

VERITROOPER Scout Report — SEC 10K multi — VERITROOPER v10

SEC 10K multi — VERITROOPER v10
26 August 2026



VERITROOPER v10 accuracy & robustness audit. Generated 2026-08-26 17:30:00.

Audit summary

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

Field	Value
Run identifier	VT-20260826-SEC-001 (prefix on the delivered filenames)
System evaluated	SEC 10K multi — VERITROOPER v10
Model under test (MUT)	gpt-5.6-sol
Diagnostic reviewer	gpt-5.6-sol *(same model as the MUT)*
First-pass validator	deterministic rule check (code, not a model)
Independent verifier	gemini (Gemini)
Questions generated	889 (889 scored · 0 bad tests · 0 connection error(s), excluded symmetrically)
Question set generated	2026-08-26 15:24:46
Evaluation started	2026-08-26 15:27:11
Evaluation finished	2026-08-26 17:29:18

Audit scope — point-in-time snapshot. This audit is a single point-in-time measurement of **gpt-5.6-sol** answering **889 purpose-generated questions** drawn from **SEC 10K multi — VERITROOPER v10** (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 data set × this configuration at this point in time**. It does **not** measure other deployments, later model or prompt versions, a different data set, 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	889 (of 889; 0 bad tests + 0 connection error(s) excluded symmetrically)
Final verified accuracy	baseline 72.89% → pipeline 88.08% (Δ +15.19pp)
Answers scored incorrect	baseline 241 → pipeline 106
Results fingerprint (SHA-256)	f336a4ee6c66bbee...
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. Data Handling & Isolation Posture

(Use the PDF bookmarks / outline panel to jump to any section.)

1. Executive Summary

Model under test: gpt-5.6-sol · **Scored items:** 889 · **Generated:** 2026-08-26 17:30:00
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	72.89%	88.08%
Answers scored incorrect	241	106
Improvement		+15.19pp
Baseline errors recovered		156 of 241 (65%)
New regressions (baseline correct, pipeline wrong)		21

What it means

- The VERITROOPER pipeline recovered **156 of 241** answers the vanilla-RAG baseline got wrong (65% baseline-error recovery). It also introduced **21 regressions** where the baseline had been correct; net final verified accuracy was 72.9% → 88.1%.
- Weakest area for the raw model: **Negative (hallucination traps)** at 60% baseline (n=200) — the pipeline lifts it to 87%.
- The most common baseline failure mode was **other** (58 of 194 confirmed errors) — the class of error the pipeline configuration is designed to eliminate.

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 21 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=889); 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: **889 questions** (of 889 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	72.89% (648 / 889)	88.08% (783 / 889)
Answers scored incorrect	241	106
Δ vs baseline		+15.19pp
Failure-Recovery Rate (pipeline recovers what % of baseline failures)		64.7%
95% confidence interval (Wilson)	69.9–75.7%	85.8–90.0%

The 95% confidence interval reflects sampling uncertainty on 889 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)	FAIL

Acceptance criteria & results:

Criterion	Threshold	Baseline	Pipeline
Final verified accuracy	≥ 95%	72.89% FAIL	88.08% FAIL
Answers scored incorrect	= 0	241 FAIL	106 FAIL
Unsupported-question accuracy	≥ 95%	60.00% FAIL	87.00% FAIL
Scored sample size	≥ 30	889 PASS	889 PASS
Connection errors (execution failures)	≤ 2	0 PASS	0 PASS

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	627
Recovered (baseline wrong → pipeline correct)	156
Regressed (baseline correct → pipeline wrong)	21
Both wrong	85
Total scored	889

Of the 177 questions where the arms disagreed, 156 favoured the pipeline and 21 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%	87.00%	+27.00pp	76.2%
Precision	235	65.11%	87.66%	+22.55pp	72.0%
Cross-Reference	181	61.88%	76.80%	+14.92pp	46.4%
Conditional	64	92.19%	96.88%	+4.69pp	60.0%
Cause & Effect	112	98.21%	99.11%	+0.89pp	50.0%
Exception	82	100.00%	98.78%	-1.22pp	—
Calculation	15	80.00%	66.67%	-13.33pp	0.0%
All Categories	889	72.89%	88.08%	+15.19pp	64.7%

Adjudication reconciliation

What each arm's automated sweep found, and what review then made of it. "Cleared by verification" are answers the validator flagged but the diagnostic reviewer judged correct — this is reported so you can see where judgement disagreed with the code, and those answers **still count against the model** in the final accuracy, because the deterministic check decides the score. 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	603	750
Flagged for review	286	139
— cleared by verification (false flags)	45	33
— confirmed model errors	241	106
Final correct / scored	648 / 889	783 / 889
Bad tests excluded (symmetric)	0	0
Connection errors excluded (no answer returned)	0	0
Total generated	889	889

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	379
Questions per passage (mean / median)	2.3 / 2
Most / least on a single passage	10 / 1
Passages with only one question	51
Share of questions from the top 5 passages	5%
Adversarial (hallucination-trap) questions	200 of 889 (22%)

- **Category coverage:** Precision (235), Negative (hallucination traps) (200), Cross-Reference (181), Cause & Effect (112), Exception (82), Conditional (64), Calculation (15).

