29CFR\_full\_2024\_1138\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a critical component of the PSDI safety system fails, can the press remain in service after recertification by the employer alone, without revalidation?
- ✓ **Ground truth:** No; the failure of a critical component requires removal of the safety system from service until it is recertified AND revalidated.
- ✗ **Model answered:** {"name": "recertification_by_employer", "parameters": {"component": "critical"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model confused critical with non-critical component
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 3 — `OSHA\_29CFR\_full\_2024\_2632\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If data on a skin sensitizer are not sufficient for sub-categorization, into which category must the substance be classified?
- ✓ **Ground truth:** It shall be classified in Category 1.
- ✗ **Model answered:** {"name": "classify_skin_sensitizer", "parameters": {"sub_categorization": "insufficient"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not follow instruction for insufficient data
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 4 — `OSHA\_29CFR\_full\_2024\_neg\_2602\_03` · MEDIUM confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** How many toxicologically active ingredients must mixtures A and B share for the interpolation principle in A.0.5.1.4 to apply?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "None", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** bad test
- **Doctor root cause:** [verifier] The evidence does answer the requirement qualitatively: A.0.5.1.4 applies only to mixtures “with identical ingredients” and C having “the same toxicologically active ingredients as mixtures A and B,” so A and B must share all/same toxicologically active ingredients rather than the answer being wholly unanswerable.
- **Verifier verdict:** bad test (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 5 — `OSHA\_29CFR\_full\_2024\_3305\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference

- **Question:** Using subpart (ii) and subpart (i), if a stairway-type ladder rises from a bottom step set at the maximum allowed 24 inches above the scaffold supporting level, what is the maximum vertical interval required between rest platforms?
- ✓ **Ground truth:** 12 feet (3.7 m), as specified in subpart (ii) for maximum vertical intervals between rest platforms, regardless of the 24-inch bottom step maximum in subpart (i).
- ✗ **Model answered:** {"name": "max", "parameters": {"a": "12", "b": "3"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 6 — `OSHA\_29CFR\_full\_2024\_neg\_1194\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How far from bulk sulfur or baled paper storage may welding be conducted safely?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "relocation of combustibles", "parameters": {"distance": "35"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated answer not in evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 7 — `OSHA\_29CFR\_full\_2024\_neg\_1599\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum cord- and plug-connected load permitted for a receptacle on a 40-ampere branch circuit supplying two or more receptacles?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "Table S-4", "parameters": {"Circuit rating (amperes)": "40"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated answer not in evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 8 — `OSHA\_29CFR\_full\_2024\_0564\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** When an LP-Gas-fueled vehicle is being repaired in a garage, may its container shutoff valve be open?
- ✓ **Ground truth:** Only when fuel is required for engine operation; otherwise the shutoff valve must be closed.
- ✗ **Model answered:** {"name": "check_container_shutoff_valve", "parameters": {"status": "closed"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not fully understand the conditions under which the container shutoff valve may be open
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 9 — `OSHA\_29CFR\_full\_2024\_2809\_11` · HIGH confidence

TABLE MISREAD

- **Severity:** MEDIUM — a misapplied or over-extrapolated rule — plausible but wrong
- **Category:** Precision
- **Question:** What is the street address (number and street) of the American National Standards Institute (ANSI) given under Availability of Referenced Material?
- ✓ **Ground truth:** 1430 Broadway, New York, NY 10018
- ✗ **Model answered:** {"name": "get_street_address", "parameters": {"organization": "American National Standards Institute (ANSI)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model read from wrong table row
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 10 — `OSHA\_29CFR\_full\_2024\_3020\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Cross-Reference
- **Question:** Under subpart (vii) and subpart (viii), in an area where a major spill may occur and salvage equipment must be on hand, what two distinct requirements apply regarding salvage drums and spill containment?
- ✓ **Ground truth:** U.S. Department of Transportation specified salvage drums or containers and suitable quantities of proper absorbent must be kept available and used (vii), and a spill containment program-part of the employer's safety and health program required in paragraph (b)-must be implemented to contain and isolate the entire volume of the hazardous substance being transferred (viii).
- ✗ **Model answered:** {"name": "salvage_drums_and_spill_containment_requirements", "parameters": {"subpart_vii": "True", "subpart_viii": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports the ground truth—(vii) requires DOT-specified salvage drums/containers and absorbent be kept available and used, and (viii) requires a spill containment program to contain/isolate the full volume—but the model only answered True for each subpart and did not provide the two required requirements.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 11 — `OSHA\_29CFR\_full\_2024\_3312\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If work near an insulated power line of 200 volts is being performed, what is the minimum distance required?
- ✓ **Ground truth:** 3 feet (0.9 m), since the line is less than 300 volts.

- **✗ Model answered:** {"name": "get_minimum_distance", "parameters": {"voltage": "200", "line_type": "insulated"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 12 — `OSHA\_29CFR\_full\_2024\_4398\_00` · HIGH confidence

MISINTERPRETED

- **Severity:** MEDIUM — a misapplied or over-extrapolated rule — plausible but wrong
- **Category:** Cause & Effect
- **Question:** What is the consequence when an ALJ approves a request to withdraw objections to the Assistant Secretary's findings and/or order, and there are no other pending objections?
- **✓ Ground truth:** The Assistant Secretary's findings and/or order will become the final order of the Secretary.
- **✗ Model answered:** {"name": "consequence_of_withdrawal", "parameters": {"event": "ALJ approves withdrawal of objections and no other pending objections", "result": "Assistant Secretary"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly states that under those conditions “the Assistant Secretary's findings and/or order will become the final order of the Secretary,” while the model's result of merely “Assistant Secretary” is incomplete and does not answer the consequence.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Uncertainty violation: evidence hedges but response uses 'proves'

Failure 13 — `OSHA\_29CFR\_full\_2024\_1114\_01` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the maximum nominal a.c. supply voltage permitted for a.c. control circuits and solenoid valve coils?
- **✓ Ground truth:** 120-volt a.c.
- **✗ Model answered:** {"name": "get_nominal_voltage", "parameters": {"voltage_type": "a.c.", "circuit_type": "control"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly states a.c. control circuits and solenoid valve coils are powered by 'not more than a nominal 120-volt a.c. supply,' but the model output only a function-call-like JSON without the voltage, so it did not answer the question.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 14 — `OSHA\_29CFR\_full\_2024\_1336\_00` · MEDIUM confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** Band wheels must be completely encased or guarded, but under what condition is part of the wheel allowed to be exposed?
- **✓ Ground truth:** A portion of the upper wheel immediately around the point where the blade leaves the wheel may be exposed, to permit the operator to observe movement of equipment.

- **✗ Model answered:** {"name": "band_wheels_encased", "parameters": {"condition": "to permit operator to observe movement of equipment"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 15 — `OSHA\_29CFR\_full\_2024\_1964\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under paragraph (l)(2)(i), what legend must the employer post at entrances to regulated areas, and what alternative legend could be used in its place prior to June 1, 2016 under paragraph (l)(2)(iii)?
- **✓ Ground truth:** Under (l)(2)(i) the required legend is: 'DANGER VINYL CHLORIDE MAY CAUSE CANCER AUTHORIZED PERSONNEL ONLY'. Prior to June 1, 2016, under (l)(2)(iii), the employer could instead use: 'CANCER-SUSPECT AGENT AREA AUTHORIZED PERSONNEL ONLY'.
- **✗ Model answered:** {"name": "get_legends", "parameters": {"paragraph": "(l)(2)(i)", "alternative_paragraph": "(l)(2)(iii)", "date": "June 1, 2016"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly gives the two legends—"DANGER VINYL CHLORIDE MAY CAUSE CANCER AUTHORIZED PERSONNEL ONLY" and "CANCER-SUSPECT AGENT AREA AUTHORIZED PERSONNEL ONLY"—but the model only returned a tool-call-like parameter object and did not provide either legend.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 16 — `OSHA\_29CFR\_full\_2024\_2190\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If the employer monitored within one year prior to the standard's effective date and that monitoring satisfies all other requirements, must the employer still conduct new initial monitoring?
- **✓ Ground truth:** No; the employer may rely on such earlier monitoring results to satisfy the initial monitoring requirement.
- **✗ Model answered:** {"name": "monitoring_required", "parameters": {"prior_monitoring": "True", "satisfies_requirements": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 17 — `OSHA\_29CFR\_full\_2024\_1202\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** For transmission pipeline field shop operations described in paragraph (d)(1)(ii), which additional paragraph requirement applies that is NOT required under the general transmission pipeline requirements of paragraph (d)(1)(i)?

- ✓ **Ground truth:** Paragraph (d)(1)(i) general requirements observe paragraphs (b) and (c) and § 1910.254. Field shop operations in (d)(1)(ii) additionally require paragraph (a) and § 1910.253. So paragraph (a) (and § 1910.253) is the added requirement.
- ✗ **Model answered:** {"name": "get_additional_requirement", "parameters": {"paragraph": "(d)(1)(ii)", "general_requirement": "(d)(1)(i)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model stopped before applying dependent rule
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 18 — `OSHA\_29CFR\_full\_2024\_0455\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a metallic floorplate on which a container stands during filling is electrically connected to the fill stem, must the nozzle and container also be separately electrically interconnected to comply with the grounding section?
- ✓ **Ground truth:** No, that arrangement is deemed to have complied with the grounding provisions.
- ✗ **Model answered:** {"name": "answer_grounding_question", "parameters": {"electrically_connected": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 19 — `OSHA\_29CFR\_full\_2024\_0517\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Considering the internal-combustion engine requirement in subparagraph (g)(2)(ii) and the fuel oil flow requirement in subparagraph (g)(2)(iii), what two fire-related safety measures are required regarding engines and fuel oil flow?
- ✓ **Ground truth:** Internal-combustion engines used for electric power generation must be located outside the mixing plant building, or be properly ventilated and isolated by a firewall, with exhaust systems located so spark emission cannot be a hazard; and suitable means must be provided to prevent the flow of fuel oil to the mixer in case of fire, with an automatic spring-loaded shutoff valve with fusible link installed in gravity flow systems.
- ✗ **Model answered:** {"name": "get_fire_safety_measures", "parameters": {"engine_requirement": "internal-combustion engine requirement in subparagraph (g)(2)(ii)", "fuel_oil_flow_requirement": "fuel oil flow requirement in subparagraph (g)(2)(iii)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 20 — `OSHA\_29CFR\_full\_2024\_0455\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually

- **Category:** Conditional
- **Question:** If a pit has any part within a Division 1 or 2 classified area but is provided with mechanical ventilation, how must the pit be classified?
- ✓ **Ground truth:** It shall be classified Division 2 (rather than Division 1).
- ✗ **Model answered:** {"name": "classify_pit", "parameters": {"has_mechanical_ventilation": "True", "is_within_division_1_or_2": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 21 — `OSHA\_29CFR\_full\_2024\_3162\_04` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Conditional
- **Question:** If a proposed location for an air receiver would be underground or inaccessible, is that installation allowed?
- ✓ **Ground truth:** No, under no circumstances shall an air receiver be buried underground or located in an inaccessible place.
- ✗ **Model answered:** {"name": "is_installation_allowed", "parameters": {"location": "underground or inaccessible"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly answers “Under no circumstances shall an air receiver be buried underground or located in an inaccessible place,” so the answer is no, but the model only emitted a function-style request with the location and did not provide the required no/not allowed answer.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 22 — `OSHA\_29CFR\_full\_2024\_0227\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Cross-Reference
- **Question:** Combining the UL listings, how many of the materials available from Underwriters Laboratories are IBR approved for § 1910.106(b)(1)(iii)(a)(1)?
- ✓ **Ground truth:** All three (UL 58-61, UL 80-63, and UL 142-68) are IBR approved for § 1910.106(b)(1)(iii)(a)(1).
- ✗ **Model answered:** {"name": "count", "parameters": {"iterable": "[58, 61, 80, 63, 142, 68]"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence lists three UL materials—UL 58-61, UL 80-63, and UL 142-68—each IBR approved for § 1910.106(b)(1)(iii)(a)(1), while the model attempted to count six numeric fragments rather than the three materials.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 23 — `OSHA\_29CFR\_full\_2024\_0333\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional

- **Question:** If a platform uses more than two suspension ropes in an intermittent stabilization system, which ropes require a stabilizer attachment?
- ✓ **Ground truth:** Only the two building-side suspension ropes at the platform ends require a stabilizer attachment.
- ✗ **Model answered:** {"name": "intermittent_stabilization_system", "parameters": {"number_of_suspension_ropes": "more than 2"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 24 — `OSHA\_29CFR\_full\_2024\_3130\_05` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** Category 4 flammable liquids include liquids having flashpoints above what temperature?
- ✓ **Ground truth:** above 140 °F (60 °C)
- ✗ **Model answered:** {"name": "get_flashpoint_category", "parameters": {"category": "4"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed wrong flashpoint category
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 25 — `OSHA\_29CFR\_full\_2024\_2565\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under what statute is a 'consumer product' defined for purposes of subpart (v), and under what statute is a 'hazardous substance' defined for purposes of subpart (6)(ii)?
- ✓ **Ground truth:** A 'consumer product' is defined under the Consumer Product Safety Act (15 U.S.C. 2051 et seq.) in subpart (v), while a 'hazardous substance' in subpart (6)(ii) is defined under the Comprehensive Environmental Response, Compensation and Liability Act (CERCLA) (42 U.S.C. 9601 et seq.).
- ✗ **Model answered:** {"name": "get_statute_definition", "parameters": {"term": "consumer product", "subpart": "v"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model only partially answered the question
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 26 — `OSHA\_29CFR\_full\_2024\_3327\_02` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the maximum clearance allowed above the walking/working surface for a securely fastened toeboard?
- ✓ **Ground truth:** 1/4 inch (0.7 cm)
- ✗ **Model answered:** {"name": "max_clearance", "parameters": {"clearance": "0"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed incorrect value
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 27 — `OSHA\_29CFR\_full\_2024\_3641\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** For an NPE system, given the air change specification and the pressure differential specification, what two numeric minimums must be maintained inside the enclosure?
- ✓ **Ground truth:** At least 4 air changes per hour and a minimum of -0.02 column inches of water pressure differential relative to outside pressure.
- ✗ **Model answered:** {"name": "NPE_system_specifications", "parameters": {"air_changes": "4", "pressure_differential": "0"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 28 — `OSHA\_29CFR\_full\_2024\_0325\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an employer has 8 employees and an OSHA standard in this part requires a fire prevention plan, must the plan be in writing?
- ✓ **Ground truth:** No, an employer with 10 or fewer employees may communicate the plan orally to employees.
- ✗ **Model answered:** {"name": "is_fire_prevention_plan_required_to_be_in_writing", "parameters": {"number_of_employees": "8"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed incorrect threshold value
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 29 — `OSHA\_29CFR\_full\_2024\_0455\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If only Category 3 flammable liquids with a flashpoint at or above 100 °F or Category 4 flammable liquids are handled, is special Class I electrical equipment required?
- ✓ **Ground truth:** No, ordinary electrical equipment is satisfactory, though care must be used in locating it to prevent hot metal from falling into open equipment.
- ✗ **Model answered:** {"name": "electrical_equipment_required", "parameters": {"category": "3", "flashpoint": "100"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** (none)

Failure 30 — `OSHA\_29CFR\_full\_2024\_0075\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an employer files a notice of intention to contest under § 1903.17, is the employer still required to keep the citation posted?
- ✓ **Ground truth:** Yes, the posting responsibility continues unless and until the Review Commission issues a final order vacating the citation.
- ✗ **Model answered:** {"name": "posting_requirement", "parameters": {"event": "filing a notice of intention to contest under \u00a7 1903.17"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly answers yes—“filing ... a notice of intention to contest ... shall not affect his posting responsibility ... unless and until the Review Commission issues a final order vacating the citation”—but the model only returned a vague structured label/event and did not provide that answer.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 31 — `OSHA\_29CFR\_full\_2024\_4138\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Cross-Reference
- **Question:** Comparing Option (1) physical attachment and Option (4)(B), how is the connection to the flotation device characterized in each?
- ✓ **Ground truth:** Option (1) physical attachment connects the crane/derrick physically to the barge, pontoons, vessel or other means of flotation using methods like crossed-cable systems, bolting, welding, or strapping with chains. Option (4)(B) requires that the wire rope is attached physically to the vessel/flotation device. Both require a physical attachment to the vessel/flotation device.
- ✗ **Model answered:** {"name": "characterize_connection", "parameters": {"option": "Option (1) physical attachment"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports the ground truth: Option (1) says “The crane/derrick is physically attached to the barge, pontoons, vessel or other means of flotation,” while Option (4)(B) says “The wire rope is attached physically to the vessel/flotation device”; the model only named Option (1) and did not compare or address Option (4)(B).
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 32 — `OSHA\_29CFR\_full\_2024\_3987\_01` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** The operator must not leave the controls while the load is suspended. Under what conditions does this prohibition NOT apply (i.e., when may the operator leave)?
- ✓ **Ground truth:** Only when ALL of the following are met: (i) the operator remains adjacent to the equipment and is not engaged in any other duties; (ii) the load is to be held suspended for a period exceeding normal

lifting operations; (iii) the competent person determines it is safe and implements measures to restrain the boom hoist and telescoping, load, swing, and outrigger/stabilizer functions; and (iv) barricades or caution lines and notices are erected to prevent all employees from entering the fall zone, with no employees permitted in the fall zone.

- **✗ Model answered:** {"name": "leaving_controls_while_load_suspended", "parameters": {"condition": "operator remains adjacent to the equipment and is not engaged in any other duties"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model stopped before applying dependent rule
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 33 — `OSHA\_29CFR\_full\_2024\_1848\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Based on the permissible exposure limit in Section I.D and the respirator requirements in Section III.A, at what airborne fiber concentration must you begin wearing a respirator, and what type of respirator is required if concentrations reach 3 f/cc?
- **✓ Ground truth:** You must wear a respirator when exposure exceeds the PEL of 0.1 f/cc. Since 3 f/cc exceeds 2 f/cc, an air-supplied, positive-pressure, full facepiece respirator must be used (HEPA air-purifying respirators are only allowed where concentrations do not exceed 2 f/cc).
- **✗ Model answered:** {"name": "get_respirator_requirements", "parameters": {"airborne_fiber_concentration": "3"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 34 — `OSHA\_29CFR\_full\_2024\_4450\_03` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the Statutes at Large citation (124 Stat.) for the Health Care and Education Reconciliation Act of 2010, Public Law 111-152?
- **✓ Ground truth:** 124 Stat. 1029
- **✗ Model answered:** {"name": "get_statutes_at_large_citation", "parameters": {"public_law": "111-152", "statutes_at_large": "124"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence explicitly gives the citation as “Health Care and Education Reconciliation Act of 2010, Public Law 111-152, 124 Stat. 1029,” while the model returned only “124,” omitting the required page number 1029.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 35 — `OSHA\_29CFR\_full\_2024\_2261\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional

- **Question:** When a healthcare professional is evaluating an employee after an exposure incident, is the employer required to provide the source individual's blood test results?
- ✓ **Ground truth:** Yes, but only if available.
- ✗ **Model answered:** {"name": "evaluate_exposure_incident", "parameters": {"provide_source_individual_blood_test_results": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not include the condition that the source individual's blood test results are only required if available
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 36 — `OSHA\_29CFR\_full\_2024\_1848\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the disposal requirement in Section IV.A.4 with the packaging requirement in Section IV.E, how must asbestos fireproofing material removed from a building be disposed of?
- ✓ **Ground truth:** Asbestos fireproofing or insulating material removed from buildings is a waste under IV.A.4, and per IV.E it must be disposed of in leak-tight 6-mil thick plastic bags, plastic-lined cardboard containers, or plastic-lined metal containers, and sealed while wet before drying out.
- ✗ **Model answered:** {"name": "asbestos_disposal_requirements", "parameters": {"disposal_requirement": "Section IV.A.4", "packaging_requirement": "Section IV.E"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 37 — `OSHA\_29CFR\_full\_2024\_2508\_03` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a filter removes 99.95 percent of mono-dispersed particles of 0.3 micrometers in diameter, does it qualify as a HEPA filter?
- ✓ **Ground truth:** No, a HEPA filter must be at least 99.97 percent efficient in removing such particles.
- ✗ **Model answered:** {"name": "is_HEPA_filter", "parameters": {"efficiency": "99", "particle_size": "0"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed ratio instead of quoting stated figure
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 38 — `OSHA\_29CFR\_full\_2024\_0163\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — a missing or incomplete answer — the model withheld an available answer (a gap, not a confidently wrong statement)
- **Category:** Conditional
- **Question:** If a party wishes to rely on an objection later in a proceeding, but that objection was never made before the presiding hearing examiner, can it be relied upon?
- ✓ **Ground truth:** No, only objections made before the presiding hearing examiner may be relied upon

subsequently in a proceeding.

- **✗ Model answered:** {"name": "objection_reliance", "parameters": {"made_before_examiner": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states, "Only objections made before the presiding hearing examiner may be relied upon subsequently," so an objection never made before the examiner cannot be relied upon; the model merely restated that it was not made before the examiner and did not answer the reliance question.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 39 — `OSHA\_29CFR\_full\_2024\_2437\_05` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Conditional
- **Question:** Pending a final medical determination, may the employer return an employee who has been on removal status for the preceding six months due to MDA exposure?
- **✓ Ground truth:** No; the employer shall await a final medical determination.
- **✗ Model answered:** {"name": "employer_options_pending_final_medical_determination", "parameters": {"option": "return", "removal_status_duration": "6 months"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '6' not in evidence

Failure 40 — `OSHA\_29CFR\_full\_2024\_1434\_04` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** May an employer use a method of calculating incident heat energy not included in appendix E?
- **✓ Ground truth:** Yes, if the chosen method reasonably predicts the incident energy to which the employee would be exposed.
- **✗ Model answered:** {"name": "Note 1 to paragraph", "parameters": {"I": "8", "method": "calculating incident heat energy not included in appendix E"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 41 — `OSHA\_29CFR\_full\_2024\_3038\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an employer evacuates all employees from the worksite when an emergency occurs and does not permit any employees to assist in handling the emergency, are they still required to comply with the emergency response program requirements of paragraph (p)(8)?
- **✓ Ground truth:** No, they are exempt from paragraph (p)(8) provided they have an emergency action plan complying with § 1926.35.

- **✗ Model answered:** {"name": "emergency_response_program", "parameters": {"employer_evacuates_employees": "True", "employees_assist_in_emergency": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 42 — `OSHA\_29CFR\_full\_2024\_1292\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under the general machine guarding provisions, which requires enclosure regardless of location and which only requires enclosure when located within reach from platforms or within 8 feet 6 inches from the floor?
- **✓ Ground truth:** Per paragraph (c)(2), all gears shall be completely enclosed regardless of location. Per paragraph (c)(3), sprockets and V-belt drives must be completely enclosed only when located within reach from platforms or passageways or located within 8 feet 6 inches from the floor.
- **✗ Model answered:** {"name": "get_machine_guarding_requirements", "parameters": {"location": "regardless of location", "component": "gears"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model only partially answered the question
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 43 — `OSHA\_29CFR\_full\_2024\_0325\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — a misapplied or over-extrapolated rule — plausible but wrong
- **Category:** Conditional
- **Question:** If an employer has 15 employees, does the option to communicate the fire prevention plan orally apply?
- **✓ Ground truth:** No, the oral option only applies to employers with 10 or fewer employees, so this plan must be in writing, kept in the workplace, and made available to employees.
- **✗ Model answered:** {"name": "is_oral_communication_applicable", "parameters": {"number_of_employees": "15"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model misapplied employee number threshold
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 44 — `OSHA\_29CFR\_full\_2024\_0882\_04` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** The energy control program required by paragraph (c)(1) consists of three components. Which paragraphs of subsection (c) address the second and third of those components?
- **✓ Ground truth:** Paragraph (c)(1) lists energy control procedures, employee training, and periodic inspections. Employee training is addressed in paragraph (c)(7) (Training and communication) and periodic inspections are addressed in paragraph (c)(6) (Periodic inspection).

- **✗ Model answered:** {"name": "get_paragraphs", "parameters": {"component": "employee training"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 45 — `OSHA\_29CFR\_full\_2024\_2239\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the retention requirement and the availability requirement, for how long must employee medical records be maintained, and to whom must they be provided upon request?
- **✓ Ground truth:** The medical records must be maintained for at least 40 years, or for the duration of employment plus 20 years, whichever is longer (paragraph (iii)). These records must be provided upon request to employees, designated representatives, and the Assistant Secretary in accordance with 29 CFR 1910.1020(a)-(e) and (g)-(i) (paragraph (4)(ii) / (ii) of Availability).
- **✗ Model answered:** {"name": "get_retention_and_availability_requirements", "parameters": {"record_type": "employee medical records"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not fully synthesize the retention and availability requirements from the evidence passage
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 46 — `OSHA\_29CFR\_full\_2024\_4294\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining paragraph (c) and paragraph (d)(2), what minimum information about the reporting individual's identity may or may not be reflected in the establishment's log if confidentiality is requested?
- **✓ Ground truth:** The log (per (d)(2)) requires date, time, code/reference/file number, location, description, classification, and action taken, but does NOT list the employee's name; under (c), upon request no person shall disclose the name of the individual making the report or names of employees referred to, except to authorized representatives of the Secretary. So the name is kept confidential and is not among the required log items.
- **✗ Model answered:** {"name": "get_log_info", "parameters": {"requested_confidentiality": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 47 — `OSHA\_29CFR\_full\_2024\_2444\_02` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the detector zone temperature in the GC analysis conditions for MDA?
- **✓ Ground truth:** 335 degrees C
- **✗ Model answered:** {"name": "get_detector_zone_temperature", "parameters": {"zone": "detector"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 48 — `OSHA\_29CFR\_full\_2024\_3641\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Considering both the glove bag work practice for adjacent material and the negative pressure glove box work practice for adjacent ACM, how must loose or damaged material near the enclosure be handled in each case?
- **Ground truth:** In both cases, the loose friable or damaged material adjacent to the operation shall be wrapped and sealed in two layers of six mil plastic (or otherwise rendered intact) before beginning the job.
- **Model answered:** {"name": "handle_adjacent_material", "parameters": {"material_condition": "loose", "enclosure_type": "glove bag"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 49 — `OSHA\_29CFR\_full\_2024\_0742\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** Combining the requirements for who may make repairs and how repairs must be conducted, what two standards govern a general repair to a respirator?
- **Ground truth:** Per subpart (i), repairs may be made only by persons appropriately trained to perform such operations using only the manufacturer's NIOSH-approved parts; and per subpart (ii), repairs shall be made according to the manufacturer's recommendations and specifications for the type and extent of repairs to be performed.
- **Model answered:** {"name": "perform_repairs", "parameters": {"standard_1": "Repairs or adjustments to respirators are to be made only by persons appropriately trained to perform such operations and shall use only the respirator manufacturer", "standard_2": "Repairs shall be made according to the manufacturer"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence requires repairs be by trained persons using only the manufacturer's NIOSH-approved parts and be made according to the manufacturer's recommendations/specifications, but the model's answer is truncated to 'use only the respirator manufacturer' and 'according to the manufacturer,' omitting the operative parts and recommendations/specifications details.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '1' not in evidence

Failure 50 — `OSHA\_29CFR\_full\_2024\_4446\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — a misapplied or over-extrapolated rule — plausible but wrong
- **Category:** Conditional
- **Question:** If a timely petition for review is filed and the case was heard before an ALJ, who transmits the

record to the appropriate court?

- ✓ **Ground truth:** The ARB or the ALJ, as the case may be, transmits the record (including the record of proceedings before the ALJ).
- ✗ **Model answered:** {"name": "transmit_record", "parameters": {"entity": "ALJ"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not fully capture the transmitting entity
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 51 — `OSHA\_29CFR\_full\_2024\_neg\_0484\_04` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum number of shifts' worth of flammable liquid supply permitted to be kept near spraying operations before bulk storage rules apply?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "None", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** bad test
- **Doctor root cause:** [verifier] The ground truth says UNANSWERABLE, but the evidence directly answers that liquids kept near spraying operations 'should ordinarily not exceed a supply for 1 day or one shift,' i.e. one shift's worth, so the test label is unsupported.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 52 — `OSHA\_29CFR\_full\_2024\_3959\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If equipment operations would breach the Table A minimum approach distance, but the employer determines it IS feasible to do the work without breaching that distance, is the operation permitted under the exception in this section?
- ✓ **Ground truth:** No; the exception requires the employer demonstrate it is infeasible to do the work without breaching the minimum approach distance (paragraph (a)), so if it is feasible, the operation closer than Table A is prohibited.
- ✗ **Model answered:** {"name": "is_operation_permitted", "parameters": {"feasible_to_do_work_without_breaching_distance": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 53 — `OSHA\_29CFR\_full\_2024\_4294\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under paragraph (d)(3), agency inspections for imminent danger conditions must be conducted within 24 hours; if instead the Secretary determines an inspection is warranted for such a report under paragraph (e)(3), what response time applies?
- ✓ **Ground truth:** The same response time of within 24 hours, because under (e)(3) the Secretary shall

observe the same response times as required of the agencies under the Executive Order and § 1960.28(d)(3), which sets 24 hours for imminent danger conditions.

- **✗ Model answered:** {"name": "get_response_time", "parameters": {"paragraph": "e", "subparagraph": "3", "condition": "imminent danger"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 54 — `OSHA\_29CFR\_full\_2024\_neg\_1599\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum load in amperes for a receptacle on a 50-ampere branch circuit?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** {"name": "get_maximum_load", "parameters": {"circuit_rating": "50"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated answer not in evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 55 — `OSHA\_29CFR\_full\_2024\_3312\_04` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a scaffold must be closer to power lines than the specified minimum distance because the clearance is necessary for performance of work, when is that permitted?
- **✓ Ground truth:** Only after the utility company or electrical system operator has been notified and has deenergized the lines, relocated them, or installed protective coverings to prevent accidental contact.
- **✗ Model answered:** {"name": "check_clearance_exception", "parameters": {"distance": "necessary for performance of work"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 56 — `OSHA\_29CFR\_full\_2024\_1593\_00` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** For doors and covers of enclosures used solely as pull boxes, splice boxes, or junction boxes, what is the exception to the requirement that they be locked, bolted, or screwed on?
- **✓ Ground truth:** Underground box covers that weigh over 45.4 kg (100 lb) are excepted from the requirement to be locked, bolted, or screwed on.
- **✗ Model answered:** {"name": "exception_to_locking_requirement", "parameters": {"weight": "100"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 57 — `OSHA\_29CFR\_full\_2024\_4456\_00` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** When a timely objection is filed, are all provisions of the preliminary order stayed?
- **Ground truth:** No. All provisions are stayed except for the portion requiring preliminary reinstatement, which is not automatically stayed and remains effective immediately upon the respondent's receipt of the findings and preliminary order.
- **Model answered:** {"name": "is_stayed", "parameters": {"provision": "preliminary order", "objection": "timely objection"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 58 — `OSHA\_29CFR\_full\_2024\_2920\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** According to paragraphs (viii) and (ix), what must be provided to a worker who reports a small skin abrasion, and who treats it?
- **Ground truth:** Per paragraph (viii), any small skin abrasions, cuts, rash, or open sores which are found or reported shall be treated by a properly designated person so that chances of exposures to the chemicals are removed. Per paragraph (ix), the worker must also have access to sufficient washing facilities, including soap, individual towels, and hot water, provided on the basis of at least one basin with a hot water faucet for every 10 employees.
- **Model answered:** {"name": "treatment_for_skin_abrasion", "parameters": {"treatment": "treated by a properly designated person"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not fully synthesize the rules from multiple paragraphs
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 59 — `OSHA\_29CFR\_full\_2024\_2784\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an OSHA regulated substance has no action level, what triggers application of paragraphs (d) and (g)(1)(ii)?
- **Ground truth:** Routinely exceeding the permissible exposure limit (in the absence of an action level), for a substance with exposure monitoring and medical surveillance requirements.
- **Model answered:** {"name": "apply_paragraphs", "parameters": {"action_level": "None", "exposure_limit": "permissible exposure limit"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states that when there is no action level, paragraphs (d) and

(g)(1)(ii) apply only where the permissible exposure limit is “routinely exceeded” for a substance with exposure monitoring and medical surveillance requirements, but the model merely named the permissible exposure limit and omitted the required routine exceedance trigger.

