

Shield 1.1: continued training of Shield Large

Technical report, 29 October 2026

Contents

	Abstract	3
1	Introduction	3
2	Background	3
3	Method	4
4	Evaluation methodology	7
5	Results	8
6	Development history	15
7	Limitations	18
8	Reproducibility statement	18
	References	19
A	Results for every dataset	19

Abstract

Shield Large, ZeroLeaks' most accurate prompt-injection detector, combines a Qwen3-1.7B classifier (its language model) with a smaller encoder, Shield Base. We retrained the language model on a newly labeled corpus focused on the text agents read, without regressing any benchmark category. Continued training on new data alone made the model forget jailbreak and non-English attacks. Three changes fixed this: replay of the original model's own training data, a teacher-agreement guard that distills from the original model except where a trusted label disagrees, and serving an equal-weight average of the original and retrained weights, a setting fixed in advance of the evaluation reported here. A pre-registered rule required the new model to beat Shield 1 Large in all five benchmark categories and by at least 0.005 overall, on both training seeds; four candidates failed it before Shield 1.1 passed. Shield Large's mean balanced accuracy rises from 0.904 to 0.915 on the 35 headline datasets of the Shield 1 benchmark, and the share of benign texts flagged falls from 5.4% to 4.1%. At a false-alarm rate of at most 1%, recall rises from 62.5% to 64.4%; at 5%, from 83.9% to 86.1%. A matched control shows that 53% of the gain comes from running the language model unquantized rather than as the 4-bit build Shield 1 serves; training's share is concentrated in agent attacks and multilingual text. The benchmark was used to choose between training rounds, so these are development results; a fresh held-out test will follow.

1 Introduction

Prompt-injection detectors sit where text enters an agent's context: tool results, retrieved pages, MCP tool descriptions, skill files and user messages. Shield 1, released in October 2026, offered four hosted tiers; its most accurate, Shield Large, scored 0.904 mean balanced accuracy on a benchmark of 35 datasets published in 2025 and 2026. Its weakest category was agent attacks, instructions hidden in what an agent reads, at 0.810.

We set out to improve Shield Large on agent text without giving up anything elsewhere. The constraint matters: in a security product, a gain on average that costs a category of attacks is a regression for every customer who depends on that category. We therefore fixed, before training, a rule that the new model had to clear in every category.

This report describes what we trained, how we measured it, what failed along the way, and what the result does and does not show. Appendix A gives results for every dataset.

2 Background

2.1 Shield Large

Shield Large scores a text with two classifiers. The language model is Qwen3-1.7B [8] fine-tuned with LoRA [2] as a binary classifier over windows of 512 tokens (stride 384, at most 16 windows); Shield Base is a multilingual encoder over windows of 256 tokens. Each takes its maximum over windows. The combined score is 0.6 times the language model's probability plus 0.4 times Shield

Base's, and a text is flagged when the score is at least 0.5 or when deterministic rules match. Shield 1.1 changes only the language model's weights; the combination, threshold, rules and windows are unchanged, and every score in this report is for that full combination unless stated otherwise.

2.2 The deployed build

Shield 1's hosted Shield Large serves its language model as a 4-bit build. All Shield 1 Large numbers in this report are for that deployed build, the comparator of record. Our evaluation runtime runs the language model unquantized, as a bfloat16 language model with its classification head and adapter kept in float32, on an NVIDIA L4. We therefore also score Shield 1's language model unquantized in the same runtime, which separates the effect of precision from the effect of training.

2.3 Label definition

The new corpus uses a label definition (v12) under which an attack is an instruction that tries to redirect an AI system across a trust boundary; a harmful topic alone is not an attack, and quoted attacks or security writing are benign in context. This definition is narrower than the benchmark's jailbreak category and Shield 1's behavior, which count jailbreak wrappers around harmful requests as attacks. The development history below explains why this mattered.

3 Method

3.1 Starting point

We continued from Shield 1 Large's language model rather than training a new one. We recovered the original training checkpoint and checked it against the deployed model's exported parameters and tokenizer before any training, then trained a fresh LoRA adapter of rank 32 together with the classification head.

3.2 Data

No customer data and no free-tier traffic went into training.

Accepted corpus. 254,006 training records (212,262 benign, 41,744 attacks), rendered as 252,461 training examples. It combines Shield's earlier public training sources relabeled under the new definition, public datasets first published in 2025 or 2026, a curated set of older injection collections, and 39,618 generated agent records. Each generated attack is an instruction inserted into a real benign text from a training split (skill files, tool outputs, MCP descriptions, web pages, agent traces, system prompts, chat and code), paired with a clean twin and a hard twin that puts a harmless instruction, or the attack quoted as data, at the same position. Destinations in generated payloads are on reserved `.invalid` domains.

