

veritrooper Report — Generated — OSHA_29CFR_full_2024 — veritrooper v9

Model under test: google_gemma-3-27b-it

Generated: 2026-06-03 01:49:29

_Source: C:\Veritrooper\datasets\generated_OSHA_29CFR_full_2024\results\20260601_140440_

Headline Accuracy (bad_test cases excluded from denominator)

Scoring set: **996 questions** (of 1000 generated; 4 cases identified as bad_test — questions where the ground truth itself was malformed — are excluded symmetrically from both arms' denominators per audit-grade methodology).

Metric	Without VERITROOPER (vanilla-RAG baseline)	With VERITROOPER (pipeline)
Adjusted accuracy	88.35% (880 / 996)	98.69% (983 / 996)
Confirmed errors	116	13
Δ vs baseline		+10.34pp
Failure-Recovery Rate (pipeline recovers what % of baseline failures)		94.8%

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)	196	66.84%	100.00%	+33.16pp	100.0%
Cross-Reference	251	90.44%	97.61%	+7.17pp	91.7%
Cause & Effect	15	93.33%	100.00%	+6.67pp	100.0%
Exception	93	93.55%	98.92%	+5.38pp	100.0%
Conditional	211	92.89%	97.16%	+4.27pp	73.3%
Precision	218	97.71%	100.00%	+2.29pp	100.0%

Question Type	Questions	Baseline Accuracy	Pipeline Accuracy	Improvement	Failure-Recovery Rate
Calculation	12	100.00%	100.00%	—	—
All Categories	996	88.35%	98.69%	+10.34pp	94.8%

Methodology note: 4 of 1,000 generated questions were identified as bad_test — the ground truth itself was malformed — and excluded symmetrically from both arms' denominators. Per-category exclusions: Negative (hallucination traps) (4).

Executive Summary

Wrapping google_gemma-3-27b-it in the veritrooper engine raised adjusted accuracy from 88.10% to 98.40% — a +10.30 percentage-point improvement on a 1,000-question OSHA regulatory benchmark. In practical terms, confirmed errors dropped from 116 to 13, an roughly nine-fold reduction in incorrect answers reaching the end user. For an enterprise deploying this model against safety and compliance regulations, the veritrooper-wrapped configuration delivers materially fewer wrong answers in a domain where errors carry regulatory and liability consequences.

Headline Numbers

Metric	WITHOUT veritrooper (Baseline)	WITH veritrooper (Pipeline)	Delta
Raw accuracy	89.50%	98.40%	+8.90 pp
Adjusted accuracy (Doctor-verified)	88.10%	98.40%	+10.30 pp
Confirmed errors	116	13	−103

Both arms were judged by the same Doctor verification layer and audited by an independent frontier-model verifier, so the comparison applies a consistent standard to each side.

Without veritrooper — Baseline LLM Performance

What the LLM Did On Its Own

With vanilla RAG only, google_gemma-3-27b-it answered the majority of questions correctly but produced 116 confirmed errors out of 1,000. The dominant failure mode was answering questions that had no supportable answer in the source material — the model fabricated or over-reached rather than declining. For instance, when asked the specific street address of the American National Standards Institute, the model answered "25 West 43rd Street" when the correct answer was "1430 Broadway, New York, NY 10018"; and when asked the spacing of flag signals during switching operations, the model produced a confident, detailed answer when the correct response was that the question was unanswerable from the provided material. In

several cases the model also mishandled aggregate or computed questions — for example, listing per-state compliance officer counts when the question asked for a single combined total approved on a specific date.

Failure Patterns (Baseline)

The Doctor identified the following clusters in baseline failures:

- **Other** — the dominant cluster by a wide margin, largely consisting of the model supplying an answer where the correct response was that the question could not be answered from the evidence.
- **Computed vs. quoted** — the model quoting source text rather than performing the arithmetic or aggregation the question required, or vice versa.
- **Wrong focus** — the model answering an adjacent or partial point rather than the specific item asked.
- **Misinterpretation** — the model misreading the scope or intent of the regulatory text.
- **Knowledge gap** — the model lacking the specific factual detail required and substituting an incorrect value.

Training Targeting Recommendations (Baseline)

These are the fixes the customer would have to pursue WITHOUT veritrooper:

1. **Strengthen refusal behavior on unanswerable questions.** The largest share of baseline errors came from the model answering when the correct response was "unanswerable." Targeted fine-tuning and prompt rules should reward declining when the evidence does not support a definitive answer.
 2. **Separate quote-retrieval from computation.** Add training that distinguishes when a question requires aggregating, summing, or selecting a single combined value versus quoting verbatim text, to address the computed-vs-quoted failures.
 3. **Tighten question-focus alignment.** Train the model to answer the precise item asked (e.g., a specific combined total, a specific address) rather than reproducing nearby but non-responsive material.
 4. **Improve factual grounding on specific values.** Where knowledge-gap errors produced wrong concrete details (addresses, citations, numbers), reinforce verbatim grounding to the retrieved source rather than substitution from prior assumptions.
-

With veritrooper — Pipeline Performance

What the LLM Did With veritrooper

Wrapped by veritrooper, google_gemma-3-27b-it reached 98.40% adjusted accuracy with only 13 confirmed errors. Two of the residual cases were excluded as malformed tests (bad_test) — Pipeline Failures 4 and 5 — where the underlying question's ground truth was itself defective; these are excluded symmetrically from both arms and do not count against the model. Among the genuine residual errors, the recurring theme is logical reasoning over conditional regulatory language: the model occasionally misapplied "may" versus "must" or mishandled multi-condition tests. For example, when asked whether a horizontal overhead belt must be guarded for its entire length, the model concluded "No" on the grounds that the belt failed the speed threshold, missing that meeting any single one of the three conditions (here the center-to-center distance and the belt width) triggers full-length guarding. A small number of residual errors also involved the model declining to answer questions that were in fact answerable from the evidence.

Training Targeting Recommendations (Pipeline)

The residual errors are few but cluster around conditional-logic reasoning and are worth targeting:

- 1. **Reinforce multi-condition ("any one of") logic.** Train the model to recognize that when a rule lists multiple triggering conditions joined by "or," satisfying any single condition is sufficient — addressing the overhead-belt guarding error and similar cases.
- 2. **Sharpen "may" versus "must" distinctions.** The fire-prevention-plan case shows the model inverting a permissive provision; reinforce parsing of permissive ("may") versus mandatory ("must") regulatory language.
- 3. **Calibrate refusals toward answerable multi-paragraph questions.** A few residual cases declined to synthesize an answer that was supported across two provisions; train the model to combine evidence from multiple cited paragraphs when both are present.

Independent Verification

An independent frontier-model verifier (gpt) audited the Doctor's verdicts on both arms using the same standard. In the pipeline phase, 13 Doctor verdicts were overridden and none required human review; in the baseline phase, 36 were overridden and none required human review. Because the same second-tier audit was applied to both sides, neither arm received preferential treatment, and the reported delta reflects a consistently judged comparison.

Bottom Line

With veritrooper, google_gemma-3-27b-it produced 13 confirmed errors versus 116 without it — a +10.30 percentage-point gain in adjusted accuracy on regulatory content. For an enterprise where incorrect or fabricated answers on OSHA regulations carry compliance risk, the veritrooper-wrapped configuration is the recommended deployment, with the small residual error set pointing to a focused, well-bounded training effort around conditional-logic reasoning.

Complete Failure Listing — With VERITROOPER (Pipeline)

16 confirmed errors. 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
conditional	6	6	0	0
cross_reference	6	6	0	0
negative	3	0	3	0
exception	1	1	0	0
TOTAL	16	13	3	0

Failure 1 — `OSHA_29CFR_full_2024_0325_00` · HIGH confidence · Misinterpreting may as must

- **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:** Yes, the plan must be in writing because the regulation states an employer with 10 or fewer employees may communicate the plan orally, implying those with more than 10 must have it in writing. Since this employer has 8 employees, it falls within the group that may communicate orally, but is not required to.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model misinterpreted 'may' as 'must' and incorrectly inferred a requirement for a written plan based on the exception for smaller employers.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Direction mismatch: GT negative, response affirmative

Failure 2 — `OSHA_29CFR_full_2024_2204_01` · HIGH confidence · Unknown root cause

- **Category:** cross_reference
- **Question:** Using the normal total white blood count value and the adjustment noted for cigarette smokers, what could the upper white count value be for a cigarette smoker relative to the standard normal?
- **Ground truth:** The normal total white blood count is approximately 7,200/mm³ plus or minus 3,000 (so up to about 10,200/mm³). For cigarette smokers, the upper range may be 2,000 cells higher than normal for the laboratory, so the upper value could be about 12,200/mm³.
- **Model answered:** The normal total white blood count is approximately 7,200/mm³ plus or minus 3,000. For cigarette smokers, the upper range may be 2,000 cells higher than normal. Therefore, the upper white count value for a cigarette smoker could be 9,200/mm³.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Fabricated number: '9,200' not in evidence