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 60 — `OSHA\_29CFR\_full\_2024\_1317\_01` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** How does the definition of 'running line' distinguish it from other ropes?
- ✓ **Ground truth:** A running line means any moving rope as distinguished from a stationary rope such as a guyline.
- ✗ **Model answered:** {"name": "get_definition", "parameters": {"term": "running line"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 61 — `OSHA\_29CFR\_full\_2024\_4010\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an auditor determines that the written and practical tests are valid and reliable but the auditor is an employee of the employer, can the auditor's approval satisfy the test development requirements?
- ✓ **Ground truth:** No, because paragraph (e)(1)(ii)(B) requires that the auditor is not an employee of the employer.
- ✗ **Model answered:** {"name": "auditor_approval", "parameters": {"valid_and_reliable": "True", "auditor_employed_by_employer": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 62 — `OSHA\_29CFR\_full\_2024\_0644\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Using paragraph (p)(1), which paragraph governs emergency planning and response, and which governs the development of the process hazard analysis?
- ✓ **Ground truth:** Emergency planning and response is governed by paragraph (n), and the development of the process hazard analysis is governed by paragraph (e).
- ✗ **Model answered:** {"name": "get_paragraphs", "parameters": {"paragraph_number": "p", "subparagraph_number": "1"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model read from wrong paragraph
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 63 — `OSHA\_29CFR\_full\_2024\_2444\_05` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the required accuracy to a 95 percent confidence level for the MDA monitoring method at the select PEL?
- ✓ **Ground truth:** plus or minus 25 percent
- ✗ **Model answered:** {"name": "get_required_accuracy", "parameters": {"confidence_level": "95", "pel": "select PEL"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 64 — `OSHA\_29CFR\_full\_2024\_3038\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an employer's permit-required contingency plan fully addresses certain subjects, must the emergency response plan duplicate those subjects?
- ✓ **Ground truth:** No, the plan need not duplicate them, provided the contingency plan is made part of the emergency response plan.
- ✗ **Model answered:** {"name": "answer", "parameters": {"question": "If an employer"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 65 — `OSHA\_29CFR\_full\_2024\_3428\_01` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** The headache ball, hook, or load shall not be used to transport personnel - what is the stated exception?
- ✓ **Ground truth:** Except as provided in paragraph (c)(4) of this section (using cranes or derricks to hoist employees on a personnel platform).
- ✗ **Model answered:** {"name": "exception_to_headache_ball_hook_or_load_usage", "parameters": {"exception": "when work under this subpart is being conducted, provided that all provisions of \u00a71926.1431 (except for \u00a71926.1431(a)) are met"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 66 — `OSHA\_29CFR\_full\_2024\_2476\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the definitions of 'emergency' and 'employee exposure,' if maintenance personnel control an incidental MC release per paragraph (f) while not wearing respiratory protection, is this an emergency, and would the airborne MC they would be exposed to count as employee exposure?
- **Ground truth:** It is not considered an emergency, because an incidental release controlled by employees such as maintenance personnel in accordance with the leak/spill provisions of paragraph (f) is excluded from the emergency definition. However, the airborne MC they would be exposed to (or would be exposed to absent respiratory protection) still constitutes employee exposure.
- **Model answered:** {"name": "is_emergency", "parameters": {"release_controlled": "True", "respiratory_protection_used": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 67 — `OSHA\_29CFR\_full\_2024\_3172\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** For a horizontal belt where both runs are 36 inches from the floor, what guarding is required, and how does this differ from the general rule for runs that are 7 feet or less from the floor?
- **Ground truth:** Under paragraph (e)(1)(i), where both runs are 42 inches or less from the floor, the belt shall be fully enclosed. Since 36 inches is 42 inches or less, the belt must be fully enclosed. This differs from the general rule for runs 7 feet or less from the floor, where the guard need only extend to at least 15 inches above the belt or to a standard height.
- **Model answered:** {"name": "guarding_required", "parameters": {"height": "36", "general_rule_height": "7"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 68 — `OSHA\_29CFR\_full\_2024\_1695\_04` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining paragraph (a)(5)(i) and (a)(5)(ii), what handling restrictions apply when plugging in an energized cord connector that is wet from being immersed in water?
- **Ground truth:** Employees' hands may not be wet when plugging and unplugging energized equipment (per (a)(5)(i)), and the energized plug and receptacle connection may be handled only with insulating protective equipment because a wet connection could provide a conducting path to the employee's hand (per (a)(5)(ii)).
- **Model answered:** {"name": "restriction", "parameters": {"condition": "wet from being immersed in water", "handling": "only with insulating protective equipment"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 69 — `OSHA\_29CFR\_full\_2024\_4618\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Conditional
- **Question:** If an observer wishes to watch monitoring that requires entry into an area where protective clothing is required, what three obligations fall on the employer?
- **Ground truth:** The employer must provide the observer with the required personal protective clothing or equipment, assure its use, and require the observer to comply with all other applicable safety and health procedures.
- **Model answered:** {"name": "observation_procedures", "parameters": {"i": "3"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model provided a fabricated number
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '3' not in evidence

Failure 70 — `OSHA\_29CFR\_full\_2024\_3390\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** For a precast concrete member that is 150 feet long, paragraph (ii) governs the maximum distance. How does this maximum compare to the general maximum given in paragraph (i)?
- **Ground truth:** For a 150-foot member, half the length is 75 feet, but the cap is 60 feet (whichever is less), so the maximum is 60 feet under paragraph (ii). This is greater than the 25-foot general maximum given in paragraph (i).
- **Model answered:** {"name": "compare_distances", "parameters": {"length": "150"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 71 — `OSHA\_29CFR\_full\_2024\_4138\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Both Option (2) corraling and Option (4) centerline cable system aim to prevent shifting, but they impose different standards on whether any movement is tolerated. What does each require regarding movement during operation?
- **Ground truth:** Option (2) corraling requires that the corraling system does not allow the equipment to shift by any amount of shifting in any direction. Option (4) centerline cable system requires that the crane/derrick is secured from movement during operation (per subpart (E)), though its attachment method must also allow the crew to move the crane/derrick longitudinally for repositioning.
- **Model answered:** {"name": "get_movement_requirement", "parameters": {"option": "2"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 72 — `OSHA\_29CFR\_full\_2024\_1593\_03` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** What is the exception that allows equipment operating at 600 volts or less to be installed within a high-voltage enclosure, room, or vault without separation?
- ✓ **Ground truth:** Switches or other equipment operating at 600 volts or less, and serving only equipment within the high-voltage vault, room, or enclosure, may be installed in the high-voltage enclosure if accessible to qualified persons only.
- ✗ **Model answered:** {"name": "allow_installation", "parameters": {"voltage": "600", "location": "high-voltage vault", "access": "qualified persons only"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 73 — `OSHA\_29CFR\_full\_2024\_0478\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Precision
- **Question:** For spraying operations other than electrostatic spraying, what is the minimum average air velocity required over the open face of the booth?
- ✓ **Ground truth:** 100 linear feet per minute
- ✗ **Model answered:** {"name": "get_minimum_average_air_velocity", "parameters": {"operation_type": "non-electrostatic spraying"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states the air velocity for spraying operations except electrostatic spraying must be "not less than 100 linear feet per minute," but the model only emitted a function call with the operation type and did not provide the required value.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 74 — `OSHA\_29CFR\_full\_2024\_2920\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Based on paragraph (v) and paragraph (vii), what protections are required for an employee who manually adds chemicals to a tank containing a liquid that may irritate the skin if splashed?
- ✓ **Ground truth:** Per paragraph (v), the employee must wear either tight-fitting chemical goggles or an effective face shield because of the danger of splashing during manual additions. Per paragraph (vii), there must also be a nearby supply of clean cold water (with a quick opening valve and at least 48 inches of hose not smaller than three-fourths inch), or alternatively deluge showers and eye flushes, so liquids can be quickly washed off the skin or clothing.
- ✗ **Model answered:** {"name": "get_protections", "parameters": {"action": "manually adds chemicals to a tank", "hazard": "liquid that may irritate the skin if splashed"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not fully synthesize rules from multiple regulatory sections
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 75 — `OSHA\_29CFR\_full\_2024\_4269\_00` · MEDIUM confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** According to the purpose and scope, public employees were not assured protections under the Act prior to a State plan because of a definition in one section, but a different section provides the route for ensuring their protection. Which sections are these, and what does each establish?
- **Ground truth:** Section 3(5) excludes State and local government agencies from the definition of 'employer,' which is why such public employees had not heretofore been assured any protections under the Act. Section 18 (and specifically section 18(b)) is the only provision under which such public employees are ensured protection through an approved State plan, permitting States to submit plans for any occupational safety and health issue for which a Federal standard has been promulgated under section 6.
- **Model answered:** {"name": "get_sections", "parameters": {"section1": "3(5)", "section2": "18"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports §3(5) as excluding State and local government agencies from the definition of employer and §18/§18(b) as the State-plan route for protection, but the model only named the sections and did not answer what each establishes.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 76 — `OSHA\_29CFR\_full\_2024\_1434\_02` · MEDIUM confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an employee performs work within reaching distance of exposed energized parts while wearing a wristwatch that does not increase the hazards associated with contact, must the employer ensure it is removed or rendered nonconductive?
- **Ground truth:** No; the requirement to remove or render nonconductive does not apply if such articles do not increase the hazards associated with contact with the energized parts.
- **Model answered:** {"name": "ensure_conductive_articles_removal", "parameters": {"increase_hazards": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 77 — `OSHA\_29CFR\_full\_2024\_1264\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the beater guard requirement in paragraph (d)(1) and the feed roll requirement in paragraph (d)(3), what is the protective purpose served by guards on opening and picking machinery in each case?

- ✓ **Ground truth:** The beater covers (d)(1) must prevent contact with the beater, while the feed roll guard (d)(3) must be designed to prevent the operator from reaching the nip while the machinery is in operation. Both serve to prevent operator contact with hazardous moving parts.
- ✗ **Model answered:** {"name": "protective_purpose", "parameters": {"guard_type": "beater guard", "protective_purpose": "prevent contact with the beater"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model stopped before applying dependent rule
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 78 — `OSHA\_29CFR\_full\_2024\_3423\_00` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the minimum factor by which jacking equipment must be capable of supporting the load being lifted during jacking operations?
- ✓ **Ground truth:** two and one-half times
- ✗ **Model answered:** {"name": "jacking_equipment_support_factor", "parameters": {"factor": "2"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states, "Jacking equipment shall be capable of supporting at least two and one-half times the load," so the required factor is 2.5, not the model's 2.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 79 — `OSHA\_29CFR\_full\_2024\_3038\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an employer arranges in advance for an outside fully-trained emergency response team to respond in a reasonable period, must all employees still be trained to the full degree specified?
- ✓ **Ground truth:** No, provided all employees who may come to the incident first have sufficient awareness training and have been instructed to call the designated outside team.
- ✗ **Model answered:** {"name": "train_all_employees", "parameters": {"arrangements_made": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 80 — `OSHA\_29CFR\_full\_2024\_0694\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Considering both the nonpotable water rules in (n)(2) and the potable water rules in (n)(1), what must be done to distinguish unsafe water outlets and how must the two water systems relate?
- ✓ **Ground truth:** Outlets for nonpotable water, such as firefighting water, must be identified to clearly indicate the water is unsafe and not to be used for drinking, washing, or cooking; and there must be no cross-connection, open or potential, between a system furnishing potable water and a system furnishing nonpotable water.

- **✗ Model answered:** {"name": "identify_water_outlets", "parameters": {"potable_water": "false", "nonpotable_water": "true"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not fully address the question
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 81 — `OSHA\_29CFR\_full\_2024\_2138\_05` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** Under paragraph (l)(3)(iii), if both the initial and the medical-exam monitoring show CdU exceeding 15 µg/g Cr and CdU also exceeding 3 µg/g Cr, what must the physician do?
- **✓ Ground truth:** The physician shall medically remove the employee from exposure to cadmium at or above the action level.
- **✗ Model answered:** {"name": "physician_determination", "parameters": {"CdU": "15", "CdB": "5"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 82 — `OSHA\_29CFR\_full\_2024\_2155\_08` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the 'high level' threshold for beta-2-microglobulin associated with greater chance of developing kidney diseases?
- **✓ Ground truth:** greater than 1000 micrograms per gram of creatinine (µg/g Cr)
- **✗ Model answered:** {"name": "get_high_level_threshold", "parameters": {"marker": "beta-2-microglobulin"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed incorrect threshold
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 83 — `OSHA\_29CFR\_full\_2024\_3312\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If conductive material is being handled on a scaffold near an uninsulated power line of less than 50 kv, what is the minimum required clearance?
- **✓ Ground truth:** 10 feet (3.1 m).
- **✗ Model answered:** {"name": "get_minimum_clearance", "parameters": {"voltage": "less than 50 kv", "line_type": "uninsulated"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 84 — `OSHA\_29CFR\_full\_2024\_2437\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an examining physician determines that an employee's abnormal liver function tests are NOT associated with MDA exposure, can the employee still be temporarily removed?
- **Ground truth:** Yes, if the physician also determines the abnormalities may be exacerbated as a result of occupational exposure to MDA.
- **Model answered:** {"name": "temporary_removal_due_to_final_medical_determination", "parameters": {"reason": "abnormal liver function tests not associated with MDA exposure"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 85 — `OSHA\_29CFR\_full\_2024\_1096\_00` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the maximum opening that work rests shall be kept adjusted to from the wheel on offhand grinding machines?
- **Ground truth:** one-eighth inch
- **Model answered:** {"name": "get_maximum_opening", "parameters": {"machine_type": "offhand grinding machines"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed incorrect value
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 86 — `OSHA\_29CFR\_full\_2024\_0767\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** Based on the test duration rules in paragraph (b), if a fit test protocol consists of four standard exercises plus the grimace exercise, what is the total time spent performing exercises?
- **Ground truth:** Four standard exercises at one minute each equals 4 minutes (240 seconds), plus the grimace exercise at 15 seconds, for a total of 4 minutes and 15 seconds (255 seconds).
- **Model answered:** {"name": "calculate_total_time", "parameters": {"num_standard_exercises": "4", "grimace_duration": "15", "standard_duration": "60"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '60' not in evidence

Failure 87 — `OSHA\_29CFR\_full\_2024\_0333\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** Under the system utilizing angulated roping and building face rollers, is it acceptable where the suspended equipment is 140 feet above ground level?
- ✓ **Ground truth:** No; the system is acceptable only where the suspended portion does not exceed 130 feet above a safe surface or ground level.
- ✗ **Model answered:** {"name": "is_acceptable", "parameters": {"system": "angulated roping and building face rollers", "height": "140"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 88 — `OSHA\_29CFR\_full\_2024\_3020\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining subpart (vi) and subpart (ix), what must happen before a drum that cannot be moved without rupture is emptied?
- ✓ **Ground truth:** Before any movement of drums or containers, all employees exposed to the transfer operation must be warned of the potential hazards associated with the contents (vi); and a drum that cannot be moved without rupture, leakage, or spillage must be emptied into a sound container using a device classified for the material being transferred (ix).
- ✗ **Model answered:** {"name": "warn_employees", "parameters": {"drum_contents": "hazardous substance"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 89 — `OSHA\_29CFR\_full\_2024\_neg\_2625\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the reference number or citation of the footnote marked '1' regarding appropriate animal tests?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_footnote_reference", "parameters": {"footnote_number": "1"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated answer not in evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 90 — `OSHA\_29CFR\_full\_2024\_4301\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If some employees of an establishment are covered under collective bargaining arrangements and others are not, who must the nonmanagement members represent?
- ✓ **Ground truth:** Members shall be representative of both groups.
- ✗ **Model answered:** {"name": "get_nonmanagement_members_representation", "parameters": {"collective_bargaining": "some", "other_employees": "not"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 91 — `OSHA\_29CFR\_full\_2024\_1848\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Using the recordkeeping requirements in Section V.C and the latency information in Section II.A, how long must medical records be kept, and how does this relate to when disease symptoms typically appear?
- ✓ **Ground truth:** Medical records must be kept for the period of employment plus thirty (30) years (V.C), which accommodates the fact that symptoms of asbestos-related diseases generally do not appear for 20 or more years after initial exposure (II.A).
- ✗ **Model answered:** {"name": "calculate_recordkeeping_requirements", "parameters": {"medical_records": "true", "latency": "20"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 92 — `OSHA\_29CFR\_full\_2024\_0694\_01` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the toilet facility table in (n)(3)(i), how many toilet seats and urinals are required for a worksite with 160 employees versus one with 250 employees?
- ✓ **Ground truth:** For 160 employees (more than 20, fewer than 200): one toilet seat and one urinal per 40 employees = 4 toilet seats and 4 urinals. For 250 employees (more than 200): one toilet seat and one urinal per 50 employees = 5 toilet seats and 5 urinals.
- ✗ **Model answered:** [{"name": "calculate_toilet_facilities", "parameters": {"employees": "160"}}; {"name": "calculate_toilet_facilities", "parameters": {"employees": "250"}}]
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states 160 employees use the 'one toilet seat and one urinal per 40 employees' rule and 250 uses 'per 50 employees,' yielding 4 and 5 respectively, while the model only emitted tool-call placeholders without the required numbers.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 93 — `OSHA\_29CFR\_full\_2024\_4502\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under paragraph (a) materials are redacted consistent with the Privacy Act, yet paragraph (d) addresses confidentiality of certain persons. How do these provisions treat the complainant differently regarding confidentiality protection?
- **Ground truth:** Paragraph (a) redactions are made consistent with the Privacy Act of 1974 and other confidentiality laws, but the complainant (or counsel) receives an unredacted copy of those materials. Paragraph (d) requires investigations to protect the confidentiality of any person who provides information on a confidential basis other than the complainant-so the complainant is expressly excluded from that confidential-informant protection.
- **Model answered:** {"name": "compare_provisions", "parameters": {"provision1": "paragraph (a)", "provision2": "paragraph (d)", "regarding": "confidentiality protection"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 94 — `OSHA\_29CFR\_full\_2024\_4031\_00` · HIGH confidence

CALCULATION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Calculation
- **Question:** A crane has a rated capacity of 60,000 pounds for the current radius and configuration. When using a suspended personnel platform, what is the maximum total load (platform loaded, including hook, load line and rigging) permitted, except during proof testing?
- **Ground truth:** 30,000 pounds
- **Model answered:** {"name": "calculate_max_total_load", "parameters": {"rated_capacity": "60000"}}
- **Doctor diagnosis:** **Dr. Pi** model error
- **Doctor root cause:** [verifier] The evidence says the total load “must not exceed 50 percent of the rated capacity,” so 50% of 60,000 is 30,000 pounds; the model only returned a function-call-like JSON with rated_capacity 60000 and did not provide the required maximum load.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 95 — `OSHA\_29CFR\_full\_2024\_3703\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Comparing Structure No. 10 and Structure No. 11, both have a count of 0 - what distinct reasons are given for each?
- **Ground truth:** Structure No. 10 has a count of 0 because the fiber is outside the graticule, while Structure No. 11 has a count of 0 because the fiber crosses the graticule twice.
- **Model answered:** {"name": "get_structure_explanation", "parameters": {"structure_no": "10"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 96 — `OSHA\_29CFR\_full\_2024\_2138\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** Under paragraph (l)(3)(ii)(C), if the physician determines that medical removal is not necessary after a full medical examination, what must the employer provide until levels fall to or below the thresholds?
- **✓ Ground truth:** The employer must provide biological monitoring semiannually and annual medical examinations until CdU falls to ≤ 3 $\mu\text{g/g}$ Cr, $\beta 2\text{-M}$ to ≤ 300 $\mu\text{g/g}$ Cr, and CdB to ≤ 5 $\mu\text{g/lwb}$.
- **✗ Model answered:** {"name": "provide_until_levels_fall", "parameters": {"CdU": "3", " $\beta 2\text{-M}$ ": "300", "CdB": "5"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence says the employer must “provide biological monitoring ... on a semiannual basis” and “provide annual medical examinations,” but the model only listed the threshold values and omitted the required provisions.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 97 — `OSHA\_29CFR\_full\_2024\_0333\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** For an intermittent stabilization system anchor, what is the minimum design wind load per anchor if two anchors share the wind load?
- **✓ Ground truth:** 300 pounds (1334 n).
- **✗ Model answered:** {"name": "calculate_design_wind_load", "parameters": {"anchors": "2"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 98 — `OSHA\_29CFR\_full\_2024\_2164\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Comparing the control-limit criterion for CDB/CDU calibration standards with the control-limit criterion for B2MU control samples, what is the measurement error threshold in each case that signals the run has exceeded control limits?
- **✓ Ground truth:** For CDB/CDU calibration, a standard more than $\pm 15\%$ of its theoretical value means the run has exceeded control limits; for B2MU, if the measurement error of one or more control samples is more than 15%, the run exceeds control limits. Both use a 15% threshold.
- **✗ Model answered:** {"name": "control_limit_criterion", "parameters": {"criterion": "CDB/CDU calibration standards"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 99 — `OSHA\_29CFR\_full\_2024\_3208\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If electric equipment is located where it would be exposed to physical damage, what requirement applies to the enclosures or guards?
- ✓ **Ground truth:** They must be arranged and of such strength as to prevent such damage.
- ✗ **Model answered:** {"name": "enclosures_or_guards_requirement", "parameters": {"location": "exposed to physical damage"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states that in such locations, "enclosures or guards shall be so arranged and of such strength as to prevent such damage," while the model only identified the location/requirement label and did not provide the required arrangement or strength.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 100 — `OSHA\_29CFR\_full\_2024\_1084\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the guard sizing rule and the prohibition in paragraph (m)(1), what minimum guard diameter applies to a circular shaper, and what type of device is never acceptable as a substitute?
- ✓ **Ground truth:** The diameter of circular shaper guards shall be not less than the greatest diameter of the cutter, and in no case is a warning device of leather or other material attached to the spindle acceptable.
- ✗ **Model answered:** {"name": "get_minimum_guard_diameter", "parameters": {"diameter": "greatest diameter of the cutter"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence answers both parts—"The diameter of circular shaper guards shall be not less than the greatest diameter of the cutter" and "In no case shall a warning device of leather or other material attached to the spindle be acceptable"—but the model only provided the guard diameter and omitted the prohibited device.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 101 — `OSHA\_29CFR\_full\_2024\_0517\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the equipment requirements in subparagraph (g)(2)(iii) and the composition provisions in subparagraph (g)(2)(iv), what must mixing and packaging equipment be constructed of, and which substances are prohibited from use?
- ✓ **Ground truth:** Mixing and packaging equipment shall be constructed of materials compatible with the fuel-ammonium nitrate composition. Prohibited substances include hydrocarbon liquid fuel with flashpoint lower than No. 2 diesel fuel oil (125 °F minimum), crude oil and crankcase oil, and peroxides and chlorates.
- ✗ **Model answered:** {"name": "get_equipment_requirements", "parameters": {"subparagraph": "(g)(2)(iii)", "composition_provisions": "(g)(2)(iv)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 102 — `OSHA\_29CFR\_full\_2024\_1332\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the log hauls and slips control requirement with the log decks barricade requirement, what positioning protections are required for operators/sawyers relative to the logs?
- ✓ **Ground truth:** Per paragraph (vi)(e), log haul controls shall be located so the operator is at all times in the clear of logs, machinery, lines, and rigging (and where control is by lever exposed to incoming logs, the lever shall operate the haul only when moved toward the log slip or log pond); per paragraph (3)(iii), a barricade or other positive stop of sufficient strength to stop any log shall be erected between the sawyer's stand and the log deck.
- ✗ **Model answered:** {"name": "log_hauls_and_slips_control", "parameters": {"requirement": "operators/sawyers must be positioned where they will, at all times, be in the clear of logs, machinery, lines, and rigging"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 103 — `OSHA\_29CFR\_full\_2024\_1288\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a rope washer is designed so that water cannot splash the operator, floor, or working surface, is a splash guard still required?
- ✓ **Ground truth:** No, splash guards are required on rope washers unless the machine is designed to prevent splashing.
- ✗ **Model answered:** {"name": "check_splash_guard_requirement", "parameters": {"machine_type": "rope washer", "splash_prevention_design": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 104 — `OSHA\_29CFR\_full\_2024\_4551\_01` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** Under § 1990.106(a)(1), within how many days must the institutes' answer to the Secretary's review request be provided?
- ✓ **Ground truth:** one hundred twenty (120) days
- ✗ **Model answered:** {"name": "get_answer", "parameters": {"section": "1990.106(a)(1)", "question": "within how many days must the institutes"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not compute the time period

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 105 — `OSHA\_29CFR\_full\_2024\_0485\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** In an unsprinklered building where sprinklers are installed only to protect spraying areas, must the installation conform to the requirements of § 1910.159?
- ✓ **Ground truth:** Yes, the installation shall conform to those standards insofar as they are applicable.
- ✗ **Model answered:** {"name": "conformance_requirement", "parameters": {"building_type": "unsprinklered", "sprinkler_installation": "only to protect spraying areas"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly answers yes—“the installation shall conform to such standards insofar as they are applicable”—but the model only returned a structured label/parameters and did not state the required yes/insofar-applicable answer.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 106 — `OSHA\_29CFR\_full\_2024\_neg\_4455\_00` · MEDIUM confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the filing fee for the complainant to request a hearing after receiving the findings?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "None", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated answer not in evidence
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 107 — `OSHA\_29CFR\_full\_2024\_3020\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Cross-Reference
- **Question:** Per subpart (viii), where is the spill containment program required to originate, and what does it reference in another part of the section?
- ✓ **Ground truth:** The spill containment program is part of the employer's safety and health program required in paragraph (b) of this section, and it must be implemented to contain and isolate the entire volume of the hazardous substance being transferred where major spills may occur.
- ✗ **Model answered:** {"name": "answer_origin_reference", "parameters": {"origin": "the employer", "reference": "paragraph (b) of this section"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] Evidence says the spill containment program is “part of the employer's safety and health program required in paragraph (b),” so answering its origin as merely “the employer” omits the operative program and is not equivalent to the ground truth.
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** (none)

Failure 108 — `OSHA\_29CFR\_full\_2024\_4215\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Whose interests were served before the revised benchmarks were approved, and what was the same party's role in the original requirement for benchmarks?
- **Ground truth:** The AFL-CIO was served with notice before approval of the revised benchmarks, and the AFL-CIO was the plaintiff in the 1978 Court Order (AFL-CIO v. Marshall) that originally required the compliance staffing benchmarks.
- **Model answered:** {"name": "get_party_info", "parameters": {"party_name": "AFL-CIO", "role": "service"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports that there was “service on the AFL-CIO” before approval and that the original requirement came from the 1978 Court Order in “AFL-CIO v. Marshall,” making AFL-CIO the first-named party/plaintiff; the model identified AFL-CIO but gave only “service” as its role and omitted/misstated the court-case role.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 109 — `OSHA\_29CFR\_full\_2024\_4188\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under the screw conveyor guarding rule, if a screw conveyor is located under a gin stand and the gin is about to be started, what guarding requirement applies and what additional action must occur per the warning device provision?
- **Ground truth:** A screw conveyor under a gin stand is considered guarded by location (so no additional cover or grating guard is required), but per the warning device provision, an audible signal of sufficient volume to be heard by employees must be sounded each time before starting the gin.
- **Model answered:** {"name": "guarding_requirement", "parameters": {"location": "under a gin stand", "action": "starting the gin"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states that “Screw conveyors under gin stands shall be considered guarded by location” and that an audible signal “shall be sounded each time before starting the gin,” but the model only restated the location/action parameters and did not provide either required rule application.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 110 — `OSHA\_29CFR\_full\_2024\_3959\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If the work is covered by subpart V of this part, does the minimum clearance distance determination under paragraph (c)(1) apply?
- **Ground truth:** No; for subpart V work, paragraph (c)(1) does not apply, and instead the minimum approach distances established under § 1926.960(c)(1)(i) apply.

- **✗ Model answered:** {"name": "apply_minimum_clearance_distance_determination", "parameters": {"covered_by_subpart_V": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 111 — `OSHA\_29CFR\_full\_2024\_4502\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under paragraph (a), what specific rights must OSHA notify the respondent of when notifying it of the filing of the complaint, and what are the two response timeframes associated with those rights set out in paragraphs (b) and (f)?
- **✓ Ground truth:** OSHA must notify the respondent of its rights under paragraphs (b) and (f) of the section and § 1987.110(e). Under paragraph (b), the respondent may, within 20 days of receipt of the notice, submit a written statement and affidavits/documents and request a meeting. Under paragraph (f), the respondent must present its evidence within 10 business days of OSHA's notification (or as soon thereafter as OSHA and the respondent agree).
- **✗ Model answered:** {"name": "get_notification_rights", "parameters": {"rights": "paragraph (b) and (f)", "response_timeframes": "[20 days, 10 business days]"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 112 — `OSHA\_29CFR\_full\_2024\_1349\_01` · HIGH confidence

OVERCONFIDENT

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Cross-Reference
- **Question:** Combining the rules on how the chain saw must be started and how it must be held during operation, what is required regarding the operator's hands?
- **✓ Ground truth:** When starting, drop starting is prohibited and the saw must be started on the ground or where otherwise firmly supported (per (vi)); during operation, the chain saw shall be held with the thumbs and fingers of both hands encircling the handles unless the employer demonstrates a greater hazard is posed by keeping both hands on the saw (per (viii)).
- **✗ Model answered:** {"name": "operator_hands_requirement", "parameters": {"hands_position": "encircling the handles", "specific_situation_exception": "the employer demonstrates that a greater hazard is posed by keeping both hands on the chain saw"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The ground truth is supported by evidence because (vi) says the saw must be started “on the ground or where otherwise firmly supported” and drop starting is prohibited, while (viii) requires “thumbs and fingers of both hands encircling the handles,” but the model omitted the starting requirement.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Uncertainty violation: evidence hedges but response uses 'demonstrates that'

Failure 113 — `OSHA\_29CFR\_full\_2024\_0494\_02` · HIGH confidence

MISINTERPRETED

- **Severity:** MEDIUM — a misapplied or over-extrapolated rule — plausible but wrong
- **Category:** Conditional
- **Question:** If an undercoating operation uses materials not more hazardous than kerosene but is conducted in a garage lacking adequate natural or mechanical ventilation, does the exemption apply?
- ✓ **Ground truth:** No, because adequate natural or mechanical ventilation is required for the exemption, so the operation is subject to all spray finishing requirements.
- ✗ **Model answered:** {"name": "is_exempt", "parameters": {"is_hazardous": "False", "has_ventilation": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model misinterpreted exemption conditions
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 114 — `OSHA\_29CFR\_full\_2024\_0882\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the durability requirement in paragraph (c)(5)(ii)(A)(3) and the training limitation in paragraph (c)(7)(ii)(D), what is required regarding tags used in corrosive environments such as areas where acid and alkali chemicals are handled?
- ✓ **Ground truth:** Tags shall not deteriorate when used in corrosive environments such as areas where acid and alkali chemicals are handled and stored, and employees shall be trained that tags and their means of attachment must be made of materials which will withstand the environmental conditions encountered in the workplace.
- ✗ **Model answered:** {"name": "get_tag_requirements", "parameters": {"environment": "corrosive", "chemicals": "acid and alkali"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 115 — `OSHA\_29CFR\_full\_2024\_4301\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** At an establishment where employees are not represented under collective bargaining arrangements, how are nonmanagement members determined?
- ✓ **Ground truth:** They shall be determined through procedures devised by the agency which provide for effective representation of all employees.
- ✗ **Model answered:** {"name": "determine_nonmanagement_members", "parameters": {"representation_type": "not represented under collective bargaining arrangements"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 116 — `OSHA\_29CFR\_full\_2024\_1748\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the requirements of paragraph (g)(2) and paragraph (g)(3), what must the employer do regarding the room both during and after an aerosol-generating procedure on a person with suspected or confirmed COVID-19?
- ✓ **Ground truth:** During the procedure, per (g)(2), the employer must ensure it is performed in an existing AIIR if available; after the procedure is completed, per (g)(3), the employer must clean and disinfect the surfaces and equipment in the room or area where the procedure was performed.
- ✗ **Model answered:** {"name": "answer_question", "parameters": {"question": "Combining the requirements of paragraph (g)(2) and paragraph (g)(3), what must the employer do regarding the room both during and after an aerosol-generating procedure on a person with suspected or confirmed COVID-19?"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 117 — `OSHA\_29CFR\_full\_2024\_3092\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Per paragraph (a)(1), how is the container requirement different for highly viscous materials versus ordinary flammable liquids of 5 gallons or less, and per paragraph (a)(2), what storage areas are prohibited for both?
- ✓ **Ground truth:** Highly viscous (extremely hard to pour) materials may be used and handled in original shipping containers rather than approved safety cans or DOT approved containers required for ordinary flammable liquids of 5 gallons or less; under (a)(2) both are prohibited from storage in areas used for exits, stairways, or normally used for the safe passage of people.
- ✗ **Model answered:** {"name": "container_requirement_difference", "parameters": {"material_type": "highly viscous", "quantity": "5 gallons or less"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports the GT—ordinary flammable liquids of 5 gallons or less require approved safety cans or DOT-approved containers, highly viscous materials may be used/handled in original shipping containers, and both may not be stored in exits, stairways, or safe-passage areas—but the model only returned a partial function-style label/parameters and omitted the required comparison and prohibited storage areas.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 118 — `OSHA\_29CFR\_full\_2024\_3020\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** According to subpart (v) and subpart (vi), what organizational and warning steps apply to drum movement operations?
- ✓ **Ground truth:** Site operations must be organized to minimize the amount of drum or container movement (v), and prior to any movement of drums or containers, all exposed employees must be warned of the

potential hazards associated with the contents (vi).

- **✗ Model answered:** {"name": "organize_and_warn", "parameters": {"step": "drum movement operations"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 119 — `OSHA\_29CFR\_full\_2024\_0494\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — a misapplied or over-extrapolated rule — plausible but wrong
- **Category:** Conditional
- **Question:** If an automobile undercoating spray operation is conducted in a garage with adequate ventilation but uses materials more hazardous than kerosene with solvents having a flash point of 90 °F, is it exempt from spray finishing requirements?
- **✓ Ground truth:** No, it does not conform to the exemption provisions and is therefore subject to all spray finishing requirements of this section.
- **✗ Model answered:** {"name": "is_exempt_from_spray_finishing_requirements", "parameters": {"has_adequate_ventilation": "true", "is_kerosene_or_safer": "false", "flash_point": "90"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model misapplied exemption criteria
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 120 — `OSHA\_29CFR\_full\_2024\_neg\_4485\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many days does a third-party commercial carrier have to deliver a complaint for it to count as timely?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** {"name": "get_days_for_timely_filing", "parameters": {"delivery_method": "third-party commercial carrier"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated answer not in evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 121 — `OSHA\_29CFR\_full\_2024\_4294\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Per paragraph (d)(4), within how many days must an agency notify an employee that no hazard is found, and within how many days must a completed health-violation inspection report be made available to that employee?
- **✓ Ground truth:** The agency must notify within 15 days if it finds no reasonable grounds and won't inspect; an inspection/investigation report must be made available within 15 days for safety violations or within 30 days for health violations.