Labels. Two frontier language models from different providers labeled every text with the same prompt, and a text was accepted only when both agreed; generated attacks also needed both evidence spans to overlap the inserted passage. On a reference set labeled by blind AI reviewers, the agreement rule admitted 193 correct labels out of 194, while keeping 67% of attacks, so it trades coverage for precision. The audits were done by blind AI reviewers as well: AI agents that re-labeled samples using only the label definition, without the accepted labels or the labeling models' answers. Audits of the generated data passed the pre-set bar of 95%: 150 of 150 for the first batch and 148 of 150 for the second, which had one AI reviewer instead of two, a documented deviation. An audit of the relabeled earlier corpus confirmed 413 of 415. The 2025 and 2026 public sources had no separate audit. An alternative rule that also admitted texts where one labeling model abstained reached only 136 of 150 and was not used.

Screening. Benchmark datasets, other splits of them and their source families were excluded. Training text was screened against every benchmark text by exact and normalized match, word eight-gram containment and MinHash similarity.

3.3 Replay

Continued training on the accepted corpus alone forgot attack styles the corpus barely covers (see Development history). We therefore mix in a replay of the data Shield 1 Large's language model was trained on: 118,581 rows (51,176 attacks, 67,405 benign), sampled across strata of jailbreak, multilingual, classic injection and agent-document attacks, and benign text that resembles each of them, plus 13,584 jailbreak and multilingual attack rows re-admitted in a later round. Replay rows keep their source labels; rows that an earlier independent audit disputed were dropped. Replay is screened against the benchmark like the rest of the data. In the final recipe, replay attack rows in the jailbreak and multilingual strata carry 1.5 times their earlier weight, with benign replay rescaled to keep the classes balanced. Rehearsal of earlier data is a standard defense against forgetting [5].

3.4 Loss and the teacher-agreement guard

Each example contributes cross-entropy with its label plus 0.25 times the KL divergence from the original model's cached prediction, in the spirit of distillation [3] and learning without forgetting [4], with origin and class weights that balance attack and benign mass. Distilling from the original model everywhere has a cost: the original calls 16,590 of the 210,730 accepted benign examples attacks, and plain distillation would pull each of them toward a false alarm. The guard resolves each example:

- teacher and label agree: train on both;
- they disagree and the label is trusted: train on the label alone;
- they disagree and the label is untrusted: leave the example out.

Trusted labels are those both labeling models agreed on, and replay labels confirmed by an independent audit. Untrusted labels are unaudited replay rows, the re-admitted jailbreak rows (so they rehearse only the original model's own attack verdicts) and, from the final round, accepted benign texts on chat surfaces. That last rule leaves out 3,490 disputed chat examples and keeps the 13,100 disputed agent-surface examples, which carry most of the agent and false-alarm gains.

Training runs one pass over 384,697 examples (48,886,732 tokens) in 1,487 steps of about 32,768 tokens, learning rate 1e-5 with linear decay, on one NVIDIA L4 per seed, for seeds 17 and 42.

3.5 Weight-space blend

The served language model is an equal-weight average of the original and the retrained model: the LoRA update is scaled by 0.5 and the classification head is interpolated by the same amount. Interpolating between a pretrained and a fine-tuned model is known to recover robustness lost in fine-tuning while keeping much of the gain [6, 7]. We pre-registered this one setting for a single evaluation of R11, kept it after that result, and did not tune it on the benchmark, so it was fixed in advance of the R12 evaluation reported here.

3.6 Serving

The benchmark scores in this report were measured in the evaluation runtime using a bfloat16 language model with its classification head and adapter kept in float32. Shield 1 Large serves a 4-bit build. Shield 1.1 serves the language model in float32, without quantization. The served build is checked against the evaluated scores before release.

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 both.

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 1. 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 at this load it completes 1.8 times as many of them per second (25.2 against 13.9). Shield Tiered answers the texts it escalates in a little over half the time. The longest texts take roughly the same time, but these figures 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 second Shield 1.1 run gave a median of 12.3 seconds (9 of 39 completed).

The unquantized model is larger, so a new server takes longer to become ready: 196 seconds after the container starts, and 197 in a second run, against 141 for Shield 1. That affects how quickly capacity can be added, not request latency. During startup, the serving container's memory, not

counting reclaimable file cache, peaked at 53% of its limit (Shield 1: 40%), with no out-of-memory events, and GPU memory use peaked at 37% of the NVIDIA L4's memory (Shield 1: 10%). 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.

4 Evaluation methodology

4.1 Benchmark and metric

The benchmark is the one published with Shield 1: 35 headline datasets (22,001 texts) in five categories (agent attacks, prompt injection, jailbreaks, multilingual, false alarms), plus 8 datasets that are scored but carry no weight. Each dataset scores balanced accuracy; categories average their datasets, and the overall score averages the five categories, so a large dataset cannot dominate. We also pool attack recall and the benign flag rate over all headline texts. Benchmark datasets were kept out of training, and the training text was screened against every benchmark text (see Data).

4.2 Pre-registered replacement rule

On 15 October 2026, before any teacher or student run, we froze the comparison: for each training seed, Shield Large with the new language model must score above the deployed Shield 1 Large in every category, and at least 0.005 higher overall. The rule compares point estimates computed with exact rational arithmetic; it does not test significance, and the analysis marks the release as not statistically qualified because datasets share sources and are not independent samples.