Failure 3 — `OSHA_29CFR_full_2024_0333_00` · HIGH confidence · Unknown root cause

- **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:** Yes, tie-in guides may be eliminated for not more than 75 feet of the uppermost elevation of the building if angulated roping is employed and it is infeasible due to exterior building design.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Direction mismatch: GT negative, response affirmative

Failure 4 — `OSHA_29CFR_full_2024_neg_1599_01` · HIGH confidence · Unanswerable test

- **Category:** negative
- **Question:** What is the minimum wattage rating required for a heavy-duty lampholder of the medium type?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The evidence states a heavy-duty lampholder shall have a rating of not less than 660 watts if of the admedium type. Therefore, the minimum wattage rating required for a heavy-duty lampholder of the admedium type is 660 watts.
- **Doctor diagnosis:** bad_test
- **Doctor root cause:** evidence contains the answer; UNANSWERABLE label was wrong
- **Verifier verdict:** bad_test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 5 — `OSHA_29CFR_full_2024_neg_0484_04` · HIGH confidence · Unanswerable test

- **Category:** negative
- **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:** The quantity of flammable liquids kept near spraying operations should not exceed a supply for 1 day or one shift before bulk storage rules apply.
- **Doctor diagnosis:** bad_test
- **Doctor root cause:** evidence contains the answer; UNANSWERABLE label was wrong
- **Verifier verdict:** bad_test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 6 — `OSHA_29CFR_full_2024_4084_02` · HIGH confidence · Paraphrasing issue

- **Category:** conditional
- **Question:** When is the upper-most tie-in, braces, floor supports, and floor wedges required to be inspected for loose or dislodged components?
- **Ground truth:** Where the tower crane is supported by the structure.
- **Model answered:** The upper-most tie-in, braces, floor supports and floor wedges are required to be inspected for loose or dislodged components as stated in item (ii) of the evidence. The frequency of this inspection is not specified in the provided evidence.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** [verifier] The evidence directly says the components are inspected “where the tower crane is supported by the structure,” while the model instead said the frequency was not specified and did not provide that supported condition.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 7 — `OSHA_29CFR_full_2024_1593_04` · HIGH confidence · Unknown root cause

- **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:** The stated exception to the minimum clear working space requirements is for underground box covers that weigh over 45.4 kg (100 lb), which do not need to be locked.

- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 8 — `OSHA_29CFR_full_2024_4313_01` · HIGH confidence · Unknown root cause

- **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:** The evidence does not state what type of training paragraph (a) provides to employees generally that paragraph (b) does not explicitly require for employee representatives. Both paragraphs discuss safety and health training, but do not detail differences in the type of training provided.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** [verifier] The evidence explicitly says paragraph (a) training must 'inform employees of the agency occupational safety and health program, with emphasis on their rights and responsibilities,' while paragraph (b) lists introductory/specialized courses for representatives without explicitly requiring that rights-and-responsibilities program emphasis, so the model wrongly claimed the difference was not stated.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 9 — `OSHA_29CFR_full_2024_1332_03` · HIGH confidence · Unknown root cause

- **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:** The evidence does **not** contain enough information to answer the question. While it mentions chains in both sections, it does not specify provisions that address guarding or controlling them to prevent employees from being struck.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 10 — `OSHA_29CFR_full_2024_3172_00` · HIGH confidence · Unknown root cause

- **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:** No, according to paragraph (e)(2)(ii), the belt must be guarded for its entire length if it travels 1,800 feet per minute, has a center-to-center distance between pulleys of 10 feet or more, or is 8 inches or more in width. This belt travels 1,000 feet per minute, and is 10 inches in width, and has a center-to-center distance of 12 feet, but does not meet the speed requirement.

- **Doctor diagnosis:** model_error

- **Doctor root cause:** unknown

- **Verifier verdict:** model_error (confidence: high)

- **Validator flags:** Direction mismatch: GT affirmative, response negative

Failure 11 — `OSHA_29CFR_full_2024_neg_2602_03` · HIGH confidence · Unanswerable test

- **Category:** negative

- **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:** Mixtures A and B must have identical ingredients for the interpolation principle in A.0.5.1.4 to apply. Additionally, mixture C must have the same toxicologically active ingredients as mixtures A and B.

- **Doctor diagnosis:** bad_test

- **Doctor root cause:** evidence contains the answer; UNANSWERABLE label was wrong

- **Verifier verdict:** bad_test (confidence: high)

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

Failure 12 — `OSHA_29CFR_full_2024_0882_04` · HIGH confidence · Unknown root cause

- **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:** The second component is addressed in paragraph (c)(2) and the third component is addressed in paragraph (c)(6).

- **Doctor diagnosis:** model_error

- **Doctor root cause:** unknown

- **Verifier verdict:** model_error (confidence: high)

- **Validator flags:** (none)

Failure 13 — `OSHA_29CFR_full_2024_1434_07` · HIGH confidence · Unknown root cause

- **Category:** conditional

- **Question:** If the incident heat energy estimated under paragraph (I)(8)(ii) is exactly 2.0 cal/cm², must the outer layer of clothing be flame resistant under condition (D)?

- **Ground truth:** No; condition (D) requires flame resistance only when the estimated incident heat energy

exceeds 2.0 cal/cm2.

- **Model answered:** Yes, the employer shall ensure that the outer layer of clothing worn by an employee is flame resistant if the incident heat energy estimated under paragraph (l)(8)(ii) exceeds 2.0 cal/cm2, which includes exactly 2.0 cal/cm2.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Direction mismatch: GT negative, response affirmative

Failure 14 — `OSHA_29CFR_full_2024_2016_01` · HIGH confidence · Unknown root cause

- **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:** According to Appendix B, paragraphs (b)(1) and (c)(13) of the regulation define the scope of coverage and the meaning of 'toxic substance or harmful physical agent'.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 15 — `OSHA_29CFR_full_2024_1288_01` · HIGH confidence · Unknown root cause

- **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:** Yes, a splash guard is still required unless the rope washer is designed to prevent water or liquid from splashing the operator, the floor, or working surface.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Direction mismatch: GT negative, response affirmative

Failure 16 — `OSHA_29CFR_full_2024_4047_01` · HIGH confidence · Unknown root cause

- **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:** Yes, except over water, employees occupying the personnel platform must be provided and use a personal fall arrest system attached to a structural member within the platform; when working over or near water, the requirements of § 1926.106 apply.

- **Doctor diagnosis:** model_error
 - **Doctor root cause:** unknown
 - **Verifier verdict:** model_error (confidence: high)
 - **Validator flags:** Direction mismatch: GT negative, response affirmative
-

Complete Failure Listing — Without VERITROOPER (Baseline)

119 confirmed errors. 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
—	119	116	3	0
TOTAL	119	116	3	0

Failure 1 — `OSHA_29CFR_full_2024_neg_1259_04` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** How far apart must flag signals be spaced when protecting men during switching operations?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Warning flags or signs shall be placed in the center of the track at least 50 feet away from the cars and a derail set to protect workmen in the car.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a specific distance (50 feet) and details about derail sets, which are not present in the provided evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 2 — `OSHA_29CFR_full_2024_neg_4455_09` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the specific interest rate under 26 U.S.C. 6621 that applies to back pay calculations?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The interest rate applicable to underpayment of taxes under **26 U.S.C. 6621(a)(2)** is used to calculate interest on back pay.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model confidently stated a specific interest rate based on a citation, despite the evidence only stating the source of the rate (26 U.S.C. 6621) and not the rate itself. The evidence does not provide the actual interest rate.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 3 — `OSHA_29CFR_full_2024_2809_11` · HIGH confidence · Model error

- **Category:** —
- **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:** 25 West 43rd Street
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 4 — `OSHA_29CFR_full_2024_neg_4303_02` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** Which agency office is responsible for compiling the accident, injury, and illness data referenced in paragraph (a)?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The Office of Federal Agency Programs is responsible for compiling the accident, injury, and illness data referenced in paragraph (a). (§ 1960.70)
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a specific office and section number not mentioned in the provided evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 5 — `OSHA_29CFR_full_2024_neg_3806_04` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the BeLPT and which laboratory must perform it under this standard?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The BeLPT is the measurement of blood lymphocyte proliferation in a laboratory test when lymphocytes are challenged with a soluble beryllium salt. It must be analyzed in a laboratory certified under the College of American Pathologists/Clinical Laboratory Improvement Amendments (CLIA) guidelines.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated details about BeLPT and CLIA certification not present in the provided evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 6 — `OSHA_29CFR_full_2024_0742_02` · HIGH confidence · Model error

- **Category:** —
- **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:** This question cannot be answered from the provided evidence. The evidence discusses alarms in relation to reversing vehicles (subpart (ii) of paragraph (p)(1)), but does not mention alarm repair or personnel requirements during such repairs, nor does it differentiate between subparts (i) and (iii) in this

context.

