

Model card addendum: Shield 1.1

Shield Large with a retrained language model. An addendum to the Shield 1 model card,
29 October 2026

Contents

	Summary	3
1	What changed	4
2	Intended use	6
3	Training data	7
4	Training procedure	9
5	Evaluation	10
6	Limitations	17
7	Safety	18
8	Versioning	18
A	Results for every dataset	19
B	Provenance	20

Summary

This addendum to the Shield 1 model card covers Shield 1.1, an update to Shield Large. Shield Large is the most accurate of ZeroLeaks' four hosted Shield tiers for detecting prompt injection and jailbreak attempts in text that a language model application or agent is about to read. Shield 1.1 replaces the language model inside Shield Large with a retrained one and runs it unquantized. Nothing else changes: the API, the model ID (`shield-large`), the combination with Shield Base, the threshold, the rules and the price are the same. Shield (small) and Shield Base are unchanged; Shield Tiered gets Shield 1.1 on the texts it escalates to Shield Large.

The addendum reports only what changed. For the rest, see the Shield 1 model card: the tiers are in section 1.1, intended use in section 1.2, the benchmark's datasets, sources and licenses in section 2.1, the metrics in section 2.2, how each comparison detector was run in section 2.3, adaptive attacks in section 3.4, latency and cost in section 4, and deployment guidance in section 5.

Headline result. On the Shield 1 benchmark (35 public datasets in five categories, first published in 2025 or 2026), Shield Large's mean balanced accuracy rises from 0.904 with the Shield 1 language model as deployed to 0.915 with Shield 1.1, and every category improves. Pooled over the 22,001 texts of the headline sets, Shield 1.1 flags 84.8% of attacks (84.5% before) and 4.1% of benign texts (5.4% before).

Disclosure. This benchmark (43 sets, 35 headline, five categories) was also used to choose between training rounds; a fresh held-out test follows. The benchmark is therefore not unseen data for Shield 1.1.

Where the gain comes from. Running the Shield 1 language model unquantized, with no retraining, already scores 0.910: 53% of the overall gain comes from running unquantized and the rest from training. Training's contribution is concentrated in agent attacks and multilingual text.

Where it is behind. Shield 1 Large scores higher than Shield 1.1 on 8 of the 35 headline datasets, most of all on the Lakera b3 agent benchmark (0.684 to 0.652). GPT-6.1 Sol as a judge is the better detector for agent attacks and prompt injection: it scores 0.933 on agent attacks and 0.988 on prompt injection, against 0.819 and 0.968 for Shield 1.1. It needs a language-model call for every text. NVIDIA NemoGuard leads the false-alarm category, at 1.000 against 0.956 for Shield 1.1, but it catches only 2.0% of attacks, against 84.8%. Agent attacks remain the weakest category (0.819).

1 What changed

1.1 Changes from Shield 1

	Shield 1 Large	Shield 1.1
Language model	Shield 1 language model	Retrained from it: continued training with replay and a teacher-agreement guard, served as an equal-weight average with the original
Language model serving	4-bit build	float32, unquantized
Combination, threshold, rules	0.6 × language model + 0.4 × Shield Base; 0.5; rules	Unchanged
Overall score	0.904	0.915
Attacks flagged, pooled	84.5%	84.8%
Benign texts flagged, pooled	5.4%	4.1%
Headline datasets better, same, worse than Shield 1 Large	(reference)	25, 2, 8
Weights published	No	No

Table 1. Shield 1 Large and Shield 1.1 at a glance.

The change is largest on false alarms (+0.022) and smallest on jailbreaks (+0.006). On agent attacks, running unquantized alone would have cost 0.012; training more than made up for it (+0.021).

1.2 Lineage

Shield 1.1 is a release of Shield Large. Its language model descends directly from Shield 1 Large's: we recovered the original training checkpoint, verified it against the deployed model's exported parameters, and continued training it. Its internal identifier is `L7a` continuation R12, seed 42, lambda 0.5.

Tier	Model ID	What it is	In Shield 1.1
Shield	<code>shield</code>	Small multilingual encoder (S15e) with rules	Unchanged
Shield Base	<code>shield-base</code>	Larger multilingual encoder (SB1) with rules	Unchanged
Shield Large	<code>shield-large</code>	0.6 × language model (Qwen3-1.7B classifier) + 0.4 × Shield Base, with rules	Retrained language model; same combination, threshold and rules
Shield Tiered	<code>shield-tiered</code>	Shield first; uncertain texts go to Shield Large	First stage unchanged; escalated texts use the new Shield Large

Table 2. Shield tiers and what Shield 1.1 changes.

1.3 Architecture

Unchanged from Shield 1 (Shield 1 model card, section 1.1), except for the language model's weights. Shield Large scores a text with two models. Its language model is a Qwen3-1.7B model fine-tuned with LoRA as a binary classifier, reading overlapping windows of 512 tokens with a stride of 384, up to 16 windows per text. Shield Base is a multilingual encoder classifier reading windows of 256 tokens with a stride of 192. Each

model takes its highest score across its windows. Shield Large's score is 0.6 times the language model's probability plus 0.4 times Shield Base's. A text is flagged when that score is at least 0.5, or when Shield's deterministic rules (normalization, decoding of common encodings, and known attack patterns) match.

The new language model weights are an equal-weight average of Shield 1's language model and the retrained one, applied to every trained parameter, the classification head included.

1.4 Inputs and outputs

Unchanged. Input is text, such as a user message, a retrieved document, a web page, a tool result, an MCP tool description or an agent skill. Output is a binary verdict in one `prompt_injection` category with a score, plus coverage metadata telling the caller whether any part of a long input was not scored. Prompt injection and jailbreaks share the one label. The hosted API accepts the same request sizes as before.