- **✗ Model answered:** {"name": "answer_question", "parameters": {"paragraph": "d", "subparagraph": "4"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 122 — `OSHA\_29CFR\_full\_2024\_0881\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an energy isolating device has neither a hasp nor a built-in locking mechanism, when is it still considered capable of being locked out?
- **✓ Ground truth:** When lockout can be achieved without the need to dismantle, rebuild, or replace the device or permanently alter its energy control capability.
- **✗ Model answered:** {"name": "is_capable_of_being_locked_out", "parameters": {"has_hasps": "false", "has_locking_mechanism": "false", "lockout_without_dismantling": "true"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 123 — `OSHA\_29CFR\_full\_2024\_4069\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a tower crane manufacturer does not specify the maximum wind speed for operations, what wind speed limit applies?
- **✓ Ground truth:** The wind must not exceed the speed determined by a qualified person.
- **✗ Model answered:** {"name": "get_wind_speed_limit", "parameters": {"manufacturer_specified": "false"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed default instead of quoting stated figure
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 124 — `OSHA\_29CFR\_full\_2024\_3208\_01` · MEDIUM confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If electric equipment operates at exactly 50 volts, must its live parts be guarded against accidental contact?
- **✓ Ground truth:** Yes, live parts operating at 50 volts or more must be guarded against accidental contact.
- **✗ Model answered:** {"name": "guarding_of_live_parts", "parameters": {"voltage": "50"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states that live parts operating at “50 volts or more shall be guarded,” so the answer is yes, but the model only returned a function-call-like JSON label with voltage 50 and did not actually answer the yes/no obligation.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 125 — `OSHA\_29CFR\_full\_2024\_0075\_04` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an employer fails to comply with the posting requirements of paragraphs (a) and (b), what consequence applies?
- ✓ **Ground truth:** The employer shall be subject to citation and penalty in accordance with § 1903.15(d).
- ✗ **Model answered:** {"name": "get_consequence", "parameters": {"action": "failing to comply with the posting requirements of paragraphs (a) and (b)"}}
- **Doctor diagnosis:** Dr. Default model error
- **Doctor root cause:** [verifier] The evidence directly states that an employer failing to comply with paragraphs (a) and (b) 'shall be subject to citation and penalty in accordance with § 1903.15(d),' while the model only emitted a tool-call-like request and did not provide that consequence.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 126 — `OSHA\_29CFR\_full\_2024\_2155\_05` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** According to appendix A Table B, within how many days must a medical examination be administered for an employee in category B?
- ✓ **Ground truth:** 90 days
- ✗ **Model answered:** {"name": "get_required_actions", "parameters": {"category": "B"}}
- **Doctor diagnosis:** Dr. Default model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 127 — `OSHA\_29CFR\_full\_2024\_3801\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under paragraph (h)(2)(iii), an employee with reasonably expected exposure above the TWA PEL or STEL may not remove required personal protective clothing from the workplace unless it has been cleaned per paragraph (h)(3)(ii). What method of removing beryllium does (h)(3)(ii) prohibit?
- ✓ **Ground truth:** Removing beryllium by blowing, shaking, or any other means that disperses beryllium into the air is prohibited; the clothing cannot be removed from the workplace unless it has been cleaned without such air-dispersing methods.
- ✗ **Model answered:** {"name": "answer_question", "parameters": {"paragraph": "(h)(3)(ii)", "question": "What method of removing beryllium does (h)(3)(ii) prohibit?"}}
- **Doctor diagnosis:** Dr. Default model error
- **Doctor root cause:** [verifier] The evidence directly states in (h)(3)(ii) that beryllium must not be removed from required PPE "by blowing, shaking, or any other means that disperses beryllium into the air," while the model merely restated the question and provided no answer.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 128 — `OSHA\_29CFR\_full\_2024\_3312\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a scaffold is not a mobile scaffold and was not designed by a registered professional engineer for movement, may it be moved horizontally while employees are on it?
- ✓ **Ground truth:** No, it shall not be moved horizontally while employees are on it.
- ✗ **Model answered:** {"name": "is_scaffold_movement_allowed", "parameters": {"is_mobile_scaffold": "false", "is_designed_for_movement": "false"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 129 — `OSHA\_29CFR\_full\_2024\_3130\_07` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the minimum liquid capacity (in U.S. gallons) for a container to be considered a portable tank?
- ✓ **Ground truth:** more than 60 U.S. gallons
- ✗ **Model answered:** {"name": "get_minimum_liquid_capacity", "parameters": {"capacity": "60"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 130 — `OSHA\_29CFR\_full\_2024\_1084\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under paragraph (m), for a double-spindle wood shaper, what guarding and control requirements must both be met regarding the cutting heads and the spindle controls?
- ✓ **Ground truth:** The cutting heads must be enclosed with a cage or adjustable guard designed to keep the operator's hand away from the cutting edge (with circular shaper guards no less than the greatest diameter of the cutter), as required by (m)(1); and, per (m)(3), the double-spindle shaper must be provided with a spindle starting and stopping device for each spindle.
- ✗ **Model answered:** {"name": "answer_question", "parameters": {"paragraph": "(m)", "machine_type": "double-spindle wood shaper"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 131 — `OSHA\_29CFR\_full\_2024\_0040\_00` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** Under § 1902.35, for at least how long after publication of the certification in the Federal Register will evaluation for the purposes of an 18(e) determination continue?
- ✓ **Ground truth:** at least one year
- ✗ **Model answered:** {"name": "get_evaluation_period", "parameters": {"section": "1902.35", "certification_event": "publication of the certification in the Federal Register"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not fully understand the evaluation period
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 132 — `OSHA\_29CFR\_full\_2024\_4047\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If employees are occupying a personnel platform but the platform is being used over water, are they required to use a personal fall arrest system attached to a structural member within the platform?
- ✓ **Ground truth:** No; the personal fall arrest system requirement applies except over water, and when working over or near water the requirements of § 1926.106 apply instead.
- ✗ **Model answered:** {"name": "fall_protection", "parameters": {"over_water": "true"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 133 — `OSHA\_29CFR\_full\_2024\_4269\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Paragraph (a) states the purpose of adapting private-employee plan requirements to public employees only; paragraph (b) notes considerations in doing so. What two adjustments does paragraph (b) identify as flowing from differences between public and private employment?
- ✓ **Ground truth:** Paragraph (b) identifies that (1) a system of monetary penalties applicable to violations may not always be the most appropriate method of achieving compliance for public employers, and (2) the lack of Federal enforcement authority over public employers requires certain adjustments of private employer plan procedures when adapting them to plans covering only public employees.
- ✗ **Model answered:** {"name": "identify_adjustments", "parameters": {"paragraph": "b", "differences": "public and private employment"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model stopped before applying dependent rule
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 134 — `OSHA\_29CFR\_full\_2024\_0881\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an affected employee's duties are expanded to include performing servicing or maintenance covered under this section, what is that employee's status?
- ✓ **Ground truth:** That affected employee becomes an authorized employee.
- ✗ **Model answered:** {"name": "employee_status", "parameters": {"duties": "performing servicing or maintenance covered under this section"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 135 — `OSHA\_29CFR\_full\_2024\_0694\_03` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Cross-Reference
- **Question:** Given the general toilet requirement in (n)(3)(i) and the exception in (n)(3)(iv), would a mobile crew of 30 employees be required to provide toilet facilities on site if they have transportation readily available to nearby toilets?
- ✓ **Ground truth:** No. Although 30 employees would normally require one toilet seat and one urinal per 40 employees under the table, the requirements for sanitation facilities do not apply to mobile crews having transportation readily available to nearby toilet facilities.
- ✗ **Model answered:** {"name": "toilet_facilities_required", "parameters": {"number_of_employees": "30", "transportation_available": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports the ground truth because paragraph (n)(3)(iv) states the sanitation-facility requirements 'shall not apply to mobile crews having transportation readily available to nearby toilet facilities,' but the model only returned a function-like JSON with inputs and no 'No' conclusion.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 136 — `OSHA\_29CFR\_full\_2024\_4205\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Given the dates of initial and final approval in paragraphs (a) and (b), how many years elapsed between the Tennessee State plan's initial approval and its final approval?
- ✓ **Ground truth:** Initial approval was July 5, 1973, and final approval was July 22, 1985, so approximately 12 years elapsed.
- ✗ **Model answered:** {"name": "years_between", "parameters": {"start_date": "July 5, 1973", "end_date": "July 22, 1985"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports about 12 years elapsed between July 5, 1973 and July 22, 1985, while the model only returned the two dates as parameters and did not provide the elapsed-years answer.
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** (none)

Failure 137 — `OSHA\_29CFR\_full\_2024\_4313\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining paragraphs (a) and (b) of § 1960.59, what type of training does paragraph (a) provide to employees generally that paragraph (b) does not explicitly require for employee representatives?
- ✓ **Ground truth:** Paragraph (a) requires that general employee training inform employees of the agency occupational safety and health program, with emphasis on their rights and responsibilities. Paragraph (b), which addresses employee representatives, focuses instead on introductory and specialized courses to enable them to ensure safe working conditions and assist in conducting workplace safety and health inspections, and does not explicitly include the rights-and-responsibilities program emphasis stated in paragraph (a).
- ✗ **Model answered:** {"name": "training_type", "parameters": {"type": "general employee training", "paragraph": "paragraph (a)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports the GT: paragraph (a) says training shall inform employees of the agency occupational safety and health program with emphasis on rights and responsibilities, while paragraph (b) does not; the model's answer merely says "general employee training" and omits the distinguishing training content asked for.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 138 — `OSHA\_29CFR\_full\_2024\_3390\_01` · HIGH confidence

TABLE MISREAD

- **Severity:** MEDIUM — a misapplied or over-extrapolated rule — plausible but wrong
- **Category:** Cross-Reference
- **Question:** Combining paragraph (i) and paragraph (ii), what is the minimum distance a control line must be erected from the leading edge, and does it differ between general work and erecting precast concrete members?
- ✓ **Ground truth:** The minimum distance is 6 feet (1.8 m) in both cases. Paragraph (i) requires not less than 6 feet for general use, and paragraph (ii) also requires not less than 6 feet when erecting precast concrete members, so the minimum does not differ.
- ✗ **Model answered:** {"name": "get_distance", "parameters": {"paragraph": "i"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model read from wrong table row
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 139 — `OSHA\_29CFR\_full\_2024\_0003\_03` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What Statute at Large citation is given for the Intergovernmental Personnel Act of 1970 in § 1902.3(h)?
- ✓ **Ground truth:** 84 Stat. 1915
- ✗ **Model answered:** {"name": "get_statute_at_large_citation", "parameters": {"section": "1902.3(h)", "act":

"Intergovernmental Personnel Act of 1970"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 140 — `OSHA\_29CFR\_full\_2024\_4398\_03` · HIGH confidence

CAUSE-EFFECT ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cause & Effect
- **Question:** What is the effect of OSHA's approval of a settlement reached by the respondent and the complainant?
- ✓ **Ground truth:** It demonstrates OSHA's consent and achieves the consent of all three parties.
- ✗ **Model answered:** {"name": "settlement_effect", "parameters": {"settlement_approval": "OSHA"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states that "OSHA's approval of a settlement reached by the respondent and the complainant demonstrates OSHA's consent and achieves the consent of all three parties," while the model merely returned that the settlement approval was OSHA, not the effect asked for.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 141 — `OSHA\_29CFR\_full\_2024\_0742\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Cross-Reference
- **Question:** Under the provisions governing respirator repairs, if a technician needs to adjust a regulator or alarm, what two conditions must be satisfied regarding who performs the work and what parts are used?
- ✓ **Ground truth:** Per subpart (iii), reducing and admission valves, regulators, and alarms shall be adjusted or repaired only by the manufacturer or a technician trained by the manufacturer; and per subpart (i), only the respirator manufacturer's NIOSH-approved parts designed for the respirator may be used.
- ✗ **Model answered:** {"name": "get_repair_conditions", "parameters": {"component": "regulator or alarm"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports the ground truth—regulators/alarms may be adjusted or repaired only by the manufacturer or a manufacturer-trained technician, using only the manufacturer's NIOSH-approved parts—but the model answer merely emits a tool/function call and does not state those required conditions.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 142 — `OSHA\_29CFR\_full\_2024\_3162\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an air receiver is installed after the effective date of these regulations, what code must it be constructed in accordance with?
- ✓ **Ground truth:** The 1968 edition of the A.S.M.E. Boiler and Pressure Vessel Code Section VIII.
- ✗ **Model answered:** {"name": "get_construction_code", "parameters": {"installation_date": "after the effective date of these regulations"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states that new air receivers installed after the effective date "shall be constructed in accordance with the 1968 edition of the A.S.M.E. Boiler and Pressure Vessel Code Section VIII," but the model only returned a function-call-like JSON and did not provide that code.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 143 — `OSHA\_29CFR\_full\_2024\_3172\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under what specific conditions must a vertical belt running over a lower pulley more than 7 feet above the floor be guarded at the bottom, per paragraph (e)(4) read with the cross-referenced conditions?
- **✓ Ground truth:** Per paragraph (e)(4), it must be guarded at the bottom like horizontal overhead belts if the conditions in paragraphs (e)(2)(ii)(a) and (c) apply - that is, if it is located over passageways or work places and traveling 1,800 feet or more per minute (condition (a)), and if the belt is 8 inches or more in width (condition (c)).
- **✗ Model answered:** {"name": "guard_vertical_belt", "parameters": {"height": "more than 7 feet", "conditions": "conditions are as stated in paragraphs (e)(2)(ii)(a) and (c) of this section"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 144 — `OSHA\_29CFR\_full\_2024\_3038\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a current employee's previous work experience shows they have had training equivalent to the initial 24-hour training, must they still complete the initial training?
- **✓ Ground truth:** No, they are considered as meeting the initial training requirements, but they must still receive eight hours of refresher training annually.
- **✗ Model answered:** {"name": "current_employee_training", "parameters": {"previous_experience": "equivalent to initial 24-hour training"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 145 — `OSHA\_29CFR\_full\_2024\_4060\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** When hoisting an employee in a shaft operation, if the employer can demonstrate that use of a personnel platform is infeasible, what must be used?
- **✓ Ground truth:** A boatswain's chair must be used.
- **✗ Model answered:** {"name": "get_hoisting_requirement", "parameters": {"operation": "shaft", "platform_feasibility": "False"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly states that when a personnel platform is infeasible, "a boatswain's chair must be used," while the model only produced a function-call-like object with the scenario parameters and did not provide that required answer.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 146 — `OSHA\_29CFR\_full\_2024\_2160\_02` · HIGH confidence

CAUSE-EFFECT ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cause & Effect
- **Question:** What were the consequences for analytical bias and OAE resulting from the average recovery of 94.2% of the spiked filter samples in the AAS-HGA analysis?
- ✓ **Ground truth:** The analytical bias was -5.8% and the OAE was ±14.2%.
- ✗ **Model answered:** {"name": "calculate_consequences", "parameters": {"recovery": "94", "analysis_type": "AAS-HGA"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed incorrect analytical bias and OAE
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 147 — `OSHA\_29CFR\_full\_2024\_3423\_01` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** During lifting, within how many inches must all slab support points be kept to maintain the slab in a level position?
- ✓ **Ground truth:** 1/2 inch
- ✗ **Model answered:** {"name": "get_support_point_tolerance", "parameters": {"tolerance": "0"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 148 — `OSHA\_29CFR\_full\_2024\_1332\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under the log hauls and slips section, which two separate provisions address protecting employees from logs that are moving or traveling too far?
- ✓ **Ground truth:** Paragraph (vi)(f) requires a positive stop installed on all log hauls to prevent logs from traveling too far ahead in the mill, and paragraph (vi)(g) requires overhead protection for employees working below logs being moved to the log deck.
- ✗ **Model answered:** {"name": "get_provisions", "parameters": {"section": "log hauls and slips", "topic": "protecting employees from logs"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** (none)

Failure 149 — `OSHA\_29CFR\_full\_2024\_2959\_00` · MEDIUM confidence

UNIT/FORMAT MISMATCH

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Precision
- **Question:** What is the airborne concentration of lead that defines the action level, calculated as an 8-hour time-weighted average?
- ✓ **Ground truth:** 30 micrograms per cubic meter of air (30 µg/m³)
- ✗ **Model answered:** {"name": "get_action_level", "parameters": {"concentration": "30", "time_weighted_average": "8"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states the action level is an airborne concentration of lead of "30 micrograms per cubic meter of air (30 µg/m³)" as an 8-hour TWA, while the model only gave the bare numbers 30 and 8 without the required concentration units.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 150 — `OSHA\_29CFR\_full\_2024\_4065\_00` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** Under § 1926.1434, modifications or additions affecting capacity or safe operation are generally prohibited - under what circumstance are they permitted?
- ✓ **Ground truth:** Where the requirements of paragraphs (a)(1), (a)(2), (a)(3), (a)(4), or (a)(5) are met (manufacturer review and approval, manufacturer refusal to review, unavailable manufacturer, manufacturer fails to complete review within 120 days, or multiple manufacturers for marine work sites).
- ✗ **Model answered:** {"name": "get_permitted_circumstances", "parameters": {"section": "1926.1434", "modification_type": "capacity or safe operation"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not provide enough detail about the permitted circumstances
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 151 — `OSHA\_29CFR\_full\_2024\_2155\_11` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Precision
- **Question:** Below what amount of beta-2-microglobulin in the urine is normally not reabsorbed into the blood and released in urine?
- ✓ **Ground truth:** less than 300 µg/g Cr
- ✗ **Model answered:** {"name": "get_beta_2_microglobulin_normal_level", "parameters": {"level_type": "normal"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly answers the question: "A very small amount (less than 300 µg/g Cr in the urine) ... is not reabsorbed," while the model only emitted a function-call-like JSON with no amount or unit.
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** (none)

Failure 152 — `OSHA\_29CFR\_full\_2024\_0219\_02` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the publication date of ISO 13943:2000 (E/F), Fire Safety-Vocabulary?
- ✓ **Ground truth:** April, 15, 2000
- ✗ **Model answered:** {"name": "get_publication_date", "parameters": {"standard": "ISO 13943:2000 (E/F)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence explicitly gives ISO 13943:2000 (E/F) as 'First Edition, April, 15, 2000,' but the model only emitted a function-call-like request with the standard name and did not provide the publication date.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 153 — `OSHA\_29CFR\_full\_2024\_4031\_02` · HIGH confidence

CALCULATION ERROR

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Calculation
- **Question:** When hoisting personnel without a personnel platform, the equipment's rated capacity for the radius and configuration is 17,800 pounds. What is the maximum total load (including hook, load line, rigging and any other equipment that imposes a load) allowed, except during proof testing?
- ✓ **Ground truth:** 8,900 pounds
- ✗ **Model answered:** {"name": "calculate_max_load", "parameters": {"rated_capacity": "17800"}}
- **Doctor diagnosis:** **Dr. Pi** model error
- **Doctor root cause:** [verifier] The evidence says the load “must not exceed 50 percent” of the 17,800-pound rated capacity, i.e. 8,900 pounds; the model only issued a calculate_max_load call with rated_capacity 17800 and did not provide the required maximum load.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 154 — `OSHA\_29CFR\_full\_2024\_0478\_01` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** For electrostatic spraying operations, what is the minimum air velocity allowed over the open face of the booth?
- ✓ **Ground truth:** 60 linear feet per minute
- ✗ **Model answered:** {"name": "get_minimum_air_velocity", "parameters": {"operation_type": "electrostatic spraying"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly states electrostatic spraying may use air velocity 'not less than 60 linear feet per minute,' but the model only returned a function-call-like JSON with the operation type and did not provide the minimum velocity.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 155 — `OSHA\_29CFR\_full\_2024\_2444\_01` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the injector zone temperature in the GC analysis conditions for MDA?
- ✓ **Ground truth:** 235 degrees C
- ✗ **Model answered:** {"name": "get_GC_analysis_conditions", "parameters": {"zone": "injector"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 156 — `OSHA\_29CFR\_full\_2024\_3172\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** When passage of persons between the upper and lower runs of a horizontal belt is possible and is regarded as necessary, what guarding combination is required, and how does this requirement change in powerplants?
- ✓ **Ground truth:** Under paragraph (e)(1)(iii)(b), there shall be a platform over the lower run guarded on either side by a railing completely filled in with wire mesh or other filler, or by a solid barrier, and the upper run shall be guarded to prevent contact by the worker or objects he carries. However, in powerplants only the lower run of the belt need be guarded.
- ✗ **Model answered:** {"name": "guarding_combination", "parameters": {"passage_regarded_as_necessary": "True", "location": "powerplants"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 157 — `OSHA\_29CFR\_full\_2024\_3305\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Based on subpart (iii) and subpart (iv), what is the difference between the minimum step width of a non-mobile stairway-type ladder and a mobile one, and what tread requirement applies to both?
- ✓ **Ground truth:** The non-mobile minimum step width is 16 inches (41 cm) and the mobile minimum is 11 1/2 inches (30 cm), a difference of 4 1/2 inches; both must have slip-resistant treads on all steps and landings per subpart (iv).
- ✗ **Model answered:** {"name": "calculate_difference", "parameters": {"a": "16", "b": "11"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 158 — `OSHA\_29CFR\_full\_2024\_3420\_01` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** Under § 1926.702(e), concrete pumping systems using discharge pipes must be provided with pipe supports designed for what overload percentage?
- ✓ **Ground truth:** 100 percent
- ✗ **Model answered:** {"name": "answer_section_question", "parameters": {"section_number": "1926.702", "question": "What overload percentage must pipe supports be designed for in concrete pumping systems using discharge pipes under \u00a7 1926.702(e)?"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 159 — `OSHA\_29CFR\_full\_2024\_1227\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the requirements of paragraphs (v) and (vi), what must be done before making hot repairs to one generator located in a room containing additional generators?
- ✓ **Ground truth:** The carbide charge and feed mechanism of the generator being repaired must be completely removed, all its acetylene expelled by completely flooding the shell with water, and it must be disconnected from the piping system (paragraph (v)); additionally, all the other generators and piping in the room must be purged of acetylene (paragraph (vi)).
- ✗ **Model answered:** {"name": "completely flooding the generator shell with water", "parameters": {"generator": "the generator to be repaired"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model stopped before applying dependent rule
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 160 — `OSHA\_29CFR\_full\_2024\_0914\_00` · HIGH confidence

CAUSE-EFFECT ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cause & Effect
- **Question:** In extremely cold climates, what is the permitted consequence for hose storage when standard installation could damage equipment?
- ✓ **Ground truth:** Where such installation may result in damaged equipment in extremely cold climates, the hose may be stored in another location provided it is readily available and can be connected when needed.
- ✗ **Model answered:** {"name": "hose_storage", "parameters": {"climate": "extremely cold", "installation": "standard", "equipment_damage": "true"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 161 — `OSHA\_29CFR\_full\_2024\_3327\_04` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** Under the Effective Date Note, until what date were certain requirements of § 1926.451(b)(2)(i) stayed?
- ✓ **Ground truth:** Nov. 25, 1997, or until further rulemaking has been completed, whichever is later
- ✗ **Model answered:** {"name": "date_stayed_until", "parameters": {"date": "Nov. 25, 1997"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 162 — `OSHA\_29CFR\_full\_2024\_neg\_1769\_05` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** By what time of day must the employer provide records following a request?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "provide_records", "parameters": {"time": "end of the next business day"}}
- **Doctor diagnosis:** **Dr. No** bad test
- **Doctor root cause:** [verifier] The evidence explicitly answers the question: “By the end of the next business day after a request, the employer must provide,” so the model’s answer is supported and the ground truth UNANSWERABLE is wrong.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 163 — `OSHA\_29CFR\_full\_2024\_3280\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** A tank builders' scaffold and a top plate bracket scaffold are both described in the evidence. Based on the definition of 'Supported scaffold,' which of these two relies on a platform held up by a rigid support like brackets, and what is the specific bracket arrangement for each?
- ✓ **Ground truth:** Both are supported scaffolds because they use brackets (a rigid support). The tank builders' scaffold has a platform resting on brackets directly attached to a cylindrical tank or to devices attached to such a tank, while the top plate bracket scaffold is supported by brackets that hook over or are attached to the top of a wall.
- ✗ **Model answered:** {"name": "get_scaffold_info", "parameters": {"scaffold_type": "tank builders"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model only described one scaffold type and missed the other
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 164 — `OSHA\_29CFR\_full\_2024\_2565\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually

- **Category:** Cross-Reference
- **Question:** The term 'hazardous substance' appears in both subpart (v) and subpart (6)(ii). Which different statutes define the term in each, and which agency's regulations apply in each case?
- **Ground truth:** In subpart (v), 'hazardous substance' is defined by the Federal Hazardous Substances Act (15 U.S.C. 1261 et seq.) with regulations issued by the Consumer Product Safety Commission. In subpart (6)(ii), 'hazardous substance' is defined by CERCLA (42 U.S.C. 9601 et seq.) with regulations issued by the Environmental Protection Agency.
- **Model answered:** {"name": "get_statute_and_agency", "parameters": {"subpart": "(v)", "term": "hazardous substance"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 165 — `OSHA\_29CFR\_full\_2024\_1332\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Which two provisions—one from the log hauls section and one from the log decks section—address guarding or controlling chains so employees are not struck?
- **Ground truth:** Paragraph (vi)(c) requires that log haul gear and bull chain drive mechanism shall be guarded (and (vi)(d) requires substantial troughs for the return strand of log haul chains over passageways), while paragraph (3)(iv) prohibits loose chains from overhead canting devices or other equipment from hanging over the log deck in such manner as to strike employees.
- **Model answered:** {"name": "log_hauls", "parameters": {"provision": "vi", "subsection": "d"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model stopped before applying dependent rule
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 166 — `OSHA\_29CFR\_full\_2024\_4448\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** Combining paragraph (a)(2) and paragraph (b), if a complainant files a district court action because there was no final decision within 210 days of filing the complaint, what subsequent filing obligation arises and how soon?
- **Ground truth:** Because there was no final decision of the Secretary within 210 days of the filing of the complaint (a)(2), the complainant may bring the action; then within seven days after filing the federal complaint (b), they must file a file-stamped copy with the Assistant Secretary, ALJ, or ARB, depending on where the proceeding is pending.
- **Model answered:** {"name": "filing_obligation", "parameters": {"days": "7"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports that within seven days the complainant must file a copy of the file-stamped federal complaint with the Assistant Secretary, ALJ, or ARB depending on where pending, but the model supplied only the number of days and omitted the actual subsequent filing obligation.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '7' not in evidence

Failure 167 — `OSHA\_29CFR\_full\_2024\_4450\_02` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the Statutes at Large citation (124 Stat.) for the Patient Protection and Affordable Care Act, Public Law 111-148?
- ✓ **Ground truth:** 124 Stat. 119
- ✗ **Model answered:** {"name": "get_statutes_at_large_citation", "parameters": {"public_law": "111-148"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not compute the Statutes at Large citation correctly
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 168 — `OSHA\_29CFR\_full\_2024\_4010\_02` · MEDIUM confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If the auditor determines there is a significant deficiency in the program, what is triggered?
- ✓ **Ground truth:** The employer must ensure certain follow-up actions (as listed under the Deficiencies provision).
- ✗ **Model answered:** {"name": "determine_action", "parameters": {"deficiency": "significant deficiency"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 169 — `OSHA\_29CFR\_full\_2024\_4228\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** When a developmental step is missed and the State has an operational status agreement with OSHA under § 1954.3, what additional assurance must the State provide?
- ✓ **Ground truth:** The State must provide an assurance that the missed step will not affect the effectiveness of State enforcement in any issues for which the State program has been deemed to be operational.
- ✗ **Model answered:** {"name": "get_additional_assurance", "parameters": {"agreement": "§ 1954.3", "status": "operational"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 170 — `OSHA\_29CFR\_full\_2024\_2019\_01` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What minimum percentage of beryllium by weight must materials contain to define a beryllium

work area when processed during listed operations?

- ✓ **Ground truth:** 0.1 percent by weight
- ✗ **Model answered:** {"name": "get_beryllium_work_area_percentage", "parameters": {"percentage": "0"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 171 — `OSHA\_29CFR\_full\_2024\_1317\_00` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** According to the definition of 'lumber hauling truck,' which two types of industrial trucks are explicitly excluded from being a lumber hauling truck?
- ✓ **Ground truth:** A lift truck or a carrier - a lumber hauling truck is an industrial truck other than a lift truck or a carrier, used for the transport of lumber.
- ✗ **Model answered:** {"name": "get_excluded_truck_types", "parameters": {"truck_type": "lumber hauling truck"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 172 — `OSHA\_29CFR\_full\_2024\_2508\_01` · HIGH confidence

MISINTERPRETED

- **Severity:** MEDIUM — a misapplied or over-extrapolated rule — plausible but wrong
- **Category:** Conditional
- **Question:** If air monitoring data reflects workplace conditions with a lower exposure potential than the employer's current operations, does it qualify as objective data under this section?
- ✓ **Ground truth:** No, objective data must reflect workplace conditions closely resembling or with a higher exposure potential than the employer's current operations.
- ✗ **Model answered:** {"name": "is_objective_data", "parameters": {"data": "air monitoring data", "exposure_potential": "lower"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model misinterpreted the definition of objective data
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 173 — `OSHA\_29CFR\_full\_2024\_4047\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If personnel are being hoisted during a pile driving operation, may lifts be made on another of the equipment's load lines?
- ✓ **Ground truth:** Yes; the prohibition on lifts on other load lines while personnel are hoisted does not apply in pile driving operations.
- ✗ **Model answered:** {"name": "check_lifts_during_pile_driving", "parameters": {"operation": "pile driving"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 174 — `OSHA\_29CFR\_full\_2024\_3703\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Using the field diameter calibration requirement and the worked example for field area, what is the smallest field diameter D (in mm) still within the acceptable range, and does the worked example use a value within that range?
- **Ground truth:** The acceptable range is $100 \pm 2 \mu\text{m}$, so the smallest acceptable diameter is $98 \mu\text{m} = 0.098 \text{ mm}$. The worked example uses $D = 100 \mu\text{m} = 0.1 \text{ mm}$, which is within the acceptable range.
- **Model answered:** {"name": "min", "parameters": {"a": "0", "b": "0"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 175 — `OSHA\_29CFR\_full\_2024\_2815\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Within 1926.550, the paragraphs (a)(1), (a)(6), and (a)(11) each carry an OMB control number. Are all three the same?
- **Ground truth:** No. 1926.550(a)(1) is 1218-0115, 1926.550(a)(6) is 1218-0113, and 1926.550(a)(11) is 1218-0054 - all three are different.
- **Model answered:** {"name": "check_omb_control_numbers", "parameters": {"paragraphs": "'(a)(1)', '(a)(6)', '(a)(11)'"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 176 — `OSHA\_29CFR\_full\_2024\_1695\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under paragraph (a)(2), a flexible cord set is found to have a damaged outer jacket that might expose an employee to injury. Combining the visual inspection provisions, what must happen before any employee may use that cord set again?
- **Ground truth:** The defective or damaged item shall be removed from service, and no employee may use it until repairs and tests necessary to render the equipment safe have been made (per (a)(2)(ii)), even though such defects are exactly the kind identified during the required pre-shift visual inspection under (a)(2)(i).
- **Model answered:** {"name": "remove_from_service", "parameters": {"item": "flexible cord set", "reason": "damaged outer jacket that might expose an employee to injury"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 177 — `OSHA\_29CFR\_full\_2024\_4398\_01` · HIGH confidence

CAUSE-EFFECT ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cause & Effect
- **Question:** What results when the ARB approves a request to withdraw a petition for review of an ALJ decision, and there are no other pending petitions for review of that decision?
- ✓ **Ground truth:** The ALJ's decision will become the final order of the Secretary.
- ✗ **Model answered:** {"name": "determine_outcome_of_withdrawal", "parameters": {"entity": "ARB", "request_type": "petition for review", "decision_type": "ALJ decision", "pending_requests": "none"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 178 — `OSHA\_29CFR\_full\_2024\_1114\_02` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the maximum nominal d.c. supply voltage permitted for d.c. control circuits?
- ✓ **Ground truth:** 240-volt d.c.
- ✗ **Model answered:** {"name": "get_nominal_voltage", "parameters": {"voltage_type": "d.c.", "circuit_type": "control"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence explicitly says 'All d.c. control circuits shall be powered by not more than a nominal 240-volt d.c. supply,' but the model returned only a function-call-like JSON with parameters and did not provide the voltage answer.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 179 — `OSHA\_29CFR\_full\_2024\_4251\_01` · MEDIUM confidence

CONDITIONAL LOGIC

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Conditional
- **Question:** If the administrative law judge determines that a final decision could substantially affect an interested person, may that person be granted party status?
- ✓ **Ground truth:** Yes, the administrative law judge may, at his discretion, grant the right to participate as a party.
- ✗ **Model answered:** {"name": "determine_party_status", "parameters": {"person_affected": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports the ground truth—"Other interested persons may, at the discretion of the administrative law judge" be granted party participation if substantially affected—but the model only output a function-like parameter indicating the person is affected, not the required yes/at-discretion answer.

- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 180 — `OSHA\_29CFR\_full\_2024\_4205\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the revised compliance staffing benchmarks in paragraph (c), what is the total number of compliance officers (safety plus health) proposed for Tennessee?
- ✓ **Ground truth:** 22 safety plus 14 health compliance officers equals 36 total compliance officers.
- ✗ **Model answered:** {"name": "add", "parameters": {"a": "22", "b": "14"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states revised benchmarks of "22 safety and 14 health compliance officers," so the supported total is 36, but the model only returned an add operation with operands rather than the required total.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 181 — `OSHA\_29CFR\_full\_2024\_1206\_01` · MEDIUM confidence

WRONG CALC/QUOTE

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Cross-Reference
- **Question:** Combining the general inside-building storage rule in (b)(2)(ii) and the fuel-gas cylinder storage limit in (b)(3), what is the minimum distance from highly combustible materials and the maximum total gas capacity allowed for stored fuel-gas cylinders inside a building?
- ✓ **Ground truth:** Cylinders must be stored at least 20 feet (6.1 m) from highly combustible materials, and fuel-gas cylinders (except those in actual use or attached ready for use) are limited to a total gas capacity of 2,000 cubic feet (56 m³) or 300 pounds (135.9 kg) of liquefied petroleum gas.
- ✗ **Model answered:** {"name": "calculate_storage_requirements", "parameters": {"distance_from_combustible_materials": "20", "total_gas_capacity": "2000"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports the ground truth—inside storage must be “at least 20 feet” from highly combustible materials and fuel-gas cylinders are limited to “2,000 cubic feet ... or 300 pounds ... of liquefied petroleum gas”—but the model only gave bare numbers 20 and 2000, omitting the LPG alternative and units.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 182 — `OSHA\_29CFR\_full\_2024\_0772\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the threshold check solution preparation in paragraph (a)(5) with the nebulizer usage guidance in paragraph (a)(13), how much sodium saccharin and how much liquid are involved per nebulizer filling?
- ✓ **Ground truth:** The threshold check solution is made by dissolving 0.83 gram of sodium saccharin USP in 100 ml of warm water, and correct use of the nebulizer means approximately 1 ml of that liquid is used at a time in the nebulizer body.