- **Doctor diagnosis:** model_error
- **Doctor root cause:** model failed to identify the relevant information about alarm repairs in subpart (iii)
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 7 — `OSHA_29CFR_full_2024_neg_2625_07` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the reference number or citation of the footnote marked '1' regarding appropriate animal tests?
- **Ground truth:** UNANSWERABLE
- **Model answered:** A.4.2.1.3.1
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a citation number
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 8 — `OSHA_29CFR_full_2024_neg_0478_00` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the required testing frequency for inspecting filter rolls to ensure proper replacement of filter media?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Filter rolls shall be inspected to insure proper replacement of filter media. (5)(i)
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model confidently answered a question the evidence does not answer, quoting the passage verbatim but failing to recognize it doesn't specify a frequency for inspection.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 9 — `OSHA_29CFR_full_2024_neg_4485_06` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What penalty does an employer face if found to have retaliated against a seaman under SPA?
- **Ground truth:** UNANSWERABLE
- **Model answered:** If found to have retaliated against a seaman under SPA, an employer may be liable for: back pay with interest, compensation for any special damages, litigation costs, expert witness fees, and reasonable attorney fees. (§ 1986.106(c))
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated details about penalties for retaliation under SPA, which are not present in the provided evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 10 — `OSHA_29CFR_full_2024_4215_00` · HIGH confidence · Fabricated number

- **Category:** —

- **Question:** How many total compliance officers (safety and health combined) were approved as revised benchmarks, and on what date did the Assistant Secretary approve them?
- **Ground truth:** A total of 10 compliance officers (7 safety and 3 health) were approved on August 11, 1994.
- **Model answered:** Here's a breakdown of the approved compliance officer numbers and approval dates, based on the provided text:
 - * **Maryland:** 54 officers, July 18, 1985
 - * **Tennessee:** 36 officers, July 22, 1985
 - * **Kentucky:** 37 officers, June 13, 1985
 - * **Alaska:** 9 officers, September 28, 1984
 - * **Arizona:** 15 officers, June 20, 1985
 - * **New Mexico:** 10 officers, August 11, 1994
 - * **Virginia:** 59 officers, January 17, 1986
 - * **Iowa:** 29 officers, July 2, 1985
 - * **Minnesota:** 43 officers, July
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Fabricated number: '37' not in evidence

Failure 11 — `OSHA_29CFR_full_2024_3999_00` · HIGH confidence · Model error

- **Category:** —
- **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:** An operator-in-training may operate the equipment in the prohibited circumstances listed in (b)(3) only if the operator's trainer determines that the operator-in-training's skills are sufficient for high-skill work in multiple-lift rigging operations.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** [verifier] The provided evidence only states the exception as "except as provided in paragraph (b)(3)(v)," and does not include any detail about trainer determinations or multiple-lift rigging, so the model invented unsupported specifics while the ground truth matches the evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 12 — `OSHA_29CFR_full_2024_neg_2913_09` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What penalty applies if ventilation ductwork is improperly connected to a chimney used for products of combustion?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Ductwork used for ventilating any process covered in this standard shall not be connected to ducts ventilating any other process or any chimney or flue used for conveying any products of combustion.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 13 — `OSHA_29CFR_full_2024_4240_00` · HIGH confidence · Direction

reversed

- **Category:** —
- **Question:** When the context clearly requires otherwise, do the definitions in § 1955.2 still apply?
- **Ground truth:** No. The definitions apply 'unless the context clearly requires otherwise.'
- **Model answered:** Yes, the definitions in § 1955.2 apply "unless the context clearly requires otherwise."
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Direction mismatch: GT negative, response affirmative

Failure 14 — `OSHA_29CFR_full_2024_0325_00` · HIGH confidence · Direction reversed

- **Category:** —
- **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:** Yes, the fire prevention plan must be in writing. The evidence states: "A fire prevention plan must be in writing, be kept in the workplace, and be made available to employees for review. However, an employer with 10 or fewer employees may communicate the plan orally to employees." Since the employer has 8 employees, they fall within the exception allowing for oral communication, but the plan must still be developed and available, even if communicated orally.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model misinterprets the exception for employers with 10 or fewer employees, stating the plan 'must' be in writing despite the evidence allowing for oral communication.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Direction mismatch: GT negative, response affirmative

Failure 15 — `OSHA_29CFR_full_2024_0913_01` · HIGH confidence · Model error

- **Category:** —
- **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:** According to the evidence:
 - * **General employees** must be instructed in the recognition and avoidance of unsafe conditions and regulations "applicable to his work environment" (§ 1926.21 (b)(2)).
 - * **Fire brigade members** shall be provided with training at least annually and those performing interior structural fire fighting shall be provided with an education session or training at least quarterly (§ 1908.6 (c)(2)).

Therefore, general employees must be instructed at least annually, and those performing interior structural fire fighting must be provided with an education session or training at least quarterly.

- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 16 — `OSHA_29CFR_full_2024_neg_1037_02` · HIGH confidence · Hallucination

- **Category:** —
 - **Question:** How many proof tests must a sling manufacturer perform before a sling is approved for sale?
 - **Ground truth:** UNANSWERABLE
 - **Model answered:** The evidence states a sling manufacturer must perform a **proof test** to verify construction and workmanship before a sling is approved for sale. It also states that all welded end attachments must be proof tested at twice their rated capacity prior to initial use. Additionally, new or repaired metal mesh slings, including handles, shall be proof tested at a minimum of 1 1/2 times their rated capacity.
- Therefore, a sling manufacturer must perform **at least one** proof test, and potentially **more depending on the sling type and attachments**.
- **Doctor diagnosis:** model_error
 - **Doctor root cause:** unknown
 - **Verifier verdict:** model_error (confidence: high)
 - **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 17 — `OSHA_29CFR_full_2024_0362_01` · HIGH confidence · Model error

- **Category:** —
- **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:** The provided text does **not** contain information about whether wire rope is permitted for use as the emergency stop rope. Therefore, the question cannot be answered from the evidence.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 18 — `OSHA_29CFR_full_2024_neg_1037_00` · HIGH confidence · Hallucination

- **Category:** —
- **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:** An angle of loading of five degrees or less from the vertical may be considered a vertical angle of loading.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 19 — `OSHA_29CFR_full_2024_neg_0484_02` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** How often, in days or months, must pressure hoses and couplings be inspected?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Pressure hoses and couplings must be inspected at "regular intervals appropriate to this service" and tested with the hose extended, using the "inservice maximum operating pressures." Additionally,

arc welding and cutting cables must be inspected monthly.

- **Doctor diagnosis:** 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_2625_05` · HIGH confidence · Hallucination

- **Category:** —
- **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:** Epidemiological studies are given greater weight than case reports in the weight-of-evidence approach. Specifically, "reliable, good quality human data shall generally have precedence over other data," and epidemiological studies are mentioned as a preferred source of evidence for Category 1A classification. Case reports are mentioned as information to be considered along with epidemiological studies.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a weighting scheme not present in the provided evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 21 — `OSHA_29CFR_full_2024_neg_1769_04` · HIGH confidence · Hallucination

- **Category:** —
- **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:** 10. Employers with 10 or fewer employees on the effective date of this section are not required to comply with the COVID-19 log requirements.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a rule about employer size and COVID-19 log requirements that is not present in the provided evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 22 — `OSHA_29CFR_full_2024_neg_4403_07` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** Who within the Department of Labor is designated to conduct the investigations referenced in this part?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Compliance Safety and Health Officer.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a specific role (Compliance Safety and Health Officer) not mentioned in the provided evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 23 — `OSHA_29CFR_full_2024_2204_01` · HIGH confidence · Fabricated number

- **Category:** —
- **Question:** Using the normal total white blood count value and the adjustment noted for cigarette smokers, what could the upper white count value be for a cigarette smoker relative to the standard normal?
- **Ground truth:** The normal total white blood count is approximately 7,200/mm³ plus or minus 3,000 (so up to about 10,200/mm³). For cigarette smokers, the upper range may be 2,000 cells higher than normal for the laboratory, so the upper value could be about 12,200/mm³.
- **Model answered:** The normal total white blood count is approximately 7,200/mm³ plus or minus 3,000. For cigarette smokers, the upper range may be 2,000 cells higher than normal. Therefore, the upper white count value for a cigarette smoker could be 9,200/mm³ (7,200 + 2,000).
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model computed the upper limit incorrectly by adding 2000 to the center of the normal range, rather than the upper bound of the normal range (7200 + 3000).
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Fabricated number: '9,200' not in evidence

Failure 24 — `OSHA_29CFR_full_2024_0333_00` · HIGH confidence · Direction reversed

- **Category:** —
- **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:** Yes, tie-in guides may be eliminated for not more than 75 feet of the uppermost elevation of the building if angulated roping is employed and it is infeasible due to exterior building design.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Direction mismatch: GT negative, response affirmative

Failure 25 — `OSHA_29CFR_full_2024_neg_1738_03` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** How frequently must common areas like break rooms be cleaned under this section?
- **Ground truth:** UNANSWERABLE
- **Model answered:** High-touch surfaces and equipment in all areas, including common areas, must be cleaned **at least once a day**. Additionally, areas must be cleaned and disinfected when a COVID-19 positive person has been in the workplace within the last 24 hours.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 26 — `OSHA_29CFR_full_2024_neg_3653_08` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What exposure level triggers respirator use during the Class IV asbestos work described in subparagraph (vi)?