Passage-level counts above are exact. Whole-data-set coverage (% of ALL source passages represented) is not computed here — the data set's total passage count is not recorded in this run; it requires the source 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: 76 min 3 s.
- Answer calls: 889 baseline + 889 pipeline = 1778.
- Compute locality: MUT **gpt-5.6-sol** (hosted API) · Diagnostic reviewer **gpt-5.6-sol** (hosted API) · Verifier **gemini** (hosted API).

Estimated

- Adjudication calls: ~544 diagnostic reviewer + ~423 independent-verifier review(s) (inferred from the result records, not directly metered).
- Answer-phase tokens: ~2,638,218 (chars÷4 over system prompt + retrieved evidence + question + answer).
- Answer-phase API cost: ~\$26.38 (~\$0.03 per scored question, answer phase only — excludes diagnostic-reviewer / verifier adjudication; at gpt-5.6-sol 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 diagnostic-reviewer / 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

With vanilla RAG, the tested LLM answered most questions correctly, but a meaningful minority of answers contained confirmed errors. The failures included unsupported answers to questions that should have been treated as unanswerable, incorrect extraction of values from tables, conflicts with the required numeric answer, and responses that added information beyond the requested result. Malformed questions were excluded from the assessment rather than treated as model mistakes.

Failure Patterns (Baseline)

- **Answered an unanswerable question:** The model supplied a calculated answer even though the available evidence did not support the requested conclusion.

- **Table misread:** The model selected the wrong value from a financial table, indicating difficulty aligning the requested period, row, or column with the relevant figure.
- **Numeric answer conflict:** The model returned a value that did not match the required value, including cases involving totals and financial measures.
- **Computed versus quoted value confusion:** The model mixed calculated, quoted, pre-tax, after-tax, or signed values rather than returning the specific measure requested.
- **Scope and relevance errors:** Some responses included unsupported interpretation or unnecessary contextual material instead of limiting the answer to the requested fact.
- **Evidence-use failure:** In some cases, the response said the evidence did not specify an answer even though the required information was available in the source material.

Baseline Remediation Recommendations

- Strengthen the system prompt so the model must distinguish clearly between values quoted directly from evidence and values it calculates.
- Require the model to return "UNANSWERABLE" when the requested answer cannot be established from the retrieved material, rather than inferring or reconstructing a result.
- Require explicit row, column, period, unit, and label alignment before the model answers questions based on financial tables.
- Add an application-level numeric consistency check that compares the proposed answer with the cited source value and verifies units, signs, tax treatment, and reporting period.
- Require concise answers that address only the requested fact unless explanation is explicitly requested.
- Determine whether each evidence-use failure resulted from a retrieval miss or from the model not using retrieved evidence. Adjust retrieval only where the relevant passage was not supplied.
- Route unsupported calculations, ambiguous table lookups, and material financial totals to human review before release.
- Use targeted prompt examples and regression tests covering unanswerable questions, table extraction, period selection, and computed-versus-quoted values.

With VERITROOPER — Pipeline Performance

What the LLM Did With VERITROOPER

When wrapped by VERITROOPER, the tested LLM had a small minority of confirmed residual errors, and the broader baseline patterns were not present as confirmed model failures in this audit panel. The remaining confirmed issue was a numeric conflict involving the total value of asset-backed securities. Other residual cases were malformed or excluded tests and therefore were not treated as model errors.

Pipeline Remediation Recommendations

- Configure the model to show the source components used for financial totals before producing the final answer.
- Add an application-level reconciliation step that checks whether an aggregate matches the relevant table entries, reporting period, and units.
- Require a citation to the exact table row or disclosure supporting any material total.
- Route unresolved aggregate-value conflicts to human review and sign-off.
- Maintain a focused regression test for asset-backed securities totals and similar table-based aggregation questions.

Independent Verification

An independent verifier audited the diagnostic decisions for both the pipeline and baseline arms under the same standard. This symmetric review supports an honest comparison by applying equivalent scrutiny to answers from both configurations.

