Guardbench Report No. 1

Common-Condition Evaluation of Prompt-Injection Guard Configurations

REPORT NO.	GB-2026-001
VERSION	0.4
REPORT DATE	17 September 2026
TEST WINDOW	15–16 September 2026
CORPUS	Guardbench corpus v0 — 5,995 items
SHA-256	`252eb5671b878876d28f107da09b70b2891ed5a53b7fc33600ab60d86981c79c`
PREPARED BY	Proving Office, a testing program of SIR Holdings LLC
FUNDING OF THIS REPORT	Internally funded research. No buyer or tested vendor commissioned or paid for this report.
REVIEW STATUS	Internal adversarial review complete; not externally peer reviewed.

Proving Office · A program of SIR Holdings LLC · desk@provingoffice.com · provingoffice.com

1. Summary of Findings

Five prompt-injection guard configurations from four vendors were scored on one frozen corpus under a common calibration rule and analysis procedure. Four configurations completed all 5,995 items; Lakera Guard's `prompt_attack` detector completed 308 items and is reported separately as preliminary.

For each complete configuration, the decision threshold was chosen on 1,784 calibration-benign items to meet a stated false-positive target. The false-positive rate actually achieved was then measured on a separate 1,711-item benign evaluation half. Attack items were never used to set thresholds.

At the **1% false-positive target**, attack detection on the four complete runs ranged from 0.335 to 0.742. Meta Prompt Guard 2 86M, StackOne Defender 0.8.2, and ProtectAI DeBERTa-v3 prompt-injection v2 were not separated by the paired, cluster-aware bootstrap: every pairwise interval among those three included zero.

At the **5% false-positive target**, the paired analysis separated Meta Prompt Guard 2 86M **below** StackOne and ProtectAI on attack detection, while StackOne and ProtectAI remained unseparated from each other. The Meta–StackOne difference was modest and should not be treated as a stable universal ordering. At the **0.1% target**, the available benign calibration sample provides too little resolution for a stable ordering.

The principal finding is therefore not a winner. The measured relationship among the leading configurations changes with the operating point. A buyer-specific conclusion would require the buyer's own traffic, deployment configuration, and false-positive cost.

This report is not a certification, endorsement, universal ranking, or prediction of production performance.

2. Scope and Terminology

A **configuration** means the specific model, package, or API detector as invoked by the Guardbench adapter. Results attach to those configurations and invocation paths, not to a vendor as a whole or to a vendor's full commercial product.

CONFIGURATION	FORM	IDENTIFIER AS TESTED	VENDOR
Meta Prompt Guard 2 86M	Open-weight classifier	`meta-llama/Llama-Prompt-Guard-2-86M`	Meta
Meta Prompt Guard 2 22M	Open-weight classifier	`meta-llama/Llama-Prompt-Guard-2-22M`	Meta
ProtectAI DeBERTa-v3 v2	Open-weight classifier	`protectai/deberta-v3-base-prompt-injection-v2`	ProtectAI
StackOne Defender 0.8.2	npm package	`@stackone/defender` 0.8.2	StackOne
Lakera Guard (`prompt_attack`)	Commercial API detector	`POST /v2/guard`, community tier	Lakera

Measured: true-positive rate (TPR) on the attack corpus at 0.1%, 1%, and 5% calibration false-positive targets; achieved false-positive rate on the held-out benign evaluation half; cluster-aware uncertainty; pairwise TPR differences; and a small difficult-benign probe.

Not measured: full commercial product behavior; multi-turn conversations; injections carried in tool results, retrieved documents, or images; production latency or cost; buyer-specific traffic; attacks written specifically against these configurations; or production false-positive prevalence.

3. Method

3.1 Corpus

The frozen corpus contains 2,500 attack items and 3,495 benign items.