4.3 Matched control

Because the evaluation runtime differs from the deployed build in precision, each evaluation also scores Shield 1's language model unquantized in the same runtime. Deployed versus unquantized measures precision and runtime together; unquantized versus Shield 1.1 measures training.

4.4 Development panel and selection rule

After the replacement rule was met, we stopped using the benchmark to choose settings. A development panel of 13,754 texts in 104 datasets (6,611 attacks, 7,143 benign) was built from development and held-out sources, excluding every benchmark dataset, other splits of benchmark sources, and every text overlapping the benchmark or training data. A selection rule written before any development result required, for both seeds, every category at or above the unquantized original, then took the setting with the best mean overall gain, preferring the smaller blend weight within 0.001. The blend weight in use had been fixed in advance of the R12 evaluation, so the panel could confirm it but did not choose it. The shipped seed was chosen afterwards on the same panel, by its development gain alone, under a rule written on 27 October 2026, after both seeds' benchmark scores were known.

4.5 Disclosure

The benchmark informed every training change from the first failed round onward. It is therefore development data for Shield 1.1, not a held-out test. This benchmark (43 sets, 35 headline, five categories) was also used to choose between training rounds; a fresh held-out test follows.

5 Results

5.1 Main result

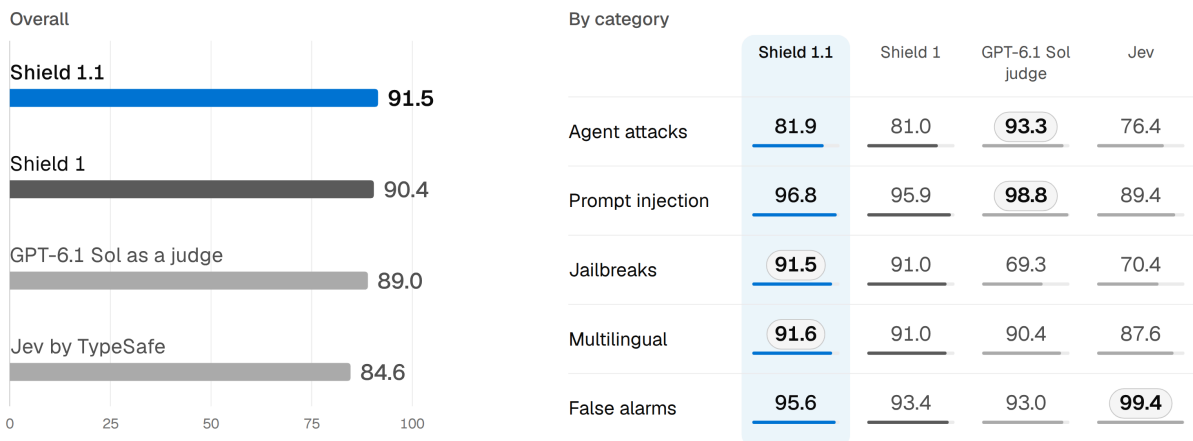
Category	Shield 1 Large (4-bit)	Shield 1 Large, unquantized	Shield 1.1	Total change	From running unquantized	From training
Agent attacks	0.810	0.798	0.819	+0.009	-0.012	+0.021
Prompt injection	0.959	0.968	0.968	+0.009	+0.009	0.000
Jailbreaks	0.910	0.918	0.915	+0.006	+0.008	-0.003
Multilingual	0.910	0.911	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 2. Balanced accuracy by category (higher is better) for Shield Large with each language model. Differences are computed from unrounded scores.

The overall gain is +0.010. The unquantized original accounts for +0.006 of it (53%) and training for +0.005. The split differs sharply by category. Running unquantized accounts for most of the gain on prompt injection, jailbreaks and false alarms, but costs agent attacks 0.012. Training adds +0.021 on agent attacks, +0.004 on multilingual text, 0.000 on prompt injection and +0.002 on false alarms. On jailbreaks (-0.003), training is slightly negative against the unquantized original, though Shield Large remains above Shield 1 Large there.

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.



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.

