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 2026-08-04 15:54:07.

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

Field	Value
Model / AI system evaluated	gpt-5.5
Material / data set evaluated	Generated — SEC_10K_multi — veritrooper v9
Number of test items	1000
Task type	generation

Provider identification (Annex IV §1) — supplied by the provider at system registration:

Field	Value
Provider — legal name	VERITROOPER LLC
Provider — registered address	14 Lily Avenue, Rouses Point, NY 12979, United States
Contact point for the technical documentation	brianb@veritrooper.com

Provider-supplied fields still required (complete before filing — VERITROOPER does not generate these, and will not invent them):

- Official system name: _____
- Official system version: _____
- Intended purpose (Annex IV §1): _____

These are captured once, at system registration, and filled in automatically on every record afterwards. Set them in the console under Connections → Provider identity, or write them to C:\ProgramData\Veritrooper\provider.json.

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 lexical retrieval (vanilla RAG): a **reproducible reference baseline**, fully specified in §2.2 so a third party can rebuild it.
- **Pipeline** — the model evaluated through the VERITROOPER engine.

2.2 Reference baseline — exact configuration

Stated so the baseline arm can be reproduced independently. Read from the retrieval code that ran, not from documentation about it.

Setting	Value
Retrieval function	Okapi BM25 (Robertson et al.)
BM25 k1	1.5
BM25 b	0.75
IDF variant	log((N - df + 0.5) / (df + 0.5) + 1)
Tokenizer	regex [A-Za-z0-9_]+, lower-cased, tokens of length 1 discarded
Chunking	fixed-size, 500 tokens, 50-token overlap, whitespace-split
Passages retrieved per question (k)	5
Query	the test question verbatim, tokenized identically to the corpus
Index	rebuilt from the corpus in §1 at run time; no external index

No stop-word list, no stemming, no query expansion and no re-ranking are applied. The baseline is deliberately a competent, ordinary retrieval path — not a weak one — because the measured difference is only meaningful against a baseline a real team would plausibly deploy.

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): gpt-5.5
- Verification model (Doctor): —
- Independent Verifier: Gemini (frontier model)
- Question categories exercised: Cross-Reference, Conditional, Precision, Exception, Negative (hallucination traps), Cause & Effect, Calculation.

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. Every verdict behind it is decided in code by a deterministic validator and pre-filter, which is why a re-run of the same answers reproduces the same number exactly. Two independent model adjudicators — the Doctor, and a Verifier from a different vendor — review every verdict on BOTH arms under the same standard, and their findings are recorded against each item. Those findings are advisory: they are published so a reader can see where judgement disagreed with the code, and they do not move the headline. Where a reviewer judged an answer correct that the deterministic check scored wrong, the answer still counts against the model and the disagreement is reported.
- **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	82.60% (826 / 1000)	96.10% (961 / 1000)
Answers scored incorrect	174	39
Δ vs baseline		**+13.50pp**
Failure-Recovery Rate (pipeline recovers what % of baseline failures)		85.6%
95% confidence interval (Wilson)	80.1–84.8%	94.7–97.1%

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	60.00%	96.50%	+36.50pp	91.2%
Cross-Reference	243	71.19%	88.07%	+16.87pp	75.7%
Precision	315	93.33%	99.68%	+6.35pp	100.0%
Cause & Effect	78	98.72%	100.00%	+1.28pp	100.0%
Calculation	9	88.89%	88.89%	—	100.0%
Conditional	47	100.00%	100.00%	—	—
Exception	108	99.07%	99.07%	—	0.0%
All Categories	**1,000**	**82.60%**	**96.10%**	**+13.50pp**	**85.6%**

4.1 Test data (Annex IV §2(g) — description of the validation data)

Property	Value
Question set	not included in this package
Source material	Generated — SEC_10K_multi — veritrooper v9
Question generation	generated from the source material by the VERITROOPER generator; not drawn from any public benchmark
Sampling of the generated set	all generated items scored
Ground-truth provenance	derived from the source material at generation time and stored with each item; the material is the authority, not the model
Items generated	1,000
Items scored	1,000
Excluded — malformed question (bad test item)	0 — excluded from BOTH arms' denominators, so neither model is penalised for a defective question
Excluded — infrastructure/connection error	0 — the model was never asked, so no verdict exists

Distribution by category — what the question set actually covers:

Category	Items	Share
precision	315	31.5%
cross_reference	243	24.3%
negative	200	20.0%
exception	108	10.8%
cause_effect	78	7.8%
conditional	47	4.7%
calculation	9	0.9%

Nothing was excluded from scoring on this run.