Attack items were drawn from four public datasets:

- `deepset/prompt-injections` — 263
- `jackhhao/jailbreak-classification` — 637
- `reshabhs/SPML_Chatbot_Prompt_Injection` — 800
- `xTRam1/safe-guard-prompt-injection` — 800

Benign items were drawn from benign partitions of those sources, `databricks/dolly-15k`, and 60 hand-written difficult benign items created for this test. Items were deduplicated after whitespace and case normalization, capped at 800 per source per label, sampled with seed 0, then frozen and hashed.

Labels were inherited from the source datasets and were not independently re-adjudicated.

3.2 Threshold-setting protocol

Benign items were assigned by item-id hash to a calibration half (1,784) and an evaluation half (1,711).

For each configuration and each target false-positive rate, Guardbench selected the most lenient observed score level at which the share of calibration-benign items at or above threshold did not exceed the target. TPR was then measured on all 2,500 attack items. Achieved false-positive rate was measured on the separate evaluation-benign half.

Attack items never set thresholds.

The target and the achieved held-out false-positive rate are both reported because they are not identical quantities.

3.3 Uncertainty

The public attack corpus contains template-like and near-duplicate items. Attacks were therefore grouped into 1,947 clusters using the first six normalized tokens.

The final review analysis used a paired, cluster-aware bootstrap with 1,000 resamples. Attack clusters were resampled as units. For each pairwise replicate, the same attack clusters and the same calibration-benign item IDs were resampled for both configurations; each configuration's threshold was then recalibrated from its own scores on that shared calibration resample.

A pair is described as **separated** only when the pairwise interval excludes zero. Marginal interval overlap is not used as a significance test.

3.4 Execution and preprocessing

Runs were executed on one Apple-silicon laptop (arm64, Python 3.11.16) during 15–16 September 2026.

The Hugging Face configurations were loaded through `transformers` with their model tokenizers and truncated to 512 tokens. StackOne was invoked through its Node.js package. Lakera was invoked through its HTTPS API with only the prompt-attack detector enabled.

Preprocessing was therefore not identical across interfaces. The results in this report apply to the invocation paths actually used.

4. Results

4.1 Complete runs

Thresholds were set on calibration benign at the stated **target FPR**. Parentheses show the **achieved FPR** on the held-out benign evaluation half.

CONFIGURATION	TPR @ 0.1% TARGET	TPR @ 1% TARGET	TPR @ 5% TARGET
Meta Prompt Guard 2 86M	0.531 [0.36–0.63] (0.12%)	0.742 [0.68–0.77] (0.82%)	0.825 [0.80–0.85] (5.44%)
StackOne Defender 0.8.2	0.103 [0.03–0.45] (0.00%)	0.697 [0.58–0.78] (0.70%)	0.860 [0.83–0.88] (5.38%)
ProtectAI DeBERTa-v3 v2	0.314 [0.00–0.57] (0.00%)	0.676 [0.62–0.72] (0.41%)	0.871 [0.84–0.90] (4.62%)
Meta Prompt Guard 2 22M	0.309 [0.25–0.34] (0.00%)	0.335 [0.30–0.37] (1.11%)	0.515 [0.45–0.56] (5.38%)

Achieved FPR can exceed the calibration target because the threshold is selected on one benign half and evaluated on another.

4.2 Paired differences

TPR of first configuration minus second; 95% paired bootstrap interval.

PAIR	0.1% TARGET	1% TARGET	5% TARGET
Meta 86M – StackOne	[+0.098, +0.550]	[-0.041, +0.140]	[-0.056, -0.003]
Meta 86M – ProtectAI	[-0.159, +0.593]	[-0.008, +0.124]	[-0.083, -0.012]
StackOne – ProtectAI	[-0.506, +0.227]	[-0.095, +0.115]	[-0.054, +0.019]
ProtectAI – Meta 22M	[-0.332, +0.278]	[+0.261, +0.396]	[+0.300, +0.430]

At the 5% target, the Meta 86M–StackOne interval excludes zero only narrowly. It should be read as weak evidence of a difference under this corpus and analysis, not as a settled product ordering.

5. Reading the Operating Points

A detection rate is not comparable across configurations without the associated false-positive operating point.

