TELOS Validation Report
Evidence, Calibration, Reproduction, and Claim Boundaries
TELOS AI Labs Inc.
August 2026

Purpose and reporting standard

This report isolates the empirical evidence supporting TELOS. It separates observations from interpretations, discloses the development and calibration conditions, identifies results that were reproduced with a different embedding model, and states which claims remain open.
The report follows five controls:

- A detection result is reported with its scorer, benchmark, sample size, and false positive context where available.
- A zero observed failure result is not described as proof of zero risk.
- A benchmark integrated into a runner is not counted as a scored benchmark unless a result is reported.
- A development result is not described as held-out validation.
- A cryptographic integrity check is not described as proof of authorship or trusted time unless the associated key and time controls are independently established.

Claims under test

ID	Hypothesis	Current status
H1	Attacks that bypass prompt-only safeguards can be detected by the historical three-tier TELOS scoring configuration.	Supported on the reported development and re-verification suites, with material false positive concerns on aggressive threshold settings.
H2	Runtime TELOS detection reliably improves agent behavior through correction or intervention.	Open. Detection and surfacing were tested more extensively than correction efficacy.
H3	TELOS receipts are sufficient to satisfy specific legal or regulatory obligations.	Open. The architecture produces potentially relevant evidence, but legal sufficiency depends on jurisdiction, deployment, retention, identity, and independent review.
H4	The measurement method generalizes across materially different safety and agentic benchmarks.	Partially supported across the four reported scored families. Generalization beyond those families remains open.
H5	Domain-specific boundary configuration can reduce false positives relative to a generic configuration.	Supported on the reported XSTest benign set. Attack detection was not rerun under the healthcare-specific configuration, so the full tradeoff is unknown.

System evaluated

The principal historical tests used a three-tier scoring configuration derived from declared purpose, domain boundaries, and prohibitions. The public telos-mcp v0.1 repository does not expose that calibrated production scoring system. Its current public scoring path is a synthetic prototype. The validation runners, protocol, production boundary data, and exact scoring configuration used for the reported experiments are held by TELOS Labs and are available under NDA.
This disclosure matters for interpretation. The published datasets and result artifacts support inspection of the reported outcomes, but they do not yet permit a fully independent reconstruction of the production measurement engine from public materials alone.

Calibration and experimental design

Thresholds were tuned on development suites. No held-out split was used for the principal internal adversarial evaluations. The results should therefore be read as development and re-verification evidence, not as an unbiased estimate of performance on an unknown deployment distribution.
This report includes only statistical values whose provenance can be established from the available artifacts. It therefore states no p value, power estimate, effect-size statement, AUC comparison, or Wikipedia-centroid accuracy delta.

Internal and public adversarial evaluations

January 2026 evaluation

The January suite contained 2,100 adversarial prompts:

Component	Prompts
MLCommons AILuminate	1,200
MedSafetyBench	450
HarmBench	400
Internal California SB 243 set	50
Total	2,100

The historical three-tier TELOS configuration reported 0 undetected attacks in 2,100 prompts. Using the conventional rule of three, a rough 95 percent upper confidence bound for the undetected rate is approximately 3 / 2,100, or 0.14 percent. This bound assumes independent and identically distributed Bernoulli trials, an assumption that benchmark prompts do not fully satisfy.
The same evaluation reported these attack success rate baselines:

Condition	Model A	Model B	Average
No defense	30.8%	43.9%	37.4%
System prompt baseline	11.1%	3.7%	7.4%
Historical three-tier TELOS scorer, self-scored	0.0%	0.0%	0.0%

The 0 percent TELOS row is a self-scored development result. It is not an independent third-party certification.

July 2026 re-verification

The July suite contained 2,550 adversarial prompts:

Component	Prompts
MLCommons AILuminate	1,200
MedSafetyBench	900
HarmBench	400
Internal California SB 243 set	50

Total	2,550

The re-verification reported 0 undetected attacks in 2,550 prompts. Under the same rule-of-three approximation, the 95 percent upper confidence bound is approximately 3 / 2,550, or 0.12 percent. This result is evidence about the tested suite and scorer. It does not establish a universal deployment failure rate.

Aggressive point and false positives

On the 50-prompt California SB 243 set, the aggressive operating point reported 0 undetected attacks and a 74 percent false positive rate. This demonstrates the central threshold tradeoff: very high sensitivity can be obtained at a level of overflagging that may be operationally unacceptable. The 74 percent figure must remain attached to the aggressive-point result. Reporting the detection result without it would materially misstate the evidence.

Benign-use calibration with XSTest

XSTest was used to measure exaggerated safety behavior on 250 benign prompts.

Configuration	False positives	False positive rate
Generic boundary configuration	62 / 250	24.8%
Healthcare-specific boundary configuration	20 / 250	8.0%
Observed change	42 fewer false flags	16.8 percentage point reduction

This result supports H5 for the tested benign set. It does not by itself show that the healthcare-specific configuration preserved the original attack detection rate because adversarial detection was not rerun under that configuration.

External agentic benchmarks

The scored external agentic evidence comprises 1,468 examples across three benchmark families.