Bottom Line

For the tested panel, the VERITROOPER-wrapped configuration is the stronger deployment choice because it left fewer answers requiring post-generation correction and the baseline failure patterns were largely absent from the pipeline arm. Human review should remain in place for material financial totals and unresolved table-based calculations, particularly where an incorrect aggregate could affect a business decision.

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:

- **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: Misinterpreted, Out of scope, Table misread, Unit/format mismatch, Wrong calc/quote, Wrong period · 74 cases this run.)

- **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: Meta leak, Misinterpreted, Out of scope, Table misread, Unit/format mismatch, Wrong period · 18 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: Math error, Misinterpreted, Out of scope, Unit/format mismatch, 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: Misinterpreted, Out of scope, Table misread, Unit/format mismatch · 8 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 · 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: Out of scope, Table misread, Wrong entity · 5 cases this run.)

- **Direction/polarity check** — Compare the answer's polarity and direction (yes/no, increase/decrease, permitted/prohibited) against the source statement; flag reversals. (Addresses: deterministic:direction conflict · 4 cases this run.)

- **Evidence-coverage check** — Before answering, verify the retrieved context actually contains a span that supports the question; when coverage is below a set threshold, route to broader / multi-query retrieval or abstain rather than answering from a thin context. (Addresses: Table misread · 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.)

- **Entity-match check** — Require the answer to reference the same entity named in the question (exact or alias match); flag cross-entity substitutions. (Addresses: Out of scope · 1 case this run.)

- **Abstention guard** — When evidence coverage is below a set threshold, require the model to abstain or hedge; flag confident answers made over weak or absent evidence. (Addresses: Certainty mismatch · 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-5.6-sol
Evaluated material / knowledge base	SEC 10K multi — VERITROOPER v10
Task type	generation
Evaluation engine	VERITROOPER v10
Test items	889
Card generated	2026-08-26 17:30:00

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

Field	Value
Test items	889
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: gpt-5.6-sol) and independently audited by gemini (Gemini), 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-5.6-sol	the model being audited	—
First-pass check (Validator)	deterministic rule check	code, not a model	yes — not a model
Diagnosis (diagnostic reviewer)	gpt-5.6-sol	verification LLM	no — the same model as the MUT
Independent verification (Verifier)	gemini (Gemini)	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.

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

Failure modes across the 194 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 — 58 answers
- deterministic:answered unanswerable — 52 answers
- deterministic:gt number conflict — 18 answers
- knowledge gap — 18 answers
- misinterpretation — 13 answers
- table misread — 11 answers

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 v10
Run started	2026-08-26 15:27:11
Run finished	2026-08-26 17:29:18
Sampling temperature	0.1
Max tokens	1024
Results content hash (SHA-256)	f336a4ee6c66bbee...

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.

Limit of this evidence: replayability and repeated-run consistency show that the record and procedure can be checked; they do not by themselves establish broad validity. The questions are generated from the supplied material, scored denominators can differ when malformed items or connection failures are excluded, and configured model roles can introduce model-to-model variation. Generalization beyond the tested question set requires separate representative validation.

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 (2026-08-26 17:29:18)
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	889
Categories	7
Refusal / unanswerable (hallucination-trap) items	200
Source material	data/SEC_10K_multi.txt

Items per category:

Question type	Items
Precision	235
Negative (hallucination traps)	200
Cross-Reference	181
Cause & Effect	112
Exception	82
Conditional	64
Calculation	15

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 v10
Generated data file	data/SEC_10K_multi_generated_20260826_152446.json
Generation temperature	0.1
Sampling strategy	full
Use limitation	Generated by VERITROOPER as a data card documenting the evaluation test set. 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.

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; AI Act application dates re-verified 2026-08-04). 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 on AI (**Regulation (EU) 2026/1744**, in force 27 July 2026) 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-20260826-SEC-001_Results_Report.pdf · VT-20260826-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-20260826-SEC-001_Results_Report.pdf · VT-20260826-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-20260826-SEC-001_Results_Report.pdf · VT-20260826-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-20260826-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-20260826-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-20260826-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-20260826-SEC-001_System_Card.pdf · VT-20260826-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-20260826-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-20260826-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-20260826-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-20260826-SEC-001_System_Card.pdf · VT-20260826-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-20260826-SEC-001_Results_Report.pdf · VT-20260826-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-20260826-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-20260826-SEC-001_System_Card.pdf · VT-20260826-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. 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-5.6-sol	endpoint locality not captured — confirm on your network
Verification model (diagnostic reviewer)	gpt-5.6-sol	endpoint locality not captured — confirm on your network
Independent verifier	gemini (Gemini)	endpoint locality not captured — confirm on your network

Cross-vendor verification: **on** for this run.

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.