

VERITROOPER Scout Report — Generated — OSHA_29CFR_full_2024 — veritrooper v9



Generated — OSHA_29CFR_full_2024 — veritrooper v9
13 July 2026

VERITROOPER v9 accuracy & robustness audit. Generated 2026-07-13 23:17:28.

Audit summary

What this report is and how it was produced — at a glance.

Field	Value
Run identifier	VT-20260601-OSHA-001 (prefix on the delivered filenames)
System evaluated	Generated — OSHA_29CFR_full_2024 — veritrooper v9
Model under test (MUT)	Meta-Llama-3.1-70B-Instruct
Diagnostician (Doctor)	Meta-Llama-3.1-70B-Instruct *(same model as the MUT)*
First-pass validator	deterministic rule check (code, not a model)
Independent verifier	GPT (frontier model)
Questions generated	1000 (1000 scored · 0 bad tests · 0 connection error(s), excluded symmetrically)
Question set generated	2026-06-01 10:47:38
Evaluation started	2026-06-01 17:38:37
Evaluation finished	2026-06-01 19:57:24
Report location	C:\VTRunRecords\OSHA_llama-3.1-70b

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

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

Results at a glance

Field	Value
Scored items	1000 (of 1000; 0 bad tests + 0 connection error(s) excluded symmetrically)
Final verified accuracy	baseline 79.9% → pipeline 95.8% (Δ +15.9pp)
Confirmed model errors	baseline 201 → pipeline 42
Results fingerprint (SHA-256)	8c9ce82426d374dd...
Human sign-off	not yet on record for these results

One consolidated evidence document. The full per-question Q&A ledger and the machine-readable exports (evidence.json, model_card.yaml, otel_spans.json) accompany this file as companions. Testing evidence only — does not constitute or confer conformity.

Contents

- 1. Executive Summary
- 2. Audit Report — Results & Analysis
- 3. System Card
- 4. Data Card
- 5. Framework Crosswalk
- 6. Performance-Gap & Coverage Diagnostic
- 7. EU AI Act Annex IV — Testing Record
- 8. Data Handling & Isolation Posture

(Use the PDF bookmarks / outline panel to jump to any section.)

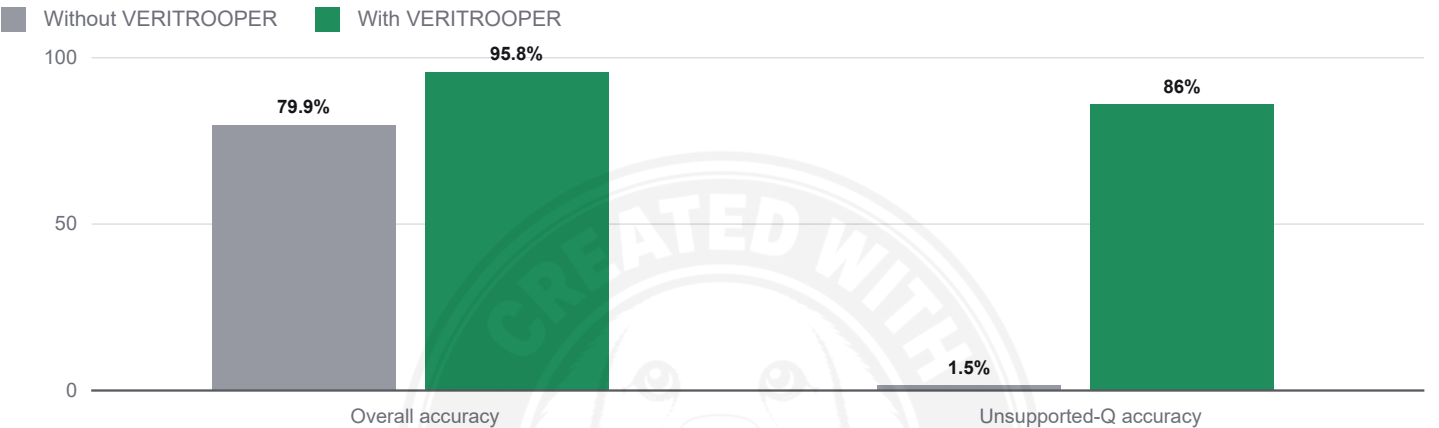
1. Executive Summary

Model under test: Meta-Llama-3.1-70B-Instruct · **Scored items:** 1000 · **Generated:** 2026-07-13 23:17:27
Human sign-off: not yet recorded for these results.
Bottom line. With VERITROOPER, the system answered more accurately and correctly refused unsupported questions.