1.5 Serving

Shield 1's hosted Shield Large used a 4-bit build of its language model. Shield 1.1 serves the language model in float32, without quantization. The benchmark results in this addendum were measured on an NVIDIA L4 using a bfloat16 language model with its classification head and adapter kept in float32, so the served build's scores can differ slightly from the measured ones (see Limitations, Serving fidelity).

We measured latency for both versions on one staging server, not in production: an AWS g6.xlarge with one NVIDIA L4, running the same server image for Shield 1 and Shield 1.1. Each model ran under load for 60 seconds, with four requests in flight sent over HTTP from the server itself, so the figures include the HTTP server and queueing but no network, TLS or gateway time. The hosted API adds network and gateway time to every request.

Model and text	Shield 1 p50	Shield 1 p95	Shield 1.1 p50	Shield 1.1 p95
Shield (small), short text	17	20	18	20
Shield Base, short text	51	56	52	58
Shield Large, short text	286	297	158	171
Shield Large, longest text	10,694	10,878	10,214	12,612
Shield Tiered, escalated text	327	345	192	212

Table 3. Latency per request on the staging server, in milliseconds (lower is better). The short text is a single short sentence; the long one is the longest text the API accepts; the escalated text is a short sentence that Shield (small) is unsure about, so Shield Tiered sends it to Shield Large. Shield (small) and Shield Base are unchanged, so their rows show the run-to-run variation.

Shield Large answers short texts in about half the time it took with Shield 1, and Shield Tiered answers the texts it escalates in a little over half the time. The longest texts take roughly the same time on both versions, but they rest on few requests: at this load most of them timed out in the server's queue (9 of 40 completed for Shield 1.1, 6 of 42 for

Shield 1). A new server takes longer to start with the unquantized model (196 seconds to ready, against 141), which affects how quickly capacity can be added, not request latency. Shield 1.1 runs on the same server type as Shield 1 Large.

The Shield 1 model card's latency figures (section 4) were measured on a different build and GPU, as model time without the HTTP server. To compare the two versions, use the figures above.

1.6

Access and weights

Shield 1.1 is available through the hosted API as `shield-large`, included in every paid ZeroLeaks plan. Its weights are not published. The small Shield model's weights are available for noncommercial use under CC BY-NC 4.0; that model is unchanged.

2

Intended use

Unchanged from Shield 1 (Shield 1 model card, sections 1.2 and 5). Shield 1.1 is for screening text before an agent or LLM application reads it: tool results, retrieved documents, web pages, emails, MCP tool descriptions, agent skills and user messages. It flags attempts to redirect a model across a trust boundary: overriding its instructions, changing its task, extracting its instructions or secrets, or making it act for someone other than its user. Jailbreak attempts count as attacks. It is one layer in a defense that also limits what the agent can do, and its verdict is a signal for the application's policy (block, ask for approval, strip, log), not a permission decision.

Out of scope:

- Granting or denying permissions. A clean verdict does not authorize an action that the application's own controls would deny.
- Content moderation for harmful topics. A harmful request with no attempt to override instructions is not, by itself, an attack under the label definition.
- Attacks that live in the structure of a tool-use conversation rather than in any one text, such as a misused function call with no instruction in the text.
- Attacks that depend on the application's system prompt, which Shield does not see.
- Images, audio, or text the caller does not pass in. Shield scores only the text it receives, and long texts can be partly unscored (check `shield.coverage`).
- Sole protection for high-impact actions such as payments, outbound messages or code execution. Use Shield Large on every text there, not Shield Tiered, and keep human approval where the stakes call for it.

3

Training data

3.1

Sources

Shield 1.1's language model was trained on two populations, mixed in every training step. No customer data and no free-tier traffic went into its training.

The accepted corpus (254,006 training records: 212,262 benign, 41,744 attacks; 252,461 training examples after rendering). It combines:

- Shield's earlier public training sources, relabeled under the current label definition. Attack examples from before 2025 were kept only where both labeling models agreed.
- Public datasets first published in 2025 or 2026: 118,376 items were labeled and 81,413 accepted.
- A curated set of older prompt-injection collections, used in the training split only, where recent data lacked classic injections.
- Generated agent data (39,618 records). Each attack is an edit of a real benign text from a training split (a skill file, tool output, MCP tool description, web page, agent trace, system prompt, chat message or code), with an instruction inserted in one of several languages, often late in a long text. Each attack comes with a clean twin (the untouched text) and a hard twin (the same position holding a harmless instruction to a person, or the attack quoted as data with a request not to follow it). Every destination in a generated payload is on a reserved `.invalid` domain.

Replay of Shield 1's own training data (132,236 rows). To stop the model forgetting what it already knew, we sampled from the data Shield 1 Large's language model was trained on: 118,581 rows across attack and benign strata (51,176 attacks, 67,405 benign), weighted toward jailbreaks, non-English attacks and the benign text that resembles them, plus 13,584 jailbreak and multilingual attack rows re-admitted in a later round (see Training procedure).

3.2

Labeling

The label definition (v12) calls a text an attack when it contains an instruction that tries to redirect an AI system across a trust boundary. A harmful topic or a request for sensitive information is not enough. Quoted attacks, security writing, instructions addressed to people and legitimate developer instructions are benign when their context supports it. An attack attempt stays an attack even if a model refuses it.

Two frontier language models from different providers labeled every text in the accepted corpus with the same prompt and gave an evidence span for each attack. This addendum does not name them. A text was accepted only when both labeling models gave the same label; disagreements were dropped, never relabeled, and so were texts whose new label conflicted with their source's. For generated attacks, both labeling models had to label the text an attack, their evidence spans had to overlap each other and the inserted passage, and long texts entered only through windows that both models had labeled. Twins were accepted only when both labeling models called them benign.

We validated the agreement rule before using it, on a reference set labeled by blind AI reviewers: of the texts it admitted, 193 of 194 were labeled correctly, against a bar of 99%. The rule is strict. It kept 67% of the reference set's attacks and 94% of its benign texts, so agreement filtering favors clear cases. That is one reason we added generated attacks in the formats agents read.