- **✗ Model answered:** {"name": "calculate_nebulizer_filling", "parameters": {"sodium_saccharin": "0", "liquid": "1"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 183 — `OSHA\_29CFR\_full\_2024\_0899\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining paragraph (c)(3) and the example it provides, what training schools should the program be similar to for a general fire brigade versus one in the oil refinery industry?
- **✓ Ground truth:** For a general fire brigade, the program should be similar to schools such as the Maryland Fire and Rescue Institute, Iowa Fire Service Extension, West Virginia Fire Service Extension, Georgia Fire Academy, New York State Department Fire Prevention and Control, Louisiana State University Firemen Training Program, or Washington State's Fire Service Training Commission for Vocational Education. For the oil refinery industry, the program should be similar to those conducted by Texas A & M University, Lamar University, Reno Fire School, or the Delaware State Fire School.
- **✗ Model answered:** {"name": "get_training_schools", "parameters": {"industry": "general"}}; {"name": "get_training_schools", "parameters": {"industry": "oil refinery"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly lists the required schools for both general brigades and oil refineries, but the model only emitted tool calls with industries and did not provide the listed schools, so it did not answer the question.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 184 — `OSHA\_29CFR\_full\_2024\_3059\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a spray booth has a frontal area of 8 square feet, is a metal deflector or curtain required at the upper outer edge?
- **✓ Ground truth:** No, the requirement applies only to booths with a frontal area larger than 9 square feet.
- **✗ Model answered:** {"name": "is_metal_deflector_required", "parameters": {"frontal_area": "8"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states the requirement applies only to spray booths with a frontal area “larger than 9 square feet,” so the supported answer for 8 square feet is no; the model merely emitted a function/tool call with the parameter 8 and did not provide that answer.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 185 — `OSHA\_29CFR\_full\_2024\_neg\_0478\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)

- **Question:** What is the maximum allowable air velocity over the open face of a booth during standard (non-electrostatic) spraying operations?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "max", "parameters": {"a": "100", "b": "60"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated answer not in evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 186 — `OSHA\_29CFR\_full\_2024\_4247\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Cross-Reference
- **Question:** Under § 1955.13, if a party believes an administrative law judge is disqualified and wishes to remove him, what two requirements from paragraph (a) must be met regarding timing and supporting documentation, and who rules on the request?
- ✓ **Ground truth:** The motion to disqualify and remove the administrative law judge must be filed prior to the time the administrative law judge files his decision, and it must be supported by affidavits setting forth the alleged ground for disqualification. The Chief Administrative Law Judge shall rule upon the motion.
- ✗ **Model answered:** {"name": "answer_question", "parameters": {"question": "Under § 1955.13, if a party believes an administrative law judge is disqualified and wishes to remove him, what two requirements from paragraph (a) must be met regarding timing and supporting documentation, and who rules on the request?", "requirements": "timing and supporting documentation", "ruling_authority": "Chief Administrative Law Judge"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence specifies the motion must be filed “prior to the time the administrative law judge files his decision” and “supported by affidavits setting forth the alleged ground,” but the model only restated the requested categories (“timing and supporting documentation”) while correctly naming the Chief Administrative Law Judge, so it omitted required substantive details.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 187 — `OSHA\_29CFR\_full\_2024\_neg\_0484\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** At what specific pressure (in psi) must the relief valve on air-pressure containers be set to operate?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_pressure", "parameters": {"psi": "the requirements of the Code for Unfired Pressure Vessels, Section VIII of the ASME Boiler and Pressure Vessel Code-1968"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated answer not in evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 188 — `OSHA\_29CFR\_full\_2024\_1434\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an employer follows the guidance in appendix E to estimate available heat energy, will OSHA consider the employer in compliance with paragraph (l)(8)(ii)?
- ✓ **Ground truth:** Yes; OSHA will deem employers following the guidance in appendix E to be in compliance with paragraph (l)(8)(ii).
- ✗ **Model answered:** {"name": "OSHA_compliance", "parameters": {"guidance_followed": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 189 — `OSHA\_29CFR\_full\_2024\_4450\_00` · HIGH confidence

TABLE MISREAD

- **Severity:** MEDIUM — a misapplied or over-extrapolated rule — plausible but wrong
- **Category:** Precision
- **Question:** On what date was the Patient Protection and Affordable Care Act, Public Law 111-148, signed into law?
- ✓ **Ground truth:** March 23, 2010
- ✗ **Model answered:** {"name": "get_date_signed_into_law", "parameters": {"public_law": "111-148"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence explicitly states Public Law 111-148 “was signed into law on March 23, 2010,” while the model only returned a tool/function call with the public law number and did not provide the date.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 190 — `OSHA\_29CFR\_full\_2024\_2632\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a substance shows a high potency in animals but only a low frequency of occurrence in humans, which sub-category applies?
- ✓ **Ground truth:** Sub-category 1A, since high potency in animals presumes potential to produce significant sensitization.
- ✗ **Model answered:** {"name": "classify_skin_sensitizer", "parameters": {"human_frequency": "low", "animal_potency": "high"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 191 — `OSHA\_29CFR\_full\_2024\_3990\_01` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** Per § 1926.1420(c), the operator's reception of signals must be by what type of system?
- ✓ **Ground truth:** A hands-free system
- ✗ **Model answered:** {"name": "answer_question", "parameters": {"section": "1926.1420(c)", "query": "type of system for operator"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 192 — `OSHA\_29CFR\_full\_2024\_3130\_02` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** For Category 1 flammable liquids, what is the maximum boiling point?
- ✓ **Ground truth:** at or below 95 °F (35 °C)
- ✗ **Model answered:** {"name": "get_flammable_liquid_category", "parameters": {"category": "1"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 193 — `OSHA\_29CFR\_full\_2024\_3280\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Using the definitions of 'Suspension scaffold' and 'Two-point suspension scaffold,' explain what makes a two-point suspension scaffold (swing stage) a specific type of suspension scaffold.
- ✓ **Ground truth:** A suspension scaffold is one or more platforms suspended by ropes or other non-rigid means from an overhead structure. A two-point suspension scaffold (swing stage) fits this because it consists of a platform supported by hangers (stirrups) suspended by two ropes from overhead supports, and it is additionally equipped with means to raise and lower the platform to desired work levels.
- ✗ **Model answered:** {"name": "explain_suspension_scaffold", "parameters": {"scaffold_type": "Two-point suspension scaffold (swing stage)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not fully understand the definitions of suspension scaffold and two-point suspension scaffold
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 194 — `OSHA\_29CFR\_full\_2024\_2250\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually

- **Category:** Conditional
- **Question:** A facility uses Universal Precautions for all specimens. A specimen container is being moved from one lab to another lab within the same facility. Is labeling or color-coding required?
- **Ground truth:** No, the labeling/color-coding exemption applies while specimens remain within the facility, provided containers are recognizable as containing specimens.
- **Model answered:** {"name": "labeling_or_color_coding_required", "parameters": {"universal_precautions": "True", "within_facility": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 195 — `OSHA\_29CFR\_full\_2024\_4205\_02` · MEDIUM confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Using the final approval date in paragraph (b) and the date the Assistant Secretary approved the revised staffing requirements in paragraph (c), what do you notice about these two dates?
- **Ground truth:** Both occurred on the same date, July 22, 1985 - the final approval of the plan and the approval of the revised compliance staffing benchmarks coincided.
- **Model answered:** {"name": "get_date_difference", "parameters": {"date1": "July 22, 1985", "date2": "July 22, 1985"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports that both dates are July 22, 1985, but the model only emitted a tool-call-style date-difference request with the two dates and did not actually answer the question by noting that they coincided.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 196 — `OSHA\_29CFR\_full\_2024\_3285\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If rulemaking on roof bracket scaffold minimum width was not completed by November 25, 1997, was the 12-inch width requirement still stayed?
- **Ground truth:** Yes, the stay lasts until November 25, 1997 or until rulemaking is completed, whichever is later.
- **Model answered:** {"name": "check_rulemaking_status", "parameters": {"completion_date": "November 25, 1997", "rulemaking_status": "not completed"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports 'Yes' because the stay lasts until 'November 25, 1997 or until rulemaking ... has been completed, whichever is later,' but the model only produced a tool-call-like JSON and did not answer whether the requirement was still stayed.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 197 — `OSHA\_29CFR\_full\_2024\_4301\_04` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** At the national level, if employees are not represented by organizations having exclusive recognition on an agency basis or national consultation rights, how are nonmanagement members determined?
- ✓ **Ground truth:** They shall be determined through procedures devised by the agency which provide for effective representation of all employees.
- ✗ **Model answered:** {"name": "determine_nonmanagement_members", "parameters": {"level": "national", "representation": "not represented"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 198 — `OSHA\_29CFR\_full\_2024\_2019\_00` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the action level concentration of airborne beryllium, calculated as an 8-hour TWA?
- ✓ **Ground truth:** 0.1 micrograms per cubic meter of air ($\mu\text{g}/\text{m}^3$)
- ✗ **Model answered:** {"name": "get_action_level", "parameters": {"concentration": "0", "time": "8"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed incorrect value
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 199 — `OSHA\_29CFR\_full\_2024\_neg\_1599\_01` · MEDIUM confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum wattage rating required for a heavy-duty lampholder of the medium type?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_minimum_wattage_rating", "parameters": {"type": "medium"}}
- **Doctor diagnosis:** **Dr. No** bad test
- **Doctor root cause:** [verifier] The evidence gives a determinate rule: a heavy-duty lampholder is at least 660 W only if it is the “admedium type” and “not less than 750 watts if of any other type,” so a medium type not identified as admedium would fall under the 750 W category; the GT of UNANSWERABLE is therefore not supported.
- **Verifier verdict:** bad test (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 200 — `OSHA\_29CFR\_full\_2024\_3801\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference

- **Question:** Combining paragraph (h)(1) and paragraph (h)(2)(ii), what triggers the requirement to provide PPE and what document governs both its use and its removal?
- ✓ **Ground truth:** PPE must be provided where airborne exposure exceeds or can reasonably be expected to exceed the TWA PEL or STEL (h)(1), and both its use and its removal are governed by the written exposure control plan required by paragraph (f)(1) (h)(1) and (h)(2)(ii).
- ✗ **Model answered:** {"name": "answer", "parameters": {"question": "Combining paragraph (h)(1) and paragraph (h)(2)(ii), what triggers the requirement to provide PPE and what document governs both its use and its removal?"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 201 — `OSHA\_29CFR\_full\_2024\_4366\_03` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** If a timely petition for review is filed, when does the ALJ's decision NOT automatically become the final order of the Secretary?
- ✓ **Ground truth:** Unless the ARB, within 30 days of the filing of the petition, issues an order notifying the parties that the case has been accepted for review.
- ✗ **Model answered:** {"name": "paragraph_b_exception", "parameters": {"timely_petition_for_review": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 202 — `OSHA\_29CFR\_full\_2024\_1434\_06` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an employee exposed to electric arc hazards wears nylon clothing, but the employer demonstrates the fabric was treated to withstand the conditions the employee may encounter, is the clothing prohibited?
- ✓ **Ground truth:** No; the prohibition does not apply if the employer demonstrates the fabric has been treated to withstand the conditions or that the employee wears it in a manner that eliminates the hazard.
- ✗ **Model answered:** {"name": "check_prohibited_clothing", "parameters": {"clothing_material": "nylon", "treated_to_withstand_conditions": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 203 — `OSHA\_29CFR\_full\_2024\_4301\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional

- **Question:** At an establishment where employees are represented under collective bargaining arrangements, how are nonmanagement committee members appointed?
- ✓ **Ground truth:** They shall be appointed from among those recommended by the exclusive bargaining representative.
- ✗ **Model answered:** {"name": "appoint_nonmanagement_members", "parameters": {"method": "collective bargaining arrangements"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 204 — `OSHA\_29CFR\_full\_2024\_4031\_01` · HIGH confidence

CALCULATION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Calculation
- **Question:** A boom-attached personnel platform is used on equipment whose rated capacity for the radius and configuration is 24,500 pounds. What is the maximum total weight of the loaded personnel platform allowed, except during proof testing?
- ✓ **Ground truth:** 12,250 pounds
- ✗ **Model answered:** {"name": "calculate_max_weight", "parameters": {"rated_capacity": "24500"}}
- **Doctor diagnosis:** **Dr. Pi** model error
- **Doctor root cause:** [verifier] The evidence says the loaded boom-attached platform 'must not exceed 50 percent' of 24,500 pounds, which is 12,250 pounds, but the model only emitted a function call with the rated capacity and did not provide the required maximum weight.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 205 — `OSHA\_29CFR\_full\_2024\_neg\_1599\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum branch-circuit conductor ampacity required for a cord- and plug-connected arc welder receptacle?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "Table S-4", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only says arc-welder receptacle ratings may be 'not less than the minimum branch-circuit conductor ampacity' but does not state that ampacity, so the GT UNANSWERABLE is supported, while the model answered 'Table S-4' rather than refusing.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 206 — `OSHA\_29CFR\_full\_2024\_1695\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Both paragraph (a)(2)(ii) (portable equipment) and paragraph (c)(2) (test instruments) describe

what to do upon finding a defect that might expose an employee to injury. What common requirement do both impose?

- ✓ **Ground truth:** In both cases the defective or damaged item shall be removed from service, and no employee may use it until repairs and tests necessary to render the equipment safe have been made.
- ✗ **Model answered:** {"name": "common_requirement", "parameters": {"paragraphs": "['(a)(2)(ii)', '(c)(2)']"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 207 — `OSHA\_29CFR\_full\_2024\_4251\_00` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an interested person submits a petition to participate as a party 35 days after the notice was published in the Federal Register, does it meet the timing requirement of paragraph (b)(1)?
- ✓ **Ground truth:** No, because the petition must be submitted within 30 days after the notice has been published in the Federal Register.
- ✗ **Model answered:** {"name": "check_timing_requirement", "parameters": {"days_after_notice": "35", "required_days": "30"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states a petition must be submitted “within 30 days after the notice ... has been published,” so a 35-day submission does not meet the requirement, while the model only returned extracted parameters and did not answer no.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 208 — `OSHA\_29CFR\_full\_2024\_neg\_1037\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum angle of loading permitted for a basket hitch configuration before it must be re-rated?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_angle_of_loading", "parameters": {"configuration": "basket hitch"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated answer not in evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 209 — `OSHA\_29CFR\_full\_2024\_4294\_04` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining paragraph (d)(1) and paragraph (d)(4), who receives copies of reports and of no-inspection notifications at the establishment level?
- ✓ **Ground truth:** Under (d)(1), a copy of each report received is sent to the appropriate establishment safety

and health committee; under (d)(4), a copy of each no-inspection notification is provided to the appropriate certified safety and health committee where established under Executive Order 12196.

- **✗ Model answered:** {"name": "get_recipients_of_copies", "parameters": {"paragraph": "d", "items": ["reports", "no-inspection notifications"]}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 210 — `OSHA\_29CFR\_full\_2024\_neg\_3653\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum efficiency rating, expressed as a percentage, required for the HEPA filters provided for air-purifying respirators?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** HEPA filters are required to be 99.97% efficient.
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated answer not in evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '99.97%' not in evidence

Failure 211 — `OSHA\_29CFR\_full\_2024\_1332\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Considering the dry deck storage requirement to minimize accidental rolling and the log decks 'Stops' provision, what two distinct kinds of devices/arrangements does the evidence require to prevent logs from rolling in these respective areas?
- **✓ Ground truth:** For dry deck storage (paragraph (v)(c)), logs shall be arranged to minimize the chance of accidentally rolling from the deck; for log decks (paragraph (3)(ii) 'Stops'), log decks shall be provided with adequate stops, chains, or other safeguards to prevent logs from rolling down the deck onto the carriage or its runway.
- **✗ Model answered:** {"name": "log_decks_Stops", "parameters": {"stops": "adequate stops, chains, or other safeguards", "location": "log decks"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports two requirements—dry deck logs must be 'arranged to minimize the chance of accidentally rolling' and log decks need 'adequate stops, chains, or other safeguards'—but the model only gave the log-deck stops/safeguards and omitted the dry-deck arrangement.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 212 — `OSHA\_29CFR\_full\_2024\_3641\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference

- **Question:** Both the glove bag work practices and the negative pressure glove box work practices specify a minimum staffing level. How many persons are required to perform each operation?
- **Ground truth:** At least two persons are required for Class I glovebag removal operations, and at least two persons are required to perform the negative pressure glove box removal.
- **Model answered:** {"name": "get_minimum_staffing_level", "parameters": {"operation": "glove bag work practices"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 213 — `OSHA\_29CFR\_full\_2024\_4450\_01` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** On what date was the Health Care and Education Reconciliation Act of 2010, Public Law 111-152, signed into law?
- **Ground truth:** March 30, 2010
- **Model answered:** {"name": "get_date_signed_into_law", "parameters": {"public_law": "111-152"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence explicitly states Public Law 111-152 was “signed into law on March 30, 2010,” while the model only returned a function-call JSON with the public law number and did not answer the date.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 214 — `OSHA\_29CFR\_full\_2024\_2160\_00` · HIGH confidence

CAUSE-EFFECT ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cause & Effect
- **Question:** Why was the standard deviation of the 0.2 ng/mL cadmium standard (0.75 abs-sec), rather than the reagent blank, used to calculate the detection limits by the IUPAC method?
- **Ground truth:** Because the near lack of a blank signal (five of the six reagent blank peak area readings were zero) does not satisfy a strict interpretation of the IUPAC method for determining the detection limits.
- **Model answered:** {"name": "calculate_detection_limits", "parameters": {"std_dev": "0", "slope": "51"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not understand the context of the IUPAC method for determining detection limits
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 215 — `OSHA\_29CFR\_full\_2024\_2444\_00` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the column zone temperature in the GC analysis conditions for MDA?
- **Ground truth:** 220 degrees C
- **Model answered:** {"name": "get_gc_condition", "parameters": {"condition": "column zone temperature"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 216 — `OSHA\_29CFR\_full\_2024\_2016\_03` · MEDIUM confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Using the 1980 RTECS listing counts described in Appendix B, how many of the 168,096 total listings are NOT names of different chemicals with associated toxicity data?
- **✓ Ground truth:** 122,940. Of the 168,096 total listings, 45,156 are names of different chemicals with their associated toxicity data; the remaining 122,940 are synonyms ($168,096 - 45,156 = 122,940$).
- **✗ Model answered:** {"name": "subtract", "parameters": {"a": "168096", "b": "45156"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 217 — `OSHA\_29CFR\_full\_2024\_4060\_02` · MEDIUM confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** When hoisting an employee in a chimney operation, may a boatswain's chair be used if the employer has not demonstrated that a personnel platform is infeasible?
- **✓ Ground truth:** No; the employee must be in a personnel platform unless the employer can demonstrate that use of a personnel platform is infeasible.
- **✗ Model answered:** {"name": "is_boatswains_chair_allowed", "parameters": {"feasible": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence says a boatswain's chair is used only when the employer can demonstrate a personnel platform is infeasible, so the answer is No; the model's structured output with only feasible: False does not convey that prohibition and, if read as platform feasibility being false, points to the exception rather than the stated condition.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 218 — `OSHA\_29CFR\_full\_2024\_2160\_01` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cause & Effect
- **Question:** What were the consequences for analytical bias and OAE resulting from the average recovery of 118.2% at 0.1 times and 104.0% in the range of 0.5 to 2.0 times the TWA PEL target concentration in the flame AAS analysis?
- **✓ Ground truth:** The analytical bias found over the tested concentration range was +4.0% and the OAE was ±6.0%.
- **✗ Model answered:** {"name": "calculate_consequences", "parameters": {"analytical_bias_1": "118.2", "analytical_bias_2": "104.0", "oe_1": "6.0", "oe_2": "6.0"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed ratio instead of quoting stated figure
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 219 — `OSHA\_29CFR\_full\_2024\_4366\_01` · MEDIUM confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** Although an order of reinstatement is effective during ARB review, under what circumstance can that effectiveness be stayed?
- **Ground truth:** Unless the ARB grants a motion by the respondent to stay that order based on exceptional circumstances.
- **Model answered:** {"name": "circumstance_for_stay", "parameters": {"circumstance": "exceptional circumstances"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 220 — `OSHA\_29CFR\_full\_2024\_3987\_03` · MEDIUM confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** If a warning (tag-out or do not operate) sign is on the equipment or starting control, when may the operator activate the switch or start the equipment?
- **Ground truth:** Only after the sign has been removed by a person authorized to remove it, or after the operator has verified that (A) no one is servicing, working on, or otherwise in a dangerous position on the machine, and (B) the equipment has been repaired and is working properly.
- **Model answered:** {"name": "activate_switch_or_start_equipment", "parameters": {"warning_sign": "tag-out or do not operate", "authorized_removal": "True", "verification": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence allows activation only after authorized removal of the sign or verification that no one is in danger and the equipment is repaired/working properly; the model's terse JSON omits the required verified facts and appears to require/merely flag both "authorized_removal" and "verification" rather than stating the alternative conditions.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 221 — `OSHA\_29CFR\_full\_2024\_2239\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** According to the record-keeping requirements, both a copy of the physician's written opinion and the cytologic examination slide and written description for the most recent 10 years must be kept. Under which subparts are these two items respectively required?
- **Ground truth:** The copy of the physician's written opinion is required under paragraph (i)(b), and the cytologic examination slide and written description for the most recent 10 years is required under paragraph

(ii)(g).

- **✗ Model answered:** {"name": "get_record_keeping_requirements", "parameters": {"item": "copy of the physician"}}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 222 — `OSHA\_29CFR\_full\_2024\_3110\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under the marine service station rules, if a tank is installed on a pier (rather than on shore) and is located at an elevation producing gravity head on the dispensing unit, what two requirements must be met regarding maximum stored quantity and the tank outlet?
- **✓ Ground truth:** Per paragraph (k)(2)(ii), the quantity stored on the pier may not exceed 1,100 gallons (4,163.5 L) aggregate capacity, and per paragraph (k)(2)(iv), the tank outlet must be equipped with a pressure control valve positioned adjacent to and outside the tank block valve, adjusted so liquid cannot flow by gravity in case of piping or hose failure.
- **✗ Model answered:** {"name": "get_marine_service_station_requirements", "parameters": {"tank_location": "pier", "elevation": "producing gravity head"}}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 223 — `OSHA\_29CFR\_full\_2024\_4269\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the policy statement referenced from section 2(b) with the exclusion in section 3(5), why does the Act's general policy not automatically protect State and local government employees?
- **✓ Ground truth:** Although section 2(b) sets forth the policy of assuring every working man and woman safe and healthful working conditions, State and local government agencies are excluded from the definition of 'employer' in section 3(5), so public employees are not automatically covered and can only be protected under an approved State plan via section 18.
- **✗ Model answered:** {"name": "get_policy_exclusion_reason", "parameters": {"policy_statement": "section 2(b)", "exclusion_section": "section 3(5)", "employee_type": "State and local government employees"}}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly answers that although §2(b) states a policy of assuring safe working conditions, "State and local government agencies are excluded from the definition of 'employer' in section 3(5)" and public employees are protected only under an approved State plan, while the model merely returned parameters and did not state that reason.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** According to paragraph (i), one option for clean-up is to meet all requirements of paragraphs (b) through (o). What alternative does paragraph (ii) provide when the clean-up is done on plant property using plant or workplace employees?
- **Ground truth:** Under paragraph (ii), as an alternative to meeting all requirements of paragraphs (b) through (o), such employees shall have completed the training requirements of 29 CFR 1926.35, 1926.59, and 1926.103, and other appropriate safety and health training made necessary by the tasks they are expected to perform, such as personal protective equipment and decontamination procedures; additionally, all equipment used shall be in serviceable condition and inspected prior to use.
- **Model answered:** {"name": "get_alternative_clean_up_option", "parameters": {"option": "clean-up is done on plant property using plant or workplace employees"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not fully capture the alternative clean-up option provided by paragraph (ii)
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Complete Failure Listing — Without VERITROOPER (Baseline)

363 confirmed model errors + 6 excluded bad tests (malformed question, not counted against the model). Each entry shows the full question, the model's full answer, the ground truth, the Doctor's diagnosis, and (where available) the Verifier's independent audit verdict.

Failure shape at a glance — Without VERITROOPER (Baseline)

Category	Failures	Model error	Bad test	Review recommended
Negative (hallucination traps)	195	189	6	0
Cross-Reference	75	75	0	0
Conditional	61	61	0	0
Precision	15	15	0	0
Exception	13	13	0	0
Cause & Effect	7	7	0	0
Calculation	3	3	0	0
TOTAL	369	363	6	0

Severity mix (confirmed errors): 195 HIGH, 161 MEDIUM, 7 LOW. Severity is a domain-agnostic heuristic from the error type — a confidently-wrong answer outranks a missing one; re-weight for your risk context.

Failure 1 — `OSHA\_29CFR\_full\_2024\_neg\_0429\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many feet higher than flammable liquid storage should a bulk oxygen system be located?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** {"name": "elevation", "parameters": {"elevation": "higher"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 2 — `OSHA\_29CFR\_full\_2024\_neg\_4403\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Who within the Department of Labor is designated to conduct the investigations referenced in this part?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** {"name": "Compliance Safety and Health Officer", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 3 — `OSHA\_29CFR\_full\_2024\_0676\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** When may a question of employee exposure to hazardous concentrations require monitoring under paragraph (h)?
- **✓ Ground truth:** Where there may be a question of employee exposure to hazardous concentrations of hazardous substances, to assure proper selection of engineering controls, work practices, and personal protective equipment.
- **✗ Model answered:** {"name": "monitoring", "parameters": {"h": "paragraph (h)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states monitoring is required “where there may be a question of employee exposure to hazardous concentrations of hazardous substances” to assure proper selection of controls/PPE, but the model only returned a generic JSON label for monitoring under paragraph (h) and did not answer the when/why asked.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 4 — `OSHA\_29CFR\_full\_2024\_neg\_2602\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum allowable concentration variation between batches before a new classification under A.0.5.1.2 becomes mandatory?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "A.0.5.1.2", "parameters": {"concentration variation": "significant variation"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 5 — `OSHA\_29CFR\_full\_2024\_2250\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** Equipment may have become contaminated with blood, but the employer can demonstrate that decontamination of that equipment is not feasible. Must it still be decontaminated before servicing?
- ✓ **Ground truth:** No, decontamination is required unless the employer can demonstrate that decontamination of such equipment is not feasible.
- ✗ **Model answered:** {"name": "decontaminate_equipment", "parameters": {"contaminated": "True", "feasible": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 6 — `OSHA\_29CFR\_full\_2024\_neg\_2902\_09` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the required color of the warning signs marking abrasive-blasting enclosures?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_color", "parameters": {"tag_type": "Warning"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence contains no requirement for the color of warning signs marking abrasive-blasting enclosures, so the supported answer is UNANSWERABLE; the model did not refuse or state that the evidence lacks the answer, but instead attempted to call get_color for a warning tag.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 7 — `OSHA\_29CFR\_full\_2024\_neg\_4403\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the filing fee for submitting a discrimination complaint under this part?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_filing_fee", "parameters": {"complaint_type": "discrimination"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 8 — `OSHA\_29CFR\_full\_2024\_neg\_2913\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How much open space, in inches, must be maintained between the duct and a combustible roof at the point of penetration?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "open_space", "parameters": {"inches": "18"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 9 — `OSHA\_29CFR\_full\_2024\_1138\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a critical component of the PSDI safety system fails, can the press remain in service after recertification by the employer alone, without revalidation?
- ✓ **Ground truth:** No; the failure of a critical component requires removal of the safety system from service until it is recertified AND revalidated.
- ✗ **Model answered:** {"name": "recertification_by_employer", "parameters": {"failure_type": "critical component"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed incorrect recertification procedure
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 10 — `OSHA\_29CFR\_full\_2024\_0772\_00` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference

- **Question:** According to paragraph (a)(5), the threshold check solution can be prepared from the fit test solution referenced in which other paragraph, and how is it diluted?
- ✓ **Ground truth:** It is prepared by putting 1 ml of the fit test solution (referenced in (b)(5)) into 100 ml of distilled water.
- ✗ **Model answered:** {"name": "prepare_threshold_check_solution", "parameters": {"paragraph_number": "5", "dilution_factor": "100"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model confused preparation method
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 11 — `OSHA\_29CFR\_full\_2024\_neg\_0429\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What minimum ventilation rate is required for a noncombustible building housing a bulk oxygen system?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_ventilation_rate", "parameters": {"building_type": "noncombustible", "system_type": "bulk oxygen"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 12 — `OSHA\_29CFR\_full\_2024\_neg\_0484\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What temperature must electrically powered spray liquid heaters be limited to during operation?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "temperature_limit", "parameters": {"temperature_limit": "199.4 \u00b0F (93 \u00b0C)"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 13 — `OSHA\_29CFR\_full\_2024\_neg\_1769\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum number of employees an employer can record before the COVID-19 log requirement no longer applies?
- ✓ **Ground truth:** UNANSWERABLE

- **✗ Model answered:** {"name": "COVID-19 log requirement exception", "parameters": {"number of employees": "10"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 14 — `OSHA\_29CFR\_full\_2024\_neg\_4297\_03` · MEDIUM confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What advance notice period is required for OSHA's 'announced' inspections under this section?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** {"name": "get_advance_notice_period", "parameters": {"inspection_type": "announced"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence authorizes 'announced or unannounced inspections' but gives no advance notice period, so the supported answer is UNANSWERABLE; the model's function call to get an advance notice period is not equivalent to declining based on missing evidence.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 15 — `OSHA\_29CFR\_full\_2024\_neg\_1194\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What grounding voltage threshold triggers the requirement to protect personnel from shock when floors are wet down?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** {"name": "protect_personnel_from_shock", "parameters": {"voltage": "30"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 16 — `OSHA\_29CFR\_full\_2024\_neg\_2902\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum width permitted for aisles and walkways near abrasive-blasting operations?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** {"name": "get_aisle_width", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only says aisles and walkways must be kept clear of abrasive

creating a slipping hazard, and gives no maximum width; the ground truth UNANSWERABLE is supported, while the model's function call get_aisle_width is not a refusal or equivalent unanswerable answer.

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 17 — `OSHA\_29CFR\_full\_2024\_neg\_1738\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What temperature threshold qualifies as a 'fever' among the listed COVID-19 symptoms?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "COVID-19 symptoms", "parameters": {"Fever or chills": "100.4 \u00b0F"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 18 — `OSHA\_29CFR\_full\_2024\_0913\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under what condition is the educational program of paragraph (g)(1) required, and how does its triggering condition differ from the training in paragraph (g)(3)?
- **Ground truth:** The educational program in (g)(1) is required where the employer has provided portable fire extinguishers for employee use in the workplace, while the training in (g)(3) is required for employees designated to use fire fighting equipment as part of an emergency action plan.
- **Model answered:** {"name": "educational_program", "parameters": {"condition": "when it can be shown that their work activities could result in exposure to blood or OPIM"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model confused the triggering condition for the educational program with the training in paragraph (g)(3)
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 19 — `OSHA\_29CFR\_full\_2024\_neg\_2625\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Within what timeframe after exposure must asthma symptoms appear to count as evidence of respiratory hypersensitivity?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "A.4.2.1.2.3", "parameters": {"(iii)": "within hours"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 20 — `OSHA\_29CFR\_full\_2024\_neg\_2602\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many toxicologically active ingredients must mixtures A and B share for the interpolation principle in A.0.5.1.4 to apply?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "A.0.5.1.4", "parameters": {"n": "1"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 21 — `OSHA\_29CFR\_full\_2024\_1157\_01` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** When are sprocket wheels and chains exempt from the requirement that they be enclosed?
- ✓ **Ground truth:** When they are more than seven (7) feet above the floor or platform (the enclosure requirement does not apply to manually operated sprockets either).
- ✗ **Model answered:** {"name": "sprocket_wheels_and_chains_exempt", "parameters": {"height": "7 feet"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence says sprocket wheels and chains need not be enclosed only when they are “more than seven (7) feet above the floor or platform,” and it separately excludes “manually operated sprockets,” while the model’s answer merely says “height: 7 feet,” omitting the more-than condition, floor/platform reference, and manual-sprocket exception.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 22 — `OSHA\_29CFR\_full\_2024\_0772\_04` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Using paragraphs (a)(3) and (a)(12), what must the test subject do during the screening test and what additional action is required once a taste response is elicited?
- ✓ **Ground truth:** During the test the subject must breathe through a slightly open mouth with tongue extended and report when a sweet taste is detected. Once a taste response is elicited, the subject must take note of the taste for reference in the fit test.
- ✗ **Model answered:** {"name": "taste_threshold_screening", "parameters": {"action": "don the test enclosure", "additional_action": "take note of the taste for reference in the fit test"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 23 — `OSHA\_29CFR\_full\_2024\_3092\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the container requirements in paragraph (a)(1) and the storage restrictions in paragraph (a)(2), if a worker handles 4 gallons of flammable liquid, what type of container is required and in what areas may it not be stored?
- **Ground truth:** Since 4 gallons is 5 gallons or less, approved safety cans or DOT approved containers must be used (per (a)(1)), and per (a)(2) it shall not be stored in areas used for exits, stairways, or normally used for the safe passage of people.
- **Model answered:** {"name": "container_requirements", "parameters": {"quantity": "4", "container_type": "closed containers", "storage_restrictions": "outside of an inside storage room or storage cabinet in a building or in any one fire area of a building"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence requires for 4 gallons that 'Approved safety cans or Department of Transportation approved containers shall be used' and says flammable liquids shall not be stored in 'exits, stairways, or normally used for the safe passage of people,' while the model instead gave 'closed containers' and unrelated storage restrictions not in the cited text.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 24 — `OSHA\_29CFR\_full\_2024\_neg\_1194\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How far from bulk sulfur or baled paper storage may welding be conducted safely?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "welding_safety_distance", "parameters": {"x": "35"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 25 — `OSHA\_29CFR\_full\_2024\_neg\_1599\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum cord- and plug-connected load permitted for a receptacle on a 40-ampere branch circuit supplying two or more receptacles?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "get_max_load", "parameters": {"circuit_rating": "40", "receptacle_rating": "40"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 26 — `OSHA\_29CFR\_full\_2024\_3608\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Per the requirements in subparts (i) and (iii), what is the maximum drop weight a device must withstand and within what distance must it activate after a fall?
- ✓ **Ground truth:** The device must withstand an 18-inch (41 cm) drop of a 500-pound (226 kg) weight, and it must be activated within 2 feet (.61 m) after a fall occurs.
- ✗ **Model answered:** {"name": "calculate_drop_test", "parameters": {"weight": "250", "distance": "4"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed different values for drop test
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 27 — `OSHA\_29CFR\_full\_2024\_neg\_1738\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Within how many hours of a confirmed COVID-19 positive result must the employer notify the Assistant Secretary?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "reporting_covid_19_fatalities_and_hospitalizations_to_osh", "parameters": {"hours": "8"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 28 — `OSHA\_29CFR\_full\_2024\_0564\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** When an LP-Gas-fueled vehicle is being repaired in a garage, may its container shutoff valve be open?
- ✓ **Ground truth:** Only when fuel is required for engine operation; otherwise the shutoff valve must be closed.
- ✗ **Model answered:** {"name": "LP-Gas-fueled vehicle repair in garage", "parameters": {"valve_status": "closed"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 29 — `OSHA\_29CFR\_full\_2024\_2809\_11` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the street address (number and street) of the American National Standards Institute (ANSI) given under Availability of Referenced Material?
- ✓ **Ground truth:** 1430 Broadway, New York, NY 10018
- ✗ **Model answered:** {"name": "get_address", "parameters": {"organization": "American National Standards Institute (ANSI)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 30 — `OSHA\_29CFR\_full\_2024\_neg\_3806\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What form must the employee use to provide written consent for releasing prior medical examination records?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "appendix A to this section contains a sample form which may be used to establish specific written consent for access to employee medical records", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 31 — `OSHA\_29CFR\_full\_2024\_0598\_00` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the minimum distance required between underground containers of over 2,000 gallons capacity?
- ✓ **Ground truth:** at least 5 feet
- ✗ **Model answered:** {"name": "get_minimum_distance", "parameters": {"water_capacity": "2001"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed incorrect minimum distance
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 32 — `OSHA\_29CFR\_full\_2024\_3312\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional

- **Question:** If work near an insulated power line of 200 volts is being performed, what is the minimum distance required?
- ✓ **Ground truth:** 3 feet (0.9 m), since the line is less than 300 volts.
- ✗ **Model answered:** {"name": "minimum_distance", "parameters": {"voltage": "200"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 33 — `OSHA\_29CFR\_full\_2024\_4398\_00` · HIGH confidence

CAUSE-EFFECT ERROR

- **Severity:** MEDIUM — a missing or incomplete answer — the model withheld an available answer (a gap, not a confidently wrong statement)
- **Category:** Cause & Effect
- **Question:** What is the consequence when an ALJ approves a request to withdraw objections to the Assistant Secretary's findings and/or order, and there are no other pending objections?
- ✓ **Ground truth:** The Assistant Secretary's findings and/or order will become the final order of the Secretary.
- ✗ **Model answered:** {"name": "ALJ_approves_withdrawal_of_objections", "parameters": {"result": "The Assistant Secretary"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly states, “the Assistant Secretary's findings and/or order will become the final order of the Secretary,” while the model answered only “The Assistant Secretary,” which is incomplete and does not state the consequence.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 34 — `OSHA\_29CFR\_full\_2024\_2438\_00` · HIGH confidence

CAUSE-EFFECT ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cause & Effect
- **Question:** If a removed employee files a claim for workers' compensation payments for a MDA-related disability, what must the employer do?
- ✓ **Ground truth:** The employer shall continue to provide medical removal protection benefits pending disposition of the claim.
- ✗ **Model answered:** {"name": "handle_workers_compensation_claim", "parameters": {"claim": "MDA-related disability"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model confused the process of handling workers
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 35 — `OSHA\_29CFR\_full\_2024\_neg\_1769\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What penalty does an employer face for failing to record a positive case in the COVID-19 log

within the 24-hour window?

- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "report_penalty", "parameters": {"amount": "10000"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated a penalty amount not present in the evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '10000' not in evidence

Failure 36 — `OSHA\_29CFR\_full\_2024\_neg\_1769\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What form number must employers use to record entries in the COVID-19 log?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "OSHA Forms", "parameters": {"number": "300, 300A, and 301"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 37 — `OSHA\_29CFR\_full\_2024\_neg\_2902\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Within how many minutes must dust spills outside an abrasive-blasting enclosure be cleaned up to comply with this section?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "clean_up_dust_spills", "parameters": {"minutes": "immediately"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only says dust spills 'shall be cleaned up promptly' and gives no number of minutes, while the model committed to 'immediately' instead of refusing as unanswerable.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 38 — `OSHA\_29CFR\_full\_2024\_1336\_00` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** Band wheels must be completely encased or guarded, but under what condition is part of the wheel allowed to be exposed?
- ✓ **Ground truth:** A portion of the upper wheel immediately around the point where the blade leaves the wheel may be exposed, to permit the operator to observe movement of equipment.
- ✗ **Model answered:** {"name": "band_wheels_encased_or_guarded", "parameters": {"condition": "to permit operator to observe movement of equipment"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 39 — `OSHA\_29CFR\_full\_2024\_2190\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If the employer monitored within one year prior to the standard's effective date and that monitoring satisfies all other requirements, must the employer still conduct new initial monitoring?
- **Ground truth:** No; the employer may rely on such earlier monitoring results to satisfy the initial monitoring requirement.
- **Model answered:** {"name": "initial monitoring", "parameters": {"satisfies_requirements": "True", "monitored_within_one_year": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 40 — `OSHA\_29CFR\_full\_2024\_neg\_2913\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which OSHA office is responsible for approving the design specifications of the fan plenum?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "Office of Variance Determination", "parameters": {"location": "Occupational Safety and Health Administration, U.S. Department of Labor, Room N3653, 200 Constitution Avenue, NW., Washington, DC 20210"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 41 — `OSHA\_29CFR\_full\_2024\_1202\_00` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** For transmission pipeline field shop operations described in paragraph (d)(1)(ii), which additional paragraph requirement applies that is NOT required under the general transmission pipeline requirements of paragraph (d)(1)(i)?
- **Ground truth:** Paragraph (d)(1)(i) general requirements observe paragraphs (b) and (c) and § 1910.254. Field shop operations in (d)(1)(ii) additionally require paragraph (a) and § 1910.253. So paragraph (a) (and § 1910.253) is the added requirement.
- **Model answered:** {"name": "paragraph requirement", "parameters": {"(d)(1)(ii)": "1910.253 and 1910.254"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model confused additional requirements for field shop operations
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 42 — `OSHA\_29CFR\_full\_2024\_neg\_4297\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How long after an employee's imminent danger report must the agency or committee respond before OSHA may inspect under paragraph (a)(3)?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "paragraph", "parameters": {"a": "3"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 43 — `OSHA\_29CFR\_full\_2024\_0455\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a metallic floorplate on which a container stands during filling is electrically connected to the fill stem, must the nozzle and container also be separately electrically interconnected to comply with the grounding section?
- ✓ **Ground truth:** No, that arrangement is deemed to have complied with the grounding provisions.
- ✗ **Model answered:** {"name": "is_grounding_required", "parameters": {"is_metallic_floorplate_connected": "true"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 44 — `OSHA\_29CFR\_full\_2024\_0517\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Considering the internal-combustion engine requirement in subparagraph (g)(2)(ii) and the fuel oil flow requirement in subparagraph (g)(2)(iii), what two fire-related safety measures are required regarding engines and fuel oil flow?
- ✓ **Ground truth:** Internal-combustion engines used for electric power generation must be located outside the mixing plant building, or be properly ventilated and isolated by a firewall, with exhaust systems located so spark emission cannot be a hazard; and suitable means must be provided to prevent the flow of fuel oil to the mixer in case of fire, with an automatic spring-loaded shutoff valve with fusible link installed in gravity flow systems.
- ✗ **Model answered:** {"name": "get_fire_safety_measures", "parameters": {"engine_requirement": "internal-combustion engine requirement in subparagraph (g)(2)(ii)", "fuel_oil_flow_requirement": "fuel oil flow

requirement in subparagraph (g)(2)(iii)"} }

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 45 — `OSHA\_29CFR\_full\_2024\_neg\_4303\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What form must a committee use to request inspection reports from an agency?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "request_inspection_report", "parameters": {"committee_name": "Establishment Committee", "agency_name": "Occupational Safety and Health Administration"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 46 — `OSHA\_29CFR\_full\_2024\_neg\_2902\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum thickness in millimeters required for the leather gloves worn by abrasive-blasting operators?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_minimum_thickness", "parameters": {"material": "leather", "gloves_type": "abrasive-blasting operators"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 47 — `OSHA\_29CFR\_full\_2024\_neg\_4297\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many days does the Secretary have to respond to a certified committee's request before conducting an inspection under § 1960.31(a)(2)?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "calculate_response_time", "parameters": {"days": "90"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 48 — `OSHA\_29CFR\_full\_2024\_neg\_2602\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which federal office is responsible for reviewing batching classification decisions made under A.0.5.1.2?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "Assistant Secretary", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 49 — `OSHA\_29CFR\_full\_2024\_0742\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** If an alarm is being repaired, how do the personnel requirements differ between subpart (i) and subpart (iii)?
- ✓ **Ground truth:** Subpart (i) generally requires appropriately trained persons for repairs, but subpart (iii) imposes a stricter standard for alarms, requiring they be adjusted or repaired only by the manufacturer or a technician trained by the manufacturer.
- ✗ **Model answered:** {"name": "compare_personnel_requirements", "parameters": {"subpart_i": "alarm repair", "subpart_iii": "alarm repair"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 50 — `OSHA\_29CFR\_full\_2024\_neg\_2913\_09` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What penalty applies if ventilation ductwork is improperly connected to a chimney used for products of combustion?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "violation_penalty", "parameters": {"penalty_amount": "1000", "penalty_reason": "improperly connected ventilation ductwork"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '1000' not in evidence

Failure 51 — `OSHA\_29CFR\_full\_2024\_neg\_3512\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum thickness required for the bulkheads used to contain compressed air?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "bulkhead_thickness", "parameters": {"wheel_diameter": "20", "wheel_hole_diameter": "12"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only says bulkheads containing compressed air must be tested to resist expected pressure and gives no minimum thickness, so the ground truth UNANSWERABLE is supported; the model did not refuse but returned unrelated fabricated parameters.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 52 — `OSHA\_29CFR\_full\_2024\_2565\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under what statute is a 'consumer product' defined for purposes of subpart (v), and under what statute is a 'hazardous substance' defined for purposes of subpart (6)(ii)?
- ✓ **Ground truth:** A 'consumer product' is defined under the Consumer Product Safety Act (15 U.S.C. 2051 et seq.) in subpart (v), while a 'hazardous substance' in subpart (6)(ii) is defined under the Comprehensive Environmental Response, Compensation and Liability Act (CERCLA) (42 U.S.C. 9601 et seq.).
- ✗ **Model answered:** {"name": "get_statute_definition", "parameters": {"term": "consumer product", "subpart": "v"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model only partially answered the question
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 53 — `OSHA\_29CFR\_full\_2024\_neg\_1738\_09` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the penalty for an employer who fails to designate a person as isolated during the potential transmission period?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "notify_employer_of_failure_to_correct_violation_and_propose_additional_penalty", "parameters": {"penalty_amount": "1000", "violation_description": "Failure to designate a person as isolated during the potential transmission period"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated

number: '1000' not in evidence

Failure 54 — `OSHA\_29CFR\_full\_2024\_3327\_02` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the maximum clearance allowed above the walking/working surface for a securely fastened toeboard?
- ✓ **Ground truth:** 1/4 inch (0.7 cm)
- ✗ **Model answered:** {"name": "get_clearance", "parameters": {"x": "0"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed ratio instead of quoting stated figure
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 55 — `OSHA\_29CFR\_full\_2024\_2138\_04` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** Under paragraph (l)(3)(iii), if initial monitoring shows CdU of 16 µg/g Cr, but the second set of monitoring results during the medical exam does not show any mandatory removal trigger level exceeded, must the employee be removed by the mandatory provisions?
- ✓ **Ground truth:** No; if the second set does not show a mandatory removal trigger level exceeded, the employee is not required to be removed by the mandatory provisions of this paragraph.
- ✗ **Model answered:** {"name": "paragraph", "parameters": {"l": "3", "i": "iii", "CdU": "16", "trigger_level": "15"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 56 — `OSHA\_29CFR\_full\_2024\_neg\_2902\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Who is responsible for supplying the heavy canvas or leather gloves to the operators?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "supply_gloves", "parameters": {"employer": "True"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only says operators 'shall be equipped with heavy canvas or leather gloves' and does not identify who must supply them, so the ground truth UNANSWERABLE is supported while the model incorrectly asserted 'employer'.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 57 — `OSHA\_29CFR\_full\_2024\_neg\_3653\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How often must the medical examination referenced in paragraph (h)(2)(ii) be repeated for employees using respirators?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "medical examination", "parameters": {"frequency": "every 12 months"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 58 — `OSHA\_29CFR\_full\_2024\_neg\_0484\_08` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How frequently must containers and air-storage tanks under air pressure be retested for maintenance under the ASME Code?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "retest_for_maintenance", "parameters": {"frequency": "at regular intervals"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 59 — `OSHA\_29CFR\_full\_2024\_neg\_0429\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum thickness required for the noncombustible surfacing in liquid oxygen leakage areas?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_min_thickness", "parameters": {"material": "noncombustible surfacing", "location": "liquid oxygen leakage areas"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 60 — `OSHA\_29CFR\_full\_2024\_3285\_01` · HIGH confidence

MISINTERPRETED

- **Severity:** MEDIUM — a misapplied or over-extrapolated rule — plausible but wrong

- **Category:** Conditional
- **Question:** If a platform's front edge is 16 inches from the face of the work and no guardrail or personal fall arrest system is used, does it comply with paragraph (b)(3)?
- ✓ **Ground truth:** No, because the front edge must not be more than 14 inches from the face of the work unless guardrails and/or personal fall arrest systems are used.
- ✗ **Model answered:** {"name": "check_compliance", "parameters": {"distance": "16"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model misinterpreted the requirement for platform front edge distance
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 61 — `OSHA\_29CFR\_full\_2024\_neg\_4297\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What form must a certified safety and health committee use to submit its inspection request to the Secretary?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "OSHA-20", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only says OSHA inspections may occur “in response to a request from half the membership of record of any certified safety and health committee” and does not identify any required form, so the model’s specific answer “OSHA-20” is unsupported while the ground truth UNANSWERABLE is correct.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 62 — `OSHA\_29CFR\_full\_2024\_neg\_1194\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum allowable concentration of combustible dust before welding is prohibited under the explosive atmospheres provision?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_maximum_allowable_concentration", "parameters": {"class": "Class II, Division 1", "factor": "combustible dust"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 63 — `OSHA\_29CFR\_full\_2024\_neg\_0429\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)

- **Question:** What minimum distance must a bulk oxygen system maintain from flammable gas lines?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "distance_from_exposures", "parameters": {"exposure_type": "flammable gas lines", "capacity": "less than 5000"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only says bulk oxygen systems must not be exposed to flammable gas lines and gives no minimum distance, so the supported answer is unanswerable, while the model did not refuse and instead returned an unsupported parameterized answer including an invented capacity of "less than 5000."
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 64 — `OSHA\_29CFR\_full\_2024\_neg\_4485\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the fee for filing a retaliation complaint under SPA?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_fee", "parameters": {"section": "1986.114", "complaint_type": "retaliation"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 65 — `OSHA\_29CFR\_full\_2024\_neg\_2913\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What minimum temperature rating must the automatic fire dampers be certified to withstand?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "automatic fire damper", "parameters": {"temperature_rating": "250"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '250' not in evidence

Failure 66 — `OSHA\_29CFR\_full\_2024\_3208\_02` · MEDIUM confidence

UNIT/FORMAT MISMATCH

- **Severity:** MEDIUM — a missing or incomplete answer — the model withheld an available answer (a gap, not a confidently wrong statement)
- **Category:** Conditional
- **Question:** If live parts are elevated above the floor to exclude unqualified persons, what minimum elevation is required for this to serve as adequate guarding?
- ✓ **Ground truth:** 8 feet (2.44 m) or more above the floor or other working surface.

- **✗ Model answered:** {"name": "elevation_for_guarding", "parameters": {"elevation": "8"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states guarding by elevation requires live parts to be at an "elevation of 8 feet (2.44 m) or more," while the model answered only "8" without the required unit, making the answer incomplete rather than semantically equivalent to the ground truth.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 67 — `OSHA\_29CFR\_full\_2024\_1848\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Based on the permissible exposure limit in Section I.D and the respirator requirements in Section III.A, at what airborne fiber concentration must you begin wearing a respirator, and what type of respirator is required if concentrations reach 3 f/cc?
- **✓ Ground truth:** You must wear a respirator when exposure exceeds the PEL of 0.1 f/cc. Since 3 f/cc exceeds 2 f/cc, an air-supplied, positive-pressure, full facepiece respirator must be used (HEPA air-purifying respirators are only allowed where concentrations do not exceed 2 f/cc).
- **✗ Model answered:** {"name": "respirator_requirements", "parameters": {"PEL": "0", "airborne_fiber_concentration": "3"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 68 — `OSHA\_29CFR\_full\_2024\_3282\_00` · HIGH confidence

EXCEPTION MISSED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Exception
- **Question:** What is the general capacity requirement for each scaffold and scaffold component, and under what conditions does it NOT apply?
- **✓ Ground truth:** Each scaffold and scaffold component shall be capable of supporting, without failure, its own weight and at least 4 times the maximum intended load applied or transmitted to it - except as provided in paragraphs (a)(2), (a)(3), (a)(4), (a)(5) and (g) of this section.
- **✗ Model answered:** {"name": "capacity_requirement", "parameters": {"scaffold_component": "scaffold", "capacity": "4 times the maximum intended load", "conditions": "aerial lifts"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states the requirement is to support "its own weight and at least 4 times the maximum intended load" except for paragraphs "(a)(2), (a)(3), (a)(4), (a)(5) and (g)," while the model omitted own weight and invented "aerial lifts" as the exception condition.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 69 — `OSHA\_29CFR\_full\_2024\_neg\_2129\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream

- **Category:** Negative (hallucination traps)
- **Question:** What is the cadmium action level concentration referenced in work operations under subsection (v)?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_cadmium_action_level", "parameters": {"concentration": "2.5"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 70 — `OSHA\_29CFR\_full\_2024\_4450\_03` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the Statutes at Large citation (124 Stat.) for the Health Care and Education Reconciliation Act of 2010, Public Law 111-152?
- ✓ **Ground truth:** 124 Stat. 1029
- ✗ **Model answered:** {"name": "get_statutes_at_large_citation", "parameters": {"public_law": "111-152", "act_name": "Health Care and Education Reconciliation Act of 2010"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not compute the Statutes at Large citation
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 71 — `OSHA\_29CFR\_full\_2024\_2261\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Conditional
- **Question:** When a healthcare professional is evaluating an employee after an exposure incident, is the employer required to provide the source individual's blood test results?
- ✓ **Ground truth:** Yes, but only if available.
- ✗ **Model answered:** {"name": "provide_source_individual_blood_test_results", "parameters": {"required": "true"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence requires providing the source individual's blood testing results only "if available," while the model answered simply that it is required, omitting the necessary condition reflected in the ground truth.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 72 — `OSHA\_29CFR\_full\_2024\_neg\_3512\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What air pressure value must compressed-air bulkheads be tested to in pounds per square inch?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "test_compressed_air_bulkheads", "parameters": {"pressure": "50"}}

- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only says bulkheads shall be tested to resist “the highest air pressure which may be expected to be used” and gives no psi value, while the model fabricated a specific value of 50 despite the ground truth being unanswerable.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 73 — `OSHA\_29CFR\_full\_2024\_2565\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining subparts (6)(i) and (6)(ii), which federal agency administers the regulations referenced in both exclusions, and which two statutes are involved?
- **Ground truth:** The Environmental Protection Agency administers the regulations referenced in both subparts. Subpart (6)(i) involves the Solid Waste Disposal Act as amended by the Resource Conservation and Recovery Act of 1976 (42 U.S.C. 6901 et seq.), and subpart (6)(ii) involves CERCLA (42 U.S.C. 9601 et seq.).
- **Model answered:** {"name": "Combining_subparts_6i_and_6ii", "parameters": {"agency": "Federal Occupational Safety and Health Administration", "statutes": "Williams-Steiger Occupational Safety and Health Act of 1970, Contract Work Hours and Safety Standards Act"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model confused federal agencies
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 74 — `OSHA\_29CFR\_full\_2024\_neg\_3512\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum height required for the standard railings installed along the walkways?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "standard railing height", "parameters": {"height": "42 inches"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 75 — `OSHA\_29CFR\_full\_2024\_neg\_4403\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which specific federal office or agency receives the complaints filed under the Pipeline Safety Act?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "OSHA Area Director", "parameters": {"address": "http://www.osha.gov"}}

- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 76 — `OSHA\_29CFR\_full\_2024\_1434\_04` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** May an employer use a method of calculating incident heat energy not included in appendix E?
- **Ground truth:** Yes, if the chosen method reasonably predicts the incident energy to which the employee would be exposed.
- **Model answered:** {"name": "incident_heat_energy_calculation", "parameters": {"method_included_in_appendix_E": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 77 — `OSHA\_29CFR\_full\_2024\_3038\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an employer evacuates all employees from the worksite when an emergency occurs and does not permit any employees to assist in handling the emergency, are they still required to comply with the emergency response program requirements of paragraph (p)(8)?
- **Ground truth:** No, they are exempt from paragraph (p)(8) provided they have an emergency action plan complying with § 1926.35.
- **Model answered:** {"name": "is_exempt_from_emergency_response_program", "parameters": {"evacuate_employees": "True", "permit_employee_assistance": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 78 — `OSHA\_29CFR\_full\_2024\_0325\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — a misapplied or over-extrapolated rule — plausible but wrong
- **Category:** Conditional
- **Question:** If an employer has 15 employees, does the option to communicate the fire prevention plan orally apply?
- **Ground truth:** No, the oral option only applies to employers with 10 or fewer employees, so this plan must be in writing, kept in the workplace, and made available to employees.
- **Model answered:** {"name": "check_oral_communication_eligibility", "parameters": {"number_of_employees": "15"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model misapplied the employee threshold for oral communication option
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** (none)

Failure 79 — `OSHA\_29CFR\_full\_2024\_2437\_08` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** When must the employer return an employee to former job status following temporary medical removal?
- ✓ **Ground truth:** When the employee no longer shows signs or symptoms of exposure to MDA, upon the advice of the physician, or when a subsequent final medical determination finds the employee no longer has a detected condition placing them at increased risk.
- ✗ **Model answered:** {"name": "return_to_former_job_status", "parameters": {"employee_removed_due_to_blood_lead_level": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 80 — `OSHA\_29CFR\_full\_2024\_0882\_04` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** The energy control program required by paragraph (c)(1) consists of three components. Which paragraphs of subsection (c) address the second and third of those components?
- ✓ **Ground truth:** Paragraph (c)(1) lists energy control procedures, employee training, and periodic inspections. Employee training is addressed in paragraph (c)(7) (Training and communication) and periodic inspections are addressed in paragraph (c)(6) (Periodic inspection).
- ✗ **Model answered:** {"name": "energy control program components", "parameters": {"c": "2-3"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 81 — `OSHA\_29CFR\_full\_2024\_1622\_01` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** Combining the support interval requirements for conductors smaller than No. 8 with the strap interval requirement for flexible nonmetallic tubing, what is the common maximum interval value that both share?
- ✓ **Ground truth:** Both share a maximum interval of 1.37 m (4.5 ft) - conductor supports at intervals not exceeding 1.37 m (4.5 ft) per paragraph (ii)(C), and tubing straps at intervals not exceeding 1.37 m (4.5 ft) per paragraph (iii).
- ✗ **Model answered:** {"name": "max", "parameters": {"x": "305", "y": "152", "z": "137"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed ratio instead of quoting stated figure
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** Fabricated number: '137' not in evidence

Failure 82 — `OSHA\_29CFR\_full\_2024\_neg\_2129\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which OSHA office reviews the examining physician's determination that an employee cannot function normally with a respirator?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "OSHA_review", "parameters": {"office": "Assistant Secretary"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 83 — `OSHA\_29CFR\_full\_2024\_3059\_00` · HIGH confidence

MISINTERPRETED

- **Severity:** MEDIUM — a misapplied or over-extrapolated rule — plausible but wrong
- **Category:** Conditional
- **Question:** If a spray material is known to be highly susceptible to spontaneous heating and ignition, may filter rolls be used when applying it?
- ✓ **Ground truth:** No, filters or filter rolls shall not be used when applying such a material.
- ✗ **Model answered:** {"name": "filter_roll_usage", "parameters": {"susceptible": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model misinterpreted the filter roll usage rule
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 84 — `OSHA\_29CFR\_full\_2024\_4294\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining paragraph (c) and paragraph (d)(2), what minimum information about the reporting individual's identity may or may not be reflected in the establishment's log if confidentiality is requested?
- ✓ **Ground truth:** The log (per (d)(2)) requires date, time, code/reference/file number, location, description, classification, and action taken, but does NOT list the employee's name; under (c), upon request no person shall disclose the name of the individual making the report or names of employees referred to, except to authorized representatives of the Secretary. So the name is kept confidential and is not among the required log items.
- ✗ **Model answered:** {"name": "minimum_information_about_reporting_individual", "parameters": {"paragraph_c": "paragraph (c)", "paragraph_d2": "paragraph (d)(2)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 85 — `OSHA\_29CFR\_full\_2024\_neg\_3653\_08` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What exposure level triggers respirator use during the Class IV asbestos work described in subparagraph (vi)?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "respirator_use", "parameters": {"exposure_level": "above the TWA or excursion limit"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 86 — `OSHA\_29CFR\_full\_2024\_3641\_03` · MEDIUM confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Considering both the glove bag work practice for adjacent material and the negative pressure glove box work practice for adjacent ACM, how must loose or damaged material near the enclosure be handled in each case?
- ✓ **Ground truth:** In both cases, the loose friable or damaged material adjacent to the operation shall be wrapped and sealed in two layers of six mil plastic (or otherwise rendered intact) before beginning the job.
- ✗ **Model answered:** [{"name": "wrap_and_seal_loose_ACM", "parameters": {"method": "glove bag work practice for adjacent material"}}; {"name": "wrap_and_seal_loose_ACM", "parameters": {"method": "negative pressure glove box work practice for adjacent ACM"}}]
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 87 — `OSHA\_29CFR\_full\_2024\_neg\_3512\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How frequently must intermediate bulkheads with locks be re-inspected after installation?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "inspect_intermediate_bulkheads", "parameters": {"frequency": "at least once a week"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 88 — `OSHA\_29CFR\_full\_2024\_0742\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** Combining the requirements for who may make repairs and how repairs must be conducted, what two standards govern a general repair to a respirator?
- ✓ **Ground truth:** Per subpart (i), repairs may be made only by persons appropriately trained to perform such operations using only the manufacturer's NIOSH-approved parts; and per subpart (ii), repairs shall be made according to the manufacturer's recommendations and specifications for the type and extent of repairs to be performed.
- ✗ **Model answered:** {"name": "combines_requirements_for_repairs", "parameters": {"employee": "authorized personnel", "location": "designated for such repairs"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence requires repairs be made only by appropriately trained persons using manufacturer NIOSH-approved parts and according to the manufacturer's recommendations/specifications, while the model instead gives 'authorized personnel' and an unsupported 'designated' location, omitting the governing repair standards.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 89 — `OSHA\_29CFR\_full\_2024\_3162\_03` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If the maximum allowable working pressure of a receiver is 100 psi, what is the highest pressure the safety valves may allow before they must relieve?
- ✓ **Ground truth:** 110 psi, since pressure may not exceed the maximum allowable working pressure by more than 10 percent.
- ✗ **Model answered:** {"name": "calculate_highest_pressure", "parameters": {"x": "100"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed ratio instead of quoting stated figure
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 90 — `OSHA\_29CFR\_full\_2024\_neg\_2913\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How frequently must the inspection or clean-out doors themselves be inspected for proper functioning?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "inspect_inspection_or_clean_out_doors", "parameters": {"frequency": "every 30 days"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 91 — `OSHA\_29CFR\_full\_2024\_4446\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a timely petition for review is filed and the case was heard before an ALJ, who transmits the record to the appropriate court?
- ✓ **Ground truth:** The ARB or the ALJ, as the case may be, transmits the record (including the record of proceedings before the ALJ).
- ✗ **Model answered:** {"name": "transmit_record", "parameters": {"person": "ALJ"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not fully capture the conditional transmission of the record by both ARB and ALJ
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 92 — `OSHA\_29CFR\_full\_2024\_neg\_0484\_04` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum number of shifts' worth of flammable liquid supply permitted to be kept near spraying operations before bulk storage rules apply?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "bulk_storage", "parameters": {"quantity": "1"}}
- **Doctor diagnosis:** **Dr. No** bad test
- **Doctor root cause:** [verifier] The evidence directly states the quantity near spraying operations "should ordinarily not exceed a supply for 1 day or one shift," so the model's answer of 1 shift is supported and the UNANSWERABLE ground truth is not.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 93 — `OSHA\_29CFR\_full\_2024\_3959\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If equipment operations would breach the Table A minimum approach distance, but the employer determines it IS feasible to do the work without breaching that distance, is the operation permitted under the exception in this section?
- ✓ **Ground truth:** No; the exception requires the employer demonstrate it is infeasible to do the work without breaching the minimum approach distance (paragraph (a)), so if it is feasible, the operation closer than Table A is prohibited.
- ✗ **Model answered:** {"name": "is_operation_permitted", "parameters": {"determines_infeasible": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 94 — `OSHA\_29CFR\_full\_2024\_neg\_1599\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum load in amperes for a receptacle on a 50-ampere branch circuit?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_maximum_load", "parameters": {"circuit_rating": "50"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 95 — `OSHA\_29CFR\_full\_2024\_3312\_04` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a scaffold must be closer to power lines than the specified minimum distance because the clearance is necessary for performance of work, when is that permitted?
- ✓ **Ground truth:** Only after the utility company or electrical system operator has been notified and has deenergized the lines, relocated them, or installed protective coverings to prevent accidental contact.
- ✗ **Model answered:** {"name": "exception_to_paragraph", "parameters": {"f": "6", "clearance": "necessary for performance of work"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 96 — `OSHA\_29CFR\_full\_2024\_4456\_00` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** When a timely objection is filed, are all provisions of the preliminary order stayed?
- ✓ **Ground truth:** No. All provisions are stayed except for the portion requiring preliminary reinstatement, which is not automatically stayed and remains effective immediately upon the respondent's receipt of the findings and preliminary order.
- ✗ **Model answered:** {"name": "objections_to_findings_and_preliminary_order", "parameters": {"timely_objection": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 97 — `OSHA\_29CFR\_full\_2024\_0494\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually

- **Category:** Conditional
- **Question:** If undercoating spray operations are conducted in a garage with adequate natural ventilation using solvents listed as having a flash point in excess of 100 °F, are they exempt from spray finishing requirements?
- ✓ **Ground truth:** Yes, they are exempt from the requirements pertaining to spray finishing operations.
- ✗ **Model answered:** {"name": "k", "parameters": {"Underwriters": " Laboratories in respect to fire hazard rating 30-40", "flash point": "100"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 98 — `OSHA\_29CFR\_full\_2024\_2920\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** According to paragraphs (viii) and (ix), what must be provided to a worker who reports a small skin abrasion, and who treats it?
- ✓ **Ground truth:** Per paragraph (viii), any small skin abrasions, cuts, rash, or open sores which are found or reported shall be treated by a properly designated person so that chances of exposures to the chemicals are removed. Per paragraph (ix), the worker must also have access to sufficient washing facilities, including soap, individual towels, and hot water, provided on the basis of at least one basin with a hot water faucet for every 10 employees.
- ✗ **Model answered:** {"name": "provide_treatment_and_first_aid", "parameters": {"treatment": "treatment by a properly designated person", "first_aid_supplies": "appropriate first-aid supplies"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not fully synthesize rules from multiple regulatory sections
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 99 — `OSHA\_29CFR\_full\_2024\_neg\_1738\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum negative pressure differential (in inches of water) that an AIIR must maintain?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "pressure_differential", "parameters": {"pressure": "0"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 100 — `OSHA\_29CFR\_full\_2024\_neg\_2913\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream

- **Category:** Negative (hallucination traps)
- **Question:** What minimum air velocity (in feet per minute) must be maintained through the spray booth ventilation system?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "minimum_air_velocity", "parameters": {"velocity": "100"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 101 — `OSHA\_29CFR\_full\_2024\_3468\_01` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an air quality test record relates to exposure to a toxic substance, how long must it be retained compared to a record that does not involve a toxic substance?
- ✓ **Ground truth:** Toxic substance exposure records are retained in accordance with § 1910.33, while all other air quality test records are retained only until completion of the project.
- ✗ **Model answered:** {"name": "recordkeeping", "parameters": {"toxic_substance_record": "30 years", "non_toxic_substance_record": "completion of the project"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states only that toxic-substance exposure records are retained "in accordance with § 1910.33" and all other air quality test records "until completion of the project"; the model's specific "30 years" is not supported by the provided evidence.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 102 — `OSHA\_29CFR\_full\_2024\_neg\_3806\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many days does the employer have to forward the information to the licensed physician before the examination?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "information_to_PLHCP", "parameters": {"days": "15"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 103 — `OSHA\_29CFR\_full\_2024\_neg\_1738\_08` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many square feet must an ambulatory surgical center provide per patient to qualify as an ambulatory care setting?

- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "ambulatory care setting space requirements", "parameters": {"square feet per patient": "100"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '100' not in evidence

Failure 104 — `OSHA\_29CFR\_full\_2024\_2784\_01` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an OSHA regulated substance has no action level, what triggers application of paragraphs (d) and (g)(1)(ii)?
- ✓ **Ground truth:** Routinely exceeding the permissible exposure limit (in the absence of an action level), for a substance with exposure monitoring and medical surveillance requirements.
- ✗ **Model answered:** {"name": "paragraphs (d) and (g)(1)(ii) trigger", "parameters": {"action_level": "null", "PEL": "permissible exposure limit"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states that, absent an action level, paragraphs (d) and (g)(1)(ii) apply when 'the permissible exposure limit is routinely exceeded' for a substance with exposure monitoring and medical surveillance requirements; the model only mentions a null action level and PEL without the required 'routinely exceeded' condition.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 105 — `OSHA\_29CFR\_full\_2024\_1317\_01` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** How does the definition of 'running line' distinguish it from other ropes?
- ✓ **Ground truth:** A running line means any moving rope as distinguished from a stationary rope such as a guyline.
- ✗ **Model answered:** {"name": "definition_of_running_line", "parameters": {"rope_type": "running line"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 106 — `OSHA\_29CFR\_full\_2024\_4010\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an auditor determines that the written and practical tests are valid and reliable but the auditor is an employee of the employer, can the auditor's approval satisfy the test development requirements?
- ✓ **Ground truth:** No, because paragraph (e)(1)(ii)(B) requires that the auditor is not an employee of the employer.

- **✗ Model answered:** {"name": "auditor_approval", "parameters": {"valid": "True", "reliable": "True", "auditor_employer": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model confused auditor
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 107 — `OSHA\_29CFR\_full\_2024\_neg\_1738\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What form must be used to document a close contact exposure under this section?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** {"name": "OSHA 301", "parameters": {"form type": "Injury and Illness Incident Report"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 108 — `OSHA\_29CFR\_full\_2024\_neg\_1037\_10` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which OSHA form must the employer use to record a worker as 'designated'?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** {"name": "OSHA 301", "parameters": {"form": "Injury and Illness Incident Report"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 109 — `OSHA\_29CFR\_full\_2024\_neg\_4485\_09` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which OSHA office is designated to handle complaints from seamen who reside outside the United States?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** {"name": "OSHA office responsible for enforcement activities", "parameters": {"geographical area": "outside the United States"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 110 — `OSHA\_29CFR\_full\_2024\_neg\_2625\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum exposure concentration that triggers respiratory sensitizer classification under these criteria?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "classification_criteria_for_mixtures", "parameters": {"concentration_limit": "0"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 111 — `OSHA\_29CFR\_full\_2024\_neg\_1194\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What fire-resistance rating must the shields used near combustible walls meet?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_fire_resistance_rating", "parameters": {"material": "shields used near combustible walls"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 112 — `OSHA\_29CFR\_full\_2024\_neg\_1599\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** For a 60-ampere branch circuit supplying two or more receptacles, what receptacle ratings are listed in Table S-5?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "Table S-5", "parameters": {"circuit_rating_amperes": "60"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 113 — `OSHA\_29CFR\_full\_2024\_neg\_2902\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as

authoritative and is hard to catch downstream

- **Category:** Negative (hallucination traps)
- **Question:** What penalty is imposed on an operator who fails to wear the required safety shoes?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "impose_penalty", "parameters": {"amount": "1000", "imprisonment": "6 months"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated penalty for non-existent rule
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 114 — `OSHA\_29CFR\_full\_2024\_neg\_0484\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum capacity allowed for containers supplying spray nozzles by means other than gravity flow or air pressure?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_max_capacity", "parameters": {"container_type": "closed type or provided with metal covers kept closed"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 115 — `OSHA\_29CFR\_full\_2024\_neg\_1769\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which OSHA office must employers submit their COVID-19 log to for review?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "OSHA Forms 300, 300A, and 301", "parameters": {"employer": "Assistant Secretary"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 116 — `OSHA\_29CFR\_full\_2024\_neg\_1037\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What load multiple is applied as the proof load relative to the rated capacity?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "proof_test", "parameters": {"factor": "1"}}

- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 117 — `OSHA\_29CFR\_full\_2024\_0022\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If the Assistant Secretary publishes a notice in the Federal Register on a given date, by what deadline must the applying State agency publish reasonable notice containing the same information within the State?
- ✓ **Ground truth:** No later than 5 days following the publication of the notice in the Federal Register.
- ✗ **Model answered:** {"name": "publish_notice", "parameters": {"days": "10"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed incorrect deadline
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 118 — `OSHA\_29CFR\_full\_2024\_neg\_0484\_10` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum ventilation airflow rate required in the mixing room when transferring these liquids?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_ventilation_airflow_rate", "parameters": {"room": "mixing room", "liquids": "flammable liquids or liquids with a flashpoint greater than 199.4 \u00b0F (93 \u00b0C)"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 119 — `OSHA\_29CFR\_full\_2024\_3468\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Conditional
- **Question:** In an underground drill-and-blast operation, if employees are about to re-enter the area after blasting, must the air be tested for flammable gas, and is any further testing required while they work?
- ✓ **Ground truth:** Yes, the air must be tested for flammable gas prior to re-entry after blasting, and tested continuously while employees are working underground.
- ✗ **Model answered:** {"name": "test_for_flammable_gas", "parameters": {"re_entry": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence says air must be tested 'prior to re-entry after blasting, and continuously when employees are working underground,' but the model only indicated testing for re-entry and omitted the continuous testing requirement.
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** (none)

Failure 120 — `OSHA\_29CFR\_full\_2024\_neg\_2129\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many days does the employer have to provide a medical examination after an employee reports breathing difficulty during respirator use?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "paragraph (I)(6)(iii)", "parameters": {"days": "as soon as possible"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 121 — `OSHA\_29CFR\_full\_2024\_neg\_1599\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** By what date must existing receptacle installations be brought into compliance with these rating provisions?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_installation_compliance_date", "parameters": {"year": "1990", "month": "7", "day": "23"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 122 — `OSHA\_29CFR\_full\_2024\_3172\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** For a horizontal belt where both runs are 36 inches from the floor, what guarding is required, and how does this differ from the general rule for runs that are 7 feet or less from the floor?
- ✓ **Ground truth:** Under paragraph (e)(1)(i), where both runs are 42 inches or less from the floor, the belt shall be fully enclosed. Since 36 inches is 42 inches or less, the belt must be fully enclosed. This differs from the general rule for runs 7 feet or less from the floor, where the guard need only extend to at least 15 inches above the belt or to a standard height.
- ✗ **Model answered:** {"name": "belt_guarding", "parameters": {"distance_from_floor": "36", "belt_width": "8"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 123 — `OSHA\_29CFR\_full\_2024\_neg\_0484\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum steel pipe wall thickness required for piping systems conveying these flammable liquids?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_minimum_wall_thickness", "parameters": {"material": "steel"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 124 — `OSHA\_29CFR\_full\_2024\_4486\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under paragraph (a), the respondent is notified of its rights under paragraphs (b) and (f). Combining these, what are the two separate time limits the respondent has for responding under each of those paragraphs?
- ✓ **Ground truth:** Under paragraph (b), the respondent has 20 days of receipt of the notice of the filing of the complaint to submit a written statement, affidavits, or documents and to request a meeting. Under paragraph (f), the respondent must present its evidence within 10 business days of the Assistant Secretary's notification (or as soon thereafter as agreed if the interests of justice require).
- ✗ **Model answered:** {"name": "respondent_rights", "parameters": {"b": "20 days", "f": "20 days"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 125 — `OSHA\_29CFR\_full\_2024\_3130\_08` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the maximum capacity (in gallons) of a safety can?
- ✓ **Ground truth:** not more than 5 gallons
- ✗ **Model answered:** {"name": "get_max_capacity", "parameters": {"container_type": "safety can"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 126 — `OSHA\_29CFR\_full\_2024\_3390\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually

- **Category:** Cross-Reference
- **Question:** For a precast concrete member that is 150 feet long, paragraph (ii) governs the maximum distance. How does this maximum compare to the general maximum given in paragraph (i)?
- ✓ **Ground truth:** For a 150-foot member, half the length is 75 feet, but the cap is 60 feet (whichever is less), so the maximum is 60 feet under paragraph (ii). This is greater than the 25-foot general maximum given in paragraph (i).
- ✗ **Model answered:** {"name": "compare_maxima", "parameters": {"x": "150"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 127 — `OSHA\_29CFR\_full\_2024\_4138\_00` · HIGH confidence

TABLE MISREAD

- **Severity:** MEDIUM — a misapplied or over-extrapolated rule — plausible but wrong
- **Category:** Cross-Reference
- **Question:** Both Option (2) corraling and Option (4) centerline cable system aim to prevent shifting, but they impose different standards on whether any movement is tolerated. What does each require regarding movement during operation?
- ✓ **Ground truth:** Option (2) corraling requires that the corraling system does not allow the equipment to shift by any amount of shifting in any direction. Option (4) centerline cable system requires that the crane/derrick is secured from movement during operation (per subpart (E)), though its attachment method must also allow the crew to move the crane/derrick longitudinally for repositioning.
- ✗ **Model answered:** {"name": "corraling_system_requirements", "parameters": {"system_type": "Option (2)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model read from wrong table row
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 128 — `OSHA\_29CFR\_full\_2024\_neg\_1037\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Within how many days must a sling manufacturer document a completed proof test?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "document_proof_test", "parameters": {"days": "7"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 129 — `OSHA\_29CFR\_full\_2024\_3948\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** A powered industrial truck (forklift) is configured to hoist and lower by means of a winch and

horizontally move a suspended load. Is it excluded under (c)(8)?

