



Generated — OSHA_29CFR_full_2024 — veritrooper v9
13 July 2026

Deployed-system documentation for the evaluated AI system, generated from an independent accuracy & robustness audit. Generated 2026-07-13 23:17:27.

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.

Contents (click a section to jump)

1. System identification
2. Intended use & out-of-scope uses
3. Evaluation data
4. Evaluation methodology
5. Quantitative results
6. Robustness & adversarial behaviour
7. Known limitations & failure modes
8. Remediation
9. Human oversight & sign-off
10. Reproducibility & provenance
11. Data handling & isolation posture
12. Framework alignment
13. Deployment context & risk considerations
14. Card version

1. 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

2. 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: _____

3. 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.

4. 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.

5. Quantitative results

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%

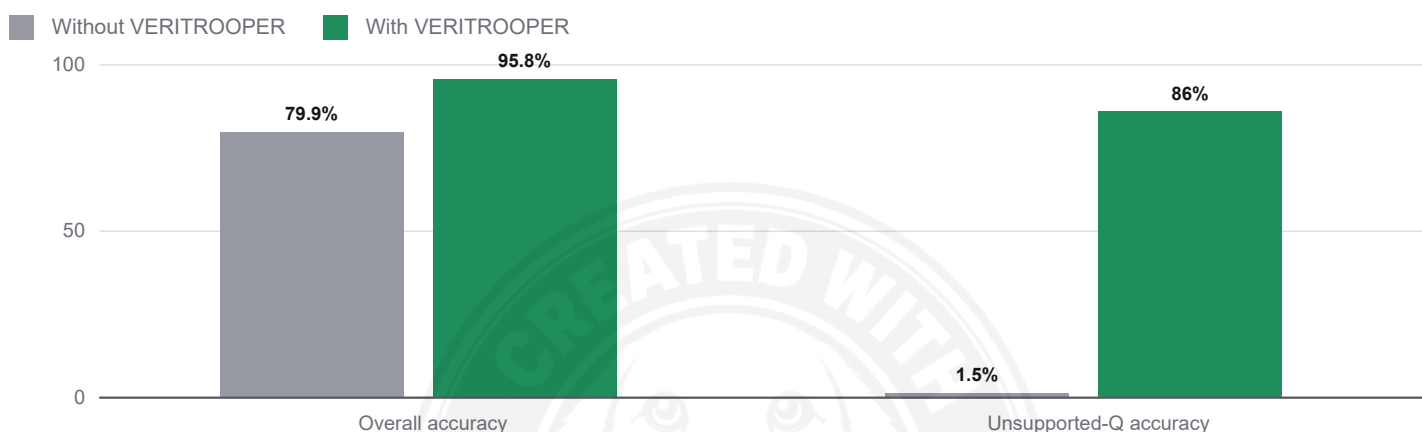
The 95% confidence interval reflects sampling uncertainty on 1000 scored items: a point estimate of 100% does not prove a 0% error rate in the field — it means no error was observed in this sample; a tighter bound needs more questions.

Per-Category Accuracy & Failure Recovery

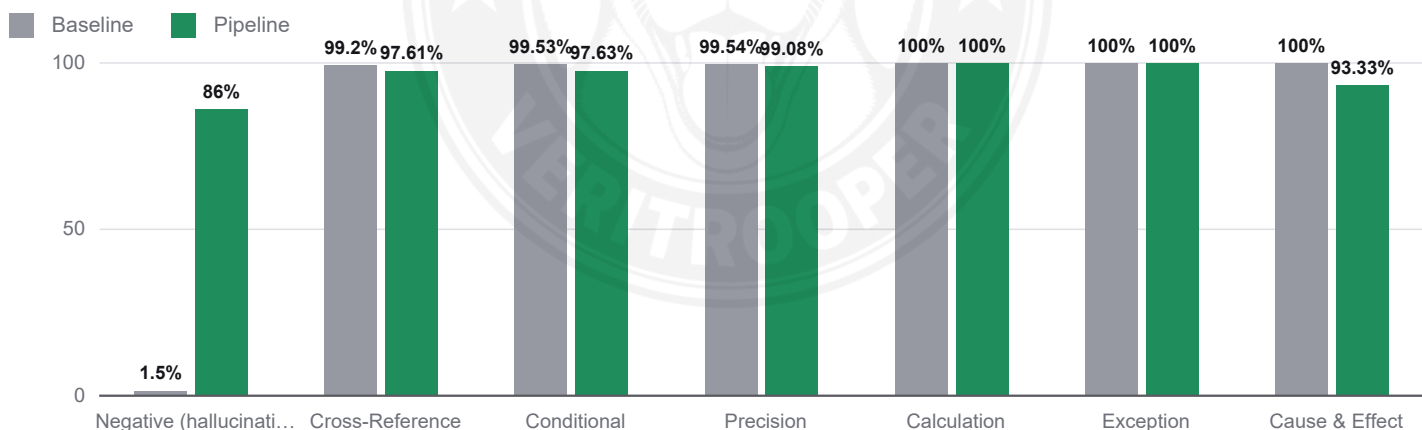
Per-category accuracy with and without VERITROOPER. **Failure-Recovery Rate** measures the fraction of vanilla-RAG baseline failures the pipeline recovers — the most direct signal of where the architecture adds value on this material. Every category is shown, including any where the pipeline matches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

Accuracy — with vs. without VERITROOPER



Per-category accuracy — baseline vs. pipeline



6. Robustness & adversarial behaviour

The set includes adversarial hallucination-trap items (the negative category, n=200) whose correct response is to refuse. On these, the raw model scored **1.50%** and the pipeline **86.00%** — the single clearest signal of whether the system answers questions its evidence cannot support.

7. 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.

8. Remediation

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.

9. 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.

10. 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.

11. Data handling & isolation posture

VERITROOPER reads the evaluated material and the model without modifying or exporting either, and writes only its own audit artifacts, locally, into this results folder. When every selected endpoint (model under test and grader) is on your network or local, the audit runs with no outbound internet. See the accompanying **Data Handling & Isolation Posture** document for this run's recorded components and the honest scope of that claim.

12. Framework alignment

This audit produces evidence that **maps to** accuracy/robustness and documentation elements of the EU AI Act, the NIST AI RMF, and ISO/IEC 42001. It does not by itself establish conformity with any of them. The accompanying **Framework Crosswalk** shows the mapping control-by-control.

13. 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: _____

14. 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)

Generated by VERITROOPER as a system card documenting measured system behaviour. 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.