Replay rows keep their source dataset's label. Where an earlier model-based audit of Shield 1's training data disagreed with a label, the row was dropped, and sources with high measured disagreement kept only rows the audit confirmed.

3.3 Audits

Blind AI reviewers did the audits: AI agents, not people, that re-labeled random samples of accepted records using only the label definition, without seeing the accepted labels or the labeling models' answers. The audit rule called for two reviewers per sample. A record counts as confirmed when at least one reviewer gives the accepted label and neither gives the opposite. Both batches of generated data passed the pre-registered bar of 95% before they were used, the second with one AI reviewer instead of two. The re-labeled earlier corpus was audited the same way. The 2025 and 2026 public sources and the curated older injection set were labeled under the validated agreement rule, but their accepted records had no separate audit.

Audited population	Sample	Confirmed	Blind AI reviewers	Result
Relabeled earlier corpus, accepted records (both labeling models agree)	415	413	two	99.5%
Synthetic agent-surface data, batch 1	150	150	two	100.0%
Synthetic agent-surface data, batch 2	150	148	one (deviation, see text)	98.7%
Rejected alternative: admit rows where one labeling model abstains	150	136	two	90.7%, below the bar, not used

Table 4. Blind label audits by AI reviewers. Generated-data samples drew 80 attack and 70 benign records.

The second batch of generated data is a documented deviation. The rule asked for two AI reviewers, but two further AI reviewers were stopped by automated safety filters while reading the attack records and were not rerun in altered form. The batch was admitted on one complete blind AI review with no contradictions, because the first batch, from the same generator, labelers and acceptance rule, passed with two reviewers. The generators of this data come from the same model families as the labeling models, which is why every batch was audited before use.

3.4 Screening

Benchmark datasets, other splits of them, and their source families were excluded from training. New and replayed training text was screened against every benchmark text with exact, normalized and near-duplicate matching (word eight-gram containment and MinHash similarity), and matches were removed. Generated attacks were built only on

texts from training splits and checked so that no inserted passage overlaps a benchmark text. The development panel described under Evaluation was screened against both the benchmark and the training data.

4 Training procedure

Starting point. Shield 1 Large's language model, recovered from its original training checkpoint and checked against the deployed model's exported parameters before any training.

Recipe. A fresh LoRA adapter of rank 32 and a trainable classification head, learning rate $1e-5$ with linear decay, one pass over 384,697 examples (48,886,732 tokens) in 1,487 steps of about 32,768 tokens, on one NVIDIA L4 per seed. Accepted and replay examples are interleaved in every step.

Loss and the teacher-agreement guard. Each example's loss is cross-entropy with its label plus 0.25 times the divergence from the original model's prediction, which acts as a teacher, weighted so that attacks and benign texts carry equal mass. A guard decides what each example learns:

- If the teacher agrees with the label, the example is trained on both, as above.
- If they disagree and the label is trusted (both labeling models agreed, or a replay label confirmed by an earlier independent audit), the example is trained on its label alone. No trusted benign text is ever pulled toward the teacher's attack verdict, which protects false alarms.
- If they disagree and the label is not independently trusted, the example is left out.

Three rules set what is trusted. Replay rows without an audit confirmation are untrusted. The re-admitted jailbreak attack rows are untrusted, so they train only where the original model already calls them attacks. Accepted benign texts on chat surfaces (user messages and system prompts) are untrusted: the original model disputes some of them because they wrap a harmful request in a jailbreak, and training those toward benign taught earlier rounds to pass jailbreaks. This last rule leaves out 3,490 of the 16,590 accepted benign examples the teacher disputes, and keeps the 13,100 disputed benign examples from agent surfaces, which carry most of the agent and false-alarm gains. Replay attack rows in the jailbreak and multilingual strata carry 1.5 times their earlier weight, with benign replay scaled to keep the classes balanced.

Blend. The served language model is the average of the original and the retrained weights (adapter scale 0.5). This setting was fixed in advance of the R12 evaluation, the round this addendum reports, and it was not tuned on the benchmark. A development panel built later confirmed it.

Seeds and selection. We trained seeds 17 and 42, and both passed the replacement rule. We ship seed 42, the one with the higher development-panel gain over the unquantized original (+0.00390 against +0.00385 overall); the two are effectively tied. The rule that

made this choice uses only the development panel, but we wrote it on 27 October 2026, after both seeds' benchmark scores were known. On the benchmark, the shipped seed scores slightly lower than seed 17 (0.9148 against 0.9153).

5 Evaluation

Unless a table or figure says otherwise, each Shield 1.1 score below is for Shield Large's full combination (language model, Shield Base and rules) with the shipped seed, measured in our evaluation runtime, and "Shield 1 Large" means the deployed 4-bit build.

5.1 Benchmark

The benchmark is the one in the Shield 1 model card, which lists every dataset with its source and license (section 2.1): 43 public datasets, of which 35 count toward the headline score (22,001 texts) in five categories. Agent attacks (14 datasets) are indirect injections in tool results, tool descriptions, agent skills, web pages and documents. Prompt injection (4) covers direct attempts to override an application's instructions or extract its secrets. Jailbreaks (4) are attempts to make a model ignore its safety training.

Multilingual (7) covers attacks and benign text outside English. False alarms (6) is benign text that detectors often mistake for attacks. Every dataset was first published in 2025 (9) or 2026 (26). Before Shield 1's scoring, each was audited (40 labels per dataset) and 8 datasets with more than 15% wrong labels were removed. The other 8 datasets in the evaluation panel (`agentbeats`, `lmarena`, `mcp_catalog`, `probes_mixed`, `quadrat`, `repo-rails`, `trace_commons` and `xl_jailbreak`) are scored but carry no weight in any score reported here. The benchmark's datasets and source families were kept out of training, and training text was screened against every benchmark text (see Screening). Because we used the benchmark to choose between training rounds, it is development data for Shield 1.1, not a held-out test (see Replacement rule).