4.2 Integrity of this record

This record is part of a sealed evidence package. The seal is verifiable by a third party without VERITROOPER and without a network:

Element	Value
Manifest	manifest.json
Manifest SHA-256	fce58bca7cc07d1f2059db18e7f64871a7e4ac9782385a8c7938e8a6c3480285
Files covered by the manifest	32
Digital signature	detached CMS/PKCS#7 (manifest.p7s)
Trusted timestamp	RFC 3161 (timestamp.tsr)
Signer fingerprint	5A:91:5B:25:B2:C3:25:FC:B1:32:97:20:CD:D8:78:E5:B7:C8:A7:76:B8:D2:72:0B:B1:87:89:2E:E3:E2:A5:D0

How to verify. Recompute the SHA-256 of each file listed in Integrity/manifest.json and compare; then verify the detached signature over the manifest with openssl cms -verify -binary -content Integrity/manifest.json -in Integrity/manifest.p7s -inform DER -certfile Integrity/signer_cert.pem and, where present, the timestamp with openssl ts -verify -in Integrity/timestamp.tsr. Integrity/VERIFY.txt in the package repeats these steps. None of this requires VERITROOPER or a network.

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), data governance (Art. 10) — documented elsewhere in the technical file.

Which parts of Article 15 this run supports — and which it does not. Article 15 requires appropriate levels of accuracy, robustness and cybersecurity, and declaration of the relevant accuracy metrics. This record is evidence for the first, partial evidence for the second, and **no evidence at all for the third**:

Article 15 property	Covered here	What was actually done
Accuracy, and declared accuracy metrics	Yes	Measured on the question set in §1–§2 and §4.1, with the metric defined in §3 and the reference baseline specified in §2.2.
Robustness	Partial	Covers inputs the question set contains, including adversarially-shaped and unanswerable items where present. It does not cover distribution shift, degraded retrieval, or malformed or hostile input generally.
Cybersecurity — resilience to attempts to alter use, outputs or performance	No	Not tested. No data-poisoning, model-poisoning, adversarial-example, prompt-injection, model-evasion or confidentiality attack was attempted, and no such resistance is claimed or implied.

Cybersecurity evidence under Article 15(5) must come from elsewhere in the technical file. Nothing in this record should be read as supplying it.

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	2026-06-02 13:48:41
Evaluation finished	2026-06-02 15:16:16
Test items	1000
Evaluation engine version	VERITROOPER v9
Source record (reproducible from stored data)	SEC-10K_gpt-5.5

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. **This record is not yet signed.**

Sign it electronically (recommended). The sign-off is recorded against the SHA-256 fingerprint of this exact result set, so the approval is provably tied to these numbers and any later change to the results visibly invalidates it. A signature on a printout has no such binding — it attests to a sheet of paper.

In the console	open this run and choose Sign off
From the command line	python vt.py signoff "C:\Users\brian\AppData\Local\Temp\claude\D--Veritrooper\3bc0f06d-99e8-4bff-8d4f-56c0ffc80108\scratchpad\rerender2\sec-10k_gpt-5.5_run_record\SEC-10K_gpt-5.5" --name "…" --role "…" --decision approved
Results fingerprint to be bound	ab193965df3f23f3ab63ce78e0b87bc9…

Re-running the record after signing replaces this section with the completed, hash-bound attestation.

Or complete by hand — if your process requires a wet signature, note that this form is bound to nothing but itself:

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