Scorecard

Metric	Without VERITROOPER (vanilla-RAG baseline)	With VERITROOPER
Final verified accuracy	79.90%	95.80%
Confirmed model errors	201	42
Improvement		+15.90pp
Failures recovered		85%

Accuracy — with vs. without VERITROOPER



What it means

- The VERITROOPER pipeline recovered **85%** of the 201 answers the vanilla-RAG baseline got wrong (79.9% → 95.8% final verified accuracy).
- Weakest area for the raw model: **Negative (hallucination traps)** at 2% baseline (n=200) — the pipeline lifts it to 86%.
- The most common baseline failure mode was **other** (206 of 361 confirmed errors) — the class of error the pipeline configuration is designed to eliminate.

Recommended action

- The tested pipeline configuration outperformed the baseline on this material and is the preferred candidate for broader validation and — subject to provider risk review and recorded human sign-off — a limited controlled pilot. It removed the confirmed error class above at no accuracy cost on this set; final release remains subject to the recorded release disposition and responsible-person sign-off.
- Consider the deterministic guardrails in the full report to catch the same failure class in your own stack — no model retraining.
- Treat these figures as indicative on this material (scored n=1000); broaden the set for statistically tight claims before an unconditional production decision.

This one-pager summarizes the run; the full report, the per-question listings, and the compliance-evidence artifacts accompany it in the same package.

2. Audit Report — Results & Analysis

Headline Accuracy (bad-test cases excluded from denominator)

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

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

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

Release disposition

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

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

Acceptance criteria & results:

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

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

Human sign-off

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

Paired outcome (same questions, both arms)

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

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

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

Per-Category Accuracy & Failure Recovery

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

including any where the pipeline matches or underperforms baseline.

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

Adjudication reconciliation

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

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

Test-set coverage

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

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

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

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

Estimated audit resource usage

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

Measured

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

API).

Estimated

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

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

Without veritrooper — Baseline LLM Performance

What the LLM Did On Its Own

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

Failure Patterns (Baseline)

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

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

Baseline Remediation Recommendations

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

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

With veritrooper — Pipeline Performance

What the LLM Did With veritrooper

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

Pipeline Remediation Recommendations

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

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

Independent Verification

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

Bottom Line

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

Deterministic Guardrails

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

Block — reject or correct automatically:

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

Flag — detect and route to review:

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

Case counts are per flagged answer and can overlap — a single answer may exhibit more than one failure pattern — so they need not sum to the number of failed questions. Patterns without a listed guard are semantic reasoning slips with no clean deterministic check; the remediation recommendations above address those.

3. System Card

About this document. A system card documents the system as deployed — the model together with its retrieval and the evaluation wrapper — not just the trained weights. VERITROOPER auto-populates the measurable sections from one audit run; sections that only the provider can supply (intended use, deployment context) are marked as such and left for the provider to complete. This card is evidence of measured behaviour; it does not certify the system.

System identification

Field	Value
Model under test	Meta-Llama-3.1-70B-Instruct
Evaluated material / knowledge base	Generated — OSHA_29CFR_full_2024 — veritrooper v9
Task type	generation
Evaluation engine	VERITROOPER v9
Test items	1000
Card generated	2026-07-13 23:17:27

Intended use & out-of-scope uses

Provider-supplied — VERITROOPER does not infer intended use. Complete before publishing this card.

- Primary intended use(s): _____
- Intended users: _____
- Out-of-scope / prohibited uses: _____

Evaluation data

The system was evaluated on a purpose-generated question set derived from the provider's material. Question categories exercised: Negative (hallucination traps), Cross-Reference, Calculation, Precision, Conditional, Exception, Cause & Effect.

Field	Value
Test items	1000
Categories	7
Refusal / unanswerable items (hallucination traps)	200
Source material	data/OSHA_29CFR_full_2024.txt

Full provenance — generation method, parameters, and the exact system prompt — is in the accompanying Data Card.