5.2 Metric

Unchanged (Shield 1 model card, section 2.2). For each dataset, balanced accuracy: the share of attacks flagged on attack-only datasets, the share of benign texts left unflagged on benign-only datasets, and the mean of the two on mixed datasets. Dataset scores are averaged within each category, and the five category scores are averaged for the overall score, so the false-alarm category weighs as much as each attack category. We also report the share of all attacks and of all benign texts flagged, pooled over the headline sets. The technical report adds recall at fixed false-alarm rates.

5.3 Replacement rule

Before the first training round (15 October 2026), we fixed the rule a candidate had to meet to replace Shield 1's language model. Under that rule, Shield Large had to score higher than Shield 1 Large in each of the five categories, and at least 0.005 higher overall, on both training seeds. The rule did not change afterwards. We evaluated five candidates

against it in sequence; the first four failed and Shield 1.1 passed. Each failure informed the next training change, which is how the benchmark came to be used in model selection. The rule compares point estimates. It does not establish statistical significance, and our analysis records the release as not yet statistically qualified because the datasets are not independent samples.

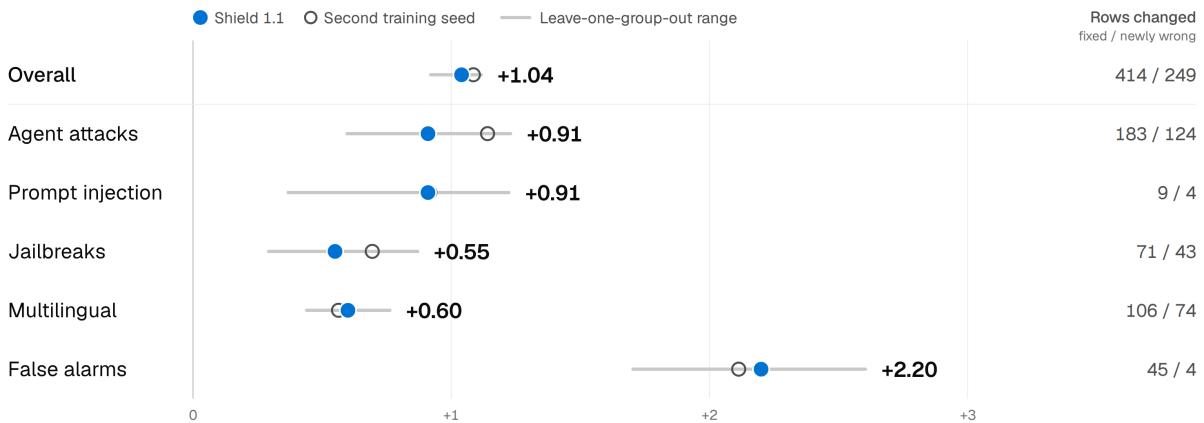
5.4 Results

Category	Shield 1 Large (4-bit)	Shield 1 Large, unquantized	Shield 1.1	Total change	From running unquantized	From training
Agent attacks	0.810	0.798	0.819	+0.009	-0.012	+0.021
Prompt injection	0.959	0.968	0.968	+0.009	+0.009	0.000
Jailbreaks	0.910	0.918	0.915	+0.006	+0.008	-0.003
Multilingual	0.910	0.911	0.916	+0.006	+0.002	+0.004
False alarms	0.934	0.954	0.956	+0.022	+0.020	+0.002
Overall	0.904	0.910	0.915	+0.010	+0.006	+0.005

Table 5. Balanced accuracy by category (higher is better). "Unquantized" is Shield 1's language model with no retraining, run unquantized (bfloat16 language model with its classification head and adapter kept in float32) in the same runtime as Shield 1.1. "From running unquantized" is unquantized minus deployed; "From training" is Shield 1.1 minus unquantized. Differences are computed from unrounded scores.

Shield 1.1 is ahead of Shield 1 in all five categories

Change from Shield 1 in points of balanced accuracy. The line runs from the lowest to the highest change when any one group of related sets is left out: a sensitivity check, not a confidence interval.



ZeroLeaks prompt-injection benchmark: 43 sets, 35 in the headline score, 5 categories. This benchmark was also used to choose between training rounds; a fresh held-out test follows.

ZeroLeaks

zeroleaks.ai

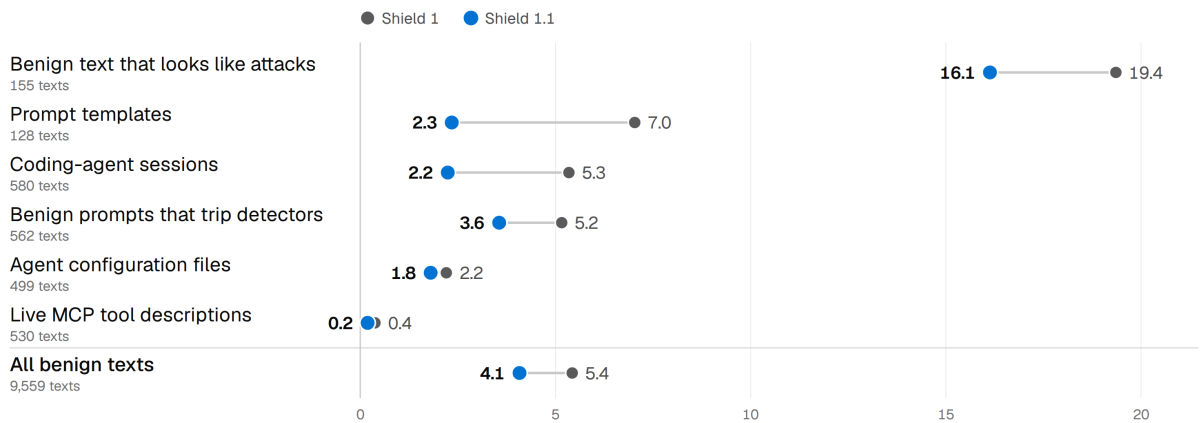
Figure 1. Change by category. Shield 1.1 minus Shield 1 per category, with the second training seed and the range when each group of related datasets is left out (a sensitivity check, not a confidence interval). Above zero means Shield 1.1 is better.