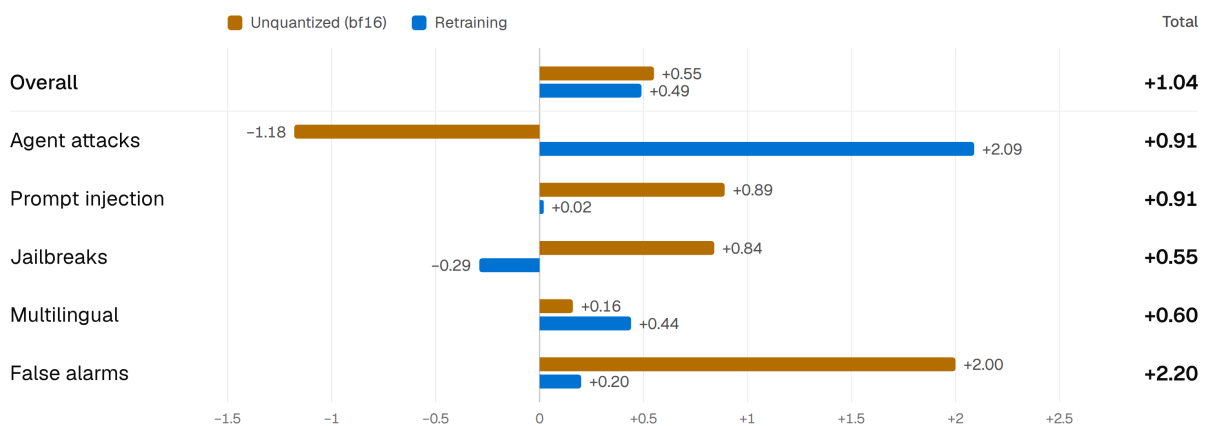
ZeroLeaks

zeroleaks.ai

Figure 1. 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.

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 2. 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.

5.2 Detection rates and paired changes

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%

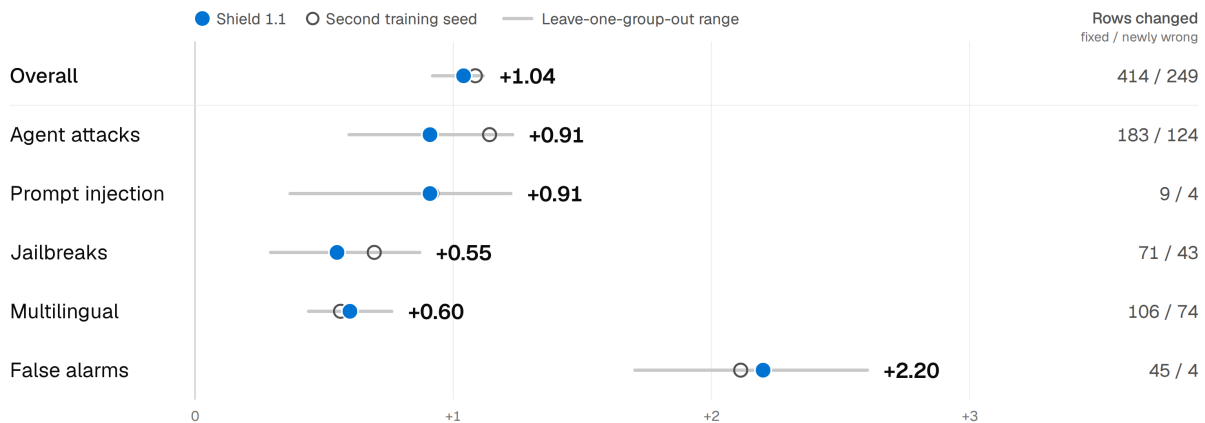
Table 3. Pooled detection rates over the headline sets: attacks flagged (higher is better) and benign texts flagged (lower is better).

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 4. Text-level changes against Shield 1 Large on the headline sets.

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 3. 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, attack recall goes from 84.5% to 84.8% and the benign flag rate from 5.4% to 4.1%, a relative reduction of 25% in false alarms. Text by text, Shield 1.1 corrects 414 of Shield 1 Large's errors and introduces 249 new ones; seed 17 gives 431 and 244.

5.3 Recall at fixed false-alarm rates

Balanced accuracy and the pooled rates above use the default threshold. A deployment that can accept only a fixed share of false alarms needs recall at that share, and recent work on injection detectors recommends reporting it [9]. For each model we keep Shield Large's combination and rules and move only the threshold on the combined score. Pooling the 9,559 benign texts of the headline sets, we take the lowest threshold at which the false-alarm rate is at most the target; the 19 benign texts that the rules flag count against the target at every threshold. We then report recall over the 12,442 attack texts at that threshold, with the false-alarm rate it achieves. The threshold is set on the same benign texts it is evaluated on, so these are descriptive operating points, not thresholds calibrated for new traffic. The deployed model's and Shield Base's probabilities are stored rounded, but no text's score lies within that rounding of either threshold, so these rates are exact for the recorded scores.

Shield Large with	Recall, false alarms at most 1%	False-alarm rate	Recall, false alarms at most 5%	False-alarm rate	Recall at the default threshold	False-alarm rate
Shield 1 language model, 4-bit (Shield 1 Large)	62.5%	0.99%	83.9%	4.99%	84.5%	5.43%
Shield 1 language model, unquantized	64.4%	0.99%	84.7%	4.99%	83.7%	4.47%
Shield 1.1	64.4%	0.99%	86.1%	4.99%	84.8%	4.08%
Shield 1.1, seed 17 (not shipped)	64.6%	0.99%	86.3%	4.99%	84.9%	4.09%

Table 5. Recall (higher is better) at fixed false-alarm rates, pooled over the headline sets. Each threshold is set on the benign texts so that the false-alarm rate is at most the target; the achieved rate is shown next to each recall. Shield Large's combination with Shield Base and its rules are kept throughout.