Evaluation methodology

- **Baseline arm** — the model with standard retrieval (vanilla RAG).
- **Pipeline arm** — the same model evaluated through the VERITROOPER engine.
- **Adjudication** — each verdict checked by a verification model (the Doctor: Meta-Llama-3.1-70B-Instruct) and independently audited by GPT (frontier model), applied to **both** arms under the same standard so neither arm gets a free pass.
- **Bad-test exclusion** — items whose own ground truth was malformed are excluded symmetrically from both arms; the excluded count is reported.

Independence of the judging steps

Who did what in this audit, and how independent each judging step is from the model under test (MUT):

Step	Performed by	Type	Independent of the MUT?
Answer under test	Meta-Llama-3.1-70B-Instruct	the model being audited	—
First-pass check (Validator)	deterministic rule check	code, not a model	yes — not a model
Diagnosis (Doctor)	Meta-Llama-3.1-70B-Instruct	verification LLM	no — the same model as the MUT
Independent verification (Verifier)	GPT (frontier model)	model / vendor	yes — an independent verifier

Note: on this run the Doctor is the same model as the MUT, so the Doctor step is not an independent second vendor. The independent cross-check is the Verifier, applied to both arms under the same standard.

Baseline configuration

Baseline retrieval configuration — a **standardized BM25 reference baseline** the pipeline is measured against (a reproducible lexical-retrieval reference, not necessarily the customer's production RAG stack):

Parameter	Value
Retrieval	BM25 lexical (Robertson et al. defaults)
Chunking	fixed-size, 500 tokens, 50-token overlap
Retrieved chunks	top-5 by BM25 score
Reranking	none
Query rewriting	none
Dense embeddings	none

Full machine-readable run configuration is in config_snapshot.json.

The full quantitative results and adjudication reconciliation, remediation guardrails, data-handling posture, and framework mapping for this system appear once — in the Results & Analysis, Data Handling & Isolation Posture, and Framework Crosswalk sections of this report — and are not repeated here.

Known limitations & failure modes

Failure modes across the 361 confirmed baseline model errors (a failed answer may carry more than one diagnostic label — or one that maps to no named pattern — so these counts need not equal the number of failed answers):

- other — 206 answers
- knowledge gap — 133 answers
- computed vs quoted — 16 answers
- misinterpretation — 3 answers
- unit format mismatch — 2 answers
- table misread — 1 answer

Results establish measured behaviour on the tested material, not field performance outside the tested distribution.

Human oversight & sign-off

No human sign-off is on record for these exact results yet. The responsible person should review and sign the audit (recorded, hash-bound, in audit_trail.jsonl) before this card is published.

Reproducibility & provenance

Field	Value
Engine version	VERITROOPER v9
Run started	2026-06-01 17:38:37
Run finished	2026-06-01 19:57:24
Sampling temperature	0.1
Max tokens	200
Results content hash (SHA-256)	8c9ce82426d374dd...

System prompt given to the model under test:

Answer the question accurately using only the provided evidence.
Keep your response concise — one or two sentences.
Use only information from the evidence provided.

The run is reproducible from the stored record (certification_report.json + round_1.json + the generated question set) in the same package.

Deployment context & risk considerations

Provider-supplied — complete before publishing this card.

- Deployment context (where/how the system is used): _____
- Human-oversight measures in production (Art. 14): _____
- Known ethical / risk considerations & mitigations: _____

Card version

Field	Value
Card revision	this audit run (2026-06-01 19:57:24)
Supersedes	prior audits of the same panel (see the drift report, if any)

4. Data Card

Motivation

This question set exists to measure the model's accuracy and robustness on the provider's own material — not on a generic public benchmark. Questions are generated from the source document so the evaluation reflects the deployment domain.

Composition

Field	Value
Total items	1000
Categories	7
Refusal / unanswerable (hallucination-trap) items	200
Source material	data/OSHA_29CFR_full_2024.txt

Items per category:

Question type	Items
Cross-Reference	251
Precision	218
Conditional	211
Negative (hallucination traps)	200
Exception	93
Cause & Effect	15
Calculation	12

Generation process

- The source document was chunked; questions were generated per chunk with code-computed ground truth where the answer is arithmetic or tabular.
- Adversarial negative items (hallucination traps) were generated to have **no supported answer** — the correct response is to refuse — and passed through a verification gate before inclusion.
- Items whose ground truth was later found malformed are quarantined as **bad tests** and excluded symmetrically from scoring rather than counted against either arm.