Pooled over the headline sets	Shield 1 Large	Shield 1 Large, unquantized	Shield 1.1
Attacks flagged	84.5%		84.8%
Benign texts flagged	5.4%		4.1%

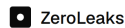
Table 6. Share of attack texts flagged (higher is better) and of benign texts flagged (lower is better), pooled over the 22,001 texts of the headline sets.

Shield 1.1 raises fewer false alarms on all six benign test sets

Share of benign texts flagged as attacks (%), lower is better. Each set holds only benign texts; the last row pools every benign text in the 35 headline sets.



ZeroLeaks prompt-injection benchmark: 43 sets, 35 in the headline score, 5 categories. This benchmark was also used to choose between training rounds; a fresh held-out test follows.



zeroleaks.ai

Figure 2. False alarms on benign datasets. Share of benign texts flagged on each of the six datasets in the false-alarm category, and pooled over every benign text in the headline sets. Lower is better.

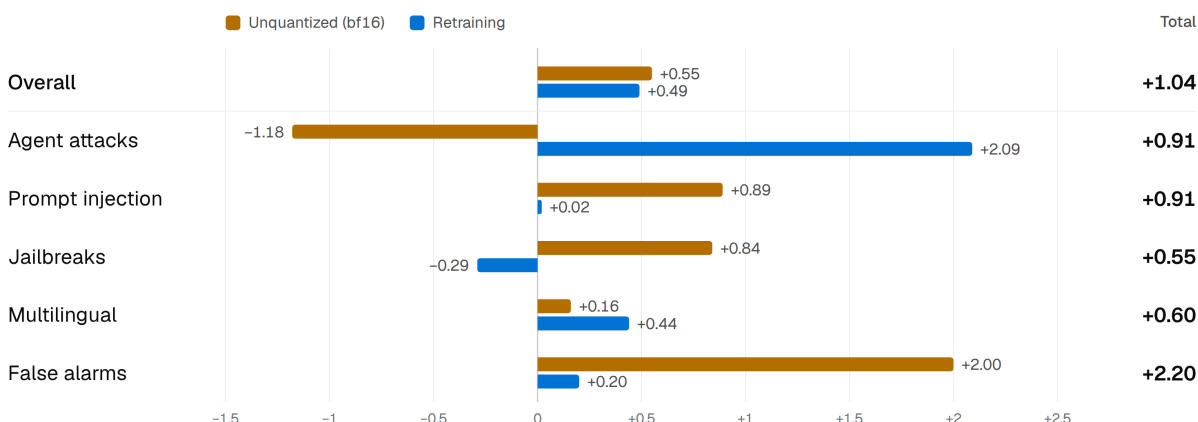
Outside the false-alarm category, one benign-only dataset moved the other way: on Turkish over-refusal prompts (multilingual), Shield 1.1 flags 4.2% of benign texts, against 3.7% for Shield 1 Large.

Texts in the headline sets	Count
Right for both models	19,301
Wrong for Shield 1 Large, right for Shield 1.1	414
Right for Shield 1 Large, wrong for Shield 1.1	249
Wrong for both models	2,037
Total	22,001

Table 7. Text-by-text comparison with Shield 1 Large on the headline sets.

About half of the gain comes from running the model unquantized

Change from Shield 1 in points of balanced accuracy. Unquantized: Shield 1's own language model run in bf16 instead of as the 4-bit build (runtime differences included). Retraining: Shield 1.1 against that unquantized model.



ZeroLeaks prompt-injection benchmark: 43 sets, 35 in the headline score, 5 categories. This benchmark was also used to choose between training rounds; a fresh held-out test follows.

ZeroLeaks

zeroleaks.ai

Figure 3. Where the gain comes from. Change from Shield 1 split into running Shield 1's own language model unquantized (in our evaluation runtime, so runtime differences are included) and retraining (Shield 1.1 against that unquantized model). Points of balanced accuracy; above zero is a gain.

5.5

Development panel

The development panel (13,754 texts in 104 datasets, 6,611 attacks and 7,143 benign) comes from Shield's development and held-out sources. It excludes every benchmark dataset and other splits of them, and every text that overlaps the benchmark or the training data. We used it, never the benchmark, to confirm the blend and to choose the seed. Some of its agent and multilingual rows are held-out splits of Shield's own data generators, so it measures in-distribution performance more than the benchmark does.

Category	Sets	Attack texts	Benign texts	Change vs unquantized original	95% interval
Agent attacks	33	1,818	1,180	+0.0070	+0.0030 to +0.0111
Prompt injection	11	2,089	911	+0.0054	+0.0003 to +0.0104
Jailbreaks	6	1,435	334	+0.0019	-0.0027 to +0.0064
Multilingual	17	1,269	1,721	+0.0036	+0.0001 to +0.0071
False alarms	37	0	2,997	+0.0016	-0.0005 to +0.0038
Overall	104	6,611	7,143	+0.0039	+0.0021 to +0.0057

Table 8. Shield 1.1 against the unquantized Shield 1 language model on the development panel, both inside Shield Large's combination. Change in balanced accuracy, where above zero favors Shield 1.1; intervals are paired 95% intervals (a paired normal approximation stratified by dataset and class, which treats texts within a dataset as independent).

