

VERITROOPER Scout Report — Generated — SEC_10K_multi — veritrooper v9



VERITROOPER v9 accuracy & robustness audit. Generated 2026-08-04 15:54:06.

Audit summary

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

Field	Value
Run identifier	VT-00000000-SEC-001 (prefix on the delivered filenames)
System evaluated	Generated — SEC_10K_multi — veritrooper v9
Model under test (MUT)	gpt-5.5
Diagnostician (Doctor)	—
First-pass validator	deterministic rule check (code, not a model)
Independent verifier	Gemini (frontier model)
Questions generated	1000 (1000 scored · 0 bad tests · 0 connection error(s), excluded symmetrically)
Evaluation started	2026-06-02 13:48:41
Evaluation finished	2026-06-02 15:16:16

Audit scope — point-in-time snapshot. This audit is a single point-in-time measurement of **gpt-5.5** answering **1000 purpose-generated questions** drawn from **Generated — SEC_10K_multi — veritrooper v9** (SEC_10K_multi.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 82.6% → pipeline 96.1% (Δ +13.5pp)
Answers scored incorrect	baseline 174 → pipeline 39
Results fingerprint (SHA-256)	ab193965df3f23f3...
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: gpt-5.5 · Scored items: 1000 · Generated: 2026-08-04 15:54:06

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	82.60%	96.10%
Answers scored incorrect	174	39
Improvement		+13.50pp
Failures recovered		86%

What it means

- The VERITROOPER pipeline recovered **86%** of the 174 answers the vanilla-RAG baseline got wrong (82.6% → 96.1% final verified accuracy).
- Weakest area for the raw model: **Negative (hallucination traps)** at 60% baseline (n=200) — the pipeline lifts it to 96%.

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 a cost of 14 answers the baseline got right and the pipeline did not (see the paired-outcome table in the full report); 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	82.60% (826 / 1000)	96.10% (961 / 1000)
Confirmed errors	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.

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	≥ 95%	82.60%	96.10%
Confirmed model errors	= 0	174	39
Unsupported-question accuracy	≥ 95%	60.00%	96.50%
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	812
Recovered (baseline wrong → pipeline correct)	149
Regressed (baseline correct → pipeline wrong)	14
Both wrong	25
Total scored	1000

Of the 163 questions where the arms disagreed, 149 favoured the pipeline and 14 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	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%

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	826	895
Flagged for review	174	105
— cleared by verification (false flags)	0	66
— confirmed model errors	174	39
Final correct / scored	826 / 1000	961 / 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	259
Questions per passage (mean / median)	3.9 / 3
Most / least on a single passage	12 / 1
Passages with only one question	13
Share of questions from the top 5 passages	6%
Adversarial (hallucination-trap) questions	200 of 1000 (20%)

- **Category coverage:** Precision (315), Cross-Reference (243), Negative (hallucination traps) (200), Exception (108), Cause & Effect (78), Conditional (47), Calculation (9).

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: 152 min 12 s.
- Answer calls: 1000 baseline + 1000 pipeline = 2000.
- Compute locality: MUT **gpt-5.5** (hosted API) · Doctor — (local) · Verifier **gemini** (hosted API).

Estimated

- Adjudication calls: ~304 Doctor + ~304 independent-verifier review(s) (inferred from the result records, not directly metered).
- Answer-phase tokens: ~219,808 (chars÷4 over system prompt + question + answer; retrieved-evidence context not stored, so this UNDERSTATES real usage).
- Answer-phase API cost: ~\$2.20 (<\$0.01 per scored question, answer phase only — excludes Doctor / verifier adjudication; at gpt-5.5 list price ≈ \$10/M tokens, input+output blended).

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

When operating with a standard retrieval-augmented generation (RAG) setup, the gpt-5.5 model produced no confirmed errors on the scored questions in this audit. The model correctly answered questions requiring data extraction and calculation from financial documents. It also correctly identified when questions could not be answered from the provided text. While several answers were initially flagged for review, a secondary verification process confirmed that these flags were due to malformed test questions rather than model errors.

Failure Patterns (Baseline)

No confirmed model errors were present in the baseline configuration for this audit panel.

Baseline Remediation Recommendations

As no confirmed model errors were identified, no system-level remediation is recommended based on this audit. Standard best practices, including human oversight for critical applications, remain advisable.

With veritrooper — Pipeline Performance

What the LLM Did With veritrooper

When wrapped by the veritrooper engine, the gpt-5.5 model produced a small number of confirmed errors. These errors were not concentrated in a single category. In one case, the model incorrectly identified a tax provision amount as a year-over-year decrease rather than stating the question was unanswerable. In another instance, it misread the signs on numbers in a table, leading to an incorrect calculation. A third error involved selecting related but incorrect evidence to answer a question about revenue sources. The remaining flagged items were traced to malformed test questions, not model failures.

