

VERITROOPER Scout — EU AI Act — Annex IV Technical Documentation: Testing & Validation Record



Article 11 / Annex IV §2(g) — validation and testing procedures, the metrics used to measure accuracy and robustness, and dated test logs. Generated 2026-07-13 23:17:27.

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.

Contents (click a section to jump)

- 1. System identification
- 2. Testing & validation methodology (Annex IV §2(g))
- 3. Metrics used (Annex IV §2(g))
- 4. Measured accuracy levels (Annex IV §2(g) results; supports §3)
- 5. Scope and limitations (supports Annex IV §3 and §4)
- 6. Test log (Annex IV §2(g) — dated)
- 7. Responsible-person sign-off (Annex IV §2(g) — "dated and signed by the responsible persons")

1. 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): _____

2. 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.

3. 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.

4. Measured accuracy levels (Annex IV §2(g) results; supports §3)

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%

5. 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.

6. 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

7. 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: _____

Generated by VERITROOPER as testing evidence in support of EU AI Act Article 15 and Annex IV §2(g). This artifact does not constitute or confer conformity. Have counsel review before external use.