Other detectors

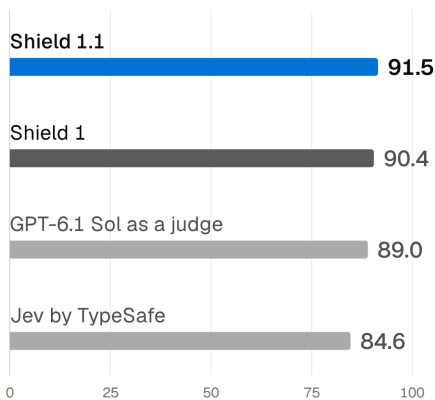
We compare Shield 1.1 with every detector that has verified predictions on all 43 sets, scored with the same code as Shield 1.1: 14 detectors, including GPT-6.1 Sol as a judge and the paid services and open models from the Shield 1 card. Each ran at its default setting on the full text, and a text the platform declined to process counts as flagged. An LLM judge answers our own prompt, which asks whether a text contains a prompt injection or jailbreak, so its score depends on that prompt. For GPT-6.1 Sol as a judge, one unparsable response counts as not flagged and 12 declined texts count as flagged. Their predictions were made between 27 September 2026 and 9 October 2026. The paid services and open models are the same runs as in the Shield 1 card, which describes how each was run (section 2.3). Left out are 26 detectors that lack predictions on some or all sets, and 7 earlier, failed or superseded runs.

Shield 1.1 has the highest overall score of the 16. GPT-6.1 Sol as a judge is the better detector for agent attacks and prompt injection: it scores 93.3% on agent attacks and 98.8% on prompt injection, against 81.9% and 96.8% for Shield 1.1. It needs a language-model call for every text. NVIDIA NemoGuard leads the false-alarm category, at 100.0% against 95.6% for Shield 1.1, but it catches only 2.0% of attacks, against 84.8%.

Shield 1.1 scores 91.5 overall, up from 90.4

Mean balanced accuracy (%), higher is better. We ran every detector on the same texts; Jev by TypeSafe scored highest of the 13 other detectors. Circled: best in each category.

Overall



By category

	Shield 1.1	Shield 1	GPT-6.1 Sol judge	Jev
Agent attacks	81.9	81.0	93.3	76.4
Prompt injection	96.8	95.9	98.8	89.4
Jailbreaks	91.5	91.0	69.3	70.4
Multilingual	91.6	91.0	90.4	87.6
False alarms	95.6	93.4	93.0	99.4

ZeroLeaks prompt-injection benchmark: 43 sets, 35 in the headline score, 5 categories. This benchmark was also used to choose between training rounds; a fresh held-out test follows.

 ZeroLeaks

zeroleaks.ai

Figure 4. Shield 1.1 against Shield 1 and other detectors. Mean balanced accuracy overall and by category for Shield 1.1, Shield 1, every verified LLM judge and the best other detector. Higher is better.

Detector	Type	Overall	Agent attacks	Prompt injection	Jailbreaks	Multilingual	False alarms	Attacks flagged	Benign flagged
Shield 1.1	ZeroLeaks	91.5%	81.9%	96.8%	91.5%	91.6%	95.6%	84.8%	4.1%
Shield 1 Large	ZeroLeaks	90.4%	81.0%	95.9%	91.0%	91.0%	93.4%	84.5%	5.4%
GPT-6.1 Sol as a judge	LLM judge	89.0%	93.3%	98.8%	69.3%	90.4%	93.0%	78.7%	3.5%
Jev by TypeSafe	Paid service	84.6%	76.4%	89.4%	70.4%	87.6%	99.4%	67.2%	0.9%
OpenAI Guardrails	Paid service	80.1%	68.6%	83.1%	82.9%	75.4%	90.3%	65.4%	7.1%
Patronus Wolf Defender	Open model	79.6%	62.1%	79.2%	81.7%	80.4%	94.7%	57.7%	5.6%
Horizon-Labs guard large	Open model	78.7%	64.8%	88.1%	73.2%	75.0%	92.4%	63.6%	10.3%
Amazon Bedrock Guardrails	Paid service	72.8%	58.0%	80.6%	69.4%	62.1%	94.0%	43.4%	2.9%
TestSavant defender large	Open model	71.0%	49.8%	69.6%	79.7%	73.8%	82.0%	60.3%	24.9%
Sheltron 68M	Open model	68.3%	55.2%	77.4%	75.7%	61.8%	71.5%	58.5%	27.9%
PIGuard	Open model	66.3%	56.7%	77.0%	75.2%	55.7%	66.8%	58.0%	27.4%
Azure AI Prompt Shields	Paid service	63.2%	51.8%	70.8%	48.2%	50.5%	94.6%	26.5%	1.3%
ProtectAI v2 (LLM Guard)	Open model	61.0%	46.0%	70.0%	67.2%	54.9%	67.1%	45.4%	28.2%
Meta Prompt Guard 2 86M	Open model	58.6%	46.5%	54.6%	51.7%	52.5%	87.7%	29.9%	2.8%
deepset DeBERTa injection	Open model	57.2%	64.7%	66.2%	78.6%	63.4%	13.0%	99.8%	91.3%
NVIDIA NemoGuard	Open model	49.0%	35.7%	37.5%	28.7%	43.1%	100.0%	2.0%	0.0%

Table 9. Every verified detector on the 35 headline sets. Balanced accuracy (higher is better); the best score in each accuracy column is in bold, with no allowance for the margin of error. The last two columns pool every headline text: attacks flagged (higher is better) and benign texts flagged (lower is better). This benchmark was also used to choose between our training rounds, which can only favor Shield 1.1 here. NVIDIA NemoGuard has the best false-alarm score (100.0%) but catches 2.0% of attacks, against 84.8%.

Shield 1.1 scores highest of the 16 detectors we tested

Mean balanced accuracy (%), higher is better. We ran every detector at its default setting on the same texts; LLM judges use our prompt. Only detectors with predictions on all 43 sets are included.