Recommended uses & limitations

- **Use:** point-in-time accuracy/robustness evaluation of a model on this material; regression testing across model or prompt changes (re-run the frozen panel).
- **Not for training.** These items are evaluation artifacts; training on them would contaminate future measurement.
- **Not a traffic sample.** The set probes capability by question type; it is not a statistical sample of real production queries, and category counts are generation-driven, not usage-weighted.

Provenance & reproducibility

Field	Value
Engine version	VERITROOPER v9
Generated data file	data/OSHA_29CFR_full_2024_generated_20260601_104738.json
Generation temperature	0.1
Sampling strategy	full

5. Framework Crosswalk

This crosswalk shows which accuracy, robustness, and documentation requirements across the **EU AI Act**, the **NIST AI RMF**, and **ISO/IEC 42001** the evidence in this package **maps to** — and which it does not. It is a pointer from requirement to evidence, to speed a review. It **does not** assert conformity with any framework; conformity is a determination for the provider and its assessor.

Regulatory posture (verified 2026-07-09). No harmonised standard for high-risk AI has yet been cited in the Official Journal, so **nothing on the market today confers a presumption of conformity** — any vendor claiming AI Act "compliance" is claiming something that cannot currently be conferred. The Digital Omnibus (Council green light 29 June 2026, OJ publication pending) defers the high-risk obligations to **2 December 2027** for stand-alone systems and **2 August 2028** for AI embedded in regulated products. The general-purpose AI obligations (Chapter V) have applied since **2 August 2025** and were **not** deferred. No regime reviewed requires an independent third-party AI audit. This package is **evidence for** the obligations below; it is not, and cannot be, compliance with them.

EU AI Act (Reg. (EU) 2024/1689)

Reference	Requirement (summary)	VERITROOPER evidence	Artifact
Art. 15(3) — Declared accuracy metrics	The levels of accuracy and the relevant accuracy metrics of a high-risk system shall be declared in the accompanying instructions of use. A binding disclosure, not a recommendation.	A named metric with its definition, the measured level, the question set it was measured on, and a reproducible verdict per question — the content that disclosure requires.	VT-20260601-OSHA-001_Results_Report.pdf · VT-20260601-OSHA-001_System_Card.pdf · certification_report.json
Art. 55(1)(a) — GPAI with systemic risk: documented adversarial testing	Model evaluation using standardised protocols and tools, including documented adversarial testing. In force since 2 Aug 2025 — NOT deferred by the Digital Omnibus.	Negative / hallucination-trap categories are adversarial by construction; every prompt, answer and verdict is retained, and the package is cryptographically signed (with an RFC 3161 trusted timestamp where a network was available at sealing).	VT-20260601-OSHA-001_Results_Report.pdf · VT-20260601-OSHA-001_All_QA_Pairs.pdf · findings.sarif.json
Art. 15 — Accuracy & robustness (cybersecurity not assessed)	Appropriate accuracy levels and robustness across the lifecycle. Art. 15 also covers cybersecurity, which this accuracy/robustness audit does not assess.	Final verified accuracy + per-category accuracy + adversarial (hallucination-trap) results, both arms.	VT-20260601-OSHA-001_Results_Report.pdf · VT-20260601-OSHA-001_System_Card.pdf
Art. 11 / Annex IV §2(g)	Validation & testing procedures, metrics, and dated test logs in the technical file.	Annex IV Testing & Validation Record with dated log and metric definitions.	VT-20260601-OSHA-001_Annex_IV_Test_Record.pdf
Art. 10 — Data & data governance	Examination of data for gaps, shortcomings and representativeness.	Performance-gap / representativeness diagnostic by question category.	VT-20260601-OSHA-001_Gap_Diagnostic.pdf
Art. 72 / Annex IV §9 — Post-market monitoring	A monitoring system tracking performance over the system's life.	Accuracy-drift report across repeated audits of a frozen question panel.	drift_report (when ≥2 audits of a panel exist)
Art. 14 — Human oversight	Human oversight of the deployed system in operation.	NOT covered by this audit. The sign-off here is QMS review of test results, not operational oversight; supply Art. 14 measures separately.	— (provider-supplied)

