VERITROOPER · DEPLOY ACCURACY ANYWHERE

Technical Proof Packet

How VERITROOPER audits any LLM running on any written data. Patent-pending architecture — USPTO Application No. 19/685,794, filed 22 May 2026.

What it is. VERITROOPER is an **audit and diagnostic layer** for any language model running on the buyer's own data — not a model, not a serving wrapper, not a guardrail. It generates a typed test set from raw source material, runs the model against ground-truth answers extracted from the same material, settles clear verdicts with deterministic code, escalates contested verdicts to an independent cross-vendor frontier verifier, and emits a reproducible per-question evidence trail plus an Annex IV-shaped artifact (opt-in) suitable for diligence and regulatory review.

1 Architecture — six stages

Deterministic where deterministic answers exist; the language model is restricted to language work.

Autonomous question generation

Typed questions — including hallucination-trap items — generated from semantically chunked source data and gated for evidence-grounding. No human-curated test set; no reuse of public benchmark questions on customer surfaces.

Deterministic per-question pipeline ENFORCED

The LLM writes prose only. Arithmetic, evidence parsing, and numeric bounds are handled by deterministic code; the validator stage applies a fixed sequence of deterministic checks before any judgment step runs.

Doctor — first-pass diagnosis

Reviews flagged cases and reclassifies. The engine ensures the diagnostic step is never weaker than the model under review. The Doctor proposes; it does not get the final word.

Independent cross-vendor verifier — final say on contested verdicts CODE-ENFORCED

Reviews only the flagged subset, not every answer. **When the model under test is itself a frontier model, vendor rotation is enforced in code** — the verifier is mechanically guaranteed to be from a different vendor than the model under test. For non-frontier models the verifier is configurable; the engine permits an air-gapped operating mode, with that trade clearly disclosed.

Symmetric production-style retrieval baseline arm

The same questions run through a standard production-style retrieval arm, scored by the same Doctor and Verifier at identical strictness. The reported delta cannot be manufactured by grading the baseline

harder.

Reporter

Emits a per-failure narrative with model-specific training and retrieval-tuning recommendations, plus reproducible certification numbers and an opt-in Annex IV-shaped artifact (positioned as conformity-supporting evidence, never as automatic compliance).

2 Why the headline number is trustworthy

The scoring rules consistently round against the product, not for it.

Nothing confirms itself

The model under test never gets the final word on its own answers. For frontier MUTs the verifier's vendor is mechanically rotated by the engine; for smaller MUTs the verifier is a frontier model. Proposing and confirming are never done by the same vendor.

Uncertainty counts against the system

No-answer, hedge, wrong polarity, and refusal-to-commit are all scored as failures. The reported number is a conservative floor of what survives independent review, not a best-case.

bad_test handling is visible

Questions whose auto-generated ground truth is itself malformed are flagged `bad_test`, removed symmetrically from both arms' denominators, and left visible in the report — separating test-generation failure from model failure rather than hiding it in the denominator.

Reproducible from records AUDIT-GRADE

Each question's record captures the verdict chain and the supporting evidence. An outside auditor re-derives the headline number from stored data without re-running the model.

Fair comparison

Baseline and pipeline arms face the same Doctor and Verifier at the same strictness. The before-and-after is measured on the identical question set and scored by the same independent standard.

3 Representative current-engine result

IRS Tax Code (Title 26), 1,002 generated questions, identical question set across every model under test, audit-grade scoring with bad_test removal and borderline-as-failure discipline.

MODEL UNDER TEST	BASELINE (PRODUCTION-STYLE RETRIEVAL)	AUDITED (EVIDENCE-GROUNDED)	RECOVERED
Claude Opus 4.8	94.36%	100.00%	**+5.64pp**
GPT-5.5	90.78%	99.70%	**+8.92pp**
Gemini 2.5 Pro	93.99%	99.40%	**+5.41pp**
Qwen 2.5 72B	86.76%	98.19%	**+11.43pp**
Llama 3.1 70B	87.69%	97.70%	**+10.01pp**
Gemma 3 27B	92.57%	97.89%	**+5.32pp**
Qwen 2.5 7B (local)	86.49%	95.90%	**+9.41pp**

95.9% – 100% audited band across all seven models tested

From a 7B local model through three flagship frontier vendors, the audited accuracy lands in a narrow band. The architecture compresses a wide model-quality spread into a tight, regulated-deployment-fit band.

Pipeline value is inversely proportional to model quality

Weaker models gain more. The lift is largest where it is needed most — a 7B model recovers nearly twice the percentage points that a frontier model does, on the same questions.

Cross-generation proof — the architecture audits models that didn't exist when it was built

GPT-5.5 (released this month) was audited on the same 1,002 questions, unmodified — standalone 93.04%, audited **99.70%**, landing at the identical 99.70% ceiling as GPT-5.4 (which started lower at 90.78%). The architectural ceiling is model-generation-independent.