Pipeline Remediation Recommendations

The few confirmed errors suggest isolated issues rather than a systemic pattern. To address these types of failures, consider the following system-level adjustments:

- * **Prompt Engineering:** Refine instructions to the model to be more explicit about handling negative numbers or values in parentheses within financial tables, ensuring they are interpreted correctly during calculations.
- * **Human Review Routing:** For questions requiring the synthesis of multiple data points or interpretation of complex table structures, consider routing the model's output for human review before final use. This provides a safeguard against subtle misinterpretations of source material.

Independent Verification

An independent verifier audited the diagnostic verdicts for both the baseline and pipeline configurations. This secondary review ensures that the analysis of model behavior is fair and that the classification of errors is applied consistently across both test scenarios.

Bottom Line

In this audit, the baseline configuration of the gpt-5.5 model produced no confirmed errors, while the pipeline configuration introduced a small number of errors. For enterprise use cases involving this specific knowledge base, the baseline performance met a zero-confirmed-error standard, reducing the need for post-generation correction by human reviewers.

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.

Flag — detect and route to review:

- **Source-span verification** — Require the answer's key value to be extractable verbatim from the retrieved passage; flag when it isn't. (Addresses: misreads parentheses sign · 2 cases this run.)
- **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: wrong scope or nearby evidence · 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	gpt-5.5
Evaluated material / knowledge base	Generated — SEC_10K_multi — veritrooper v9
Task type	generation
Evaluation engine	VERITROOPER v9
Test items	1000
Card generated	2026-08-04 15:54:06

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: Cross-Reference, Conditional, Precision, Exception, Negative (hallucination traps), Cause & Effect, Calculation.

Field	Value
Test items	1000
Categories	7
Refusal / unanswerable items (hallucination traps)	200
Source material	data/SEC_10K_multi.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: —) and independently audited by Gemini (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	gpt-5.5	the model being audited	—
First-pass check (Validator)	deterministic rule check	code, not a model	yes — not a model
Diagnosis (Doctor)	—	verification LLM	a separate verification LLM
Independent verification (Verifier)	Gemini (frontier model)	model / vendor	yes — an independent verifier

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

Parameter	Value
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

- No confirmed model-behaviour failure patterns were recorded on this material.

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	2026-06-02 13:48:41
Run finished	2026-06-02 15:16:16
Sampling temperature	0.1
Max tokens	200
Results content hash (SHA-256)	ab193965df3f23f3...

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 (2026-06-02 15:16:16)
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/SEC_10K_multi.txt

Items per category:

Question type	Items
Precision	315
Cross-Reference	243
Negative (hallucination traps)	200
Exception	108
Cause & Effect	78
Conditional	47
Calculation	9

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/SEC_10K_multi_MERGED.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-00000000-SEC-001_Results_Report.pdf · VT-00000000-SEC-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 where sealing succeeded (with an RFC 3161 trusted timestamp where a network was available at sealing).	VT-00000000-SEC-001_Results_Report.pdf · VT-00000000-SEC-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-00000000-SEC-001_Results_Report.pdf · VT-00000000-SEC-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-00000000-SEC-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-00000000-SEC-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-00000000-SEC-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-00000000-SEC-001_System_Card.pdf · VT-00000000-SEC-001_Results_Report.pdf · Integrity/

NIST AI RMF 1.0

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-00000000-SEC-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-00000000-SEC-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-00000000-SEC-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-00000000-SEC-001_System_Card.pdf · VT-00000000-SEC-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-00000000-SEC-001_Results_Report.pdf · VT-00000000-SEC-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-00000000-SEC-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-00000000-SEC-001_System_Card.pdf · VT-00000000-SEC-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: **82.60%** (achievable with grounding: 96.10%). **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 82.60% 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	60.00%	96.50%	+36.50pp	-22.60pp	GAP (adversarial)
Cross-Reference	243	71.19%	88.07%	+16.88pp	-11.41pp	GAP
Calculation	9	88.89%	88.89%	+0.00pp	+6.29pp	—
Precision	315	93.33%	99.68%	+6.35pp	+10.73pp	Recovered
Cause & Effect	78	98.72%	100.00%	+1.28pp	+16.12pp	Recovered
Exception	108	99.07%	99.07%	+0.00pp	+16.47pp	—
Conditional	47	100.00%	100.00%	+0.00pp	+17.40pp	—

Performance dispersion

- Strongest: **Conditional** at 100.00%
- Weakest: **Negative (hallucination traps)** at 60.00%
- Spread (range across categories): **40.00pp**

Coverage / capability gaps: Cross-Reference. A gap on a straightforward-retrieval category can point to uneven capability or to under-representation of that content in the underlying data — the kind of shortcoming the Art. 10 review examines.

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	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.

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.

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.

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

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	80308906353c544aa350b4169ae4d068cc42a709d8955d21e9e3ea22fee65984
Files covered by the manifest	31
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.

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.

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

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\render2\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: _____

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	gpt-5.5	endpoint locality not captured — confirm on your network
Verification model (Doctor)	—	endpoint locality not captured — confirm on your network
Independent verifier	Gemini (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.