Texas TRAIGA (HB 149) — in force 1 Jan 2026

Reference	Requirement (summary)	VERITROOPER evidence	Artifact
Affirmative defense	No audit is mandated. The Act gives an affirmative defense for substantial compliance with the NIST AI RMF Generative AI Profile (AI 600-1) via internal review, and for discovering violations through testing, including adversarial or red-team testing.	A dated, signed audit with adversarial categories and reproducible verdicts — the testing record the defense contemplates.	VT-20260601-OSHA-001_Results_Report.pdf · Integrity/manifest.json

Colorado SB26-189 (ADMT) — obligations from 1 Jan 2027

Reference	Requirement (summary)	VERITROOPER evidence	Artifact
Developer documentation + 3-year records	Developers must give deployers documentation of intended uses, known limitations, and human-review instructions; both parties retain compliance records for at least three years. The replaced 2024 Act's audit regime was dropped — no third-party audit is required.	Known-limitations evidence (failure clusters, severity, guardrail recipes) inside a cryptographically signed package (RFC 3161 timestamped where a network was available at sealing) that supports the retention duty.	VT-20260601-OSHA-001_System_Card.pdf · VT-20260601-OSHA-001_Results_Report.pdf · Integrity/

Reference	Requirement (summary)	VERITROOPER evidence	Artifact
MEASURE 2.3 — Performance measured	System performance is measured and documented in deployment conditions.	Two-arm accuracy measurement on domain material with per-category breakdown.	VT-20260601-OSHA-001_Results_Report.pdf · evidence.json
MEASURE 2.5 — Validity & reliability	Validity and reliability are evaluated and documented.	Independent adjudication of each verdict on both arms; reproducible from stored data.	VT-20260601-OSHA-001_System_Card.pdf · evidence.json
MEASURE 2.7 — Robustness / adversarial	Resilience to adversarial or out-of-distribution inputs is assessed.	Dedicated hallucination-trap category measuring refusal on unsupported questions.	VT-20260601-OSHA-001_Results_Report.pdf (Negative category)
MAP / MANAGE — documentation & traceability	Context, limitations and decisions are documented and traceable.	System card, data card, audit trail, the hash-bound human sign-off once recorded, and the whole package sealed with a standards-based digital signature (plus an RFC 3161 trusted timestamp where a network was available at sealing).	VT-20260601-OSHA-001_System_Card.pdf · VT-20260601-OSHA-001_Data_Card.pdf · audit_trail.jsonl · Integrity/

ISO/IEC 42001:2023 (AIMS)

Reference	Requirement (summary)	VERITROOPER evidence	Artifact
Clause 9 — Performance evaluation	Monitoring, measurement, analysis and evaluation of the AI system.	Measured accuracy/robustness with a repeatable, documented procedure.	VT-20260601-OSHA-001_Results_Report.pdf · VT-20260601-OSHA-001_Annex_IV_Test_Record.pdf
Annex A — AI system verification & validation	The AI system is verified and validated against requirements.	Independent two-arm V&V of answers against ground truth, adjudicated.	VT-20260601-OSHA-001_System_Card.pdf
Annex A — Documented information / impact	Documented information on AI system performance and data is maintained.	System card + data card + machine-readable evidence bundle.	VT-20260601-OSHA-001_System_Card.pdf · VT-20260601-OSHA-001_Data_Card.pdf · evidence.json

How to read this

- **Maps to** means the artifact provides evidence relevant to the requirement — a reviewer still has to accept it in context.
- Rows marked **NOT covered** flag requirements this audit does not address (notably operational human oversight); supply those separately.
- ISO/IEC 42001 clause and Annex A references are named by control area; confirm the exact clause numbers against your Statement of Applicability.

6. Performance-Gap & Coverage Diagnostic

Scope. Article 10 of Regulation (EU) 2024/1689 requires examining data for gaps, shortcomings, and representativeness. This diagnostic surfaces where the deployed system's accuracy is uneven across the tested question categories on the provider's material (the audited, achievable accuracy is shown alongside as a benchmark) — behavioural evidence supporting that examination. It identifies **capability and coverage gaps by question type**; it does **not** measure demographic or protected-attribute bias, and does not establish or rule out discrimination (that requires attribute-labelled data this test does not contain). It is evidence supporting the Art. 10 review, not a data-governance determination, and does not constitute or confer conformity.