At a false-alarm rate of at most 1%, Shield 1.1 catches 64.4% of attacks, against 62.5% for Shield 1 Large. That is 238 more attacks caught: 242 more from running the language model unquantized and 4 fewer from training. At this strict setting, then, the gain over Shield 1 Large comes from running unquantized, not from training. At 5%, recall is 86.1% against 83.9%. That is 278 more attacks caught: 100 from running the language model unquantized and 178 from training. The seed we did not ship catches 64.6% and 86.3% at the two settings.

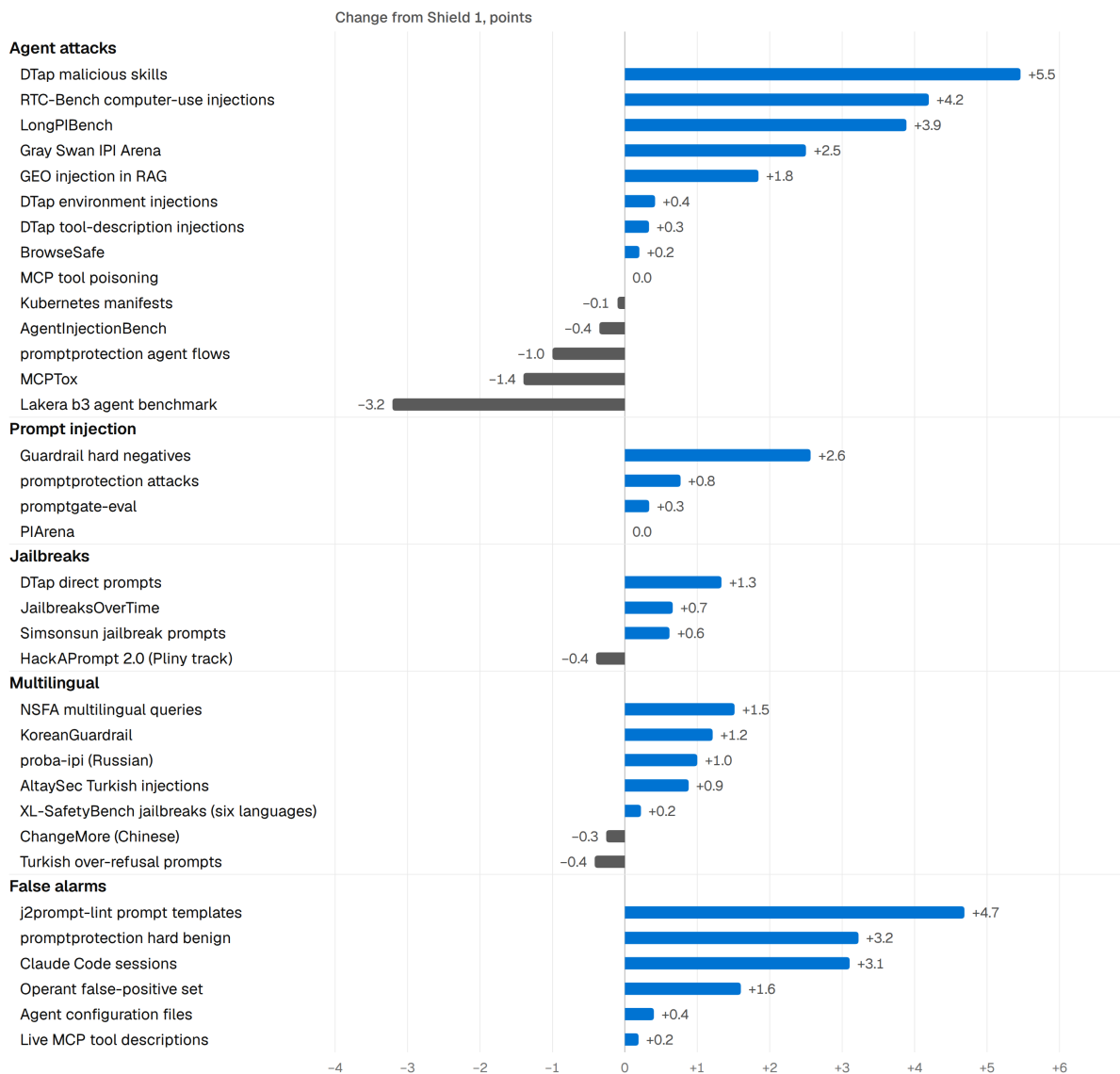
Detectors that return only a yes/no verdict, such as an LLM judge asked for a yes or no answer, have a single operating point and cannot be placed at these false-alarm rates. Some of the other detectors in the comparison do return scores; we have not computed their operating points for this report.

5.4 Per-dataset breakdown

Shield 1.1 is above Shield 1 Large on 25 headline datasets, equal on 2 and below on 8. The largest gain is on DTap malicious skills (+0.055); the largest loss is on the Lakera b3 agent benchmark (−0.032).

Shield 1.1 is higher on 25 of 35 headline sets and lower on 8

Change in balanced accuracy from Shield 1 on each headline set, points.
Blue: higher. Gray: lower. Sets are grouped by category and sorted by change.