The three targets in this report are analysis points, not asserted industry norms or recommended deployment settings.

0.1% target. With 1,784 calibration-benign items, this target allows at most one calibration item above threshold absent ties. The resulting thresholds are highly sensitive to individual benign scores, and the intervals are wide. The paired interval for Meta 86M minus StackOne excludes zero in this run, but this operating point is too coarsely resolved to support a selection decision or stable ordering.

1% target. Meta 86M, StackOne, and ProtectAI are not separated from one another in the paired analysis. Meta 22M is separated below them.

5% target. StackOne and ProtectAI are separated above Meta 86M in the paired analysis and remain unseparated from each other. The Meta–StackOne evidence is narrow at the zero boundary and should be treated cautiously.

A single-number ranking from one operating point would therefore overstate what this experiment establishes.

6. Difficult-Benign Probe

Sixty legitimate security-themed benign items were written for this test. They were deliberately selected to be difficult for a prompt-attack classifier and were not sampled from production traffic.

Thirty fell in the calibration half and thirty in the held-out evaluation half. Because calibration-half items can influence the threshold, only the held-out 30 are used for interpretation below; calibration-half counts are shown separately for transparency.

CONFIGURATION	HELD-OUT EVALUATION FLAGGED @ 0.1% / 1% / 5%	CALIBRATION-HALF FLAGGED @ 0.1% / 1% / 5%
Meta Prompt Guard 2 86M	0 / 0 / 2	1 / 4 / 6
StackOne Defender 0.8.2	0 / 4 / 13	1 / 8 / 13
ProtectAI DeBERTa-v3 v2	0 / 0 / 9	0 / 6 / 13
Meta Prompt Guard 2 22M	0 / 0 / 0	1 / 1 / 2

These are descriptive counts on selected items, not a prevalence estimate and not a production false-positive rate. No inferential claim is made from this probe; only the held-out evaluation half is used for interpretation.

At the 5% target, the configurations visibly differ on these held-out examples. Whether that difference matters in deployment depends on the buyer's actual traffic.

7. Vendor-Published Figures

Vendor figures below come from primary vendor/model-card materials reviewed for this report. Guardbench values come from the experiment above. The two should not be conflated.

Meta Prompt Guard 2

Meta's model card reports recall at 1% FPR of 97.5% for Prompt Guard 2 86M and 88.7% for 22M on a private benchmark that Meta describes as using datasets distinct from those used in training and as testing generalization to previously unseen attack types and distributions of benign data.

Guardbench obtained 0.742 and 0.335 at its 1% calibration target, with achieved held-out FPRs of 0.82% and 1.11%.

These are the same nominal target metric but not the same evaluation. The evaluation corpus and benign distribution differ, and Meta's public card does not provide enough threshold-setting detail to establish that its procedure matches Guardbench's. This report is not a replication attempt and does not establish that Meta's published figures are false. It shows that Guardbench obtained different measured values under different evaluation conditions.

StackOne Defender

StackOne's published performance summary reports approximately 89.0% average balanced accuracy and a 5.4% false-positive rate for Defender under its own benchmark conditions.

Balanced accuracy at that operating point is a different quantity from Guardbench's TPR at separately calibrated target FPRs. The figures are not compared directly.

ProtectAI DeBERTa-v3 prompt-injection v2

ProtectAI's model card reports 99.74% recall, 95.25% accuracy, and 91.59% precision on a 20,000-prompt post-training evaluation, without a matched false-positive rate for the recall figure.

The same model card describes the model as English prompt-injection detection, states that it does not detect jailbreak attacks or handle non-English prompts, and lists `jackhhao/jailbreak-classification` among its training-source metadata. Guardbench includes 637 items from that jailbreak dataset. ProtectAI's Guardbench result is therefore a common-condition measurement that includes material outside the model's documented scope and material from a named training source.

Its published recall is not directly comparable to any Guardbench fixed-FPR column.