- **Ground truth:** UNANSWERABLE
- **Model answered:** Respirators must be used during Class IV asbestos work "within regulated areas where employees who are performing other work are required to use respirators."
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 27 — `OSHA_29CFR_full_2024_2815_06` · HIGH confidence · Model error

- **Category:** —
- **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:** The OMB control number for 1926.1203 through 1926.1213 is 1218-0258, while the OMB control number for 1926.1402 through 1926.1441 is 1218-0261.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** [verifier] The evidence supports 1926.1203-1926.1213 as all 1218-0258, but within 1926.1402-1926.1441 it lists an exception—"1926.1427; OMB control No.: 1218-0270"—so the model's blanket statement that the 1926.1402-1926.1441 series is 1218-0261 is incomplete/incorrect.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 28 — `OSHA_29CFR_full_2024_neg_1599_01` · MEDIUM confidence · Bad question

- **Category:** —
- **Question:** What is the minimum wattage rating required for a heavy-duty lampholder of the medium type?
- **Ground truth:** UNANSWERABLE
- **Model answered:** 660 watts.
- **Doctor diagnosis:** bad_test
- **Doctor root cause:** [verifier] The evidence gives a wattage rule for heavy-duty lampholders—"not less than 660 watts if of the admedium type and not less than 750 watts if of any other type"—so the ground truth of UNANSWERABLE is not supported; the doctor's branch-circuit-condition rationale confuses when heavy-duty lampholders are required with the rating required once a lampholder is heavy-duty.
- **Verifier verdict:** bad_test (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 29 — `OSHA_29CFR_full_2024_2784_01` · HIGH confidence · Model error

- **Category:** —
- **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:** Exceeding the permissible exposure limit (PEL) triggers the application of paragraphs (d) and (g)(1)(ii).
- **Doctor diagnosis:** model_error
- **Doctor root cause:** [verifier] The evidence states paragraphs apply only where, absent an action level, the PEL is "routinely exceeded" for a substance with exposure monitoring and medical surveillance requirements; the model omitted the routine-exceedance qualifier and related condition, so its answer is materially incomplete.

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

Failure 30 — `OSHA_29CFR_full_2024_3327_04` · HIGH confidence · Model error

- **Category:** —
- **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:** November 25, 1997.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** [verifier] The evidence says the stay lasted “until Nov. 25, 1997, or until further rulemaking has been completed, whichever is later,” so answering only “November 25, 1997” omits the material later-of condition and is not equivalent to the ground truth.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 31 — `OSHA_29CFR_full_2024_4084_02` · HIGH confidence · Wrong focus

- **Category:** —
- **Question:** When is the upper-most tie-in, braces, floor supports, and floor wedges required to be inspected for loose or dislodged components?
- **Ground truth:** Where the tower crane is supported by the structure.
- **Model answered:** The upper-most tie-in, braces, floor supports and floor wedges are required to be inspected during **monthly** inspections (section 4(ii)).
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model focused on the inspection frequency (monthly) instead of the location where the inspection is required.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 32 — `OSHA_29CFR_full_2024_neg_1194_07` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What fire-resistance rating must the shields used near combustible walls meet?
- **Ground truth:** UNANSWERABLE
- **Model answered:** According to the evidence, fire-resistant shields or guards used near combustible walls must have a fire-resistance rating to prevent ignition. (x) Noncombustible walls. If welding is to be done on a metal wall, partition, ceiling or roof, precautions shall be taken to prevent ignition of combustibles on the other side, due to conduction or radiation, preferably by relocating combustibles. Where combustibles are not relocated, a fire watch on the opposite side from the work shall be provided. (vii) Relocation of combustibles. Where practicable, all combustibles shall be relocated at least 35 feet (10.7 m) from the work site. Where relocation is impracticable, combustibles shall be protected with flameproofed covers or otherwise shielded with metal or asbestos guards or curtains.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a requirement for a specific fire-resistance rating, which is not stated in the provided evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 33 — `OSHA_29CFR_full_2024_neg_1769_01` · HIGH confidence ·

Hallucination

- **Category:** —
- **Question:** Which OSHA office must employers submit their COVID-19 log to for review?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The Assistant Secretary.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated an answer not found in the provided text
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 34 — `OSHA_29CFR_full_2024_neg_4303_07` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the cost to the committee for obtaining the NIOSH hazard evaluation reports listed in paragraph (a)?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The evidence states that the Director of NIOSH may assist agencies "...upon request...with reimbursement." Therefore, there is a cost to the committee for obtaining NIOSH hazard evaluation reports.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model inferred a cost from the phrase 'with reimbursement' when the evidence does not state the amount or conditions of that reimbursement.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 35 — `OSHA_29CFR_full_2024_0455_02` · MEDIUM confidence · Model error

- **Category:** —
- **Question:** If a pit has any part within a Division 1 or 2 classified area and is NOT provided with mechanical ventilation, how must all areas within the pit be classified?
- **Ground truth:** All areas within the pit shall be classified Division 1.
- **Model answered:** Class I Group D Division 1.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: medium)
- **Validator flags:** (none)

Failure 36 — `OSHA_29CFR_full_2024_neg_4403_06` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the deadline for filing objections to the findings and preliminary orders?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The deadline for filing objections to the findings and preliminary orders is **30 days** from receipt of the findings and preliminary order, unless an objection and request for a hearing have been timely filed. (See § 1986.106(a) and multiple other sections).
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a deadline not found in the provided evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 37 — `OSHA_29CFR_full_2024_neg_4455_05` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the mailing address of the Chief Administrative Law Judge that must appear in the findings?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The findings will give the address of the Chief Administrative Law Judge, U.S. Department of Labor, or appropriate information regarding filing objections electronically with the Office of Administrative Law Judges.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 38 — `OSHA_29CFR_full_2024_3130_08` · HIGH confidence · Model error

- **Category:** —
- **Question:** What is the maximum capacity (in gallons) of a safety can?
- **Ground truth:** not more than 5 gallons
- **Model answered:** The evidence does not provide information about the maximum capacity of a "safety can". It discusses vaporizers and storage containers, but does not mention safety cans. Therefore, the question cannot be answered from the provided text.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** [verifier] The evidence explicitly defines a safety can as an approved closed container 'of not more than 5 gallons capacity,' so the model's claim that the text does not mention safety cans is contradicted by the evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 39 — `OSHA_29CFR_full_2024_4047_02` · HIGH confidence · Direction reversed

- **Category:** —
- **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:** No, lifts may not be made on any other of the equipment's load lines while personnel are being hoisted, except in pile driving operations (k)(11)(i).
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Direction mismatch: GT affirmative, response negative

Failure 40 — `OSHA_29CFR_full_2024_3972_00` · HIGH confidence · Model error

- **Category:** —
- **Question:** Under the monthly inspection rule, what is the minimum retention period for the inspection documentation?
- **Ground truth:** three months
- **Model answered:** The documentation for the monthly inspection must be retained according to § 1926.1412(e)(3). The evidence does not state the retention period specified in § 1926.1412(e)(3).