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. Change on each headline dataset. Balanced accuracy of Shield 1.1 minus Shield 1 on each of the 35 headline datasets. Above zero means Shield 1.1 is better.

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 6. Headline datasets where Shield 1.1 scores below Shield 1 Large, with the unquantized original for reference.

On seven of the eight regressions the unquantized original was already below the deployed build, so running unquantized explains part of each drop. Appendix A lists every dataset.

5.5 Development panel

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 7. Development-panel change in balanced accuracy for the shipped seed against the unquantized original, at the selected setting (above zero favors Shield 1.1). Paired 95% intervals use a paired normal approximation stratified by dataset and class, which treats texts within a dataset as independent.

Blend weight on the new model	Mean development change, overall	Every category at or above the original, both seeds
0.40	+0.0033	yes
0.50	+0.0039	yes
0.60	+0.0042	yes
0.70	+0.0042	yes
0.80	+0.0041	yes
1.00	+0.0048	yes

Table 8. The pre-registered blend sweep on the development panel, at the shipped combination weight, mean over seeds. Every blend weight met the eligibility condition.

The best eligible setting gave +0.0048 at a blend weight of 1.00. The rule takes the smallest weight within 0.001 of the best eligible setting; that was 0.50 (+0.0039), the weight already in use. Development gains over the unquantized original are modest: the overall interval is +0.0021 to +0.0057, and the intervals for jailbreaks and false alarms include zero.

5.6 Comparison with other detectors

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 detector with verified predictions on all 43 sets, scored with the same code. 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). The benchmark's use in model selection 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 has the highest overall score of the 16 detectors. It trails on agent attacks, where GPT-6.1 Sol as a judge scores 93.3% against 81.9%; prompt injection, where GPT-6.1 Sol as a judge scores 98.8% against 96.8%; and false alarms, where NVIDIA NemoGuard scores 100.0% against 95.6% but catches 2.0% of attacks, against 84.8%. The comparison runs date from 27 September 2026 to 9 October 2026; the model card addendum lists how each detector was run. LLM judges answer our own prompt, so their scores depend on that prompt.

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.

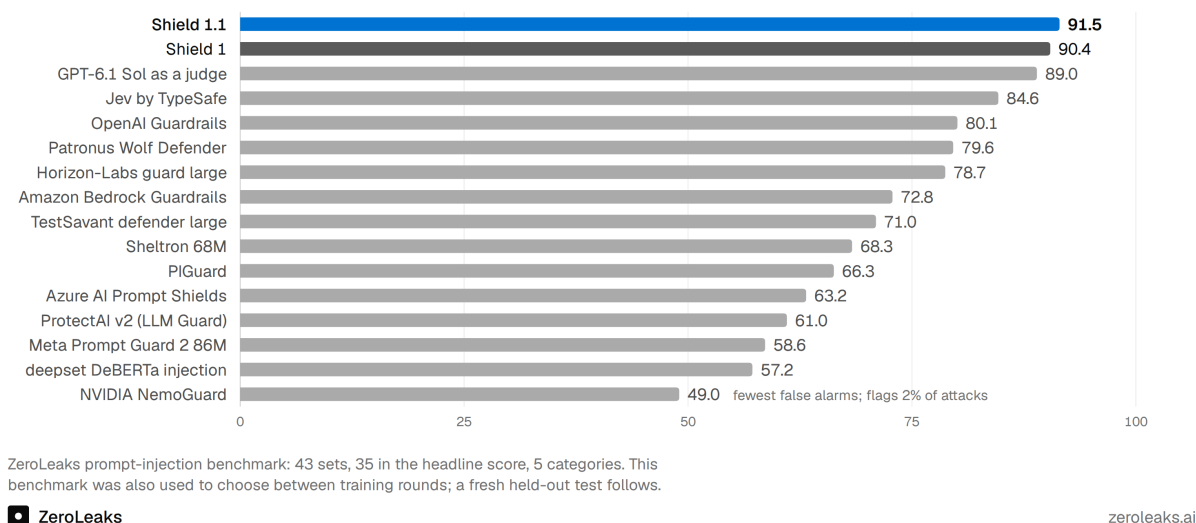


Figure 5. Every detector on the benchmark. Overall score of every detector with verified predictions on all 43 sets. Higher is better.

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 both seeds on the benchmark.

The two seeds differ by at most 0.002 in any category. We ship seed 42, chosen on the development panel (+0.00390 against +0.00385 overall), not on the benchmark. The rule that chose it uses only the 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).

6 Development history

Runs numbered before R9 ended without a complete, verified model (cloud capacity and packaging failures), so R9 is the first round with results. Every round used the same replacement rule and the same evaluation runtime.

