

GuardrailProbe Independent Benchmark — July 2026

Run ID	834bb9d2-b56d-44b6-82d7-acb630cca1fe
Timestamp	2026-07-03T08:57:56.654285+00:00
Generated	2026-07-03T09:06:33Z
Backends tested	nemo, lakera, azure_content_safety, aws_bedrock, llama_firewall, llm_guard, guardrails_ai, presidio, openai_moderation, azure_prompt_shields
Best overall	llm_guard
Worst overall	llama_firewall

TL;DR

Winner	llm_guard
Best accuracy/latency ratio	aws_bedrock
Biggest improvement vs last month	—
Biggest regression vs last month	—
Backends tested	10
Backends skipped	0
Total probes run	707
Report generated	2026-07-03 08:57 UTC
Run ID	834bb9d2-b56d-44b6-82d7-acb630cca1fe

Overall Comparison

Backend	Overall %	vs Last Month	Best Category	Worst Category	Avg Latency
nemo	84.6%	—	LLM01	LLM10	32189 ms
lakera	82.0%	—	LLM01	LLM10	750 ms
azure_content_safety	25.6%	—	LLM02	LLM01	644 ms
aws_bedrock	59.0%	—	LLM01	LLM10	412 ms
llama_firewall	11.5%	—	LLM01	LLM05	5158 ms
llm_guard	85.9%	—	LLM01	LLM10	2784 ms
guardrails_ai	2.6%	—	LLM01	LLM02	344 ms
presidio	6.4%	—	LLM02	LLM01	1618 ms
openai_moderation	100.0%	—	LLM01	LLM01	33905 ms
azure_prompt_shields	24.4%	—	LLM01	LLM09	656 ms

Per-Category Results (OWASP LLM Top 10)

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

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

Content Moderation Results

Backend	Hate	Violence	Sexual	Self-Harm	Overall CM
nemo	100%	100%	100%	100%	100%
lakera	100%	100%	100%	100%	100%
azure_content_safety	100%	100%	100%	80%	95%
aws_bedrock	100%	100%	100%	100%	100%
llama_firewall	0%	0%	20%	0%	5%
llm_guard	100%	100%	100%	100%	100%
guardrails_ai	0%	0%	0%	0%	0%
presidio	0%	0%	0%	0%	0%
openai_moderation	0%	0%	0%	0%	0%
azure_prompt_shields	0%	0%	0%	0%	0%

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
guardrails_ai	2.6%	344 ms	Moderate	Batch processing, async pipelines
aws_bedrock	59.0%	412 ms	Moderate	Batch processing, async pipelines
azure_content_safety	25.6%	644 ms	Moderate	Batch processing, async pipelines
azure_prompt_shields	24.4%	656 ms	Moderate	Batch processing, async pipelines
lakera	82.0%	750 ms	Moderate	Batch processing, async pipelines
presidio	6.4%	1618 ms	Slow	Offline analysis, compliance audits
llm_guard	85.9%	2784 ms	Slow	Offline analysis, compliance audits
llama_firewall	11.5%	5158 ms	Slow	Offline analysis, compliance audits
nemo	84.6%	32189 ms	Slow	Offline analysis, compliance audits
openai_moderation	100.0%	33905 ms	Slow	Offline analysis, compliance audits

Notable Bypasses

OWASP Category	Severity	Count
LLM04	high	1
LLM04	medium	1
LLM10	high	3
LLM10	medium	1

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 7
```

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.