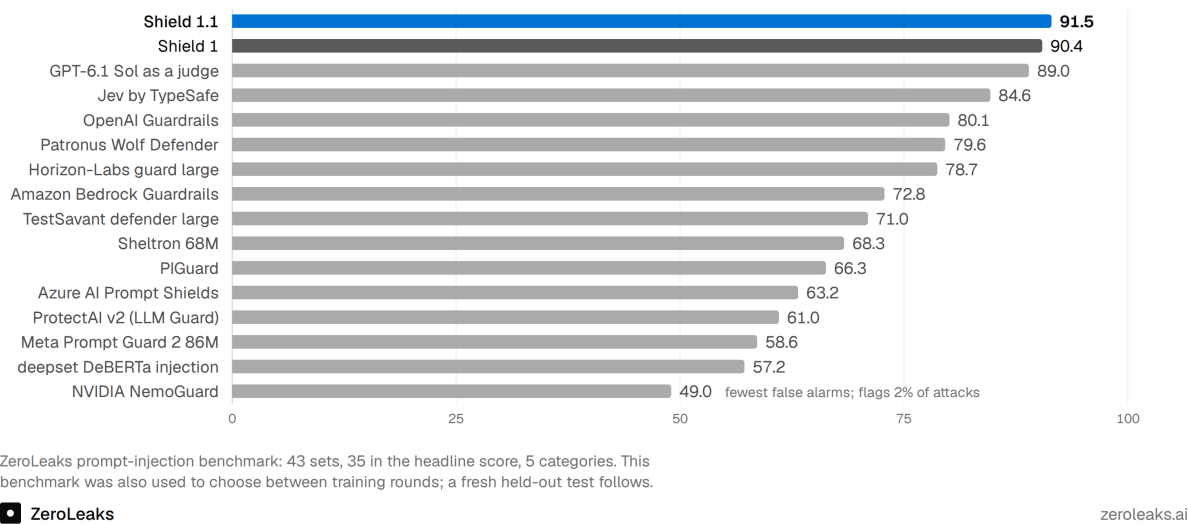


Figure 5. Every detector on the benchmark. Overall score of every detector with verified predictions on all 43 sets. Higher is better.

5.7 Seed variation

Category	Seed 17	Seed 42 (shipped)
Agent attacks	0.821	0.819
Prompt injection	0.968	0.968
Jailbreaks	0.917	0.915
Multilingual	0.915	0.916
False alarms	0.955	0.956
Overall	0.915	0.915

Table 10. Balanced accuracy (higher is better) of the two trained seeds on the benchmark, each inside Shield Large's combination.

6 Limitations

6.1 Regressions

Dataset	Category	Texts	Shield 1 Large	Shield 1.1	Change	Unquantized original
Lakera b3 agent benchmark	Agent attacks	187	0.684	0.652	-0.032	0.626
MCPTox	Agent attacks	827	0.865	0.851	-0.014	0.823
promptprotection agent flows	Agent attacks	91	0.966	0.956	-0.010	0.956
Turkish over-refusal prompts	Multilingual	240	0.963	0.958	-0.004	0.946
HackAPrompt 2.0 (Pliny track)	Jailbreaks	503	0.911	0.907	-0.004	0.917
AgentInjectionBench	Agent attacks	182	0.897	0.893	-0.004	0.890
ChangeMore (Chinese)	Multilingual	1,963	0.942	0.939	-0.003	0.940
Kubernetes manifests	Agent attacks	851	0.981	0.980	-0.001	0.980

Table 11. Headline datasets where Shield 1.1 scores below Shield 1 Large. On these datasets, Shield 1 Large is the better model.

Of these, five are agent-attack datasets (Lakera b3 agent benchmark, MCPTox, promptprotection agent flows, AgentInjectionBench and Kubernetes manifests): the category improved on average while they moved the other way. On seven of them, the unquantized original already scored below the deployed model, so part of the drop comes from running unquantized rather than from training.

6.2 Weakest datasets

Dataset	Category	Shield 1 Large	Shield 1.1
DTap malicious skills	Agent attacks	0.571	0.626
RTC-Bench computer-use injections	Agent attacks	0.604	0.646
Lakera b3 agent benchmark	Agent attacks	0.684	0.652
DTap tool-description injections	Agent attacks	0.667	0.670

Table 12. Headline datasets where Shield 1.1 scores below 0.70.

6.3 Other limitations

- **Benchmark reuse.** The benchmark guided the choice between training rounds. Treat these results as development results on public data until the fresh held-out test is published.
- **No statistical qualification.** The replacement rule compares point estimates. Confidence intervals that account for related datasets have not been computed for the benchmark.
- **Serving fidelity.** Scores were measured in our evaluation runtime with a bfloat16 language model with its classification head and adapter kept in float32. Shield 1.1 serves the language model in float32, without quantization. The served build's scores can differ slightly; they are checked against these before release.

- **Jailbreak definition.** The label definition treats a harmful request wrapped in a jailbreak, with no attempt to override instructions, as benign, while the benchmark and Shield 1 count jailbreak attempts as attacks. Shield 1.1 keeps Shield 1's behavior through replay and the guard, but on jailbreaks the retrained model scores -0.003 against the unquantized original.
- **Agent attacks.** At 0.819, agent attacks are still the weakest category; the weakest datasets are listed above.
- **Adaptive attacks.** We have not re-run adaptive or gradient-based attacks against Shield 1.1. The Shield 1 card's adaptive results (section 3.4) were measured on an earlier release and do not describe this model. An attacker who can query a detector repeatedly can get past every detector we have tested some of the time.
- **Shield Tiered.** Shield (small) is unchanged. Shield Tiered gets Shield 1.1 on the texts it escalates to Shield Large, and we have not re-measured Tiered on the benchmark for this release.
- **Languages.** Most of the training data is English, and non-English coverage is uneven across languages.
- **Context.** Shield cannot see the system prompt or the conversation, so an input that is only an attack against one application's rules can look ordinary.
- **Latency.** The latency figures come from one staging server, not from production, and the figures for the longest texts rest on few completed requests. The hosted API adds network and gateway time.