Round	Agent attacks	Prompt injection	Jailbreaks	Multilingual	False alarms	Overall	Gate
R9	+0.015 / +0.010	+0.005 / +0.010	-0.046 / -0.056	0.000 / -0.002	+0.024 / +0.026	-0.001 / -0.002	failed
R10	+0.015 / +0.013	+0.005 / +0.012	-0.036 / -0.050	-0.001 / -0.004	+0.025 / +0.025	+0.002 / -0.001	failed
R11	+0.020 / +0.017	+0.011 / +0.013	-0.029 / -0.039	-0.003 / -0.003	+0.025 / +0.026	+0.005 / +0.002	failed
R11, 50/50 blend	+0.010 / +0.006	+0.012 / +0.013	-0.003 / -0.004	+0.005 / +0.006	+0.021 / +0.022	+0.009 / +0.009	failed
R12, 50/50 blend	+0.011 / +0.009	+0.009 / +0.009	+0.007 / +0.006	+0.006 / +0.006	+0.021 / +0.022	+0.011 / +0.010	passed

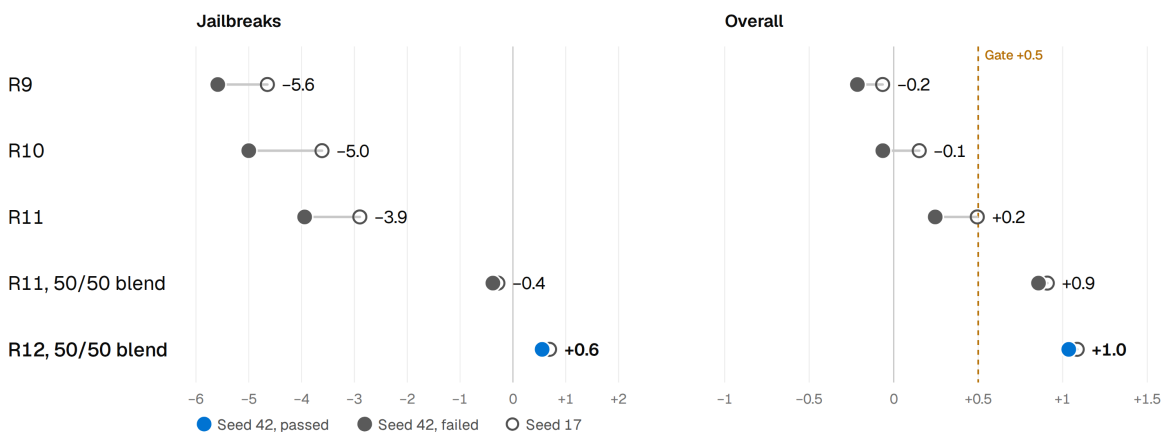
Table 11. Category and overall change against Shield 1 Large (above zero is better), seed 17 / seed 42, for each round. The two blend rows serve an equal-weight average of the original and the retrained weights.

	Agent attacks	Prompt injection	Jailbreaks	Multilingual	False alarms	Overall
Unquantized original minus Shield 1 Large	-0.012	+0.009	+0.008	+0.002	+0.020	+0.005

Table 12. The matched control, identical in every round.

Four candidates failed the gate before Shield 1.1 passed

Change from Shield 1 in balanced accuracy, points. Dot: seed 42. Ring: seed 17. The gate needs every category above zero and the overall at least +0.5 points on both seeds. Jailbreaks failed in every earlier candidate.



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 6. Candidates against the gate. Jailbreak and overall change from Shield 1 for each candidate, both seeds. The gate needs every category above zero and the overall at least 0.005 higher.

Round	What changed	What happened
R9	Continued training on the accepted corpus only, with the original model as teacher.	Agent attacks and false alarms improved; jailbreaks and non-English attacks were forgotten.
R10	Added replay of Shield 1's own training data and the teacher-agreement guard.	Replay recovered part of the forgetting and kept every gain; jailbreaks still failed. Root cause: the new label definition treats jailbreak wrappers around harmful requests as benign, and the audit had removed exactly those replay sources.
R11	Re-admitted the removed jailbreak attack sources as untrusted replay, trained only where the original model also calls them attacks.	The jailbreak gap narrowed and every gain grew; jailbreaks and multilingual still below Shield 1.
R11, 50/50 blend	Evaluated R11 once as a 50/50 weight blend with the original model, a weight pre-registered for that single evaluation.	Every category passed except jailbreaks, by a small margin. The blend weight stayed fixed at one half; the next change had to be in training.
R12, 50/50 blend	Stopped training benign chat rows against the original model's attack verdict, and gave jailbreak and multilingual replay attacks more weight.	Passed the gate on both seeds.

Table 13. What changed in each round and what it showed.

R9: new data alone forgets. Continuing on the accepted corpus with the original model as teacher improved agent attacks and false alarms, but jailbreaks fell by 0.046 and 0.056 against Shield 1 Large, and multilingual detection barely moved. The losses were concentrated in classic jailbreaks and Chinese attacks, styles the accepted corpus barely contains and the cached teacher outputs did not cover. We also found that plain distillation pulled every benign text the original model disputes toward a false alarm.