Accuracy by category (weakest first)

Overall deployed accuracy: **79.90%** (achievable with grounding: 95.80%). **Recovered** is the accuracy the deployed system is leaving on the table in that category (pipeline – baseline); weighted by N it sums to the headline delta. **vs deployed avg** is that category's baseline against the 79.90% deployed average — a signed delta, so a positive number means the category is stronger than average. **Status** describes the DEPLOYED system: categories **10pp or more below** the deployed average are flagged GAP for examination.

Question type	N	Baseline acc	Pipeline acc	Recovered	vs deployed avg	Status
Negative (hallucination traps)	200	1.50%	86.00%	+84.50pp	-78.40pp	GAP (adversarial)
Cross-Reference	251	99.20%	97.61%	-1.59pp	+19.30pp	—
Conditional	211	99.53%	97.63%	-1.90pp	+19.63pp	—
Precision	218	99.54%	99.08%	-0.46pp	+19.64pp	—
Calculation	12	100.00%	100.00%	+0.00pp	+20.10pp	—
Exception	93	100.00%	100.00%	+0.00pp	+20.10pp	—
Cause & Effect	15	100.00%	93.33%	-6.67pp	+20.10pp	—

Performance dispersion

- Strongest: **Calculation** at 100.00%
- Weakest: **Negative (hallucination traps)** at 1.50%
- Spread (range across categories): **98.50pp**

Adversarial (refusal-behaviour) gaps: Negative (hallucination traps). These are hallucination-trap categories whose questions are intentionally unsupported — the correct answer is to refuse. A low score here is a model **refusal/robustness** weakness (the model answered anyway), **not** evidence of data under-representation; the unsupported items are absent from the source by design.

Coverage

All categories have at least 5 questions.

7. EU AI Act Annex IV — Testing Record

Scope of this document. This record populates the validation-and-testing elements of the technical documentation required under Article 11 and Annex IV of Regulation (EU) 2024/1689 (the EU AI Act) — specifically Annex IV §2(g), and it supports §3 (expected accuracy levels and limitations) and §4 (appropriateness of the performance metrics). It is **one component** of the full technical documentation; the remaining elements (§1, §2(a)–(f), §5, §7–§9) are supplied by the provider.

This document is testing evidence only. It does not by itself constitute, certify, or confer conformity with the EU AI Act, and does not replace the conformity assessment or the provider's other obligations.

System identification

Field	Value
Model / AI system evaluated	Meta-Llama-3.1-70B-Instruct
Material / data set evaluated	Generated — OSHA_29CFR_full_2024 — veritrooper v9
Number of test items	1000
Task type	generation

Provider-supplied fields (complete before filing — these are not generated by VERITROOPER):

- Provider — legal name & address: _____
- Official system name & version: _____
- Intended purpose (Annex IV §1): _____

Testing & validation methodology (Annex IV §2(g))

The model was evaluated on a purpose-generated question set derived from the material in §1, under two arms scored on identical questions:

- **Baseline** — the model with standard retrieval (vanilla RAG), a standardized BM25 reference baseline.
- **Pipeline** — the model evaluated through the VERITROOPER engine.

Each verdict was checked by a verification model (the Doctor) and independently audited by a separate Verifier applied to **both** arms under the same standard, so neither arm receives a free pass.

- Evaluation engine: **VERITROOPER v9**
- Model under test (MUT): Meta-Llama-3.1-70B-Instruct
- Verification model (Doctor): Meta-Llama-3.1-70B-Instruct
- Independent Verifier: GPT (frontier model)
- Question categories exercised: Negative (hallucination traps), Cross-Reference, Calculation, Precision, Conditional, Exception, Cause & Effect.

Independence note: on this run the Doctor is the same model as the MUT, so the Doctor step is not an independent second vendor.

The independent cross-check is the Verifier above, applied identically to both arms.

Bad-test exclusion. Questions whose own ground truth was found to be malformed are excluded symmetrically from both arms' denominators, so neither model is penalised for a defective question; the count excluded is shown in §4. The set includes adversarial hallucination-trap items (the negative category) that probe robustness, not only straightforward retrieval.