- **Doctor diagnosis:** model_error
- **Doctor root cause:** [verifier] The evidence explicitly states, "This document must be retained for a minimum of three months," so the ground truth is supported and the model was wrong to say the evidence does not state the retention period.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 41 — `OSHA_29CFR_full_2024_1227_01` · HIGH confidence · Model error

- **Category:** —
- **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:** Before making hot repairs to one generator in a room with other generators, all generators and piping must be purged of acetylene.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** [verifier] The question asks to combine (v) and (vi), but the model only gave the paragraph (vi) purging requirement and omitted paragraph (v)'s required steps such as completely removing the carbide charge/feed mechanism, flooding the generator shell with water to expel acetylene, and disconnecting it from the piping system.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 42 — `OSHA_29CFR_full_2024_3948_04` · HIGH confidence · Direction reversed

- **Category:** —
- **Question:** If a knuckle-boom truck crane transfers a precast concrete panel onto a structure using a fork/cradle while equipped with an overload prevention device, does the (c)(17) material-delivery exclusion apply?
- **Ground truth:** No; the exclusion does not apply when the material handled is a prefabricated component such as a precast concrete panel.
- **Model answered:** Yes, the (c)(17)(ii) exclusion applies. The evidence states the exclusion applies when an articulating/knuckle-boom truck crane transfers building supply sheet goods or packaged materials (including precast concrete panels) from the truck crane onto a structure using a fork/cradle, provided the truck crane is equipped with a properly functioning automatic overload prevention device.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model misapplied the exclusion criteria. It focused on the presence of the overload prevention device and the use of a fork/cradle, but failed to recognize the explicit exclusion for 'prefabricated components' like precast concrete panels as stated in (c)(17)(iii)(B).
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Direction mismatch: GT negative, response affirmative

Failure 43 — `OSHA_29CFR_full_2024_neg_2625_00` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What specific sensitization frequency percentage distinguishes sub-category 1A from sub-category 1B?
- **Ground truth:** UNANSWERABLE

- **Model answered:** For **skin sensitization**, a response of **at least 30%** of animals distinguishes sub-category 1A from sub-category 1B. For **respiratory sensitization**, sub-category 1A shows a **high frequency of occurrence in humans**, while sub-category 1B shows a **low to moderate frequency of occurrence in humans**.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a 30% threshold for skin sensitization, which is not present in the provided evidence. It also correctly described the difference for respiratory sensitization, but the question asked about both.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 44 — `OSHA_29CFR_full_2024_3468_02` · HIGH confidence · Model error

- **Category:** —
- **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:** The record must have been maintained for three years after the completion of training and indicate the name and job classification of each employee trained, the date the training was completed, and the topic of the training (§ 1910.1024(o)(4)).
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model pulled information about training records instead of air quality test records
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 45 — `OSHA_29CFR_full_2024_3047_01` · HIGH confidence · Model error

- **Category:** —
- **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:** Paragraph 1910.120(g).
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model misidentified the paragraph number
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 46 — `OSHA_29CFR_full_2024_0022_03` · HIGH confidence · Model error

- **Category:** —
- **Question:** If an interested person wishes to submit written data, views, and arguments on the proposal, within what period after publication of the notice in the Federal Register must they do so?
- **Ground truth:** Within 30 days after publication of the notice in the Federal Register.
- **Model answered:** 35 days. (See § 1902.49(c) and § 1902.11(d))
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 47 — `OSHA_29CFR_full_2024_0075_04` · HIGH confidence · Model error

- **Category:** —
- **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:** This question cannot be answered from the provided text. The text does not state what happens if an employer fails to comply with the posting requirements of paragraphs (a) and (b).
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 48 — `OSHA_29CFR_full_2024_neg_2913_01` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** How frequently must the inspection or clean-out doors themselves be inspected for proper functioning?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The evidence states inspection and tests shall be made of each platform installation every 30 days, or where the work cycle is less than 30 days such inspection and/or test shall be made prior to each work cycle.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 49 — `OSHA_29CFR_full_2024_neg_3806_07` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What qualifications must the licensed physician hold to conduct the examination under this standard?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The licensed physician must ensure all medical examinations and procedures are performed by or under their supervision, and the physician must be licensed. Additionally, anyone administering pulmonary function testing must complete a training course in spirometry sponsored by an appropriate academic or professional institution.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated requirements for the physician's qualifications, as the evidence only discusses the report the physician must provide.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 50 — `OSHA_29CFR_full_2024_neg_4455_08` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What standard does the ALJ use to determine whether a complaint was frivolous or brought in bad faith?
- **Ground truth:** UNANSWERABLE
- **Model answered:** If, upon the request of the respondent, the ALJ determines that a complaint was frivolous

or was brought in bad faith, the ALJ may award to the respondent a reasonable attorney fee, not exceeding \$1,000.

- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a standard for determining frivolousness; the evidence only states the respondent can request an award of attorney fees if they allege frivolousness, but does not define the standard the ALJ uses.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 51 — `OSHA_29CFR_full_2024_neg_1037_07` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What certification must an equivalent entity hold to be recognized for performing repairs?
- **Ground truth:** UNANSWERABLE
- **Model answered:** An equivalent entity must have written certification from the manufacturer, or a nationally recognized testing laboratory, that the modification results in the equipment being safe for its intended use.
(g)
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 52 — `OSHA_29CFR_full_2024_neg_4485_04` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** How long does OSHA have to complete its investigation after a SPA complaint is filed?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The provided evidence does **not** state how long OSHA has to complete its investigation after a SPA complaint is filed. It only details the initial steps of the investigation, such as notifying parties and allowing for initial responses within 20 days of the complaint being filed.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 53 — `OSHA_29CFR_full_2024_0772_02` · HIGH confidence · Fabricated number

- **Category:** —
- **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:** 90.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Fabricated number: '90' not in evidence

Failure 54 — `OSHA_29CFR_full_2024_0362_00` · HIGH confidence · Model error

- **Category:** —
- **Question:** Regarding controller contacts carrying the main motor current, under what condition do they NOT need to be copper to carbon or equal?
- **Ground truth:** Where the circuit is broken at two or more points simultaneously (and the requirement also does not apply to the oil-immersed type).
- **Model answered:** Controller contacts carrying the main motor current do NOT need to be copper to carbon or equal if they are of the oil-immersed type.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 55 — `OSHA_29CFR_full_2024_neg_1037_05` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What load multiple is applied as the proof load relative to the rated capacity?
- **Ground truth:** UNANSWERABLE
- **Model answered:** A proof load of 1 1/2 times their rated capacity is required for new and repaired metal mesh slings, including handles. A proof load of 125 percent of the platform's rated capacity is required prior to hoisting employees on a personnel platform.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 56 — `OSHA_29CFR_full_2024_neg_2902_06` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** Who is responsible for supplying the heavy canvas or leather gloves to the operators?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The employer is responsible for equipping operators with heavy canvas or leather gloves.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a responsibility for the employer; the evidence only states operators shall be equipped with gloves, not who supplies them.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 57 — `OSHA_29CFR_full_2024_neg_0484_05` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the minimum steel pipe wall thickness required for piping systems conveying these flammable liquids?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The minimum steel pipe wall thickness required is 0.049 inch.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)

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

Failure 58 — `OSHA_29CFR_full_2024_neg_4485_00` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the fee for filing a retaliation complaint under SPA?
- **Ground truth:** UNANSWERABLE
- **Model answered:** There is no fee for filing a retaliation complaint under SPA. The evidence states, "No particular form of complaint is required. A complaint may be filed orally or in writing." (1986.103(b))
- **Doctor diagnosis:** 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_0913_03` · HIGH confidence · Model error

- **Category:** —
- **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:** The program required in 29 CFR 1910.120(g)(1) focuses on **regulatory knowledge** including a review of 29 CFR 1910.120, the core elements of an occupational safety and health program, medical surveillance, site safety and health plans, emergency response plans, and adequate illumination/sanitation.

The training required in 29 CFR 1910.120(g)(3) focuses on **truck-related and workplace-related topics** such as operating instructions, truck controls, vehicle stability, workplace surface conditions, and more.

- **Doctor diagnosis:** model_error
- **Doctor root cause:** model described program content broadly (regulatory knowledge, truck-related topics) instead of the specific focus on fire extinguisher use and firefighting hazards as stated in the evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 60 — `OSHA_29CFR_full_2024_neg_2602_06` · HIGH confidence · Hallucination

- **Category:** —
- **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:** The citation reference for the additivity formula (A.1.3.6) is **ER26MR12.061**.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a citation reference not found in the provided evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 61 — `OSHA_29CFR_full_2024_neg_3806_00` · HIGH confidence · Hallucination

- **Category:** —
- **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:** The employer shall provide the information to the physician when there is exposure to _____, and specifically when employee exposures are at or above the action level. The permissible exposure limit (PEL) is 5 micrograms per cubic meter of air (5 µgm/m³).

- **Doctor diagnosis:** model_error

- **Doctor root cause:** model hallucinated a PEL value not present in the provided evidence.

- **Verifier verdict:** model_error (confidence: high)

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

Failure 62 — `OSHA_29CFR_full_2024_0742_01` · HIGH confidence · Model error

- **Category:** —

- **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:** 1. Repairs shall be made by **authorized personnel**.

2. Repairs shall be conducted in locations **designated for such repairs**.

- **Doctor diagnosis:** 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 invented 'designated locations' and omitted the parts/specifications requirements.