- ✓ **Ground truth:** No; powered industrial trucks are excluded except when configured to hoist and lower (by winch or hook) and horizontally move a suspended load.
- ✗ **Model answered:** {"name": "check_exclusion", "parameters": {"i": "8", "condition": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 130 — `OSHA\_29CFR\_full\_2024\_neg\_4403\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What monetary penalty may be imposed on a pipeline operator found to have discriminated against an employee?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "adjusted civil monetary penalties", "parameters": {"penalty per willful violation under section 17": "161323"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 131 — `OSHA\_29CFR\_full\_2024\_4269\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** According to the purpose and scope, public employees were not assured protections under the Act prior to a State plan because of a definition in one section, but a different section provides the route for ensuring their protection. Which sections are these, and what does each establish?
- ✓ **Ground truth:** Section 3(5) excludes State and local government agencies from the definition of 'employer,' which is why such public employees had not heretofore been assured any protections under the Act. Section 18 (and specifically section 18(b)) is the only provision under which such public employees are ensured protection through an approved State plan, permitting States to submit plans for any occupational safety and health issue for which a Federal standard has been promulgated under section 6.
- ✗ **Model answered:** {"name": "get_sections", "parameters": {"section1": "section 3(5)", "section2": "section 18(b)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports section 3(5) because it excludes State and local government agencies from the definition of "employer," and section 18/18(b) because only under section 18 are public employees protected via an approved State plan, but the model only listed section numbers and omitted what each establishes as the question required.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 132 — `OSHA\_29CFR\_full\_2024\_1434\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an employee performs work within reaching distance of exposed energized parts while wearing a wristwatch that does not increase the hazards associated with contact, must the employer ensure it is removed or rendered nonconductive?
- ✓ **Ground truth:** No; the requirement to remove or render nonconductive does not apply if such articles do not increase the hazards associated with contact with the energized parts.
- ✗ **Model answered:** {"name": "remove_conductive_articles", "parameters": {"article": "wristwatch", "hazard_increase": "false"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 133 — `OSHA\_29CFR\_full\_2024\_neg\_2602\_06` · MEDIUM confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the citation reference for the additivity formula mentioned as A.1.3.6 in another regulatory code?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_citation_reference", "parameters": {"code": "A.1.3.6", "reference": "additivity formula"}}
- **Doctor diagnosis:** **Dr. No** bad test
- **Doctor root cause:** [verifier] The evidence explicitly identifies the additivity formula as “paragraph A.1.3.6 (the additivity formula),” so a citation reference of A.1.3.6 is supported and the ground truth UNANSWERABLE is not.
- **Verifier verdict:** bad test (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 134 — `OSHA\_29CFR\_full\_2024\_3423\_00` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the minimum factor by which jacking equipment must be capable of supporting the load being lifted during jacking operations?
- ✓ **Ground truth:** two and one-half times
- ✗ **Model answered:** {"name": "get_minimum_support_factor", "parameters": {"factor": "2"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states that jacking equipment must support "at least two and one-half times the load," so the required factor is 2.5, while the model answered 2.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 135 — `OSHA\_29CFR\_full\_2024\_3038\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an employer arranges in advance for an outside fully-trained emergency response team to respond in a reasonable period, must all employees still be trained to the full degree specified?
- ✓ **Ground truth:** No, provided all employees who may come to the incident first have sufficient awareness training and have been instructed to call the designated outside team.
- ✗ **Model answered:** {"name": "emergency_response_to_hazardous_substance_releases", "parameters": {"arrangements_made": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 136 — `OSHA\_29CFR\_full\_2024\_neg\_3653\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What civil penalty applies to an employer who selects filtering facepiece respirators against asbestos fibers in violation of this section?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "adjusted civil monetary penalties", "parameters": {"penalty_per_willful_violation": "161323"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 137 — `OSHA\_29CFR\_full\_2024\_2016\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Appendix B states the regulation applies to records of employees exposed to toxic substances or harmful physical agents, and references the non-mandatory nature established in paragraph (i). Given paragraph (i)'s statement about appendices and Appendix B's description of RTECS, does an employer have a legal obligation to purchase RTECS to determine if the rule applies?
- ✓ **Ground truth:** No. Paragraph (i) states the appendices do not, by themselves, create any additional obligations not otherwise imposed by the section, and Appendix B explicitly notes that the final regulation does not require that employers purchase a copy of RTECS, and many employers need not consult RTECS to ascertain whether their records are subject to the rule.
- ✗ **Model answered:** {"name": "does_employer_have_legal_obligation_to_purchase_rtecs", "parameters": {"appendix_b": "Appendix B to \u00a7 1910.1020\u2014Availability of NIOSH Registry of Toxic Effects of Chemical Substances (RTECS) (Non-Mandatory)", "paragraph_i": "(i)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** (none)

Failure 138 — `OSHA\_29CFR\_full\_2024\_neg\_4455\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the mailing address of the Chief Administrative Law Judge that must appear in the findings?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_address_of_chief_administrative_law_judge", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 139 — `OSHA\_29CFR\_full\_2024\_0362\_01` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** Under what condition is wire rope permitted for use as the emergency stop rope?
- ✓ **Ground truth:** Wire rope may be used only if it is marlin-covered; unless marlin-covered, it shall not be used.
- ✗ **Model answered:** {"name": "is_wire_rope_permitted", "parameters": {"condition": "as the emergency stop rope"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 140 — `OSHA\_29CFR\_full\_2024\_neg\_3512\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Who is responsible for certifying that the compressed-air bulkhead test was successfully completed?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "compressed-air_bulkhead_test_certification", "parameters": {"requester": "appointed physician"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only says bulkheads 'shall be tested' to prove pressure resistance and does not identify anyone responsible for certifying completion, while the model fabricated 'appointed physician' instead of treating it as unanswerable.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 141 — `OSHA\_29CFR\_full\_2024\_neg\_1037\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What certification must an equivalent entity hold to be recognized for performing repairs?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_certification", "parameters": {"entity_type": "equivalent entity", "certification_type": "written certification"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 142 — `OSHA\_29CFR\_full\_2024\_neg\_0429\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What penalty applies to operators who store oxygen in a building used for non-exclusive purposes?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "penalty_for_non_compliant_storage", "parameters": {"amount": "10000", "hazard": "oxygen storage in non-exclusive building"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated a penalty amount not present in the evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '10000' not in evidence

Failure 143 — `OSHA\_29CFR\_full\_2024\_1695\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** A circuit has been deenergized by a circuit protective device and a worker wants to reenergize it. Combining paragraph (b)(2) and the following Note, when may the circuit be reenergized without examining the circuit or connected equipment?
- ✓ **Ground truth:** Per (b)(2) the circuit may not be manually reenergized until it has been determined the equipment and circuit can be safely energized; however, per the Note, no examination is needed before reenergizing when it can be determined from the design of the circuit and the overcurrent devices that the automatic operation was caused by an overload rather than a fault condition.
- ✗ **Model answered:** {"name": "Reenergizing circuit", "parameters": {"circuit": "deenergized circuit", "condition": "overload rather than a fault condition"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 144 — `OSHA\_29CFR\_full\_2024\_3312\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If conductive material is being handled on a scaffold near an uninsulated power line of less than 50 kv, what is the minimum required clearance?
- ✓ **Ground truth:** 10 feet (3.1 m).
- ✗ **Model answered:** {"name": "clearance_distance", "parameters": {"voltage": "less than 50 kv", "power_line": "uninsulated"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 145 — `OSHA\_29CFR\_full\_2024\_2437\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an examining physician determines that an employee's abnormal liver function tests are NOT associated with MDA exposure, can the employee still be temporarily removed?
- ✓ **Ground truth:** Yes, if the physician also determines the abnormalities may be exacerbated as a result of occupational exposure to MDA.
- ✗ **Model answered:** {"name": "temporary_removal_due_to_abnormal_liver_function_tests", "parameters": {"associated_with_mda_exposure": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 146 — `OSHA\_29CFR\_full\_2024\_neg\_1769\_09` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Within how many days must an employer notify other employees of a potential workplace exposure identified through the log?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "notify_other_employees", "parameters": {"days": "1"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 147 — `OSHA\_29CFR\_full\_2024\_neg\_2129\_08` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as

authoritative and is hard to catch downstream

- **Category:** Negative (hallucination traps)
- **Question:** What form must the employer use to document the physician's respirator-use determination?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "Appendix C", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 148 — `OSHA\_29CFR\_full\_2024\_neg\_3653\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the duration of air supply required for the auxiliary positive-pressure SCBA used during Class I asbestos work above 1 f/cc?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_air_supply_duration", "parameters": {"type": "auxiliary positive-pressure SCBA", "work_class": "Class I", "asbestos_concentration": "above 1 f/cc"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 149 — `OSHA\_29CFR\_full\_2024\_3020\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining subpart (vi) and subpart (ix), what must happen before a drum that cannot be moved without rupture is emptied?
- ✓ **Ground truth:** Before any movement of drums or containers, all employees exposed to the transfer operation must be warned of the potential hazards associated with the contents (vi); and a drum that cannot be moved without rupture, leakage, or spillage must be emptied into a sound container using a device classified for the material being transferred (ix).
- ✗ **Model answered:** {"name": "emptying", "parameters": {"drum": "sound container", "device": "classified for the material being transferred"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 150 — `OSHA\_29CFR\_full\_2024\_neg\_2625\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the reference number or citation of the footnote marked '1' regarding appropriate animal

tests?

- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_footnote_reference", "parameters": {"n": "1"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 151 — `OSHA\_29CFR\_full\_2024\_3047\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Although the Note states the appendices serve as non-mandatory guidelines, which paragraph does it identify as making the use of Level A and Level B PPE protection mandatory in certain circumstances?
- ✓ **Ground truth:** Paragraph § 1926.65(g) makes mandatory in certain circumstances the use of Level A and Level B PPE protection.
- ✗ **Model answered:** {"name": "paragraph", "parameters": {"number": "1910.120(g)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model read from wrong paragraph number
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 152 — `OSHA\_29CFR\_full\_2024\_neg\_4455\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What penalty applies to a respondent who fails to comply with the immediately effective reinstatement requirement?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "find_penalty", "parameters": {"respondent_action": "fails to comply with the immediately effective reinstatement requirement"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 153 — `OSHA\_29CFR\_full\_2024\_1848\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Using the recordkeeping requirements in Section V.C and the latency information in Section II.A, how long must medical records be kept, and how does this relate to when disease symptoms typically appear?
- ✓ **Ground truth:** Medical records must be kept for the period of employment plus thirty (30) years (V.C), which accommodates the fact that symptoms of asbestos-related diseases generally do not appear for 20 or more years after initial exposure (II.A).

- **✗ Model answered:** {"name": "calculate_medical_record_keeping_time", "parameters": {"medical_record_keeping_time": "30", "latency_period": "20"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 154 — `OSHA\_29CFR\_full\_2024\_0694\_01` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the toilet facility table in (n)(3)(i), how many toilet seats and urinals are required for a worksite with 160 employees versus one with 250 employees?
- **✓ Ground truth:** For 160 employees (more than 20, fewer than 200): one toilet seat and one urinal per 40 employees = 4 toilet seats and 4 urinals. For 250 employees (more than 200): one toilet seat and one urinal per 50 employees = 5 toilet seats and 5 urinals.
- **✗ Model answered:** {"name": "calculate_toilet_facilities", "parameters": {"employees": "160"}}; {"name": "calculate_toilet_facilities", "parameters": {"employees": "250"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states for 'More than 20, fewer than 200' there is 'One toilet seat and one urinal per 40 employees' and for 'More than 200' per 50 employees, yielding 4/4 for 160 and 5/5 for 250, while the model only emitted function calls without the required numbers.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 155 — `OSHA\_29CFR\_full\_2024\_3468\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If the Secretary of Labor requests an air quality test record, where must it have been maintained and what details must it contain?
- **✓ Ground truth:** It must be maintained above ground at the worksite and include the location, date, time, substance and amount monitored.
- **✗ Model answered:** {"name": "get_air_quality_test_record", "parameters": {"location": "on-site", "details": "date, time, and results of the test"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not fully capture the required details of the air quality test record
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 156 — `OSHA\_29CFR\_full\_2024\_4502\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under paragraph (a) materials are redacted consistent with the Privacy Act, yet paragraph (d) addresses confidentiality of certain persons. How do these provisions treat the complainant differently regarding confidentiality protection?
- **✓ Ground truth:** Paragraph (a) redactions are made consistent with the Privacy Act of 1974 and other

confidentiality laws, but the complainant (or counsel) receives an unredacted copy of those materials. Paragraph (d) requires investigations to protect the confidentiality of any person who provides information on a confidential basis other than the complainant—so the complainant is expressly excluded from that confidential-informant protection.

- **✗ Model answered:** {"name": "redact_complainant_info", "parameters": {"complainant_info": "false"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 157 — `OSHA\_29CFR\_full\_2024\_neg\_2129\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What efficiency percentage must HEPA filters meet to qualify under paragraph (3)(i)(C)?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** {"name": "HEPA filter efficiency", "parameters": {"paragraph": "(3)(i)(C)", "efficiency": "99.97%"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 158 — `OSHA\_29CFR\_full\_2024\_4031\_00` · HIGH confidence

CALCULATION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Calculation
- **Question:** A crane has a rated capacity of 60,000 pounds for the current radius and configuration. When using a suspended personnel platform, what is the maximum total load (platform loaded, including hook, load line and rigging) permitted, except during proof testing?
- **✓ Ground truth:** 30,000 pounds
- **✗ Model answered:** {"name": "calculate_total_load", "parameters": {"rated_capacity": "60000"}}
- **Doctor diagnosis:** **Dr. Pi** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 159 — `OSHA\_29CFR\_full\_2024\_3938\_02` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** Under § 1926.1204(e)(1)(ii), continuous monitoring of entry conditions is required where entrants are working. Under what exception may employers instead use periodic monitoring?
- **✓ Ground truth:** Employers may use periodic monitoring in accordance with paragraph (e)(2) for monitoring an atmospheric hazard if they can demonstrate that equipment for continuously monitoring that hazard is not commercially available.
- **✗ Model answered:** {"name": "periodic monitoring", "parameters": {"paragraph": "(e)(2)"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 160 — `OSHA\_29CFR\_full\_2024\_neg\_0429\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How often must safety devices in a bulk oxygen system be inspected?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "inspect_bulk_oxygen_system", "parameters": {"frequency": "daily"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 161 — `OSHA\_29CFR\_full\_2024\_3208\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If electric equipment is located where it would be exposed to physical damage, what requirement applies to the enclosures or guards?
- ✓ **Ground truth:** They must be arranged and of such strength as to prevent such damage.
- ✗ **Model answered:** {"name": "requirement_for_enclosures_or_guards", "parameters": {"location": "where electric equipment is likely to be exposed to physical damage"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly states that in such locations, “enclosures or guards shall be so arranged and of such strength as to prevent such damage,” but the model answer only restates the location/condition and does not provide that requirement.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 162 — `OSHA\_29CFR\_full\_2024\_neg\_1194\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How frequently must combustible floors be re-wetted to maintain protection during welding operations?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_frequency_of_rewetting_combustible_floors", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only says combustible floors shall be kept wet, covered with damp sand, or protected by shields, and gives no re-wetting frequency; the model did not state that this was unanswerable but instead produced an unsupported function call.
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 163 — `OSHA\_29CFR\_full\_2024\_neg\_3806\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How long must the employer retain the written medical report after it is issued?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "retain_written_medical_report", "parameters": {"duration": "30 years"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 164 — `OSHA\_29CFR\_full\_2024\_1332\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the log hauls and slips control requirement with the log decks barricade requirement, what positioning protections are required for operators/sawyers relative to the logs?
- ✓ **Ground truth:** Per paragraph (vi)(e), log haul controls shall be located so the operator is at all times in the clear of logs, machinery, lines, and rigging (and where control is by lever exposed to incoming logs, the lever shall operate the haul only when moved toward the log slip or log pond); per paragraph (3)(iii), a barricade or other positive stop of sufficient strength to stop any log shall be erected between the sawyer's stand and the log deck.
- ✗ **Model answered:** {"name": "log_hauls_and_slips_control_requirement", "parameters": {}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not synthesize rules from multiple regulatory sections
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 165 — `OSHA\_29CFR\_full\_2024\_neg\_1037\_03` · MEDIUM confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum diameter required for the core in a cable laid rope?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_minimum_core_diameter", "parameters": {"rope_type": "cable_laid_rope"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 166 — `OSHA\_29CFR\_full\_2024\_2261\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Conditional
- **Question:** If a healthcare professional is responsible for an employee's Hepatitis B vaccination but is not evaluating an exposure incident, what is the employer required to provide them?
- ✓ **Ground truth:** A copy of this regulation.
- ✗ **Model answered:** {"name": "provide_information_to_healthcare_professional", "parameters": {"employee": "employee", "healthcare_professional": "healthcare professional responsible for the employee"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not provide the correct information to the healthcare professional
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 167 — `OSHA\_29CFR\_full\_2024\_0913\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the education requirement for general employees and the training requirement for designated equipment users, how often must each be repeated after the initial point?
- ✓ **Ground truth:** Both must be repeated at least annually thereafter - the educational program under (g)(2) at least annually after initial employment, and the equipment training under (g)(4) at least annually after initial assignment.
- ✗ **Model answered:** {"name": "training_and_education_frequency", "parameters": {"general_employee_training_frequency": "annually", "designated_equipment_user_training_frequency": "quarterly"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states both the education in (g)(2) and the designated-equipment training in (g)(4) must be provided 'at least annually thereafter,' but the model incorrectly said designated equipment user training is quarterly.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 168 — `OSHA\_29CFR\_full\_2024\_neg\_4303\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the cost to the committee for obtaining the NIOSH hazard evaluation reports listed in paragraph (a)?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "reimbursement", "parameters": {"c": "cost to the committee"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 169 — `OSHA\_29CFR\_full\_2024\_neg\_2129\_06` · MEDIUM confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What penalty does an employer face for failing to implement the respiratory protection program required in paragraph (2)(i)?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "paragraph", "parameters": {"number": "2", "i": "penalty for failing to implement the respiratory protection program"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated penalty for failing to implement the respiratory protection program
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 170 — `OSHA\_29CFR\_full\_2024\_neg\_4455\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the filing fee for the complainant to request a hearing after receiving the findings?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "calculate_filing_fee", "parameters": {"complainant_income": "50000"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated a specific number not present in the evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '50000' not in evidence

Failure 171 — `OSHA\_29CFR\_full\_2024\_0884\_02` · MEDIUM confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under the general rule of paragraph (e)(3) and its exception, who normally removes a lockout device, and under what condition may someone else remove it?
- ✓ **Ground truth:** Under the general rule of (e)(3), each lockout or tagout device shall be removed by the employee who applied it. Under the exception, when that authorized employee is not available, the device may be removed under the direction of the employer, provided specific procedures and training have been developed, documented, and incorporated into the energy control program, and the employer demonstrates equivalent safety.
- ✗ **Model answered:** {"name": "remove_lockout_device", "parameters": {"employee_who_applied": "normally", "exception": "someone else", "condition": "when the authorized employee who applied the device is not available to remove it"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 172 — `OSHA\_29CFR\_full\_2024\_3020\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Per subpart (viii), where is the spill containment program required to originate, and what does it reference in another part of the section?
- ✓ **Ground truth:** The spill containment program is part of the employer's safety and health program required in paragraph (b) of this section, and it must be implemented to contain and isolate the entire volume of the hazardous substance being transferred where major spills may occur.
- ✗ **Model answered:** {"name": "answer_question", "parameters": {"question": "Per subpart (viii), where is the spill containment program required to originate, and what does it reference in another part of the section?"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not synthesize rules from multiple regulatory sections
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 173 — `OSHA\_29CFR\_full\_2024\_neg\_4485\_08` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum period the time for filing may be tolled under applicable case law?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "time for filing a complaint equitably tolled", "parameters": {"days": "180"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 174 — `OSHA\_29CFR\_full\_2024\_3280\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Cross-Reference
- **Question:** Based on the definitions of 'Unstable objects' and 'Supported scaffold,' why would a barrel be inappropriate as part of a supported scaffold?
- ✓ **Ground truth:** A supported scaffold requires platforms supported by rigid support such as poles, legs, uprights, posts, or frames. Unstable objects, which include barrels, may become dislocated and shift and do not constitute a safe base support for scaffolds, so a barrel cannot serve as the rigid base support a supported scaffold needs.
- ✗ **Model answered:** {"name": "get_inappropriate_scaffold_part", "parameters": {"object_name": "barrel"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports the ground truth because barrels are explicitly listed as unstable objects that 'may become dislocated and shift' and 'do not constitute a safe base support for scaffolds,' while the model merely returned a function-style object naming 'barrel' and did not answer why it would be inappropriate.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 175 — `OSHA\_29CFR\_full\_2024\_3999\_00` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** Under what circumstances may an operator-in-training operate the equipment in the listed prohibited circumstances despite not being certified under paragraph (c)?
- ✓ **Ground truth:** As provided in the exception under paragraph (b)(3)(v) of this section.
- ✗ **Model answered:** {"name": "operator-in-training operation in prohibited circumstances", "parameters": {"circumstance": "multiple-lift rigging operations, except where the operator"}}
- **Doctor diagnosis:** Dr. Default model error
- **Doctor root cause:** model did not identify the exception under paragraph (b)(3)(v)
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 176 — `OSHA\_29CFR\_full\_2024\_4517\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Cross-Reference
- **Question:** Combining the time-for-filing rule in paragraph (d) with the tolling example it provides, if a complainant mistakenly files with an agency other than OSHA 150 days after the adverse action, is the complaint potentially still timely, and what is the base filing period?
- ✓ **Ground truth:** Yes, potentially. Paragraph (d) sets a 180-day filing period, and it states OSHA may consider the time tolled if a complainant mistakenly files with an agency other than OSHA within 180 days after the alleged adverse action. Since 150 days is within 180 days, tolling may apply.
- ✗ **Model answered:** {"name": "calculate_filing_deadline", "parameters": {"days_after_adverse_action": "150", "time_for_filing": "180"}}
- **Doctor diagnosis:** Dr. Default model error
- **Doctor root cause:** [verifier] The evidence states a complaint must be filed “Within 180 days” and tolling may apply if it is mistakenly filed with another agency “within 180 days,” so the correct answer is yes potentially with a 180-day base period; the model only returned calculation parameters and omitted the required timeliness/tolling answer.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 177 — `OSHA\_29CFR\_full\_2024\_4502\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under paragraph (a), what specific rights must OSHA notify the respondent of when notifying it of the filing of the complaint, and what are the two response timeframes associated with those rights set out in paragraphs (b) and (f)?
- ✓ **Ground truth:** OSHA must notify the respondent of its rights under paragraphs (b) and (f) of the section and § 1987.110(e). Under paragraph (b), the respondent may, within 20 days of receipt of the notice, submit a written statement and affidavits/documents and request a meeting. Under paragraph (f), the respondent must present its evidence within 10 business days of OSHA's notification (or as soon thereafter as OSHA and the respondent agree).
- ✗ **Model answered:** {"name": "notify_respondent_of_filing", "parameters": {"rights": "rights under paragraphs (b) and (f) of this section and \u00a7 1982.110(e)", "response_timeframes": "[20 days, 20 days]"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 178 — `OSHA\_29CFR\_full\_2024\_neg\_2913\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What gauge of sheet steel is required for the exhaust ductwork construction?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "sheet_steel_gauge", "parameters": {"g": "18"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 179 — `OSHA\_29CFR\_full\_2024\_1349\_01` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the rules on how the chain saw must be started and how it must be held during operation, what is required regarding the operator's hands?
- ✓ **Ground truth:** When starting, drop starting is prohibited and the saw must be started on the ground or where otherwise firmly supported (per (vi)); during operation, the chain saw shall be held with the thumbs and fingers of both hands encircling the handles unless the employer demonstrates a greater hazard is posed by keeping both hands on the saw (per (viii)).
- ✗ **Model answered:** {"name": "chain_saw_operation", "parameters": {"hands_required": "both hands encircling the handles during operation"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not fully synthesize the rules for chain saw operation
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 180 — `OSHA\_29CFR\_full\_2024\_0494\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an undercoating operation uses materials not more hazardous than kerosene but is conducted in a garage lacking adequate natural or mechanical ventilation, does the exemption apply?
- ✓ **Ground truth:** No, because adequate natural or mechanical ventilation is required for the exemption, so the operation is subject to all spray finishing requirements.
- ✗ **Model answered:** {"name": "k", "parameters": {"undercoating_materials": "not more hazardous than kerosene", "ventilation": "inadequate"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** (none)

Failure 181 — `OSHA\_29CFR\_full\_2024\_0882\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the durability requirement in paragraph (c)(5)(ii)(A)(3) and the training limitation in paragraph (c)(7)(ii)(D), what is required regarding tags used in corrosive environments such as areas where acid and alkali chemicals are handled?
- **Ground truth:** Tags shall not deteriorate when used in corrosive environments such as areas where acid and alkali chemicals are handled and stored, and employees shall be trained that tags and their means of attachment must be made of materials which will withstand the environmental conditions encountered in the workplace.
- **Model answered:** {"name": "requirement_combination", "parameters": {"durability_requirement": "paragraph (c)(5)(ii)(A)(3)", "training_limitation": "paragraph (c)(7)(ii)(D)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 182 — `OSHA\_29CFR\_full\_2024\_2164\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under the calibration provision and the internal QC analyses provision, what minimum frequency applies to running a standard versus running internal QC samples relative to compliance samples?
- **Ground truth:** A standard should be run about every fifth sample (to assure calibration), while internal QC samples should be run at a rate of 5% of the compliance samples or at least one set per analysis, whichever is greater.
- **Model answered:** {"name": "get_minimum_frequency", "parameters": {"provision": "calibration", "sample_type": "standard"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 183 — `OSHA\_29CFR\_full\_2024\_neg\_3653\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the assigned protection factor for the tight-fitting PAPR that must be provided when an employee chooses to use one?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "Assigned Protection Factor", "parameters": {"Type of respirator 1 2": "2. Powered Air-Purifying Respirator (PAPR); Half mask: 50; Full facepiece: 1,000; Helmet/hood: 4 25/1,000; Loose-fitting facepiece: 25"}}

- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 184 — `OSHA\_29CFR\_full\_2024\_3020\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** According to subpart (v) and subpart (vi), what organizational and warning steps apply to drum movement operations?
- ✓ **Ground truth:** Site operations must be organized to minimize the amount of drum or container movement (v), and prior to any movement of drums or containers, all exposed employees must be warned of the potential hazards associated with the contents (vi).
- ✗ **Model answered:** {"name": "subpart", "parameters": {"v": "Organizational steps", "vi": "Warning steps"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 185 — `OSHA\_29CFR\_full\_2024\_neg\_4455\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum dollar amount of compensatory damages that may be awarded in the preliminary order?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "issuance_of_findings_and_preliminary_orders", "parameters": {"a": "250,000"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 186 — `OSHA\_29CFR\_full\_2024\_0494\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an automobile undercoating spray operation is conducted in a garage with adequate ventilation but uses materials more hazardous than kerosene with solvents having a flash point of 90 °F, is it exempt from spray finishing requirements?
- ✓ **Ground truth:** No, it does not conform to the exemption provisions and is therefore subject to all spray finishing requirements of this section.
- ✗ **Model answered:** {"name": "k", "parameters": {"Underwriters' Laboratories": "false", "flash point": "90"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** (none)

Failure 187 — `OSHA\_29CFR\_full\_2024\_neg\_4455\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How much does it cost to file objections electronically with the Office of Administrative Law Judges?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_filing_fees", "parameters": {"agency": "Office of Administrative Law Judges", "filing_type": "electronic"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 188 — `OSHA\_29CFR\_full\_2024\_neg\_4485\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many days does a third-party commercial carrier have to deliver a complaint for it to count as timely?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "time_for_filing", "parameters": {"days": "180"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 189 — `OSHA\_29CFR\_full\_2024\_neg\_0484\_09` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum allowable resistance value for the bonding and grounding connections between containers during liquid transfer?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_max_resistance", "parameters": {"is_bonding": "True"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 190 — `OSHA\_29CFR\_full\_2024\_3312\_05` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If there is a storm or high winds, when may employees work on a scaffold?
- ✓ **Ground truth:** Only when a competent person has determined it is safe and the employees are protected by a personal fall arrest system or wind screens.
- ✗ **Model answered:** {"name": "work_on_scaffold_during_storm", "parameters": {"wind_speed": "less than 30 mph"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 191 — `OSHA\_29CFR\_full\_2024\_neg\_1259\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Who is responsible for invoking the blue flag policy referenced in paragraph (c)(9)(i)?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "invoke_blue_flag_policy", "parameters": {"policy": "blue flag policy", "paragraph": "(c)(9)(i)"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 192 — `OSHA\_29CFR\_full\_2024\_0881\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an energy isolating device has neither a hasp nor a built-in locking mechanism, when is it still considered capable of being locked out?
- ✓ **Ground truth:** When lockout can be achieved without the need to dismantle, rebuild, or replace the device or permanently alter its energy control capability.
- ✗ **Model answered:** {"name": "is_capable_of_being_locked_out", "parameters": {"hasp": "False", "locking_mechanism": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 193 — `OSHA\_29CFR\_full\_2024\_neg\_1738\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as

authoritative and is hard to catch downstream

- **Category:** Negative (hallucination traps)
- **Question:** How frequently must common areas like break rooms be cleaned under this section?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "cleaning_frequency", "parameters": {"frequency": "at least once a day"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 194 — `OSHA\_29CFR\_full\_2024\_neg\_3806\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What qualifications must the licensed physician hold to conduct the examination under this standard?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "licensed_physician_qualifications", "parameters": {"standard": "this standard"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only mentions a “licensed physician” and the contents/timing of the written medical report, without stating any qualifications to conduct the examination, so UNANSWERABLE is supported but the model’s function-call-like output did not answer or refuse as unanswerable.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 195 — `OSHA\_29CFR\_full\_2024\_2476\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under the definitions, an 'authorized person' may enter a regulated area as a designated representative of employees to observe monitoring procedures under which paragraph, and a 'regulated area' is demarcated based on exposure exceeding which two limits?
- ✓ **Ground truth:** An authorized person may observe monitoring and measuring procedures under paragraph (d) of this section. A regulated area is demarcated where an employee's exposure to airborne MC exceeds or can reasonably be expected to exceed either the 8-hour TWA PEL or the STEL.
- ✗ **Model answered:** {"name": "paragraph", "parameters": {"r": "paragraph (r)", "PEL": "PEL", "action level": "action level"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 196 — `OSHA\_29CFR\_full\_2024\_3017\_00` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually

- **Category:** Exception
- **Question:** Which employees does the note to paragraph (h) state are NOT required to be monitored?
- ✓ **Ground truth:** Employees engaged in site characterization operations covered by paragraph (c) of this section.
- ✗ **Model answered:** {"name": "get_employees_not_required_to_be_monitored", "parameters": {}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not identify the correct employees not required to be monitored
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 197 — `OSHA\_29CFR\_full\_2024\_4069\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a tower crane manufacturer does not specify the maximum wind speed for operations, what wind speed limit applies?
- ✓ **Ground truth:** The wind must not exceed the speed determined by a qualified person.
- ✗ **Model answered:** {"name": "get_wind_speed_limit", "parameters": {"manufacturer_wind_speed": "null"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not recognize the correct wind speed limit when manufacturer does not specify it
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 198 — `OSHA\_29CFR\_full\_2024\_neg\_1738\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many air changes per hour are required for an Airborne infection isolation room under this section?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "Airborne infection isolation room", "parameters": {"number of air changes per hour": "12"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 199 — `OSHA\_29CFR\_full\_2024\_3208\_01` · MEDIUM confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If electric equipment operates at exactly 50 volts, must its live parts be guarded against accidental contact?
- ✓ **Ground truth:** Yes, live parts operating at 50 volts or more must be guarded against accidental contact.
- ✗ **Model answered:** {"name": "guarding_of_live_parts", "parameters": {"voltage": "50"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence says live parts of equipment operating at “50 volts or more shall be guarded,” so the correct answer is yes, but the model only returned a function-like label with voltage 50 rather than actually answering the yes/no requirement.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 200 — `OSHA\_29CFR\_full\_2024\_neg\_1738\_06` · MEDIUM confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum cumulative time of aerosol-generating procedures allowed per patient before additional protections apply?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "get_aerosol_generating_procedure_time_limit", "parameters": {"procedure_type": "aerosol-generating", "patient_status": "suspected_or_confirmed_COVID-19"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence defines aerosol-generating procedures but gives no cumulative time limit or additional-protection threshold, so UNANSWERABLE is supported; the model did not state that the answer was not in evidence and instead attempted a time-limit function call with an unsupported patient status.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 201 — `OSHA\_29CFR\_full\_2024\_4188\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the requirements for guard support arch spacing and the warning device, what are the two distinct numeric/operational conditions that must be satisfied when using horizontal bar guards on a screw conveyor in a gin that is about to start?
- **Ground truth:** First, the horizontal bars must be supported by arches that are not more than 8 feet apart; second, an audible warning signal must be sounded each time before starting the gin.
- **Model answered:** {"name": "screw conveyor guard support arch spacing and warning device", "parameters": {"spacing": "not more than 8 feet apart", "warning device": "sounds an alarm that will be recognized by each employee as a warning that the conveyor will start and that can be clearly heard at all points along the conveyor where personnel may be present"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 202 — `OSHA\_29CFR\_full\_2024\_neg\_2625\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)

- **Question:** How many confirmed human cases are required before a substance can be classified in sub-category 1A?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "A.4.2.2.2.1", "parameters": {"human evidence": "positive responses at \u2264500 \u03bcg/cm 2 (Human Repeat Insult Patch Test (HRIPT), Human Maximization Test (HMT)\u2014induction threshold)"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 203 — `OSHA\_29CFR\_full\_2024\_2815\_06` · MEDIUM confidence

CROSS-REFERENCE MISS

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Cross-Reference
- **Question:** How does the OMB control number assigned to the series 1926.1203 through 1926.1213 compare to the one assigned to the series 1926.1402 through 1926.1441?
- ✓ **Ground truth:** The 1926.1203-1926.1213 series all carry OMB control No. 1218-0258, while the 1926.1402-1926.1441 series (excluding 1926.1427) carry OMB control No. 1218-0261 - they are different.
- ✗ **Model answered:** {"name": "compare_omb_control_numbers", "parameters": {"a": "1218-0258", "b": "1218-0261"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports 1926.1203-1926.1213 as 1218-0258 and most of 1926.1402-1926.1441 as 1218-0261, but it also explicitly lists “1926.1427; OMB control No.: 1218-0270,” so the model’s answer omits a material exception included in the ground truth.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 204 — `OSHA\_29CFR\_full\_2024\_neg\_3512\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What load capacity must the hanging walkways be designed to support?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_load_capacity", "parameters": {"material": "noncombustible material", "span": "full span of the arch between supports"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 205 — `OSHA\_29CFR\_full\_2024\_3130\_07` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the minimum liquid capacity (in U.S. gallons) for a container to be considered a portable tank?

