

GuardrailProbe Independent Benchmark — June 2026

Run ID	3acea672-6458-4c77-868c-22b038b3e695
Timestamp	2026-06-27T17:58:34.253698+00:00
Generated	2026-06-27T18:18:55Z
Backends tested	nemo, lakera, azure_content_safety, aws_bedrock, llama_firewall, llm_guard, guardrails_ai, presidio, openai_moderation, azure_prompt_shields
Best overall	llama_firewall
Worst overall	azure_content_safety

TL;DR

Winner	llama_firewall
Best accuracy/latency ratio	llm_guard
Biggest improvement vs last month	—
Biggest regression vs last month	—
Backends tested	10
Backends skipped	0
Total probes run	684
Report generated	2026-06-27 17:58 UTC
Run ID	3acea672-6458-4c77-868c-22b038b3e695

Overall Comparison

Backend	Overall %	vs Last Month	Best Category	Worst Category	Avg Latency
nemo	83.3%	—	LLM01	LLM10	30148 ms
lakera	83.3%	—	LLM01	LLM10	769 ms
azure_content_safety	25.6%	—	LLM02	LLM01	628 ms
aws_bedrock	59.0%	—	LLM01	LLM10	439 ms
llama_firewall	85.9%	—	LLM01	LLM10	1538 ms
llm_guard	80.0%	—	LLM01	LLM10	0 ms
guardrails_ai	2.6%	—	LLM01	LLM02	390 ms
presidio	6.4%	—	LLM02	LLM01	1447 ms
openai_moderation	100.0%	—	LLM01	LLM01	33289 ms
azure_prompt_shields	24.4%	—	LLM01	LLM09	680 ms

Per-Category Results (OWASP LLM Top 10)

Ref	Description	Winner	Score	Runner-up	Score
LLM01	Prompt Injection	nemo	100%	lakera	100%
LLM02	Insecure Output	lakera	81%	llama_firewall	81%

LLM03	Training Data Poisoning	nemo	100%	lakera	100%
LLM04	Model DoS	lakera	67%	llama_firewall	67%
LLM05	Supply Chain	nemo	100%	llama_firewall	100%
LLM06	Sensitive Info Disclosure	nemo	100%	lakera	100%
LLM07	Insecure Plugin	nemo	83%	llama_firewall	83%
LLM08	Excessive Agency	nemo	100%	lakera	100%
LLM09	Overreliance	nemo	100%	lakera	100%
LLM10	Model Theft	nemo	20%	lakera	20%

Content Moderation Results

Backend	Hate	Violence	Sexual	Self-Harm	Overall CM
nemo	—	—	—	—	—
lakera	—	—	—	—	—
azure_content_safety	—	—	—	—	—
aws_bedrock	—	—	—	—	—
llama_firewall	—	—	—	—	—
llm_guard	—	—	—	—	—
guardrails_ai	—	—	—	—	—
presidio	—	—	—	—	—
openai_moderation	—	—	—	—	—
azure_prompt_shields	—	—	—	—	—

Backend Capability Matrix

Backend	Prompt Injection	Jailbreak	Content Mod.	PII Detection	Agentic Safety
NeMo Guardrails	Primary	Yes	No	No	Yes
GuardrailsAI	Yes	Yes	No	Yes	No
Presidio	No	No	No	Primary	No
Lakera Guard	Primary	Yes	No	No	No
Custom HTTP	Yes	Yes	No	No	No
OpenAI Moderation	No	Yes	Primary	No	No
Azure Content Safety	No	No	Primary	No	No
Azure Prompt Shields	Primary	Yes	No	No	No
AWS Bedrock	Yes	Yes	Yes	No	No
Llama Firewall	Primary	Yes	No	No	No
LLM Guard	Yes	Yes	Yes	Yes	No

Primary = core strength | Yes = supported | No = not designed for this

Accuracy vs Latency Tradeoff

Backend	Overall %	Avg Latency	Latency Category	Recommended For
llm_guard	80.0%	0 ms	Ultra-fast	Real-time, high-throughput pipelines
guardrails_ai	2.6%	390 ms	Moderate	Batch processing, async pipelines
aws_bedrock	59.0%	439 ms	Moderate	Batch processing, async pipelines
azure_content_safety	25.6%	628 ms	Moderate	Batch processing, async pipelines
azure_prompt_shields	24.4%	680 ms	Moderate	Batch processing, async pipelines
lakera	83.3%	769 ms	Moderate	Batch processing, async pipelines
presidio	6.4%	1447 ms	Slow	Offline analysis, compliance audits
llama_firewall	85.9%	1538 ms	Slow	Offline analysis, compliance audits
nemo	83.3%	30148 ms	Slow	Offline analysis, compliance audits
openai_moderation	100.0%	33289 ms	Slow	Offline analysis, compliance audits

Notable Bypasses

No universal bypasses detected — or probe results not available for this regenerated report.

Backends Skipped This Month

Backend	Reason	Expected In
—	—	—

Month-over-Month Changes

First benchmark — no prior month comparison available.

How to Reproduce

```
pip install guardrailprobe
guardrailprobe run --year 2026 --month 6
```

Full guide: github.com/askuma/guardrailprobe/blob/main/METHODOLOGY.md

Independence Statement

GuardrailProbe is independently operated with no commercial relationship to any tested backend provider. All probes are open-source and methodology is publicly auditable. Test results reflect backend configurations at the time of testing only.