7

Safety

Shield is a detector, not a permission system. Use it with least-privilege tool access, allow-lists for outbound destinations, and human approval for high-impact actions. A clean verdict should never grant an action that the application's own controls deny.

Keeping Shield Large's weights unpublished limits direct white-box attacks, such as gradient-based searches, but does not prevent attacks that transfer from other models or that probe the hosted API.

Paid requests retain no submitted content, on any model. With the free tier's research acknowledgement, flagged free requests may contribute redacted excerpts to security research for up to a year; Shield 1.1 runs only on paid plans, so this does not apply to it.

The generated attacks in the training data were written against canary destinations on reserved `.invalid` domains. This addendum reproduces no working attack.

8

Versioning

- **Shield 1:** Shield Large with the Shield 1 language model as a 4-bit build, described in the Shield 1 model card.

- **Shield 1.1:** Shield Large with the retrained language model, unquantized. Released on Thursday, 29 October 2026.

The model ID stays `shield-large`; requests do not need to change. Shield 1 Large cannot be selected after the switch. We keep Shield 1 Large's language model and its server build unchanged, so we can switch `shield-large` back to Shield 1 Large if a problem appears after release. If a later round replaces this language model, we will regenerate the numbers in this addendum from the same pipeline and version the addendum.

A Results for every dataset

Dataset	Texts	Shield 1 Large	Unquantized	Shield 1.1	Change
Agent attacks					
AgentInjectionBench (agentinject)	182	0.897	0.890	0.893	-0.004
Lakera b3 agent benchmark (b3_agent)	187	0.684	0.626	0.652	-0.032
BrowseSafe (browsesafe)	1,000	0.855	0.858	0.857	+0.002
DTap environment injections (dtap_env)	1,199	0.734	0.730	0.738	+0.004
DTap malicious skills (dtap_skill)	238	0.571	0.592	0.626	+0.055
DTap tool-description injections (dtap_tool)	300	0.667	0.647	0.670	+0.003
GEO injection in RAG (geo_rag)	1,570	0.859	0.859	0.877	+0.018
Gray Swan IPI Arena (ipi_arena)	120	0.938	0.956	0.963	+0.025
Kubernetes manifests (k8s)	851	0.981	0.980	0.980	-0.001
LongPIBench (longpib)	312	0.766	0.772	0.805	+0.039
MCP tool poisoning (mcp_desc)	74	0.950	0.950	0.950	0.000
MCPTox (mcptox)	827	0.865	0.823	0.851	-0.014
promptprotection agent flows (pp_agent_flows)	91	0.966	0.956	0.956	-0.010
RTC-Bench computer-use injections (rtc)	429	0.604	0.531	0.646	+0.042
Prompt injection					
Guardrail hard negatives (guardrail_hn)	135	0.897	0.923	0.923	+0.026
PIArena (piaarena)	857	0.991	0.990	0.991	0.000
promptprotection attacks (pp_attacks)	130	0.985	0.992	0.992	+0.008
promptgate-eval (promptgate)	269	0.963	0.966	0.966	+0.003
Jailbreaks					
DTap direct prompts (dtap_user)	600	0.815	0.833	0.828	+0.013
HackAPrompt 2.0 (Pliny track) (hackaprompt2)	503	0.911	0.917	0.907	-0.004
JailbreaksOverTime (jot)	1,713	0.979	0.985	0.985	+0.007
Simsonsun jailbreak prompts (simsonsun)	1,948	0.934	0.938	0.940	+0.006
Multilingual					
AltaySec Turkish injections (altay_tr)	227	0.925	0.930	0.934	+0.009
ChangeMore (Chinese) (changemore_zh)	1,963	0.942	0.940	0.939	-0.003

Dataset	Texts	Shield 1 Large	Unquantized	Shield 1.1	Change
KoreanGuardrail (korean_guardrail)	502	0.835	0.843	0.847	+0.012
NSFA multilingual queries (nsfa_ml)	1,580	0.771	0.780	0.786	+0.015
proba-ipi (Russian) (proba_ipi)	600	0.942	0.947	0.952	+0.010
Turkish over-refusal prompts (tr_overrefusal)	240	0.963	0.946	0.958	-0.004
XL-SafetyBench jailbreaks (six languages) (xl_jailbreak_ml)	900	0.991	0.993	0.993	+0.002
False alarms					
Agent configuration files (agent_configs)	499	0.978	0.982	0.982	+0.004
j2prompt-lint prompt templates (j2prompt)	128	0.930	0.977	0.977	+0.047
Live MCP tool descriptions (mcp_tools_real)	530	0.996	0.998	0.998	+0.002
Operant false-positive set (operant_fp)	562	0.948	0.957	0.964	+0.016
promptprotection hard benign (pp_benign_hard)	155	0.806	0.839	0.839	+0.032
Claude Code sessions (swe_chat)	580	0.947	0.972	0.978	+0.031

Table 13. Balanced accuracy on each headline dataset (higher is better). "Shield 1 Large" is the deployed 4-bit build; "Unquantized" is Shield 1's language model with no retraining; all three use Shield Large's combination with Shield Base and the rules.

B Provenance

Every number in this addendum is generated, not typed. The release scores come from one numbers file, read through a single data module that checks it against the evaluation analysis it cites. Development-history and development-panel numbers come from frozen analysis files, each verified by its hash, and method facts come from one facts file that records a source for each. The launch package's build generates every number in this addendum, the launch post, the technical report and the press kit from these files.

- Shield 1.1 benchmark numbers file: sha256
18f9e985f850c0082be44341e735d5292315ddbcb90068ba2e2b0db9d8e2c1a1
- Shield 1.1 evaluation analysis: sha256
58c89e30ec9491e11a6e119c8a516e0713b02bf4c65e0b9f00f6e8564cadfe80