R10: replay and the guard. Replay recovered much of the forgetting and kept every gain, and the guard removed the false-alarm pressure. Jailbreaks still failed ($-0.036 / -0.050$). The cause was the label definition: the audit that cleaned the replay data applied the new definition, under which jailbreak wrappers around harmful requests are benign, and so it had removed exactly the jailbreak sources the benchmark counts as attacks.

R11: re-admitting jailbreak sources. We re-admitted those sources as untrusted replay, trained only where the original model also calls them attacks. Every gain grew and the jailbreak gap narrowed ($-0.029 / -0.039$), but jailbreaks and multilingual still failed.

R11 with the blend. After the R11 results, we pre-registered a single evaluation of R11 as an equal-weight average with the original. Every category except jailbreaks passed ($-0.003 / -0.004$). As pre-registered, we did not tune the blend weight further; the next change had to be in training.

R12: chat-surface trust and replay weight. Accepted benign chat texts that the original model calls attacks are often jailbreak-wrapped harmful requests; training them toward benign taught the model to pass jailbreaks, so they became untrusted. Jailbreak and multilingual replay attacks were weighted up. At the fixed blend, both seeds passed ($+0.011 / +0.010$ overall).

Approaches we did not ship. Before continuing Shield 1's own model, we trained new detectors on the accepted corpus, including some initialized from Shield's own encoders. None beat Shield 1 Large in every category, which is why Shield 1.1 continues Shield 1's language model. In the development sweep, giving the retrained model more weight in the blend produced slightly larger development gains, but the pre-registered rule kept the equal-weight setting.

7 Limitations

- **Benchmark reuse.** The benchmark guided training changes across five evaluations. A fresh held-out test is needed before these numbers can be read as generalization.
- **No significance test.** The replacement rule and the reported benchmark differences are point estimates. The development-panel intervals treat texts within a dataset as independent, which understates uncertainty when texts share sources.
- **Small margins.** The smallest category gain is +0.006 (jailbreaks), and the development intervals for jailbreaks and false alarms include zero.
- **Serving.** The scores were measured unquantized in our runtime. Small numeric differences between that runtime and the served build can change verdicts near the threshold; the served build is checked before release.
- **Regressions.** Shield 1.1 is worse than Shield 1 Large on 8 headline datasets, 5 of them in agent attacks, and 4 datasets remain below 0.70.
- **Label definition.** The corpus's definition differs from the benchmark's jailbreak category. We kept Shield 1's jailbreak behavior through replay and the guard; a definition that matches product behavior would be cleaner.
- **Not measured here.** Adaptive and gradient-based attacks and Shield Tiered's accuracy were not re-measured for this release. Latency was measured on one staging server, not in production, and the figures for the longest texts rest on few completed requests.

8 Reproducibility statement

The benchmark datasets are public, and the Shield 1 model card lists their sources and licenses; Appendix A gives every per-dataset score. The training corpus includes relabeled and generated data that we do not publish, and Shield Large's weights are not published, so the training cannot be reproduced externally. Internally, every input is pinned: the starting checkpoint, the training and replay manifests, the cached teacher predictions, each seed's training run and each evaluation are bound by hashes, and the evaluation recomputes every score with exact rational arithmetic from per-text predictions.

The model card for Shield 1.1 follows the reporting purposes of [1]. Every number in this report is generated by the launch package's build from the Shield 1.1 benchmark numbers file (sha256 18f8e985f850c0082be44341e735d5292315ddbcb90068ba2e2b0db9d8e2c1a1), which the build checks against the Shield 1.1 evaluation analysis it cites (sha256 58c89e30ec9491e11a6e119c8a516e0713b02bf4c65e0b9f00f6e8564cadfe80), and from frozen per-round analyses whose hashes are verified before reading. The same build produces the model card, the launch post and the press kit, so all four carry the same numbers.

References

- [1] M. Mitchell et al. Model Cards for Model Reporting. FAT* 2019. arXiv:1810.03993.
- [2] E. J. Hu et al. LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022. arXiv:2106.09685.
- [3] G. Hinton, O. Vinyals, J. Dean. Distilling the Knowledge in a Neural Network. 2015. arXiv:1503.02531.
- [4] Z. Li, D. Hoiem. Learning without Forgetting. ECCV 2016. arXiv:1606.09282.
- [5] D. Rolnick et al. Experience Replay for Continual Learning. NeurIPS 2019. arXiv:1811.11682.
- [6] M. Wortsman et al. Robust Fine-Tuning of Zero-Shot Models. CVPR 2022. arXiv:2109.01903.
- [7] G. Ilharco et al. Editing Models with Task Arithmetic. ICLR 2023. arXiv:2212.04089.
- [8] Qwen Team. Qwen3 Technical Report. 2025. arXiv:2505.09388.
- [9] Z. Liu. Passing the Test You Trained On: Re-evaluating Prompt-Injection Detectors for LLM Agents. 2026. arXiv:2610.03448.

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

Dataset	Texts	Shield 1 Large	Unquantized	Shield 1.1	Change
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
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 14. Balanced accuracy (higher is better) on each headline dataset for Shield Large with the deployed Shield 1 language model, the same model unquantized, and Shield 1.1.