Metrics used (Annex IV §2(g))

- **Accuracy** — proportion of test items answered correctly.
- **Final verified accuracy** — the headline accuracy, reported after a three-stage adjudication of every verdict: (1) an initial automated score, (2) Doctor adjudication that corrects false flags, and (3) independent cross-verification. It is the final, verified figure — not a hand-adjusted one.
- **Failure-Recovery Rate** — the fraction of baseline failures the pipeline resolves; a direct measure of value added on this material.
- **Per-category accuracy** — accuracy stratified by question type, including any adversarial categories, to surface robustness and uneven performance.

The verified accuracy results for this Annex IV record are shown once in the Results & Analysis section above and are incorporated here by reference.

Scope and limitations (supports Annex IV §3 and §4)

This record documents accuracy and robustness measured on the question set in §1–§2. It establishes measured accuracy on that material; it does not establish field performance outside the tested distribution, nor address the Act's non-testing requirements — risk management (Art. 9), human oversight (Art. 14), cybersecurity (Art. 15), data governance (Art. 10) — documented elsewhere in the technical file.

Appropriateness of the metrics (Annex IV §4). For a question-answering / generation system over the provider's material, the load-bearing properties are correctness against ground truth and resilience to adversarial items: accuracy and per-category accuracy measure the former; Failure-Recovery Rate measures recovered correctness against the baseline.

Test log (Annex IV §2(g) — dated)

Field	Value
Evaluation started	2026-06-01 17:38:37
Evaluation finished	2026-06-01 19:57:24
Test items	1000
Evaluation engine version	VERITROOPER v9
Source record (reproducible from stored data)	C:\VTRunRecords\OSHA_llama-3.1-70b

Responsible-person sign-off (Annex IV §2(g) — "dated and signed by the responsible persons")

Annex IV §2(g) requires test reports to be dated and signed by the responsible person. The provider's responsible person should review the record above and complete the following before it is included in the technical documentation:

- Reviewed by (name): _____
- Role / position: _____
- Date: _____
- Signature: _____

8. Data Handling & Isolation Posture

This document records the data-handling facts of this audit run and the network mode it can and cannot establish. It complements — it does not replace — your own network controls and any accreditation (e.g. ATO, FedRAMP, DoD IL). It is not a cryptographic attestation and does not certify the audited system.

Verifiable data handling

- **Read-only over source & model.** The engine reads the evaluated material and queries the model; it does not modify or export either.
- **Local artifacts only.** The only data written is this audit's own output (reports, JSON, this document), written locally into the results folder you control. It is not transmitted anywhere by the engine.
- **Bundled local grader available.** VERITROOPER ships an offline validator, so the grading step can run with no third-party judge model.

Components recorded for this run

Role	Recorded as	Locality
Model under test	Meta-Llama-3.1-70B-Instruct	endpoint locality not captured — confirm on your network
Verification model (Doctor)	Meta-Llama-3.1-70B-Instruct	endpoint locality not captured — confirm on your network
Independent verifier	GPT (frontier model)	endpoint locality not captured — confirm on your network

Network mode

Overall network isolation is **determined by the endpoints the operator selected** for the model under test and the Doctor. When every selected endpoint is on your network or local, the whole audit runs with no outbound internet; when a hosted API is selected, that component reaches out. This record does not capture each endpoint's address, so it does not assert full air-gap on its own — confirm the mode with your network monitoring (or a pre-run isolation check) for a demonstrable attestation.

What this is NOT

- Not an accreditation (ATO / FedRAMP / DoD IL) — that is the operating environment's, not this tool's.
- Not a cryptographic boundary or a zero-bytes-written claim — the engine writes its own audit artifacts locally.
- Not a certification of the audited system's security or conformity.

Framework relevance

Supports data-minimization and system-boundary considerations under the NIST AI RMF and data-handling controls under ISO/IEC 42001; confirm the specific controls against your own control set.

Generated by VERITROOPER as a consolidated audit report. This artifact is testing evidence only. It does not by itself constitute, certify, or confer conformity with any framework, and does not replace the conformity assessment or the provider's other obligations. Have counsel review before external use.