- **Verifier verdict:** model_error (confidence: high)

- **Validator flags:** (none)

Failure 63 — `OSHA_29CFR_full_2024_3047_00` · HIGH confidence · Model error

- **Category:** —

- **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:** According to paragraph (ii), when the clean-up is done on plant property using plant or workplace employees, those employees shall have completed the training requirements of 29 CFR 1910.38, 1910.134, 1910.1200, and other appropriate safety and health training.

- **Doctor diagnosis:** model_error

- **Doctor root cause:** unknown

- **Verifier verdict:** model_error (confidence: high)

- **Validator flags:** (none)

Failure 64 — `OSHA_29CFR_full_2024_0304_00` · HIGH confidence · Model error

- **Category:** —

- **Question:** What is the consequence of supporting the line improperly such that pulling on one section takes up slack in adjacent sections?

- **Ground truth:** It causes the line to fall below the limits specified in paragraph (d)(2)(ii) of this section (i.e., below the required height of not less than 34 inches and not more than 39 inches above the walking-working surface).
- **Model answered:** The rope, wire, or chain shall be attached at each stanchion in such a way that pulling on one section of the line between stanchions will not result in slack being taken up in adjacent sections before the stanchion tips over.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** [verifier] The evidence says improper support could cause slack in adjacent sections, "causing the line to fall below the limits specified" in paragraph (d)(2)(ii), while the model instead gave a different requirement involving stanchions tipping over and did not state the height-limit consequence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 65 — `OSHA_29CFR_full_2024_neg_1769_09` · HIGH confidence · Hallucination

- **Category:** —
- **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:** Within 24 hours. (See paragraph (q)(2)(ii) and (q)(3) of the provided evidence.)
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a requirement about notifying employees, conflating logging the positive test with a notification requirement not present in the evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 66 — `OSHA_29CFR_full_2024_neg_3512_05` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the minimum height required for the standard railings installed along the walkways?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The evidence states that guardrails or railings may be of any height **up to, but not more than, 45 inches**.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a height limit for railings not mentioned in the evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 67 — `OSHA_29CFR_full_2024_neg_2602_03` · HIGH confidence · Bad question

- **Category:** —
- **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:** Mixtures A and B must have "identical ingredients" for the interpolation principle in A.0.5.1.4 to apply.
- **Doctor diagnosis:** bad_test
- **Doctor root cause:** [verifier] The evidence directly states A.0.5.1.4 applies to three mixtures "with identical ingredients" and C having "the same toxicologically active ingredients as mixtures A and B," so A and B must

share all such ingredients; the ground truth UNANSWERABLE is not supported.

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

Failure 68 — `OSHA_29CFR_full_2024_0882_04` · HIGH confidence · Model error

- **Category:** —
- **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:** This question cannot be answered from the provided evidence. The evidence mentions an energy control program in paragraph (c)(1) but does not detail its components or which subsequent paragraphs address them.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 69 — `OSHA_29CFR_full_2024_neg_4403_05` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** How many days does an administrative law judge have to issue a decision after the post-hearing review?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The evidence states that after a post-hearing review, the administrative law judge shall issue a decision within **10 business days** after the date of the decision, unless a motion for reconsideration has been filed. (See § 1981.111(c) and § 1979.111(c)).
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a timeframe for ALJ decisions and cited sections not present in the provided evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 70 — `OSHA_29CFR_full_2024_1434_07` · HIGH confidence · Direction reversed

- **Category:** —
- **Question:** If the incident heat energy estimated under paragraph (l)(8)(ii) is exactly 2.0 cal/cm², must the outer layer of clothing be flame resistant under condition (D)?
- **Ground truth:** No; condition (D) requires flame resistance only when the estimated incident heat energy exceeds 2.0 cal/cm².
- **Model answered:** Yes, the outer layer of clothing must be flame resistant under condition (D) if the incident heat energy estimated under paragraph (l)(8)(ii) exceeds 2.0 cal/cm². The evidence states: "(D) The incident heat energy estimated under paragraph (l)(8)(ii) of this section exceeds 2.0 cal/cm²."
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Direction mismatch: GT negative, response affirmative

Failure 71 — `OSHA_29CFR_full_2024_neg_2129_04` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** Who pays for the powered air-purifying respirator when an employee chooses one over a negative-pressure respirator?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The employer must provide the powered air-purifying respirator (PAPR) when an employee chooses to use a PAPR instead of a negative-pressure respirator, provided it provides adequate protection to the employee.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 72 — `OSHA_29CFR_full_2024_3801_01` · HIGH confidence · Model error

- **Category:** —
- **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:** Based on the provided text:

The requirement to provide PPE is triggered when an employee works with liquids that may burn, irritate, or otherwise harm their skin (paragraph (h)(1)).

The document governing both its use and removal is not explicitly stated in the provided text. However, the text states a "written PPE program" should include policy statements, procedures, and guidelines and copies should be made available to all employees (paragraph 5).

- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 73 — `OSHA_29CFR_full_2024_neg_1599_04` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the maximum load in amperes for a receptacle on a 50-ampere branch circuit?
- **Ground truth:** UNANSWERABLE
- **Model answered:** 50 amperes.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 74 — `OSHA_29CFR_full_2024_0003_03` · HIGH confidence · Model error

- **Category:** —
- **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:** 42 U.S.C. 4201-4233.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 75 — `OSHA_29CFR_full_2024_neg_2913_00` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What minimum air velocity (in feet per minute) must be maintained through the spray booth ventilation system?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The minimum air velocity must be 100 linear feet per minute for most spraying operations, or 60 linear feet per minute for electrostatic spraying operations.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 76 — `OSHA_29CFR_full_2024_0767_00` · HIGH confidence · Fabricated number

- **Category:** —
- **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:** Based on the evidence, the total time spent performing exercises would be 135 seconds. Here's the breakdown:
 - * **Facing Forward:** 30 seconds (Exercise 1)
 - * **Bending Over:** 30 seconds (Exercise 1)
 - * **Head Shaking:** About 3 seconds (Exercise 1)
 - * **Grimace:** 15 seconds (6)
 - * **REDON 1:** Not specified, but assumed to be similar to other REDON exercises at 10 seconds (Exercise 1)
 - * **REDON 2:** Not specified, but assumed to be similar to other REDON exercises at 10 seconds (Exercise 1)

Total: 30 + 30 + 3 + 15 + 10 + 10 = 98 seconds.

However, the evidence also states that the bending over exercise is 30

- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Fabricated number: '135' not in evidence

Failure 77 — `OSHA_29CFR_full_2024_3092_01` · HIGH confidence · Wrong focus

- **Category:** —
- **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:** Based on the evidence provided:

* **Container requirement:** Paragraph (a)(1) states flammable liquids shall be stored in tanks or closed containers.

* **Storage restrictions:** Paragraph (a)(2)(b)(1) allows up to 25 gallons of Category 1 flammable liquids in containers.

Therefore, 4 gallons of flammable liquid must be stored in a **closed container**, and may not be stored outside of an **inside storage room or storage cabinet in a building or in any one fire area of a building** if it is a Category 1 flammable liquid.

- **Doctor diagnosis:** model_error

- **Doctor root cause:** [verifier] The evidence supports the ground truth: for 4 gallons, paragraph (a)(1) requires "Approved safety cans or Department of Transportation approved containers," and paragraph (a)(2) bars storage in "areas used for exits, stairways, or normally used for the safe passage of people," while the model instead gave a generic closed-container answer and invented unrelated storage-room/fire-area restrictions.

- **Verifier verdict:** model_error (confidence: high)

- **Validator flags:** (none)

Failure 78 — `OSHA_29CFR_full_2024_1964_02` · HIGH confidence · Model error

- **Category:** —

- **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:** This question cannot be answered from the provided evidence. The evidence references a "(vi) Warning sign" in a section about welding, but does not state what the sign must say, nor does it mention any restrictions on additional statements. Paragraph (l)(2)(ii) and (l)(6) are not present in the provided text.

- **Doctor diagnosis:** model_error

- **Doctor root cause:** unknown

- **Verifier verdict:** model_error (confidence: high)

- **Validator flags:** (none)

Failure 79 — `OSHA_29CFR_full_2024_neg_4403_00` · HIGH confidence · Hallucination

- **Category:** —

- **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:** Within 180 days. (See § 1981.103(d))

- **Doctor diagnosis:** model_error

- **Doctor root cause:** model hallucinated a specific timeframe and citation not present in the provided evidence.

- **Verifier verdict:** model_error (confidence: high)

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

Failure 80 — `OSHA_29CFR_full_2024_3017_00` · HIGH confidence · Model error

- **Category:** —
- **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:** Delivery people, messengers, and other employees who only enter a workplace briefly to drop off or pick up items are not required to be monitored.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model read from a different section of the document than the ground truth
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 81 — `OSHA_29CFR_full_2024_neg_3653_04` · HIGH confidence · Hallucination

- **Category:** —
- **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:** 1,000.