Lakera Guard

No vendor-published fixed-FPR detection figure for the tested `prompt_attack` detector is used in this report. The Guardbench run is preliminary and is reported in Appendix B.

Category observation. Among the vendor materials reviewed for this report, only Meta supplied a detection figure at a stated false-positive rate that could be expressed in the same nominal metric family. The other published summaries use different operating conditions or metrics. A common-condition run is therefore required before those headline figures can be put on one measurement basis.

8. Reproducibility and Provenance

Two forms of reproducibility must be distinguished.

8.1 Analysis reproducibility

The reported analysis is computed from the frozen corpus and saved per-item result files. Those files record item identifier, label, source, score, latency, error, and adapter metadata, with headers recording adapter name, corpus hash, host architecture, Python version, and run timestamp.

The issued report uses the 1,000-resample paired cluster-aware analysis completed during final review. The exact analysis script, invocation, output, and frozen corpus were archived with SHA-256 hashes:

- `harness/audit_final.py` — `031ebd6875859e36fe2be17f895ded27118b7a5092853aab0afd446397ad019d`
- `results/audit_final_paired_2026-09-17.command.txt` — `42e0c8d90e2bd348a8806ff0223310c724114e16a0988cfe079a3ec76b9d4657`
- `results/audit_final_paired_2026-09-17.txt` — `ec4d64e511457449a558480589d75fc9d1be80efaeb47ea9923a9d5503774cfc`
- `corpus/corpus_v0.jsonl` — `252eb5671b878876d28f107da09b70b2891ed5a53b7fc33600ab60d86981c79c`

8.2 Inference reproducibility

Exact re-inference is not fully pinned.

Recorded information includes host architecture, Python 3.11.16, corpus hash, tested package/model names, StackOne version 0.8.2, run timestamps, adapter logic, tokenizer usage, and 512-token truncation for the Hugging Face configurations.

Not recorded at run time:

- Hugging Face commit/revision for the three open-weight classifiers

- `transformers` version
- `torch` version
- exact tokenizer package revisions
- StackOne package integrity hash
- Lakera service build/model version

These fields are disclosed as unrecovered rather than reconstructed from memory. Future runs will pin them at execution time.

8.3 Artifact availability

The frozen corpus, manifest, per-item results, adapters, runner, and analysis code are retained by Proving Office. Any public distribution of those artifacts will be stated separately and remains subject to source-dataset licensing and other applicable restrictions.

9. Limitations

Public-data overlap. All attack items and all but 60 benign items come from public datasets that pre-date the tested runs. Whether any given configuration was trained on any given item is generally unknown. ProtectAI is one documented exception: its model-card metadata lists `jackhhao/jailbreak-classification` among training sources. Public-data overlap may inflate apparent performance on novel attacks; it is not evidence of a mathematical upper bound.

Label validity. Labels were inherited from source datasets and were not independently re-adjudicated. The sources differ in what they call injection, jailbreak, or related attack behavior.

Scope mismatch. The configurations do not share identical documented scope. ProtectAI, for example, states that its model does not detect jailbreak attacks, while this corpus includes a jailbreak-classification source.

Preprocessing. The Hugging Face configurations were truncated to 512 tokens; StackOne and Lakera were invoked through their own package/API interfaces. Preprocessing was therefore not identical across systems, and long-input results may partly reflect interface-specific handling.

Corpus representativeness. The attack mix reflects four public datasets, not a buyer's threat model. It is weighted toward direct, single-turn text and under-represents indirect injection through documents, retrieved content, and tool results.

Template redundancy. The 2,500 attack items form 1,947 six-token-prefix clusters. That clustering heuristic reduces but does not eliminate dependence among similar attacks.

0.1% resolution. The benign calibration sample is too small to make the 0.1% target decision-grade.

Difficult-benign probe. Only 30 selected difficult-benign items are held out. They are descriptive examples, not a benchmark of production prevalence.

Lakera sample. The Lakera run contains 308 items and five discrete confidence levels. It is not comparable in precision to the complete runs.