Negative-category (hallucination-trap) recovery

On a separate Air Force regulatory audit (AFSC 2A6X6, 487 questions, GPT-5.4 MUT) the negative category alone moved from 79.4% standalone to 99.0% audited — the overreach-on-unanswerable failure mode an internal assistant exhibits in the field.

The accuracy your model is leaving on the table

The baseline is production-style retrieval; the audited figure is the same model reasoning over the correct source evidence. The gap between them is recoverable accuracy — and the per-failure diagnostic traces every point of it to a specific question and root cause your team can close, through better retrieval, prompt rules, or targeted fine-tuning. The boundary is in the definitions, not a disclaimer: you capture the delta by acting on the fixes the run surfaces, not by deployment alone.

4 Reproducibility — independent-infrastructure evidence

Ten independent audits of a frozen 238-question UCMJ panel, run across two physically independent A100 servers (Qwen3-30B-A3B under test, Gemini verifier, Claude Opus 4.7 reporter).

QUANTITY	OBSERVED ACROSS 10 RUNS
Pipeline-arm sweep raw pass rate	**97.48% in 10 of 10 runs (range 0.00pp)**
Pipeline accuracy (bad_test-excluded headline)	97.03% – 97.48% (range 0.45pp, stdev 0.196pp)
Statistical 95% CI half-width at $p \approx 0.97$, N ≈ 238	± ~2.15pp
Observed rerun envelope as a fraction of metric's own CI half-width	**~21%**

Reruns are indistinguishable within the metric's own statistical precision

The observed pipeline-accuracy envelope across ten reruns (0.45pp) is approximately one fifth of the 95% confidence half-width the metric carries at this sample size. The MUT pipeline arm produced bit-identical pass/fail outcomes (232 of 238 passes, 97.48%) in every one of the ten runs.

The deterministic-validator layer is doing the stabilizing work

Where the LLM produces marginally different correct phrasings across runs, the validator stage recognizes them as equivalent and produces stable verdicts. The unbuffered production-style baseline arm shows the small underlying LLM stochasticity directly — making the architecture's buffer role visible in the data.

Full evidence trail preserved for third-party re-derivation

Each of the ten runs ships a full per-question record, certification report, configuration snapshot, sample manifest, run log, and rendered narrative reports. No reported number depends on re-running the model to verify.

5 Footprint — small, portable, vendor-neutral

Compact engine, commodity hardware, on-prem and air-gapped deployable.

Lightweight by design

A compact Python codebase with no heavy ML-framework dependency at audit time. Built to run on commodity hardware without specialized infrastructure.

Runs locally on a single consumer GPU

Validated end-to-end on an RTX 4090 (24 GB VRAM) with local open-weight models; a single cloud A100 supports 70B–72B; the frontier-API path requires no GPU at all.

Model-agnostic by construction

The language model is accessed through a standard chat-completion API. Replacing one model with another requires changing one URL. Every improvement made by any vendor automatically improves VERITROOPER's reach.

Deployable on-prem and air-gapped

Relevant for sovereign, defense, healthcare, and financial-services buyers whose data cannot egress to a cloud SaaS audit instrument.

6 What the buyer gets — a complete deliverable package from a single run

A score an auditor can reproduce, an engineering action plan a team can execute, and — on request — the evidence binder a regulator can review. One run produces all of it.

A reproducible accuracy measurement — and the full scoreboard behind it

Not just a single number: audited accuracy against ground truth, the production-style-retrieval baseline beside it, the recovered delta, a per-category accuracy breakdown, and the failure-recovery rate — the share of the model's real-world failures the audit caught and closed. Every figure reconstructs from the per-question evidence records, so an auditor or counsel can reproduce any cited number without re-running the model.

A per-failure engineering report — that is also a labeled training-target list TWO-FOR-ONE

For every wrong answer: the question, the model's answer, the correct answer, the supporting source evidence, the failure category, and a plain-English fix. Failures are clustered into recurring patterns and split by root cause — what a retrieval or prompt change can recover versus where the model itself hits its ceiling — so the team knows exactly what to change and what it will buy them. By construction, that same output is a ready-made fine-tuning / retrieval-tuning target list: a curated set of the model's real failures, each paired with the correct, sourced answer. One run gates the model *and* hands over the data to improve it.

EU AI Act conformity-supporting evidence (opt-in)

Testing record, post-market drift report, gap diagnostic, and a tamper-evident sign-off audit trail — generated from the same run. Positioned and labeled as evidence, never as automatic compliance. Off by default; one toggle at run start.

Patent. VERITROOPER's technical architecture is the subject of U.S. patent application No. **19/685,794**, filed **22 May 2026** (priority date). Utility nonprovisional, sole inventor Brian L. Barbour. Patent counsel of record on file.

Further material on request. Full whitepaper, technical supplement, and a current sample-run output package available under standard pre-engagement confidentiality. Inquiries via **veritrooper.com/contact**.