

VERITROOPER Scout — EU AI Act — Performance-Gap & Coverage Diagnostic



Article 10 contribution (capability & coverage gaps by question type). Model: Meta-Llama-3.1-70B-Instruct on Generated — OSHA_29CFR_full_2024 — veritrooper v9. Generated 2026-07-13 23:17:27.

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.

Generated by VERITROOPER as Article 10 capability & coverage-gap evidence. Capability and coverage gaps by question type only — adversarial (refusal-behaviour) gaps are model-robustness weaknesses, not data under-representation; and this is not a measure of demographic bias or discrimination, and not a data-governance determination. Does not constitute or confer conformity. Have counsel review before external use.