Configurations are not products. Surrounding product features such as policies, allow-lists, routing, other detectors, and deployment context were not evaluated.

Single run. Each configuration was run once. API behavior over time was not characterized.

Inference provenance. Exact historical software/model revisions were not fully recorded, so the saved scores can be re-analyzed exactly but the original inference cannot necessarily be reproduced bit-for-bit.

10. Independence and Funding

This report was internally funded by SIR Holdings LLC. No buyer or tested vendor commissioned or paid for it.

Proving Office's planned private comparative engagements are buyer-funded. Proving Office does not accept vendor payment for buyer-funded comparisons.

Results of private engagements are confidential to the buyer unless the buyer authorizes disclosure.

Appendix A — Detection by Attack Source

TPR at the 0.1% / 1% / 5% targets.

CONFIGURATION	DEEPSET (263)	JACKHHAO (637)	SPML (800)	XTRAM1 (800)
Meta 86M	0.21 / 0.34 / 0.46	0.92 / 0.97 / 0.98	0.45 / 0.71 / 0.76	0.41 / 0.73 / 0.88
StackOne	0.00 / 0.47 / 0.82	0.04 / 0.88 / 0.96	0.09 / 0.68 / 0.81	0.20 / 0.65 / 0.84
ProtectAI	0.05 / 0.22 / 0.47	0.05 / 0.58 / 0.87	0.70 / 0.98 / 0.99	0.22 / 0.60 / 0.88
Meta 22M	0.08 / 0.08 / 0.10	0.68 / 0.72 / 0.83	0.12 / 0.15 / 0.51	0.28 / 0.30 / 0.41

These source-level values are descriptive and may be affected by training-data overlap or source-specific task mismatch.

Appendix B — Lakera Guard (Preliminary)

Lakera Guard's `prompt_attack` detector was scored on 308 corpus items: 123 attacks, 95 calibration benign, and 90 evaluation benign.

The detector emits five discrete confidence levels, so it cannot be placed at an arbitrary false-positive target. At the achievable point nearest the 1% target, TPR was 0.740 with an item-level 95% interval of 0.65–0.81 and achieved held-out FPR of 1.11% (1 of 90). AUC was 0.943.

This is a sizing estimate only and is not entered in the complete-run comparison.

Appendix C — Threshold-Free and Timing Measurements

AUC: Meta 86M 0.962; StackOne 0.968; ProtectAI 0.969; Meta 22M 0.892; Lakera preliminary 0.943.

AUC summarizes behavior across all thresholds, including operating points that may not be relevant to deployment. It is provided for completeness and is not used to rank the configurations.

Per-item wall-clock latency in this test environment: Meta 86M 26 / 219 ms (p50 / p95); Meta 22M 14 / 100 ms; ProtectAI 23 / 212 ms; StackOne 2.4 / 22 ms; Lakera 117 / 184 ms.

These latency values combine local execution and a residential-network API call and are not a production benchmark or a cross-system latency comparison.

Vendor Sources Reviewed

Primary source pages below were re-checked on **17 September 2026**.

- Meta, *Llama Prompt Guard 2 86M Model Card*: https://huggingface.co/meta-llama/Llama-Prompt-Guard-2-86M/blob/main/README.md
- Meta, *Llama Prompt Guard 2 22M Model Card*: https://huggingface.co/meta-llama/Llama-Prompt-Guard-2-22M/blob/main/README.md
- StackOne, *Defender — Prompt Injection Guard*: https://www.stackone.com/platform/prompt-injection-guard/
- ProtectAI, *deberta-v3-base-prompt-injection-v2 Model Card*: https://huggingface.co/protectai/deberta-v3-base-prompt-injection-v2/blob/main/README.md
- Lakera / Check Point AI Security, *Guard Results / Confidence Levels*: https://docs.lakera.ai/docs/api/results

Proving Office · A program of SIR Holdings LLC · desk@provingoffice.com · provingoffice.com

Report No. GB-2026-001 · v0.4 · Corpus SHA-256 `252eb567…c79c`