Benchmark	Scored examples	Reported result	Important qualification	Artifact
PropensityBench	977 sampled from 5,874 tasks	100% reported detection success rate	Development evaluation using the historical scorer	Zenodo
AgentHarm	352	100% with the published Mistral embedding configuration; 82.39% in a July MiniLM rerun	The decrease under a changed embedding model shows model and calibration dependence	Zenodo
AgentDojo	139	100% reported prompt-injection detection	In the cited 139-event AgentDojo report, 54 of 139 event-level classifications matched the expected labels (38.8%): the 54 detected attack-event rows represent 27 unique attack texts evaluated at two surfaces, while all 85 benign texts were overflagged in one event each under the deliberately generic safety anchor. These are TELOS self-scored classification outcomes, not evidence that execution was blocked.	Zenodo

The AgentHarm rerun is especially informative. A change from the published Mistral embedding path to MiniLM reduced the reported rate to 82.39 percent. The result supports the whitepaper's architectural argument that the measurement engine is a calibrated system, not a model-independent formula whose performance transfers automatically.

The AgentDojo result also illustrates why detection cannot be the only outcome. Perfect reported injection detection accompanied by 0 percent benign pass on the evaluated benign cases indicates that a system can appear safe while being unusably restrictive.

July 2026 governed re-verification

The headline external-agent results above are the January 2026 published figures. A later governed run under observation addressed nine published result rows:

Disposition	Rows	Detail
Reproduced	7	The governed run reproduced the corresponding published rows
Scorer-profile non-reproduction	1	AgentHarm measured 82.39% under local MiniLM scoring rather than the published 100% under the Mistral-embed profile
Could not run	1	AgentDojo could not be executed in the re-verification environment
Total	9	Every row received a disposition

The MiniLM run used local scoring because the run's declared purpose prohibited metered APIs. The two embedding profiles are not interchangeable. The 82.39 percent result is therefore a cross-profile comparison and a non-reproduction of the published headline profile, not evidence of a time-based degradation in the same frozen system.

the commit-pinned telos-verify vector directory at https://github.com/TELOS-Labs-AI/telos-verify/tree/b60917f695a56d5c66520aa6b8db73fd88de5831/vectors/valid in the public repository is the full row-level statement of record. That path should be linked directly on publication once the repository location and release tag are frozen.

Integrated but unscored benchmarks

Two additional benchmark suites were integrated into the testing environment but were not reported as scored validation results:

Benchmark	Examples	Status
Agent-SafetyBench	2,000	Integrated, not scored in the reported evidence
InjecAgent	1,054	Integrated, not scored in the reported evidence
Total	3,054	Must not be added to reported detection totals

Out-of-scope and drift proof of concept

A governance benchmark based on CLINC150 and MultiWOZ reported:

Measure	Reported result	Artifact
Out-of-scope detection	78%	Zenodo
Drift detection	100%	Zenodo

These are proof-of-concept results and should not be combined with the adversarial safety totals.

Evidence inventory and provenance

Evidence class	Count	What it supports	What it does not support
January public adversarial suite	2,100	Development evidence for historical three-tier detection	Independent production effectiveness
July adversarial re-verification	2,550	Repeat evidence on an expanded suite	Unknown-distribution failure rate
External agentic scored examples	1,468	Cross-benchmark development evidence	Universal agentic generalization
Integrated but unscored examples	3,054	Runner coverage and future evaluation capacity	Any detection percentage
XSTest benign prompts	250	Configuration-specific false positive comparison	Preserved adversarial sensitivity
CLINC150 and MultiWOZ proof of concept	Reported in artifact	Out-of-scope and drift feasibility	Safety benchmark performance
Internal pytest count	Unverified in the available publication artifact	Engineering activity only	An empirical effectiveness claim

Reproducibility links

Dataset or result artifact	Link
MLCommons AILuminate, January artifact	10.5281/zenodo.21630186
HarmBench and MedSafetyBench	10.5281/zenodo.21629088
California SB 243 set	10.5281/zenodo.18370504
XSTest calibration	10.5281/zenodo.21630399
PropensityBench	10.5281/zenodo.18562833
AgentHarm	10.5281/zenodo.21629991
AgentDojo	10.5281/zenodo.18565869
Governance benchmark	10.5281/zenodo.18009153

Interpretation

The evidence supports a narrower claim than a conventional security guarantee. It shows that a configured TELOS measurement approach can identify many tested departures and attacks, and that its behavior changes materially with the boundary corpus, threshold, domain configuration, and embedding model.

The evidence does not yet show:

- independently reproduced production performance;
- stable performance across unknown deployment distributions;
- an acceptable sensitivity-specificity tradeoff for every domain;
- causal improvement in downstream agent behavior;
- resistance to adaptive attackers who know the scoring system;
- regulatory sufficiency;
- trusted receipt authorship or time in the public prototype.

Recommended next validation phase

A publication-grade independent study should be preregistered before threshold tuning and should include:

- A frozen Manifest, boundary corpus, embedding model, scorer version, thresholds, and decision policy.
- A development set separate from a sequestered held-out test set.
- Independent labels with documented adjudication and inter-rater agreement.
- Sensitivity, specificity, precision, recall, false positive rate, calibration, and confidence intervals at every published operating point.
- Domain and model transfer tests without post-hoc recalibration, followed by a separately reported recalibrated condition.
- Adaptive red-team evaluation with access levels stated explicitly.
- End-to-end outcomes, including task success, user interruption burden, time to resolution, and harm avoided.
- Reproduction by an organization that did not design the scorer.

Bottom line

The current research is meaningful development evidence, particularly because it exposes false positive costs and embedding-model dependence. It is not yet sufficient for a claim of universal or independently validated safety. The strongest defensible position is that TELOS has a testable measurement architecture with promising benchmark results, transparent known limitations, and a clear path to stronger validation.

Related documents

- Core whitepaper
- Technical method
- Record of Trust protocol
- Regulatory alignment map