- ✓ **Ground truth:** more than 60 U.S. gallons
- ✗ **Model answered:** {"name": "definition", "parameters": {"term": "portable tank", "value": "60"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 206 — `OSHA\_29CFR\_full\_2024\_neg\_1194\_08` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How long must a duct or conveyor system remain shut down after welding is completed?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "shutdown_time", "parameters": {"duration": "15"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 207 — `OSHA\_29CFR\_full\_2024\_0598\_01` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** Above what container capacity does the minimum 5-foot spacing requirement for underground containers apply?
- ✓ **Ground truth:** over 2,000 gallons
- ✗ **Model answered:** {"name": "calculate_minimum_spacing_requirement", "parameters": {"capacity": "1000"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed incorrect minimum spacing requirement
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 208 — `OSHA\_29CFR\_full\_2024\_4047\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If employees are occupying a personnel platform but the platform is being used over water, are they required to use a personal fall arrest system attached to a structural member within the platform?
- ✓ **Ground truth:** No; the personal fall arrest system requirement applies except over water, and when working over or near water the requirements of § 1926.106 apply instead.
- ✗ **Model answered:** {"name": "is_personal_fall_arrest_system_required", "parameters": {"over_water": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 209 — `OSHA\_29CFR\_full\_2024\_0564\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** May an LP-Gas-fueled vehicle be parked near an open pit?
- ✓ **Ground truth:** Not unless the pit is adequately ventilated.
- ✗ **Model answered:** {"name": "check_parking_proximity", "parameters": {"near_open_pit": "False"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not consider ventilation condition for open pits
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 210 — `OSHA\_29CFR\_full\_2024\_neg\_2902\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How often must the air compressor used for abrasive-blasting respirators be inspected for harmful contaminants?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "check_air_compressor", "parameters": {"frequency": "periodically"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 211 — `OSHA\_29CFR\_full\_2024\_0694\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Given the general toilet requirement in (n)(3)(i) and the exception in (n)(3)(iv), would a mobile crew of 30 employees be required to provide toilet facilities on site if they have transportation readily available to nearby toilets?
- ✓ **Ground truth:** No. Although 30 employees would normally require one toilet seat and one urinal per 40 employees under the table, the requirements for sanitation facilities do not apply to mobile crews having transportation readily available to nearby toilet facilities.
- ✗ **Model answered:** {"name": "is_toilet_facility_required", "parameters": {"number_of_employees": "30"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not account for the exception in (n)(3)(iv)
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 212 — `OSHA\_29CFR\_full\_2024\_neg\_1259\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream

- **Category:** Negative (hallucination traps)
- **Question:** What is the required tolerance range for the limit switch beyond the stated 30 degrees each way?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "limit_switch_tolerance", "parameters": {"angle": "45"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 213 — `OSHA\_29CFR\_full\_2024\_4205\_00` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Cross-Reference
- **Question:** Given the dates of initial and final approval in paragraphs (a) and (b), how many years elapsed between the Tennessee State plan's initial approval and its final approval?
- ✓ **Ground truth:** Initial approval was July 5, 1973, and final approval was July 22, 1985, so approximately 12 years elapsed.
- ✗ **Model answered:** {"name": "calculate_duration", "parameters": {"start_date": "July 5, 1973", "end_date": "July 22, 1985"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence gives initial approval as July 5, 1973 and final approval as July 22, 1985, supporting the ground truth of about 12 years elapsed, but the model only returned a duration-calculation function call with the dates and did not answer how many years elapsed.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 214 — `OSHA\_29CFR\_full\_2024\_0003\_03` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What Statute at Large citation is given for the Intergovernmental Personnel Act of 1970 in § 1902.3(h)?
- ✓ **Ground truth:** 84 Stat. 1915
- ✗ **Model answered:** {"name": "get_statute_at_large_citation", "parameters": {"section": "1902.3(h)", "act": "Intergovernmental Personnel Act of 1970"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 215 — `OSHA\_29CFR\_full\_2024\_4398\_03` · HIGH confidence

CAUSE-EFFECT ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cause & Effect
- **Question:** What is the effect of OSHA's approval of a settlement reached by the respondent and the complainant?
- ✓ **Ground truth:** It demonstrates OSHA's consent and achieves the consent of all three parties.

- **✗ Model answered:** {"name": "OSHA", "parameters": {"approval": "settlement reached by the respondent and the complainant"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly states that “OSHA's approval of a settlement reached by the respondent and the complainant demonstrates OSHA's consent and achieves the consent of all three parties,” but the model only restated the actor and settlement context without giving that effect.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 216 — `OSHA\_29CFR\_full\_2024\_neg\_3991\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the penalty for failing to agree on voice signals prior to beginning operations under § 1926.1421?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** {"name": "penalty_for_failure_to_agree_on_voice_signals", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence states only the requirement to agree on voice signals and contains no penalty, so the ground truth UNANSWERABLE is supported, but the model returned a function-call-like JSON rather than explicitly indicating the penalty is not in evidence.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 217 — `OSHA\_29CFR\_full\_2024\_neg\_1259\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How far apart must flag signals be spaced when protecting men during switching operations?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** {"name": "flag signals spacing", "parameters": {"distance": "50 feet"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 218 — `OSHA\_29CFR\_full\_2024\_neg\_2129\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Who pays for the powered air-purifying respirator when an employee chooses one over a negative-pressure respirator?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** {"name": "provided by the employer", "parameters": {}}

- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 219 — `OSHA\_29CFR\_full\_2024\_neg\_3653\_09` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How long must an employer retain records of respirator selection decisions made under this section?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "retain_records", "parameters": {"duration": "30 years"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 220 — `OSHA\_29CFR\_full\_2024\_neg\_1037\_09` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many wraps are required for a strand laid grommet-hand tucked sling made from braided wire rope instead of strand?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_minimum_wraps", "parameters": {"sling_type": "strand laid grommet-hand tucked", "wire_rope_type": "braided"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 221 — `OSHA\_29CFR\_full\_2024\_neg\_2625\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which regulatory agency reviews the weight-of-evidence determination for respiratory sensitizer classification?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "A.4.2.1.1.1", "parameters": {"Effects seen in either humans or animals will normally justify classification in a weight of evidence approach for respiratory sensitizers.": "Substances may be allocated to one of the two sub-categories 1A or 1B using a weight of evidence approach in accordance with the criteria given in Table A.4.1 and on the basis of reliable and good quality evidence from human cases or epidemiological studies and/or observations from appropriate studies in experimental animals.": "US Environmental Protection Agency (EPA)"}}

- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 222 — `OSHA\_29CFR\_full\_2024\_3172\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under what specific conditions must a vertical belt running over a lower pulley more than 7 feet above the floor be guarded at the bottom, per paragraph (e)(4) read with the cross-referenced conditions?
- **Ground truth:** Per paragraph (e)(4), it must be guarded at the bottom like horizontal overhead belts if the conditions in paragraphs (e)(2)(ii)(a) and (c) apply - that is, if it is located over passageways or work places and traveling 1,800 feet or more per minute (condition (a)), and if the belt is 8 inches or more in width (condition (c)).
- **Model answered:** {"name": "guard_vertical_belt", "parameters": {"distance": "7", "conditions": "conditions are as stated in paragraphs (e)(2)(ii) (a) and (c) of this section"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 223 — `OSHA\_29CFR\_full\_2024\_neg\_4297\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which OSHA office is designated to receive and process the inspection requests described in paragraph (a)(2)?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "OSHA Office", "parameters": {"office": "Area Office"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 224 — `OSHA\_29CFR\_full\_2024\_neg\_2913\_08` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum permitted length of exhaust ductwork before a booster fan must be installed?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "calculate_ductwork_length", "parameters": {"ductwork_diameter": "12", "booster_fan_interval": "20"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 225 — `OSHA\_29CFR\_full\_2024\_neg\_1259\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What color must derails be painted during switching operations?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "Safety color code for marking physical hazards", "parameters": {"(a)": "Color identification", "(1)": "Red", "(i)": "Danger"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 226 — `OSHA\_29CFR\_full\_2024\_2160\_02` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cause & Effect
- **Question:** What were the consequences for analytical bias and OAE resulting from the average recovery of 94.2% of the spiked filter samples in the AAS-HGA analysis?
- ✓ **Ground truth:** The analytical bias was -5.8% and the OAE was ±14.2%.
- ✗ **Model answered:** {"name": "calculate_consequences", "parameters": {"recovery": "94", "cv": "0"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed ratio instead of quoting stated figure
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 227 — `OSHA\_29CFR\_full\_2024\_neg\_2625\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What specific sensitization frequency percentage distinguishes sub-category 1A from sub-category 1B?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_sensitization_frequency", "parameters": {"sub_category": "1A"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 228 — `OSHA\_29CFR\_full\_2024\_neg\_2602\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What form must be submitted to document a new classification when batch toxicity changes under A.0.5.1.2?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "change_form", "parameters": {"description": "new classification due to batch toxicity change", "purpose": "documenting change in batch toxicity", "technical_basis": "A.0.5.1.2", "safety_and_health_considerations": "potential impact on human health and environment", "documentation_of_changes": "updated classification and associated documentation", "approvals_and_authorization": "authorized personnel"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 229 — `OSHA\_29CFR\_full\_2024\_2959\_00` · MEDIUM confidence

UNIT/FORMAT MISMATCH

- **Severity:** MEDIUM — a missing or incomplete answer — the model withheld an available answer (a gap, not a confidently wrong statement)
- **Category:** Precision
- **Question:** What is the airborne concentration of lead that defines the action level, calculated as an 8-hour time-weighted average?
- ✓ **Ground truth:** 30 micrograms per cubic meter of air (30 µg/m³)
- ✗ **Model answered:** {"name": "action_level", "parameters": {"x": "30"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence defines the action level as “30 micrograms per cubic meter of air (30 µg/m³),” while the model answered only “30” without the required concentration units, so it is incomplete rather than a validator parsing issue.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 230 — `OSHA\_29CFR\_full\_2024\_neg\_3806\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the BeLPT and which laboratory must perform it under this standard?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "BeLPT", "parameters": {"test": "measurement of blood lymphocyte proliferation in a laboratory test when lymphocytes are challenged with a soluble beryllium salt", "laboratory": "a laboratory certified under the College of American Pathologists/Clinical Laboratory Improvement Amendments (CLIA) guidelines"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 231 — `OSHA\_29CFR\_full\_2024\_0219\_02` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the publication date of ISO 13943:2000 (E/F), Fire Safety-Vocabulary?
- ✓ **Ground truth:** April, 15, 2000
- ✗ **Model answered:** {"name": "get_publication_date", "parameters": {"standard": "ISO 13943:2000 (E/F)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence explicitly states “ISO 13943:2000 (E/F)... First Edition, April, 15, 2000,” but the model only emitted a tool-call JSON with the standard name and did not provide the publication date.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 232 — `OSHA\_29CFR\_full\_2024\_neg\_1037\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many cross rods are required per foot of metal mesh fabric?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "cross_rod_requirement", "parameters": {"number_of_cross_rods": "1"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 233 — `OSHA\_29CFR\_full\_2024\_neg\_2913\_05` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the required spacing between connecting flanges for ductwork larger than 12 inches in diameter?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_required_spacing", "parameters": {"duct_diameter": "larger than 12 inches"}}
- **Doctor diagnosis:** **Dr. No** bad test
- **Doctor root cause:** [verifier] The ground truth is unsupported because the evidence does answer the question: circumferential joints in ductwork require that “at least every fourth joint shall be provided with connecting flanges,” so it is not UNANSWERABLE.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 234 — `OSHA\_29CFR\_full\_2024\_0133\_00` · HIGH confidence

CAUSE-EFFECT ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cause & Effect
- **Question:** Why must State-Plan States consult with and obtain approval for requirements that are more stringent than or supplemental to the Federal requirements for certain part 1904 provisions?
- ✓ **Ground truth:** Because of the unique nature of the national recordkeeping program.
- ✗ **Model answered:** {"name": "consult_and_obtain_approval", "parameters": {"requirement_type": "more stringent than or supplemental to the Federal requirements", "provisions": "industry exemptions, reporting of fatalities and hospitalizations, record retention, or employee involvement"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not understand the unique nature of the national recordkeeping program
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 235 — `OSHA\_29CFR\_full\_2024\_neg\_0429\_08` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which government office issues approvals for bulk oxygen system installations?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "local fire prevention office", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 236 — `OSHA\_29CFR\_full\_2024\_neg\_1259\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum weight capacity of a powered hand truck permitted under this section?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_maximum_weight_capacity", "parameters": {"truck_type": "powered hand truck"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 237 — `OSHA\_29CFR\_full\_2024\_0333\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional

- **Question:** If angulated roping is employed but the exterior building design makes tie-in guides feasible, may the tie-in guides be eliminated for the uppermost 75 feet?
- ✓ **Ground truth:** No; elimination is only permitted if it is infeasible due to exterior building design (and an angulation force of at least 10 pounds is maintained).
- ✗ **Model answered:** {"name": "eliminate_tie_in_guides", "parameters": {"employed": "True", "design": "False", "feet": "75"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 238 — `OSHA\_29CFR\_full\_2024\_neg\_0429\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum allowable storage temperature for liquid oxygen in these systems?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_max_storage_temperature", "parameters": {"material": "liquid oxygen"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 239 — `OSHA\_29CFR\_full\_2024\_neg\_4485\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What languages other than English does OSHA provide official complaint forms in?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "complaint forms in other languages", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 240 — `OSHA\_29CFR\_full\_2024\_4031\_02` · HIGH confidence

CALCULATION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Calculation
- **Question:** When hoisting personnel without a personnel platform, the equipment's rated capacity for the radius and configuration is 17,800 pounds. What is the maximum total load (including hook, load line, rigging and any other equipment that imposes a load) allowed, except during proof testing?
- ✓ **Ground truth:** 8,900 pounds
- ✗ **Model answered:** {"name": "calculate_total_load", "parameters": {"rated_capacity": "17800"}}
- **Doctor diagnosis:** **Dr. Pi** model error
- **Doctor root cause:** unknown

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 241 — `OSHA\_29CFR\_full\_2024\_neg\_3806\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum penalty an employer faces for failing to provide the written medical report within the 45-day deadline?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "calculate_penalty", "parameters": {"penalty_type": "failure_to_correct_violation", "days_late": "45"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 242 — `OSHA\_29CFR\_full\_2024\_neg\_1194\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum thickness required for the metal or asbestos guards used to shield combustibles when relocation is impracticable?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_minimum_thickness", "parameters": {"material": "metal or asbestos"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 243 — `OSHA\_29CFR\_full\_2024\_1593\_04` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** What is the stated exception to the minimum clear working space requirements specified in Table S-2?
- ✓ **Ground truth:** The Table S-2 minimum clear working space applies except as elsewhere required or permitted in this subpart.
- ✗ **Model answered:** {"name": "minimum_clear_working_space", "parameters": {"exception": "the location of the switchboards and control panels permits a continuous and unobstructed way of exit travel"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 244 — `OSHA\_29CFR\_full\_2024\_neg\_1769\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What fee may an employer charge for copying the COVID-19 records when requested?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "reasonable, non-discriminatory administrative costs", "parameters": {"search": "and copying expenses but not including overhead expenses"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 245 — `OSHA\_29CFR\_full\_2024\_2016\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** According to Appendix B, which two paragraphs of the regulation define the scope of coverage (what records the rule applies to) and the meaning of 'toxic substance or harmful physical agent'?
- ✓ **Ground truth:** Paragraph (b)(2) establishes that the regulation applies to all employee exposure and medical records of employees exposed to toxic substances or harmful physical agents, and paragraph (c)(13) defines the term 'toxic substance or harmful physical agent' to encompass chemical substances, biological agents, and physical stresses for which there is evidence of harmful health effects.
- ✗ **Model answered:** {"name": "specific_written_consent", "parameters": {"i": "1", "ii": "2"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 246 — `OSHA\_29CFR\_full\_2024\_neg\_1769\_08` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Who is designated as the custodian responsible for maintaining the confidentiality of the COVID-19 log?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_custodian", "parameters": {"log_type": "COVID-19 log"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 247 — `OSHA\_29CFR\_full\_2024\_3403\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Using the weight limitation in paragraph (k) and the operator responsibility in paragraph (h), who is responsible for the load's weight and what is the maximum allowable external load weight?
- ✓ **Ground truth:** Per paragraph (h), the helicopter operator is responsible for the size, weight, and manner in which loads are connected to the helicopter; and per paragraph (k), the weight of an external load shall not exceed the manufacturer's rating.
- ✗ **Model answered:** {"name": "operator_responsibility_and_weight_limitation", "parameters": {"k": "1926", "h": "operator"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 248 — `OSHA\_29CFR\_full\_2024\_1227\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the requirements of paragraphs (v) and (vi), what must be done before making hot repairs to one generator located in a room containing additional generators?
- ✓ **Ground truth:** The carbide charge and feed mechanism of the generator being repaired must be completely removed, all its acetylene expelled by completely flooding the shell with water, and it must be disconnected from the piping system (paragraph (v)); additionally, all the other generators and piping in the room must be purged of acetylene (paragraph (vi)).
- ✗ **Model answered:** {"name": "remove_carbide_charge_and_feed_mechanism", "parameters": {}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model stopped before applying dependent rule
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 249 — `OSHA\_29CFR\_full\_2024\_4486\_01` · HIGH confidence

OVERCONFIDENT

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Per paragraph (e)(1) a complaint will be dismissed unless a prima facie showing is made, but per paragraph (e)(4), what additional circumstance can prevent an investigation even after that showing, and by what evidentiary standard?
- ✓ **Ground truth:** Even if the complainant has made a prima facie showing, an investigation will not be conducted or will be discontinued if the respondent demonstrates by clear and convincing evidence that it would have taken the same adverse action in the absence of the complainant's protected activity.
- ✗ **Model answered:** {"name": "investigation_prevention", "parameters": {"circumstance": "respondent demonstrates that it would have taken the same adverse action in the absence of the complainant", "evidentiary standard": "clear and convincing evidence"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** Uncertainty violation: evidence hedges but response uses 'demonstrates that'

Failure 250 — `OSHA\_29CFR\_full\_2024\_1288\_04` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If dye kettles or vats may at any time contain hot or corrosive liquids, where may the installed pipes or drains empty?
- **Ground truth:** They must carry the contents safely away from the working area and shall not empty directly onto the floor.
- **Model answered:** {"name": "Pipes or drains for dye kettles and vats", "parameters": {"location": "a safe location away from the working area"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 251 — `OSHA\_29CFR\_full\_2024\_1292\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the lubrication requirement and the hot pipes requirement, what two protective measures are required to keep employees safe from machine hazards?
- **Ground truth:** Per paragraph (c)(5), stationary lubrication fittings inside a machine must be provided with extension piping to a point of safety so employees do not reach into dangerous parts when lubricating; and per paragraph (c)(8), exposed hot water and steam pipes must be covered with insulating material wherever necessary to protect employees from contact.
- **Model answered:** {"name": "protective_measures", "parameters": {"measure1": "automatic oiling device", "measure2": "insulating material or guarding"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence requires “extension piping to a point of safety” for lubrication fittings and “insulating material” for exposed hot water and steam pipes; the model’s “automatic oiling device” is not the same protective measure and is not supported by the evidence.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 252 — `OSHA\_29CFR\_full\_2024\_neg\_4455\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Within how many days must the respondent submit documentation to the Social Security Administration after the preliminary order issues?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "submit_documentation_to_social_security_administration", "parameters": {"days": "10"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 253 — `OSHA\_29CFR\_full\_2024\_2784\_02` · HIGH confidence

MISINTERPRETED

- **Severity:** MEDIUM — a misapplied or over-extrapolated rule — plausible but wrong
- **Category:** Conditional
- **Question:** For laboratories, does the full set of provisions of an OSHA health standard apply, or only part of it, and under what exceptions?
- ✓ **Ground truth:** Only the requirement to limit employee exposure to the specific permissible exposure limit applies, unless that standard states otherwise or unless the conditions of paragraph (a)(2)(iii) apply.
- ✗ **Model answered:** {"name": "apply_osh_a_health_standard", "parameters": {"provisions": "partially", "exceptions": "paragraphs (d) and (g)(1)(ii) of this section"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model misinterpreted the scope of OSHA health standard applicability
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 254 — `OSHA\_29CFR\_full\_2024\_3327\_04` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** Under the Effective Date Note, until what date were certain requirements of § 1926.451(b)(2)(i) stayed?
- ✓ **Ground truth:** Nov. 25, 1997, or until further rulemaking has been completed, whichever is later
- ✗ **Model answered:** {"name": "Effective Date Note", "parameters": {"date": "Nov. 25, 1997"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not include the condition for further rulemaking
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 255 — `OSHA\_29CFR\_full\_2024\_1264\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Based on the interlock requirement in paragraph (d)(1) and the feed roll guard requirement in paragraph (d)(3), which operating condition do both provisions specifically restrict?
- ✓ **Ground truth:** Both restrict access to hazardous parts while the machine is operating: the (d)(1) interlock prevents operating the machine while the cover is open (and prevents raising the cover while in motion), and the (d)(3) feed roll guard must prevent reaching the nip while the machinery is in operation.
- ✗ **Model answered:** {"name": "operating_condition", "parameters": {"condition": "operation of the machine with the cover open"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports the common restriction as access/opening/reaching hazardous parts while the machinery is in motion/operation, but the model answered only the (d)(1)-specific condition of operating with the cover open, which is not a feed-roll condition in (d)(3).
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 256 — `OSHA\_29CFR\_full\_2024\_neg\_1769\_05` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** By what time of day must the employer provide records following a request?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "provide_records", "parameters": {"time": "four business hours"}}
- **Doctor diagnosis:** **Dr. No** bad test
- **Doctor root cause:** [verifier] The evidence directly answers the question with “By the end of the next business day after a request,” so the ground truth of UNANSWERABLE is unsupported, even though the model’s “four business hours” is also hallucinated.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 257 — `OSHA\_29CFR\_full\_2024\_4240\_02` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** Is a separable portion of a plan always limited to an issue as defined in 29 CFR 1902.2(c)?
- ✓ **Ground truth:** No. There is an exception: wherever the Assistant Secretary has determined that other industrial, occupational or hazard groupings are administratively practicable, such groupings shall be considered separable portions of a plan.
- ✗ **Model answered:** {"name": "is_separable_portion_limited", "parameters": {"plan": "29 CFR 1902.2(c)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model read incorrect definition
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 258 — `OSHA\_29CFR\_full\_2024\_4448\_01` · MEDIUM confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining paragraph (a)(2) and paragraph (b), if a complainant files a district court action because there was no final decision within 210 days of filing the complaint, what subsequent filing obligation arises and how soon?
- ✓ **Ground truth:** Because there was no final decision of the Secretary within 210 days of the filing of the complaint (a)(2), the complainant may bring the action; then within seven days after filing the federal complaint (b), they must file a file-stamped copy with the Assistant Secretary, ALJ, or ARB, depending on where the proceeding is pending.
- ✗ **Model answered:** {"name": "file_stamped_complaint_filing_obligation", "parameters": {"days": "7"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence requires that within seven days the complainant must file a copy of the file-stamped complaint “with the Assistant Secretary, the ALJ, or the ARB,” but the model only gave a generic obligation name and the 7-day timing, omitting the filing recipient required by the obligation.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 259 — `OSHA\_29CFR\_full\_2024\_neg\_1259\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How often must the limit switch on the powered hand truck operating handle be inspected?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "inspect_limit_switch", "parameters": {"frequency": "at the beginning of each operator"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 260 — `OSHA\_29CFR\_full\_2024\_4450\_02` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the Statutes at Large citation (124 Stat.) for the Patient Protection and Affordable Care Act, Public Law 111-148?
- ✓ **Ground truth:** 124 Stat. 119
- ✗ **Model answered:** {"name": "get_statutes_at_large_citation", "parameters": {"public_law": "111-148", "statutes_at_large": "124"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence explicitly gives the full citation as “Public Law 111-148, 124 Stat. 119,” while the model returned only “124,” omitting the required “Stat. 119” portion.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 261 — `OSHA\_29CFR\_full\_2024\_0022\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — a missing or incomplete answer — the model withheld an available answer (a gap, not a confidently wrong statement)
- **Category:** Conditional
- **Question:** If particularized written objections are filed and the Assistant Secretary finds that substantial objections have been filed, what action must he take?
- ✓ **Ground truth:** He shall afford a formal or informal hearing on the subjects and issues involved under § 1902.13 or § 1902.14, or shall commence a proceeding under § 1902.17.
- ✗ **Model answered:** {"name": "afford an informal hearing", "parameters": {}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states that if substantial objections are filed, the Assistant Secretary 'shall afford a formal or informal hearing... or shall commence a proceeding under § 1902.17,' so the model's answer limiting this to only 'afford an informal hearing' is incomplete and too narrow.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 262 — `OSHA\_29CFR\_full\_2024\_4010\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If the auditor determines there is a significant deficiency in the program, what is triggered?
- **✓ Ground truth:** The employer must ensure certain follow-up actions (as listed under the Deficiencies provision).
- **✗ Model answered:** {"name": "audit_determination", "parameters": {"determination": "significant deficiency"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 263 — `OSHA\_29CFR\_full\_2024\_2190\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If initial monitoring reveals an employee's exposure is below the action level, may the employer discontinue monitoring for that employee unconditionally?
- **✓ Ground truth:** No; the employer may discontinue except as otherwise required by paragraph (e)(5).
- **✗ Model answered:** {"name": "termination_of_monitoring", "parameters": {"initial_monitoring_reveals_exposure_below_action_level": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 264 — `OSHA\_29CFR\_full\_2024\_4228\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** When a developmental step is missed and the State has an operational status agreement with OSHA under § 1954.3, what additional assurance must the State provide?
- **✓ Ground truth:** The State must provide an assurance that the missed step will not affect the effectiveness of State enforcement in any issues for which the State program has been deemed to be operational.
- **✗ Model answered:** {"name": "provide_assurance", "parameters": {"agreement": "operational status agreement", "section": "1954.3"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 265 — `OSHA\_29CFR\_full\_2024\_1226\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually

- **Category:** Cross-Reference
- **Question:** Under the maintenance and operation provisions, when servicing a batch-type generator whose carbide charge is exhausted, what must be done to the generating chamber before adding carbide, and according to subpart (D), what restriction applies to the mixture drained out during this process?
- ✓ **Ground truth:** Before additional carbide is added, the generating chamber must always be flushed out with water, renewing the water supply in accordance with the manufacturer's instruction card. The water-carbide residue mixture drained from the generator shall not be discharged into sewer pipes or stored in areas near open flames (only clear water from residue settling pits may be discharged into sewer pipes).
- ✗ **Model answered:** {"name": "flush_out_generating_chamber", "parameters": {"water_supply": "renewed in accordance with the instruction card furnished by the manufacturer"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 266 — `OSHA\_29CFR\_full\_2024\_2019\_01` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Precision
- **Question:** What minimum percentage of beryllium by weight must materials contain to define a beryllium work area when processed during listed operations?
- ✓ **Ground truth:** 0.1 percent by weight
- ✗ **Model answered:** {"name": "define_beryllium_work_area", "parameters": {"percentage": "0"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence explicitly defines a beryllium work area as one where materials “contain at least 0.1 percent beryllium by weight,” so the model’s answer of 0 is incorrect and the ground truth is supported.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 267 — `OSHA\_29CFR\_full\_2024\_0075\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If the citation issued to an employer is for a de minimis violation, must it be posted?
- ✓ **Ground truth:** No, notices of de minimis violations need not be posted.
- ✗ **Model answered:** {"name": "posting_of_citations", "parameters": {"citation_type": "de minimis violation"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly answers the question—“Notices of de minimis violations need not be posted”—but the model only returned a function-like object with the citation type and did not provide the required yes/no answer.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 268 — `OSHA\_29CFR\_full\_2024\_0168\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference

- **Question:** Combining the standard in paragraph (c) with the discovery-related provision in the same paragraph, under what two distinct circumstances does the hearing examiner have authority over the motion?
- **Ground truth:** Under paragraph (c), the hearing examiner may grant the motion if the pleadings, affidavits, material obtained by discovery or otherwise, or officially noticed matters show there is no genuine issue as to any material fact and a party is entitled to summary decision; alternatively, the hearing examiner may deny the motion whenever the moving party denies access to information by means of discovery to a party opposing the motion.
- **Model answered:** {"name": "motion_for_summary_decision", "parameters": {"paragraph": "c"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 269 — `OSHA\_29CFR\_full\_2024\_neg\_4485\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the telephone number for the OSHA office that handles SPA complaints?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "get_telephone_number", "parameters": {"agency": "OSHA", "complaint_type": "SPA"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 270 — `OSHA\_29CFR\_full\_2024\_4047\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If personnel are being hoisted during a pile driving operation, may lifts be made on another of the equipment's load lines?
- **Ground truth:** Yes; the prohibition on lifts on other load lines while personnel are hoisted does not apply in pile driving operations.
- **Model answered:** {"name": "k", "parameters": {"11": "i"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 271 — `OSHA\_29CFR\_full\_2024\_neg\_2602\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Within what timeframe must a new classification be completed after detecting significant batch variation under A.0.5.1.2?

- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "A.0.5.1.2", "parameters": {"duration": "before the next batch is produced"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 272 — `OSHA\_29CFR\_full\_2024\_3703\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Using the field diameter calibration requirement and the worked example for field area, what is the smallest field diameter D (in mm) still within the acceptable range, and does the worked example use a value within that range?
- ✓ **Ground truth:** The acceptable range is $100 \pm 2 \mu\text{m}$, so the smallest acceptable diameter is $98 \mu\text{m} = 0.098 \text{ mm}$. The worked example uses $D = 100 \mu\text{m} = 0.1 \text{ mm}$, which is within the acceptable range.
- ✗ **Model answered:** {"name": "Field Area", "parameters": {"D": "0"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 273 — `OSHA\_29CFR\_full\_2024\_2815\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Within 1926.550, the paragraphs (a)(1), (a)(6), and (a)(11) each carry an OMB control number. Are all three the same?
- ✓ **Ground truth:** No. 1926.550(a)(1) is 1218-0115, 1926.550(a)(6) is 1218-0113, and 1926.550(a)(11) is 1218-0054 - all three are different.
- ✗ **Model answered:** {"name": "check_omb_control_numbers", "parameters": {"citation": "1926.550", "paragraphs": "['(a)(1)', '(a)(6)', '(a)(11)']"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not provide explicit answer to the question
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 274 — `OSHA\_29CFR\_full\_2024\_neg\_3991\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many feet of distance must be specified in a voice signal before the operator may begin a boom movement?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "voice signal", "parameters": {"distance": "10"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 275 — `OSHA\_29CFR\_full\_2024\_neg\_2625\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What weighting score is assigned to epidemiological studies versus individual human case reports in the weight-of-evidence approach?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "A.0.3.3", "parameters": {"human data": "reliable, good quality human data shall generally have precedence over other data", "animal data": "positive results from well-conducted animal studies are not necessarily negated by the lack of positive human experience but require an assessment of the robustness, quality and statistical power of both the human and animal data"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 276 — `OSHA\_29CFR\_full\_2024\_neg\_0429\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum permissible service pressure at the point where oxygen enters the supply line?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "pressure_regulating_devices", "parameters": {"maximum_pressure": "250"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 277 — `OSHA\_29CFR\_full\_2024\_neg\_3653\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many days does an employer have to offer a transfer position to an employee who cannot function while using a respirator?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "time_to_offer_transfer_position", "parameters": {"days": "10"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 278 — `OSHA\_29CFR\_full\_2024\_1206\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Per paragraph (b)(1)(i), construction must follow which regulations, and per (b)(1)(iii), connections must comply with which standard?
- **Ground truth:** Construction and maintenance must follow the U.S. Department of Transportation regulations, 49 CFR parts 171-179, and connections must comply with ANSI B57.1-1965 (the American National Standard Compressed Gas Cylinder Valve Outlet and Inlet Connections).
- **Model answered:** {"name": "construction_and_connections", "parameters": {"regulations": "29 CFR 1910.119", "standard": "Regulator Connection Standards, 1958, Compressed Gas Association"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model confused regulations and standards
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 279 — `OSHA\_29CFR\_full\_2024\_neg\_2902\_08` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What form must be filed to certify that air compressor quality meets the 29 CFR 1910.134(i) requirements?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "OSHA_certification_form", "parameters": {"form_type": "29 CFR 1910.134(i) Air Compressor Quality Certification", "company_name": "ABC Corporation", "date": "2022-01-01"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated by providing a specific certification form and company name not mentioned in the evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '2022' not in evidence

Failure 280 — `OSHA\_29CFR\_full\_2024\_neg\_4455\_09` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the specific interest rate under 26 U.S.C. 6621 that applies to back pay calculations?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "calculate_interest_rate", "parameters": {"year": "2020"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 281 — `OSHA\_29CFR\_full\_2024\_4205\_01` · MEDIUM confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the revised compliance staffing benchmarks in paragraph (c), what is the total number of compliance officers (safety plus health) proposed for Tennessee?
- ✓ **Ground truth:** 22 safety plus 14 health compliance officers equals 36 total compliance officers.
- ✗ **Model answered:** {"name": "add", "parameters": {"a": "22", "b": "14"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states revised benchmarks of "22 safety and 14 health compliance officers," so the requested total is 36; the model only returned an add-function call with 22 and 14 rather than the total number.
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 282 — `OSHA\_29CFR\_full\_2024\_neg\_1259\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many inches must the carton-stitching machine guard extend beyond the stitching head?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "carton-stitching machine guard", "parameters": {"extension_length": "3"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 283 — `OSHA\_29CFR\_full\_2024\_neg\_1194\_09` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What temperature must damp sand be maintained at to qualify as adequate floor protection?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "temperature_maintenance", "parameters": {"temperature": "below 85\u00b0F"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 284 — `OSHA\_29CFR\_full\_2024\_neg\_1037\_08` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream

- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum reach measurement for an alloy steel chain sling used in a vertical hitch?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "Reach", "parameters": {"x": "vertical hitch", "y": "alloy steel chain sling", "z": "minimum"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 285 — `OSHA\_29CFR\_full\_2024\_3059\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a spray booth has a frontal area of 8 square feet, is a metal deflector or curtain required at the upper outer edge?
- **Ground truth:** No, the requirement applies only to booths with a frontal area larger than 9 square feet.
- **Model answered:** {"name": "is_metal_deflector_required", "parameters": {"frontal_area": "8"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed incorrect result
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 286 — `OSHA\_29CFR\_full\_2024\_neg\_4455\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What form number must the respondent use to request the award of attorney fees from the ALJ?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "OSHA Form 20", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated form number
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '20' not in evidence

Failure 287 — `OSHA\_29CFR\_full\_2024\_neg\_0478\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum allowable air velocity over the open face of a booth during standard (non-electrostatic) spraying operations?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "get_max_air_velocity", "parameters": {"velocity_type": "standard"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 288 — `OSHA\_29CFR\_full\_2024\_neg\_3991\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What maximum decibel level is required for voice signals to be considered effective under this section?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "effective_decibel_level", "parameters": {"dB": "90"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 289 — `OSHA\_29CFR\_full\_2024\_0734\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** LOW — a formatting or precision discrepancy
- **Category:** Cross-Reference
- **Question:** Based on subparts (i) and (ii), what are the two distinct surrogate-facepiece-based approaches—one for qualitative testing and one for quantitative testing—that can be used in place of modifying the user's actual atmosphere-supplying or powered air-purifying respirator facepiece?
- ✓ **Ground truth:** For qualitative fit testing (subpart i), an identical negative pressure air-purifying respirator facepiece with the same sealing surfaces can be used as a surrogate; for quantitative fit testing (subpart ii), a surrogate facepiece with a permanent sampling probe installed can be used to sample air from inside the facepiece.
- ✗ **Model answered:** {"name": "fit_testing_of_tight_fitting_atmosphere_supplying_respirators_and_tight_fitting_powered_air_purifying_respirators", "parameters": {"protocol": "OSHA-accepted QNFT protocol", "facepiece": "surrogate facepiece with appropriate filters"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence supports the ground truth—QLFT may use “an identical negative pressure air-purifying respirator facepiece with the same sealing surfaces” and QNFT may use “a permanent sampling probe onto a surrogate facepiece”—but the model gave only a vague/incorrect single surrogate-with-filters answer and omitted the permanent sampling probe approach.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 290 — `OSHA\_29CFR\_full\_2024\_neg\_3991\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What form must the lift director complete to record the agreed-upon voice signals?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "OSHA 301", "parameters": {"form": "Injury and Illness Incident Report"}}

- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 291 — `OSHA\_29CFR\_full\_2024\_neg\_0484\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** At what specific pressure (in psi) must the relief valve on air-pressure containers be set to operate?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "relief valve start-to-discharge pressure setting", "parameters": {"percent of regulator delivery pressure": "200", "Minimum": "300"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 292 — `OSHA\_29CFR\_full\_2024\_neg\_2902\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum compressive load rating that the safety-toe footwear must withstand under standard Z41.1-1967?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_min_compress_load", "parameters": {"standard": "Z41.1-1967"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 293 — `OSHA\_29CFR\_full\_2024\_neg\_3806\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which specific government office reviews the licensed physician's written medical report for compliance?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "OSHA_Review", "parameters": {"report": "Licensed physician"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only says the employer must ensure the employee receives a written medical report and that the PLHCP explains results; it does not mention any government office reviewing the report, while the model fabricated an OSHA-related answer instead of refusing.
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 294 — `OSHA\_29CFR\_full\_2024\_4446\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If the Secretary issues a decision upon discretionary review for which judicial review is available, within what time must a petition for review be filed?
- ✓ **Ground truth:** Within 60 days after the issuance of the final order.
- ✗ **Model answered:** {"name": "time_to_file_petition_for_review", "parameters": {"days": "10"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed incorrect time limit
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 295 — `OSHA\_29CFR\_full\_2024\_neg\_4297\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum number of members a safety and health committee must have to be considered 'certified' under this section?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "committee_certification", "parameters": {"members": "2"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 296 — `OSHA\_29CFR\_full\_2024\_neg\_1037\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many proof tests must a sling manufacturer perform before a sling is approved for sale?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "proof_testing", "parameters": {"number_of_tests": "1"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 297 — `OSHA\_29CFR\_full\_2024\_neg\_2129\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)

- **Question:** How often must fit testing be repeated under this section's respirator program?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "fit testing", "parameters": {"frequency": "at least annually"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 298 — `OSHA\_29CFR\_full\_2024\_neg\_3991\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Within how many hours of a worker being substituted must the voice signal meeting be reconvened?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "reconvene_voice_signal_meeting", "parameters": {"hours": "1"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 299 — `OSHA\_29CFR\_full\_2024\_neg\_4485\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How long does OSHA have to complete its investigation after a SPA complaint is filed?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "OSHA investigation timeline", "parameters": {"days": "180"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 300 — `OSHA\_29CFR\_full\_2024\_2632\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a substance shows a high potency in animals but only a low frequency of occurrence in humans, which sub-category applies?
- ✓ **Ground truth:** Sub-category 1A, since high potency in animals presumes potential to produce significant sensitization.
- ✗ **Model answered:** {"name": "classify_substance", "parameters": {"substance_name": "substance_name", "human_frequency": "low", "animal_potency": "high"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** (none)

Failure 301 — `OSHA\_29CFR\_full\_2024\_neg\_3512\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum spacing allowed between cleats used to skidproof ramped walkway surfaces?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_max_cleat_spacing", "parameters": {"cm": "14"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only says ramped walkway surfaces must be skidproofed "by cleats or by equivalent means" and gives no maximum cleat spacing, so the ground truth UNANSWERABLE is supported while the model fabricated "14 cm."
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 302 — `OSHA\_29CFR\_full\_2024\_neg\_1599\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What penalty applies if a receptacle is installed that does not conform to the ratings in Table S-5?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "violation_penalty", "parameters": {"violation_type": "receptacle_rating_violation"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated penalty for receptacle rating violation
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 303 — `OSHA\_29CFR\_full\_2024\_neg\_2902\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What maximum permissible concentration of dust is allowed in respirator air before it is considered a harmful quantity?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "maximum_use_concentration", "parameters": {"MUC": "1000"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated a specific number not present in the evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '1000' not in evidence

Failure 304 — `OSHA\_29CFR\_full\_2024\_2784\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** For an OSHA regulated substance with exposure monitoring and medical surveillance requirements, if the action level is routinely exceeded in a laboratory, which paragraphs apply?
- ✓ **Ground truth:** Paragraphs (d) and (g)(1)(ii) of this section shall apply.
- ✗ **Model answered:** {"name": "OSHA regulated substance with exposure monitoring and medical surveillance requirements", "parameters": {"action level exceeded": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states that when the action level is routinely exceeded, “paragraphs (d) and (g)(1)(ii) of this section shall apply,” but the model only restated the condition and did not provide the applicable paragraphs.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 305 — `OSHA\_29CFR\_full\_2024\_neg\_4455\_08` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What standard does the ALJ use to determine whether a complaint was frivolous or brought in bad faith?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "ALJ_determines_frivolous_or_bad_faith", "parameters": {"complaint": "complaint", "standard": "reasonable attorney"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 306 — `OSHA\_29CFR\_full\_2024\_neg\_4297\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What penalty applies to an agency that obstructs an OSHA inspector's entry under paragraph (b)?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "paragraph", "parameters": {"number": "8", "letter": "B"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 307 — `OSHA\_29CFR\_full\_2024\_1848\_04` · MEDIUM confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the friability information in Section I.C with the disposal waste types in Section IV.A, why would vinyl-asbestos floor tile removed during demolition be considered a disposal concern?
- **Ground truth:** Although vinyl-asbestos floor tile is nonfriable and generally does not emit fibers (I.C), building products containing asbestos removed during renovation or demolition are listed as a waste in IV.A.5, and per I.C these materials can emit fibers if cut, sawed, or broken during demolition.
- **Model answered:** {"name": "disposal_concern", "parameters": {"waste_type": "vinyl-asbestos floor tile", "friability": "non-friable", "concern": "emits airborne fibers when subjected to aggressive operations such as sanding or sawing"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 308 — `OSHA\_29CFR\_full\_2024\_neg\_1037\_11` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum slot width permitted for a female handle on a metal mesh choker sling?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "get_max_slot_width", "parameters": {"handle_type": "female", "sling_type": "metal mesh choker sling"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 309 — `OSHA\_29CFR\_full\_2024\_neg\_3991\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which specific certified language proficiency test must the operator and signal person pass to demonstrate effective communication?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "language and literacy requirements", "parameters": {"literacy": "relevant to the work", "language": "the language of the certification"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 310 — `OSHA\_29CFR\_full\_2024\_1202\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under paragraphs (d)(1)(iii) and (d)(1)(iv), what two distinct hazards require protective measures, and what is required for each?
- **Ground truth:** Under (d)(1)(iii), arc welding in wet conditions or high humidity requires special protection against electric shock. Under (d)(1)(iv), pressure testing of pipelines requires protecting workers and the public against injury from blowing out of closures or pressure restraining devices and against expulsion of trapped loose dirt.
- **Model answered:** {"name": "get_hazard_prevention_measures", "parameters": {"hazard_1": "hazard identification and awareness", "hazard_2": "notification of appropriate persons"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model confused specific hazard prevention measures
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 311 — `OSHA\_29CFR\_full\_2024\_neg\_4403\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What form number must an employee use to file a complaint under this part?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "OSHA-13", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '13' not in evidence

Failure 312 — `OSHA\_29CFR\_full\_2024\_0884\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining the inspection requirement of paragraph (e)(1) and the employee check requirement of paragraph (e)(2)(i), what two conditions concerning the work area must be ensured before energy is restored?
- **Ground truth:** Per (e)(1), the work area must be inspected to ensure nonessential items have been removed and that machine or equipment components are operationally intact; and per (e)(2)(i), the work area must be checked to ensure all employees have been safely positioned or removed.
- **Model answered:** {"name": "paragraph", "parameters": {"e": "1", "n": "2"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not synthesize rules from multiple sections
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 313 — `OSHA\_29CFR\_full\_2024\_2250\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** A facility uses Universal Precautions for all specimens. A specimen container is being moved from one lab to another lab within the same facility. Is labeling or color-coding required?
- ✓ **Ground truth:** No, the labeling/color-coding exemption applies while specimens remain within the facility, provided containers are recognizable as containing specimens.
- ✗ **Model answered:** {"name": "labeling_or_color_coding_required", "parameters": {"universal_precautions": "True", "within_facility": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 314 — `OSHA\_29CFR\_full\_2024\_4301\_04` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** At the national level, if employees are not represented by organizations having exclusive recognition on an agency basis or national consultation rights, how are nonmanagement members determined?
- ✓ **Ground truth:** They shall be determined through procedures devised by the agency which provide for effective representation of all employees.
- ✗ **Model answered:** {"name": "determine_nonmanagement_members", "parameters": {"level": "national", "representation": "not represented"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 315 — `OSHA\_29CFR\_full\_2024\_2019\_00` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What is the action level concentration of airborne beryllium, calculated as an 8-hour TWA?
- ✓ **Ground truth:** 0.1 micrograms per cubic meter of air ($\mu\text{g}/\text{m}^3$)
- ✗ **Model answered:** {"name": "action level", "parameters": {"concentration": "0"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 316 — `OSHA\_29CFR\_full\_2024\_neg\_1599\_01` · MEDIUM confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)

- **Question:** What is the minimum wattage rating required for a heavy-duty lampholder of the medium type?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_minimum_wattage_rating", "parameters": {"type": "admedium", "rating": "heavy-duty"}}
- **Doctor diagnosis:** **Dr. No** bad test
- **Doctor root cause:** [verifier] The ground truth is unsupported because the evidence does answer by category: a heavy-duty lampholder is at least 660 W only 'if of the admedium type' and 'not less than 750 watts if of any other type,' so a medium type is covered as an other type rather than being unanswerable.
- **Verifier verdict:** bad test (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 317 — `OSHA\_29CFR\_full\_2024\_neg\_1769\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which two forms of contact information must be included for each employee in the COVID-19 log?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_contact_info", "parameters": {"employee_id": "12345", "contact_info_types": ["phone number", "email address"]}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '12345' not in evidence

Failure 318 — `OSHA\_29CFR\_full\_2024\_3801\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Combining paragraph (h)(1) and paragraph (h)(2)(ii), what triggers the requirement to provide PPE and what document governs both its use and its removal?
- ✓ **Ground truth:** PPE must be provided where airborne exposure exceeds or can reasonably be expected to exceed the TWA PEL or STEL (h)(1), and both its use and its removal are governed by the written exposure control plan required by paragraph (f)(1) (h)(1) and (h)(2)(ii).
- ✗ **Model answered:** {"name": "combining_paragraphs", "parameters": {"h_1": "The employer is not required to pay for non-specialty safety-toe protective footwear (including steel-toe shoes or steel-toe boots) and non-specialty prescription safety eyewear, provided that the employer permits such items to be worn off the job-site.", "h_2": "When the employer provides metatarsal guards and allows the employee, at his or her request, to use shoes or boots with built-in metatarsal protection, the employer is not required to reimburse the employee for the shoes or boots."}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model failed to synthesize rules from multiple regulatory sections
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 319 — `OSHA\_29CFR\_full\_2024\_neg\_4303\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum number of training hours committee members must receive under paragraph (b)?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_minimum_training_hours", "parameters": {"x": "b"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 320 — `OSHA\_29CFR\_full\_2024\_neg\_3653\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum allowable fit-test failure rate before an employer must reassign an employee from asbestos respirator work?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "fit_test_failure_rate", "parameters": {"failure_rate": "10"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 321 — `OSHA\_29CFR\_full\_2024\_4366\_03` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** If a timely petition for review is filed, when does the ALJ's decision NOT automatically become the final order of the Secretary?
- ✓ **Ground truth:** Unless the ARB, within 30 days of the filing of the petition, issues an order notifying the parties that the case has been accepted for review.
- ✗ **Model answered:** {"name": "ALJ decision NOT automatically final order", "parameters": {"timely petition for review filed": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 322 — `OSHA\_29CFR\_full\_2024\_neg\_1194\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as

authoritative and is hard to catch downstream

- **Category:** Negative (hallucination traps)
- **Question:** What penalty applies to a worker who performs cutting in a sprinklered building while protection is impaired?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "prohibited_areas", "parameters": {"C": "In sprinklered buildings while such protection is impaired"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 323 — `OSHA\_29CFR\_full\_2024\_2815\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Sections 1926.403, 1926.404, 1926.405, 1926.407, and 1926.408 all fall under which single OMB control number?
- ✓ **Ground truth:** All five sections (1926.403, 1926.404, 1926.405, 1926.407, and 1926.408) are assigned OMB control No. 1218-0130.
- ✗ **Model answered:** {"name": "get_omb_control_number", "parameters": {"cfr_citations": "[1926.403, 1926.404, 1926.405, 1926.407, 1926.408]"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 324 — `OSHA\_29CFR\_full\_2024\_1434\_06` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an employee exposed to electric arc hazards wears nylon clothing, but the employer demonstrates the fabric was treated to withstand the conditions the employee may encounter, is the clothing prohibited?
- ✓ **Ground truth:** No; the prohibition does not apply if the employer demonstrates the fabric has been treated to withstand the conditions or that the employee wears it in a manner that eliminates the hazard.
- ✗ **Model answered:** {"name": "prohibits_clothing", "parameters": {"p": "nylon", "x": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 325 — `OSHA\_29CFR\_full\_2024\_neg\_1037\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the rated capacity in pounds for a cable laid grommet-hand tucked sling?

- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get Rated capacity", "parameters": {"sling_type": "cable laid grommet-hand tucked"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 326 — `OSHA\_29CFR\_full\_2024\_4301\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** At an establishment where employees are represented under collective bargaining arrangements, how are nonmanagement committee members appointed?
- ✓ **Ground truth:** They shall be appointed from among those recommended by the exclusive bargaining representative.
- ✗ **Model answered:** {"name": "appointment_of_nonmanagement_committee_members", "parameters": {"representation_type": "collective bargaining arrangements"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 327 — `OSHA\_29CFR\_full\_2024\_4031\_01` · HIGH confidence

CALCULATION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Calculation
- **Question:** A boom-attached personnel platform is used on equipment whose rated capacity for the radius and configuration is 24,500 pounds. What is the maximum total weight of the loaded personnel platform allowed, except during proof testing?
- ✓ **Ground truth:** 12,250 pounds
- ✗ **Model answered:** {"name": "calculate_total_weight", "parameters": {"rated_capacity": "24500"}}
- **Doctor diagnosis:** **Dr. Pi** model error
- **Doctor root cause:** [verifier] The evidence states the loaded boom-attached personnel platform 'must not exceed 50 percent' of rated capacity, so 50% of 24,500 is 12,250 pounds; the model output only a function call with rated_capacity 24500 and did not provide the required maximum weight.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 328 — `OSHA\_29CFR\_full\_2024\_neg\_4403\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many days does an administrative law judge have to issue a decision after the post-hearing review?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "issue_decision", "parameters": {"days": "30"}}

- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 329 — `OSHA\_29CFR\_full\_2024\_neg\_2913\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum allowable negative pressure permitted within the inlet ductwork before additional sealing measures are mandated?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_max_negative_pressure", "parameters": {"ductwork": "inlet"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 330 — `OSHA\_29CFR\_full\_2024\_neg\_0478\_00` · MEDIUM confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the required testing frequency for inspecting filter rolls to ensure proper replacement of filter media?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "filter_roll_inspection_frequency", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 331 — `OSHA\_29CFR\_full\_2024\_neg\_1599\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum branch-circuit conductor ampacity required for a cord- and plug-connected arc welder receptacle?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_minimum_ampacity", "parameters": {"circuit_rating": "15 or 20", "receptacle_rating": "15"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 332 — `OSHA\_29CFR\_full\_2024\_1695\_01` · MEDIUM confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Both paragraph (a)(2)(ii) (portable equipment) and paragraph (c)(2) (test instruments) describe what to do upon finding a defect that might expose an employee to injury. What common requirement do both impose?
- ✓ **Ground truth:** In both cases the defective or damaged item shall be removed from service, and no employee may use it until repairs and tests necessary to render the equipment safe have been made.
- ✗ **Model answered:** {"name": "requirement_imposed_by_OSHA", "parameters": {"requirement": "removal from service and repairs"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** (none)

Failure 333 — `OSHA\_29CFR\_full\_2024\_neg\_4485\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Within how many days after filing must OSHA notify the named respondent of the complaint?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "notify_respondent", "parameters": {"days": "20"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 334 — `OSHA\_29CFR\_full\_2024\_neg\_1769\_02` · MEDIUM confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** How many years must the COVID-19 log be preserved after this section is no longer in effect?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "calculate_log_preservation_years", "parameters": {"x": "3"}}
- **Doctor diagnosis:** **Dr. No** bad test
- **Doctor root cause:** [verifier] The evidence says the log “must be maintained and preserved while this section remains in effect,” which supports no required post-expiration preservation period rather than an unspecified 3-year period, so the model hallucinated but the UNANSWERABLE ground truth is also not supported.
- **Verifier verdict:** bad test (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 335 — `OSHA\_29CFR\_full\_2024\_neg\_4403\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as

authoritative and is hard to catch downstream

- **Category:** Negative (hallucination traps)
- **Question:** Within how many days after the alleged violation must an employee file a discrimination complaint under the Pipeline Safety Act?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "file_discrimination_complaint", "parameters": {"days": "180"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 336 — `OSHA\_29CFR\_full\_2024\_neg\_0484\_02` · MEDIUM confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How often, in days or months, must pressure hoses and couplings be inspected?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "inspect_pressure_hoses_and_couplings", "parameters": {"frequency": "regular intervals"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 337 — `OSHA\_29CFR\_full\_2024\_neg\_2129\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the PEL (permissible exposure limit) for cadmium mentioned in subsection (vi)?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_PEL", "parameters": {"vi": "5"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 338 — `OSHA\_29CFR\_full\_2024\_neg\_1738\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum allowable distance, in feet, that defines an AIIR's negative pressure boundary?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "calculate_distance", "parameters": {"x": "6"}}

- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 339 — `OSHA\_29CFR\_full\_2024\_4251\_00` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an interested person submits a petition to participate as a party 35 days after the notice was published in the Federal Register, does it meet the timing requirement of paragraph (b)(1)?
- ✓ **Ground truth:** No, because the petition must be submitted within 30 days after the notice has been published in the Federal Register.
- ✗ **Model answered:** {"name": "check_timing_requirement", "parameters": {"days": "35"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed time incorrectly
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 340 — `OSHA\_29CFR\_full\_2024\_neg\_4303\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How often must agencies update the Material Safety Data Sheets made available to committees?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "update_MSDS", "parameters": {"frequency": "annually"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 341 — `OSHA\_29CFR\_full\_2024\_neg\_1037\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum angle of loading permitted for a basket hitch configuration before it must be re-rated?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "get_angle_of_loading", "parameters": {"angle": "90"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '90' not in evidence

Failure 342 — `OSHA\_29CFR\_full\_2024\_neg\_3653\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum efficiency rating, expressed as a percentage, required for the HEPA filters provided for air-purifying respirators?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "HEPA_filter_efficiency", "parameters": {"percentage": "99"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated a specific efficiency rating
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '99' not in evidence

Failure 343 — `OSHA\_29CFR\_full\_2024\_0772\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Based on paragraphs (a)(7), (a)(8), and (a)(9), what is the total maximum number of squeezes performed before a subject is declared unable to taste saccharin per paragraph (a)(11)?
- ✓ **Ground truth:** Thirty squeezes - ten in the first set, ten in the second set, and ten in the third set. If saccharin is not tasted after these 30 squeezes, the subject is unable to taste saccharin and may not perform the saccharin fit test.
- ✗ **Model answered:** {"name": "calculate_total_squeezes", "parameters": {"x": "10", "y": "10", "z": "10"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 344 — `OSHA\_29CFR\_full\_2024\_neg\_4297\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How frequently are the scheduled inspections under paragraph (c) conducted during an agency's program evaluation?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "paragraph", "parameters": {"c": "as deemed necessary by the Secretary"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 345 — `OSHA\_29CFR\_full\_2024\_neg\_0484\_07` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What gauge of wire cable is required when containers are suspended rather than placed on metal supports?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "wire_cable_gauge", "parameters": {"suspension": "True"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 346 — `OSHA\_29CFR\_full\_2024\_neg\_2625\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which standardized animal test method must be used to obtain the 'positive results from an appropriate animal test' for respiratory sensitizer classification?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "A.4.2.1.3.1", "parameters": {"2": "recognized and validated animal models for the testing of respiratory hypersensitivity are not available", "3": "the potential of a substance to cause sensitization by inhalation in humans"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 347 — `OSHA\_29CFR\_full\_2024\_0643\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Based on the provisions for documenting responses and retaining reports, what must an employer document regarding each audit finding, and which reports must be kept available?
- ✓ **Ground truth:** The employer must promptly determine and document an appropriate response to each finding and document that deficiencies have been corrected, while retaining the two most recent compliance audit reports.
- ✗ **Model answered:** {"name": "documenting_responses", "parameters": {"action_taken": "what was done and why", "corrective_action": "the corrective action to be taken noted", "deficiency_identified": "each deficiency identified is addressed"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence requires both documenting “an appropriate response to each of the findings” and that “deficiencies have been corrected,” plus retaining “the two (2) most recent compliance audit reports,” but the model omitted the report-retention requirement and did not accurately state the required documentation.
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** (none)

Failure 348 — `OSHA\_29CFR\_full\_2024\_3172\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** For a horizontal overhead belt whose lower part is 8 feet above the floor, located over a passageway, with a center-to-center pulley distance of 12 feet, traveling 1,000 feet per minute, and 10 inches in width, must it be guarded for its entire length under paragraph (e)(2)(ii)?
- **Ground truth:** Yes. Although it travels less than 1,800 feet per minute (failing condition (a)), it must be guarded for its entire length because the center-to-center distance is 10 feet or more (condition (b), here 12 feet) and the belt is 8 inches or more in width (condition (c), here 10 inches). Any one of the conditions being met requires full-length guarding.
- **Model answered:** {"name": "guard_horizontal_overhead_belt", "parameters": {"distance_from_floor": "8", "location": "passageway", "pulley_distance": "12", "traveling_speed": "1000", "belt_width": "10"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 349 — `OSHA\_29CFR\_full\_2024\_1332\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Considering the dry deck storage requirement to minimize accidental rolling and the log decks 'Stops' provision, what two distinct kinds of devices/arrangements does the evidence require to prevent logs from rolling in these respective areas?
- **Ground truth:** For dry deck storage (paragraph (v)(c)), logs shall be arranged to minimize the chance of accidentally rolling from the deck; for log decks (paragraph (3)(ii) 'Stops'), log decks shall be provided with adequate stops, chains, or other safeguards to prevent logs from rolling down the deck onto the carriage or its runway.
- **Model answered:** {"name": "Log decks - Stops", "parameters": {"arrangements": "adequate stops, chains, or other safeguards"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not fully address the question about dry deck storage
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 350 — `OSHA\_29CFR\_full\_2024\_3641\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Both the glove bag work practices and the negative pressure glove box work practices specify a minimum staffing level. How many persons are required to perform each operation?
- **Ground truth:** At least two persons are required for Class I glovebag removal operations, and at least two persons are required to perform the negative pressure glove box removal.
- **Model answered:** {"name": "glove_bag_work_practices", "parameters": {"persons_required": "2"}}

- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 351 — `OSHA\_29CFR\_full\_2024\_neg\_0478\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the minimum size or capacity required for the water-filled metal container used to store discarded filter pads?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "get_minimum_size", "parameters": {"unit": "gallons"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only says discarded pads may be placed in a “water-filled metal container” and gives no minimum size or capacity, so the ground truth UNANSWERABLE is supported, while the model did not refuse or state unanswerable and instead issued a get_minimum_size call with gallons.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 352 — `OSHA\_29CFR\_full\_2024\_neg\_4303\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many days does an agency have to provide requested information to a committee under paragraph (a)?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "agency_responsibilities", "parameters": {"days": "15"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 353 — `OSHA\_29CFR\_full\_2024\_2160\_00` · HIGH confidence

CAUSE-EFFECT ERROR

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cause & Effect
- **Question:** Why was the standard deviation of the 0.2 ng/mL cadmium standard (0.75 abs-sec), rather than the reagent blank, used to calculate the detection limits by the IUPAC method?
- **Ground truth:** Because the near lack of a blank signal (five of the six reagent blank peak area readings were zero) does not satisfy a strict interpretation of the IUPAC method for determining the detection limits.
- **Model answered:** {"name": "calculate_detection_limits", "parameters": {"sd": "0", "m": "51"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** (none)

Failure 354 — `OSHA\_29CFR\_full\_2024\_neg\_3653\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which OSHA form must employers use to document the negative exposure assessment referenced for Class I asbestos work?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "OSHA Form", "parameters": {"Form ID": "None", "Form Name": "Employers must use a written opinion from a qualified person, not an OSHA form, to document the negative exposure assessment for Class I asbestos work."}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence only references situations where a negative exposure assessment is unavailable and does not mention any OSHA form or documentation method, while the model asserted 'None' and invented a 'written opinion from a qualified person' rather than simply saying it is not in evidence.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 355 — `OSHA\_29CFR\_full\_2024\_neg\_4303\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What penalty applies to an agency that fails to provide relevant information to a committee?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "agency_responsibilities", "parameters": {"a": "penalty for failing to provide relevant information"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 356 — `OSHA\_29CFR\_full\_2024\_neg\_1194\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Who in management must sign off to authorize cutting or welding in a restricted area?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "designated management representative", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 357 — `OSHA\_29CFR\_full\_2024\_neg\_1259\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the penalty for permitting a person to ride on a powered hand truck not designed for riding?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "penalty_for_riding_on_powered_hand_truck", "parameters": {"penalty": "shall not be permitted"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence states only the prohibition—"No person shall be permitted to ride on a powered hand truck unless it is so designed by the manufacturer"—and gives no penalty, so the ground truth UNANSWERABLE is supported while the model incorrectly treated the prohibition as the penalty.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 358 — `OSHA\_29CFR\_full\_2024\_2160\_01` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cause & Effect
- **Question:** What were the consequences for analytical bias and OAE resulting from the average recovery of 118.2% at 0.1 times and 104.0% in the range of 0.5 to 2.0 times the TWA PEL target concentration in the flame AAS analysis?
- ✓ **Ground truth:** The analytical bias found over the tested concentration range was +4.0% and the OAE was ±6.0%.
- ✗ **Model answered:** {"name": "calculate_consequences", "parameters": {"recovery_1": "118", "recovery_2": "104"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model computed ratio instead of quoting stated figure
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 359 — `OSHA\_29CFR\_full\_2024\_neg\_3806\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What specific airborne exposure level (in $\mu\text{g}/\text{m}^3$) triggers the requirement to provide this information to the physician?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "information_to_physician", "parameters": {"actual_or_representative_exposure_level": "5"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 360 — `OSHA\_29CFR\_full\_2024\_2784\_03` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If an OSHA health standard specifies prohibition of eye and skin contact, must a laboratory observe it even though only the permissible exposure limit requirement generally applies?
- **Ground truth:** Yes, prohibition of eye and skin contact where specified by any OSHA health standard shall be observed.
- **Model answered:** {"name": "§ 1910.1450", "parameters": {"a": "2", "i": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence directly answers yes—"Prohibition of eye and skin contact where specified by any OSHA health standard shall be observed"—but the model only returned a citation-like JSON to §1910.1450(a)(2)(i)/True rather than the required affirmative rule, and even points to the PEL-only subparagraph rather than the eye/skin-contact provision.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 361 — `OSHA\_29CFR\_full\_2024\_1964\_02` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** According to paragraphs (l)(2)(ii) and (l)(6), what must the sign at hazardous operations areas state, and what restriction applies to any additional statements placed near that sign?
- **Ground truth:** Under (l)(2)(ii) the sign must read: 'DANGER VINYL CHLORIDE MAY CAUSE CANCER WEAR RESPIRATORY PROTECTION AND PROTECTIVE CLOTHING IN THIS AREA AUTHORIZED PERSONNEL ONLY'. Under (l)(6), no statement may appear on or near it that contradicts or detracts from the effect of any required warning, information, or instruction.
- **Model answered:** {"name": "sign_at_hazardous_operations_areas", "parameters": {"text": "No Smoking or Open Flames", "restriction": "No other warning or caution signs are permitted in the area"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model did not accurately convey the required warning and restriction for the sign at hazardous operations areas
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 362 — `OSHA\_29CFR\_full\_2024\_neg\_4485\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What penalty does an employer face if found to have retaliated against a seaman under SPA?
- **Ground truth:** UNANSWERABLE
- **Model answered:** {"name": "penalty_for_retaliation", "parameters": {"amount": "litigation costs, expert witness fees, and reasonable attorney fees"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated penalty for retaliation
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 363 — `OSHA\_29CFR\_full\_2024\_neg\_4303\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** Which agency office is responsible for compiling the accident, injury, and illness data referenced in paragraph (a)?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "Office of Federal Agency Programs", "parameters": {}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 364 — `OSHA\_29CFR\_full\_2024\_neg\_4403\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the deadline for filing objections to the findings and preliminary orders?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "filing objections", "parameters": {"deadline": "30 days"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 365 — `OSHA\_29CFR\_full\_2024\_neg\_3991\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How long after agreeing on voice signals must the workers document their agreement in writing?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** {"name": "document_agreement", "parameters": {"duration": "30", "unit": "calendar days"}}
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** model hallucinated
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 366 — `OSHA\_29CFR\_full\_2024\_3987\_03` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** If a warning (tag-out or do not operate) sign is on the equipment or starting control, when may

the operator activate the switch or start the equipment?

- ✓ **Ground truth:** Only after the sign has been removed by a person authorized to remove it, or after the operator has verified that (A) no one is servicing, working on, or otherwise in a dangerous position on the machine, and (B) the equipment has been repaired and is working properly.
- ✗ **Model answered:** {"name": "activate_switch_or_start_equipment", "parameters": {"sign_removed": "True", "verified_safe": "True"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence says the operator must not start until the sign is removed by an authorized person OR until the operator verifies both safety conditions; the model's JSON with both sign_removed and verified_safe set to true does not capture the alternative OR condition or the authorized-removal requirement.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 367 — `OSHA\_29CFR\_full\_2024\_3110\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Under the marine service station rules, if a tank is installed on a pier (rather than on shore) and is located at an elevation producing gravity head on the dispensing unit, what two requirements must be met regarding maximum stored quantity and the tank outlet?
- ✓ **Ground truth:** Per paragraph (k)(2)(ii), the quantity stored on the pier may not exceed 1,100 gallons (4,163.5 L) aggregate capacity, and per paragraph (k)(2)(iv), the tank outlet must be equipped with a pressure control valve positioned adjacent to and outside the tank block valve, adjusted so liquid cannot flow by gravity in case of piping or hose failure.
- ✗ **Model answered:** {"name": "marine_service_station_tank_requirements", "parameters": {"elevation_producing_gravity": "True", "stored_quantity": "1100", "tank_outlet_requirement": "pressure control valve"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 368 — `OSHA\_29CFR\_full\_2024\_0913\_03` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** What is the content focus of the program required in (g)(1) versus the training required in (g)(3)?
- ✓ **Ground truth:** The (g)(1) educational program familiarizes employees with the general principles of fire extinguisher use and the hazards involved with incipient stage fire fighting, whereas the (g)(3) training covers the use of the appropriate fire fighting equipment for designated employees.
- ✗ **Model answered:** {"name": "content_focus", "parameters": {"x": "program required in (g)(1)", "y": "training required in (g)(3)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** unknown
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Using the transfer of records provisions, what must the successor employer do when the original employer ceases to do business, and what additional regulation governs the transfer?
- **Ground truth:** Whenever the employer ceases to do business, the successor employer shall receive and retain all records required to be maintained by paragraph (m) (paragraph (4)(i)). The employer must also comply with any additional requirements involving transfer of records set forth in 29 CFR 1910.1020(h) (paragraph (4)(ii)).
- **Model answered:** {"name": "transfer of records", "parameters": {"paragraph": "29 CFR 1910.1020(h)"}}
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The evidence states the successor employer 'shall receive and retain all records required to be maintained by paragraph (m)' and also comply with '29 CFR 1910.1020(h),' but the model only identified the regulation and omitted the required receive-and-retain duty.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

EU AI Act — Conformity Evidence Statement

This report documents accuracy and robustness testing of the model under test against a representative, purpose-generated question set, with independent verification of each verdict. The methodology and results are intended to serve as **accuracy and robustness evidence in support of Article 15 (Accuracy, Robustness and Cybersecurity)** of Regulation (EU) 2024/1689 (the EU AI Act), and as a testing-and-validation record suitable for inclusion in the technical documentation under Article 11 and Annex IV.

This document is testing evidence only. It does not by itself constitute, certify, or confer conformity with the EU AI Act, and does not replace the conformity assessment, risk-management, human-oversight, or other obligations that rest with the system's provider or deployer.