The evidence states: "Powered Air-Purifying Respirator (PAPR); Full facepiece: 1,000".

- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 82 — `OSHA_29CFR_full_2024_neg_1738_04` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What temperature threshold qualifies as a 'fever' among the listed COVID-19 symptoms?
- **Ground truth:** UNANSWERABLE
- **Model answered:** ≥ 100.4 °F
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 83 — `OSHA_29CFR_full_2024_neg_4485_02` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** Within how many days after filing must OSHA notify the named respondent of the complaint?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Within 20 days of receipt of the notice of the filing of the complaint, OSHA will notify the respondent.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 84 — `OSHA_29CFR_full_2024_neg_4403_02` · HIGH confidence ·

Hallucination

- **Category:** —
- **Question:** Which specific federal office or agency receives the complaints filed under the Pipeline Safety Act?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The OSHA Area Director responsible for enforcement activities in the geographical area where the employee resides or was employed, or any OSHA officer or employee, receives complaints filed under the Pipeline Safety Act. (§ 1981.103(c))
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a specific agency receiving complaints; the evidence only states procedures for handling complaints, not where they are filed.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 85 — `OSHA_29CFR_full_2024_1264_03` · HIGH confidence · Wrong calc/quote

- **Category:** —
- **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:** Both provisions specifically restrict operation when a component is not in place: paragraph (d)(1) restricts operation when the hopper is removed, and paragraph (d)(3) requires the guard to be in place for the feed rolls to operate.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model focused on specific components (hopper, cover) instead of the general principle of preventing access to hazardous parts during operation.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 86 — `OSHA_29CFR_full_2024_neg_3653_05` · HIGH confidence · Hallucination

- **Category:** —
- **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:** A serious violation, with a penalty not exceeding \$16,131.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 87 — `OSHA_29CFR_full_2024_4010_02` · HIGH confidence · Model error

- **Category:** —
- **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:** If the auditor determines there is a significant deficiency in the program, the following is

triggered:

1. No operator is qualified until the auditor confirms the deficiency has been corrected.
2. The program is audited again within 180 days of the correction confirmation.
3. The auditor files a documented report of the deficiency to the appropriate Regional Office of the Occupational Safety and Health Administration within 15 days.

- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 88 — `OSHA_29CFR_full_2024_neg_1037_11` · MEDIUM confidence · Hallucination

- **Category:** —
- **Question:** What is the maximum slot width permitted for a female handle on a metal mesh choker sling?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The evidence states that a female handle (choker) has a slot dimension that permits passage of a male handle, and distortion of the female handle is a reason for removal from service if the depth of the slot is increased more than 10 percent. However, the evidence does **not** state the maximum slot width permitted for a female handle.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 89 — `OSHA_29CFR_full_2024_1202_00` · HIGH confidence · Model error

- **Category:** —
- **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:** §§ 1910.253 and 1910.254.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model listed paragraphs that are required in BOTH (d)(1)(i) and (d)(1)(ii) instead of identifying the additional requirement for (d)(1)(ii).
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 90 — `OSHA_29CFR_full_2024_neg_3653_03` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** How often must the medical examination referenced in paragraph (h)(2)(ii) be repeated for employees using respirators?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The evidence states that a baseline chest x-ray should be repeated at five-year intervals and a pulmonary function test should be repeated every three years.
- **Doctor diagnosis:** 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_neg_3806_02` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What form must the employee use to provide written consent for releasing prior medical examination records?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The evidence states that appendix A to the section contains a sample form which may be used to establish specific written consent for access to employee medical records.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a form mentioned in an appendix that is not described in the provided evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 92 — `OSHA_29CFR_full_2024_2565_02` · HIGH confidence · Model error

- **Category:** —
- **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:** The Federal Occupational Safety and Health Administration (OSHA) and the State agency responsible under a Plan approved under section 18 of the Act administer the regulations referenced in both exclusions. The two statutes involved are the Williams-Steiger Occupational Safety and Health Act of 1970 and the Contract Work Hours and Safety Standards Act.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 93 — `OSHA_29CFR_full_2024_2508_02` · HIGH confidence · Direction reversed

- **Category:** —
- **Question:** If an area's airborne respirable crystalline silica concentration can reasonably be expected to exceed the PEL but the employer has not demarcated it, does it qualify as a regulated area?
- **Ground truth:** No, a regulated area must be demarcated by the employer.
- **Model answered:** Yes, according to the provided evidence, an area qualifies as a regulated area "wherever an employee's exposure to airborne concentrations of respirable crystalline silica is, or can reasonably be expected to be, in excess of the PEL." Demarcation is a requirement of a regulated area, not a prerequisite for it to be considered one.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Direction mismatch: GT negative, response affirmative

Failure 94 — `OSHA_29CFR_full_2024_4294_01` · HIGH confidence · Model error

- **Category:** —

- **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:** According to paragraph (c)(9)(i), employee names (column B) on the OSHA Form 300 (Log of Work-Related Injuries and Illnesses) are not required in submissions. Additionally, paragraph (d)(2) states that employee names are not required on the OSHA Form 301 (Injury and Illness Incident Report).

Therefore, the establishment's log may not include the employee's name.

- **Doctor diagnosis:** model_error

- **Doctor root cause:** unknown

- **Verifier verdict:** model_error (confidence: high)

- **Validator flags:** (none)

Failure 95 — `OSHA_29CFR_full_2024_neg_2129_01` · HIGH confidence · Hallucination

- **Category:** —

- **Question:** What is the PEL (permissible exposure limit) for cadmium mentioned in subsection (vi)?

- **Ground truth:** UNANSWERABLE

- **Model answered:** The PEL for cadmium is five micrograms per cubic meter of air (5 µg/m³), calculated as an eight-hour time-weighted average exposure (TWA).

- **Doctor diagnosis:** model_error

- **Doctor root cause:** unknown

- **Verifier verdict:** model_error (confidence: high)

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

Failure 96 — `OSHA_29CFR_full_2024_neg_2129_05` · HIGH confidence · Hallucination

- **Category:** —

- **Question:** How often must fit testing be repeated under this section's respirator program?

- **Ground truth:** UNANSWERABLE

- **Model answered:** Fit testing must be repeated prior to initial use of the respirator, whenever a different respirator facepiece (size, style, model or make) is used, and at least annually thereafter. Additionally, fit testing must be conducted whenever an employee reports or the employer/PLHCP/supervisor observes changes in the employee's physical condition that could affect respirator fit.

- **Doctor diagnosis:** model_error

- **Doctor root cause:** unknown

- **Verifier verdict:** model_error (confidence: high)

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

Failure 97 — `OSHA_29CFR_full_2024_1317_01` · HIGH confidence · Model error

- **Category:** —

- **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:** The provided text does not contain a definition of "running line" nor does it distinguish it from other ropes. Therefore, the question cannot be answered from the evidence provided.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** [verifier] The evidence explicitly states, "Running line... means any moving rope as distinguished from a stationary rope such as a guyline," while the model incorrectly claimed the definition was not provided.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 98 — `OSHA_29CFR_full_2024_neg_2913_02` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What gauge of sheet steel is required for the exhaust ductwork construction?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Exhaust ducts shall be constructed of steel and longitudinal joints in sheet steel ductwork shall be either lock-seamed, riveted, or welded.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 99 — `OSHA_29CFR_full_2024_1202_02` · HIGH confidence · Misinterpreted

- **Category:** —
- **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:** According to paragraphs (d)(1)(iii) and (d)(1)(iv), the two distinct hazards requiring protective measures are:
 - * **Fire:** The combination of substances being removed by ventilation must not cause a fire.
 - * **Explosion:** The combination of substances being removed by ventilation must not cause an explosion.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model misinterpreted the hazards described in the evidence; it focused on general ventilation hazards (fire/explosion) instead of the specific hazards related to arc welding and pressure testing.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 100 — `OSHA_29CFR_full_2024_neg_4485_07` · HIGH confidence · Hallucination

- **Category:** —
- **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:** Within 180 days of an alleged violation, a complaint is considered timely if filed via delivery to a third-party commercial carrier.
- **Doctor diagnosis:** model_error

- **Doctor root cause:** model confidently answered a question about a specific delivery method (third-party commercial carrier) when the evidence only states that delivery to a carrier counts as the date of filing, not how many days they have to deliver it.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 101 — `OSHA_29CFR_full_2024_neg_2129_02` · HIGH confidence · Hallucination

- **Category:** —
- **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:** As soon as possible. (l)(6)(iii) states "Whenever an employee has exhibited difficulty in breathing...the employer, as soon as possible, shall provide the employee with a periodic medical examination..."
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 102 — `OSHA_29CFR_full_2024_neg_1194_06` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What penalty applies to a worker who performs cutting in a sprinklered building while protection is impaired?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Cutting or welding is not permitted in sprinklered buildings while such protection is impaired. (vi)(B)
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 103 — `OSHA_29CFR_full_2024_3608_02` · HIGH confidence · Model error

- **Category:** —
- **Question:** Based on subparts (ii) and (iii), what must the device allow the employee to do regarding hand use, and to what value must it limit the descending velocity of an employee?
- **Ground truth:** It must permit the employee to ascend or descend without continually having to hold, push or pull any part of the device (leaving both hands free for climbing), and it must limit the descending velocity to 7 feet/sec. (2.1 m/sec.) or less.
- **Model answered:** This question cannot be answered from the provided text. The text discusses respirators and user seal checks, but does not contain information about a "device" or limitations on descending velocity.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model failed to identify the 'device' referenced in the question as the ladder safety device described in the evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 104 — `OSHA_29CFR_full_2024_neg_1194_03` · HIGH confidence ·

Hallucination

- **Category:** —
- **Question:** Who in management must sign off to authorize cutting or welding in a restricted area?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The designated management representative must sign off to authorize cutting or welding operations in areas not specifically designed for such processes.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a requirement for a 'designated management representative' to sign off, which is not mentioned in the provided evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 105 — `OSHA_29CFR_full_2024_neg_0478_01` · HIGH confidence · Hallucination

- **Category:** —
- **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:** 100 linear feet per minute.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model confidently stated a minimum velocity as a maximum velocity
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 106 — `OSHA_29CFR_full_2024_neg_2625_03` · HIGH confidence · Wrong calc/quote

- **Category:** —
- **Question:** What is the minimum exposure concentration that triggers respiratory sensitizer classification under these criteria?
- **Ground truth:** UNANSWERABLE
- **Model answered:** $\geq 0.1\%$
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a concentration threshold
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 107 — `OSHA_29CFR_full_2024_neg_4403_04` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What form number must an employee use to file a complaint under this part?
- **Ground truth:** UNANSWERABLE
- **Model answered:** No particular form is required, but a complaint must be in writing and include a full statement of the acts and omissions with pertinent dates.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a procedural detail not present in the evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 108 — `OSHA_29CFR_full_2024_0751_02` · MEDIUM confidence · Model error

- **Category:** —
- **Question:** If neither the workplace nor the respirator type has changed and the employee's knowledge appears adequate, can retraining still be required?
- **Ground truth:** Yes, if any other situation arises in which retraining appears necessary to ensure safe respirator use.
- **Model answered:** Yes, retraining can still be required. The evidence states retraining shall be administered annually, and "when...inadequacies in the employee's knowledge or use of the respirator indicate that the employee has not retained the requisite understanding or skill" (k)(5)(ii). This means even without changes to the workplace or respirator, if an employee doesn't demonstrate continued understanding, retraining is necessary.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** [verifier] The evidence supports 'yes' only via clause (iii), 'Any other situation arises in which retraining appears necessary,' but the model justified its yes with employee knowledge inadequacy, which the question explicitly excludes by saying the employee's knowledge appears adequate.
- **Verifier verdict:** model_error (confidence: medium)
- **Validator flags:** (none)

Failure 109 — `OSHA_29CFR_full_2024_neg_0429_02` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** How many feet higher than flammable liquid storage should a bulk oxygen system be located?
- **Ground truth:** UNANSWERABLE
- **Model answered:** When locating bulk oxygen systems near above-ground flammable or combustible liquid storage, it is advisable to locate the system on ground higher than the flammable or combustible liquid storage. Additionally, when it is necessary to locate the bulk oxygen system on ground that is level with or lower than adjacent flammable liquid storage, suitable protective means shall be taken to prevent accumulation of liquids within 50 feet of the liquefied hydrogen container.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 110 — `OSHA_29CFR_full_2024_neg_1259_06` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** Who is responsible for invoking the blue flag policy referenced in paragraph (c)(9)(i)?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The evidence states a **flagman** shall invoke the blue flag policy referenced in paragraph (c)(9)(i).
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a specific actor (flagman) responsible for invoking the policy, while the evidence only states the policy shall be invoked, not who invokes it.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 111 — `OSHA_29CFR_full_2024_neg_2602_04` · HIGH confidence · Hallucination

- **Category:** —
- **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:** The evidence states that a new classification is necessary if there is significant variation in a production batch such that the toxicity has changed, and employers may wish to develop a form or clearance sheet to facilitate processing changes through the management of change procedures. A typical change form may include a description and the purpose of the change, the technical basis for the change, safety and health considerations, documentation of changes for the operating procedures, maintenance procedures, inspection and testing, P&IDs, electrical classification, training and communications, pre-startup inspection, duration if a temporary change, approvals and authorization.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 112 — `OSHA_29CFR_full_2024_neg_4485_08` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the maximum period the time for filing may be tolled under applicable case law?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The evidence states that the time for filing a complaint may be tolled if a complainant mistakenly files with an agency other than OSHA "within 180 days" of an alleged adverse action. Therefore, the maximum period the time for filing may be tolled is **180 days**.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 113 — `OSHA_29CFR_full_2024_neg_3512_03` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the maximum spacing allowed between cleats used to skidproof ramped walkway surfaces?
- **Ground truth:** UNANSWERABLE
- **Model answered:** 14 inches (35 cm).
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated a specific measurement not found in the provided text.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 114 — `OSHA_29CFR_full_2024_neg_4455_06` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the maximum dollar amount of compensatory damages that may be awarded in the preliminary order?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Up to \$250,000.
- **Doctor diagnosis:** model_error

- **Doctor root cause:** model hallucinated a specific dollar amount for compensatory damages, which is not stated in the provided evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 115 — `OSHA_29CFR_full_2024_3020_02` · HIGH confidence · Wrong focus

- **Category:** —
- **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:** According to the evidence, subpart (v) states that emergency stop switches on conveyors shall be arranged so the conveyor cannot be restarted until the switch is reset, and subpart (vi) requires audible warning signals before conveyor startup.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** [verifier] The evidence states that operations must minimize drum/container movement and exposed employees must be warned of hazards before movement, while the model answered with unrelated conveyor emergency stop and startup warning requirements not present in the evidence.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 116 — `OSHA_29CFR_full_2024_3390_03` · HIGH confidence · Model error

- **Category:** —
- **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:** This question cannot be answered from the provided evidence. The evidence references paragraphs (i) and (ii) but does not state what those paragraphs contain or what distances they specify.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** [verifier] The evidence explicitly states paragraph (ii) allows 'not less than 6 feet ... nor more than 60 feet ... or half the length ... whichever is less' and paragraph (i) has a general maximum of '25 feet,' so for 150 feet the maximum is 60 feet, greater than 25; the model incorrectly refused.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** (none)

Failure 117 — `OSHA_29CFR_full_2024_4047_04` · HIGH confidence · Direction reversed

- **Category:** —
- **Question:** If an employer wishes to hoist employees while rubber-tired equipment is traveling because there is no less hazardous way to perform the work, is this permitted?
- **Ground truth:** No; the no-less-hazardous-way exception does not apply to rubber-tired equipment.
- **Model answered:** Yes, this is permitted. According to (k)(12)(i)(B), hoisting of employees while equipment is traveling is prohibited except where the employer demonstrates that there is no less hazardous way to perform the work.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** unknown
- **Verifier verdict:** model_error (confidence: high)

- **Validator flags:** Direction mismatch: GT negative, response affirmative; Uncertainty violation: evidence hedges but response uses 'demonstrates that'

Failure 118 — `OSHA_29CFR_full_2024_neg_1769_06` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What fee may an employer charge for copying the COVID-19 records when requested?
- **Ground truth:** UNANSWERABLE
- **Model answered:** An employer may charge "reasonable, non-discriminatory administrative costs (i.e., search and copying expenses but not including overhead expenses)" for additional copies of a record after the initial copy is provided without cost. They cannot charge for an initial request for new information added to a previously provided record or for an initial request from a collective bargaining agent.
- **Doctor diagnosis:** model_error
- **Doctor root cause:** model hallucinated information about fees for copying records; the evidence only discusses record-keeping requirements.
- **Verifier verdict:** model_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 119 — `OSHA_29CFR_full_2024_neg_1769_05` · HIGH confidence · Bad question

- **Category:** —
- **Question:** By what time of day must the employer provide records following a request?
- **Ground truth:** UNANSWERABLE
- **Model answered:** By the end of the next business day. (See § 1904.35(b)(2)(iii))
- **Doctor diagnosis:** bad_test
- **Doctor root cause:** [verifier] The evidence explicitly answers the question: "By the end of the next business day after a request," so the ground truth UNANSWERABLE contradicts the evidence; the model's timing answer matches despite an extraneous citation.
- **Verifier verdict:** bad_test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer