Benchmarking AI Guardrails

Understanding Guardrail Comparisons

Implementing guardrails is a foundational practice to secure and govern user and agentic interactions with LLM-based applications, extending across input and output filters to safeguards for reliability. Guardrails can be used to enforce ethical use and compliance policies, prevent data exposure and leakage, detect attempts to exploit the LLM's inherent vulnerability through a prompt injection, as well as identify known techniques to jailbreak the LLM.

For guardrails to accurately detect attacks and consistently enforce policies for content safety, for example, they need to combine speed of analysis with low latency in delivering a response. Guardrail design, as a result, is often a trade off between detection accuracy (low false positives and low false negatives) and speed (low latency and fast response times).

In our benchmarking analysis, we measured how Vijil Dome compares with leading hyperscaler native guardrails (AWS Bedrock, Azure AI, Google Cloud Model Armor, and Nvidia NemoGuard) for prompt injection detections, content moderation, and data privacy filters.

Understanding The Trust Score Uplift

Our performance comparison includes both standard metrics for performance as well as Vijil's Trust Score™, a multi-dimensional, comprehensive evaluation for reliability, security and safety.
The Trust Score captures how reliably an agent performs, how securely it resists attacks, and how safely it operates within boundaries. Reliability includes evaluating for hallucinations, logical errors, instruction drift, and whether the agent behaves consistently across similar conditions.

Evaluating models using the Trust Score methodology allows teams to assess the improvement from implementing guardrails. The Trust Score reflects relative improvements based on whether guardrails:

detect prompt injections before they happen

flag and block inappropriate content

obfuscate or censor personally identifiable information PII (if feasible)

Vijil's methodology positively acknowledges a model's refusal to respond to adversarial questions or prompts, resulting in an increase in the Trust Score.

We also measure reliability to ensure that guardrail false positives do not reduce the performance or accuracy of the model.

The Vijil Advantage:
Speed & Consistency

The results of our benchmarking comparisons demonstrates that Vijil Dome's guardrails perform best in combined performance efficiency. While delivering top-tier Trust Score improvements comparable to the highest performers, Vijil Dome maintains significantly lower latency overhead. Crucially, Vijil maintains consistent 99th percentile latency, ensuring reliability where comparable native guardrails experience significant lag spikes.

While other guardrails force a trade off between security and speed, Vijil Dome delivers high Vijil Trust Scores™ (Reliability + Security + Safety) with minimal impact to application performance.

Head to Head Performance Data

The following table benchmarks Trust Score Uplift against Latency (Speed/Overhead):

Method	Trust Score Improvement	Input Latency p50	Input Latency p99	Output Latency p50	Output Latency p99
Vijil Dome	13.67	0.15	0.22	0.16	0.19
Azure AI Safety	4.53	0.17	0.38	0.08	0.29
AWS Bedrock Guardrails	14.82	0.51	0.98	0.56	0.67
GCP Model Armor	14.16	0.11	0.58	0.13	0.37
Nvidia NemoGuard	11.51	0.322	1.79	0.43	1.67

Note on Configuration: All benchmarks utilized Llama 3.3 70B as the base model. To ensure a direct comparison of real world performance, Vijil Dome was hosted on an Nvidia T4 GPU instance and accessed via standard APIs, just like all comparable solutions. All guardrails were evaluated using default out of the box configurations to cover reliability, safety, and security, typically including prompt injection detection, content safety, and PII filters. For Nvidia NemoGuard, we specifically utilized the Llama 3.1 NemoGuard 8B ContentSafety NIM and NemoGuard JailbreakDetect NIM.

Competitive Breakdown

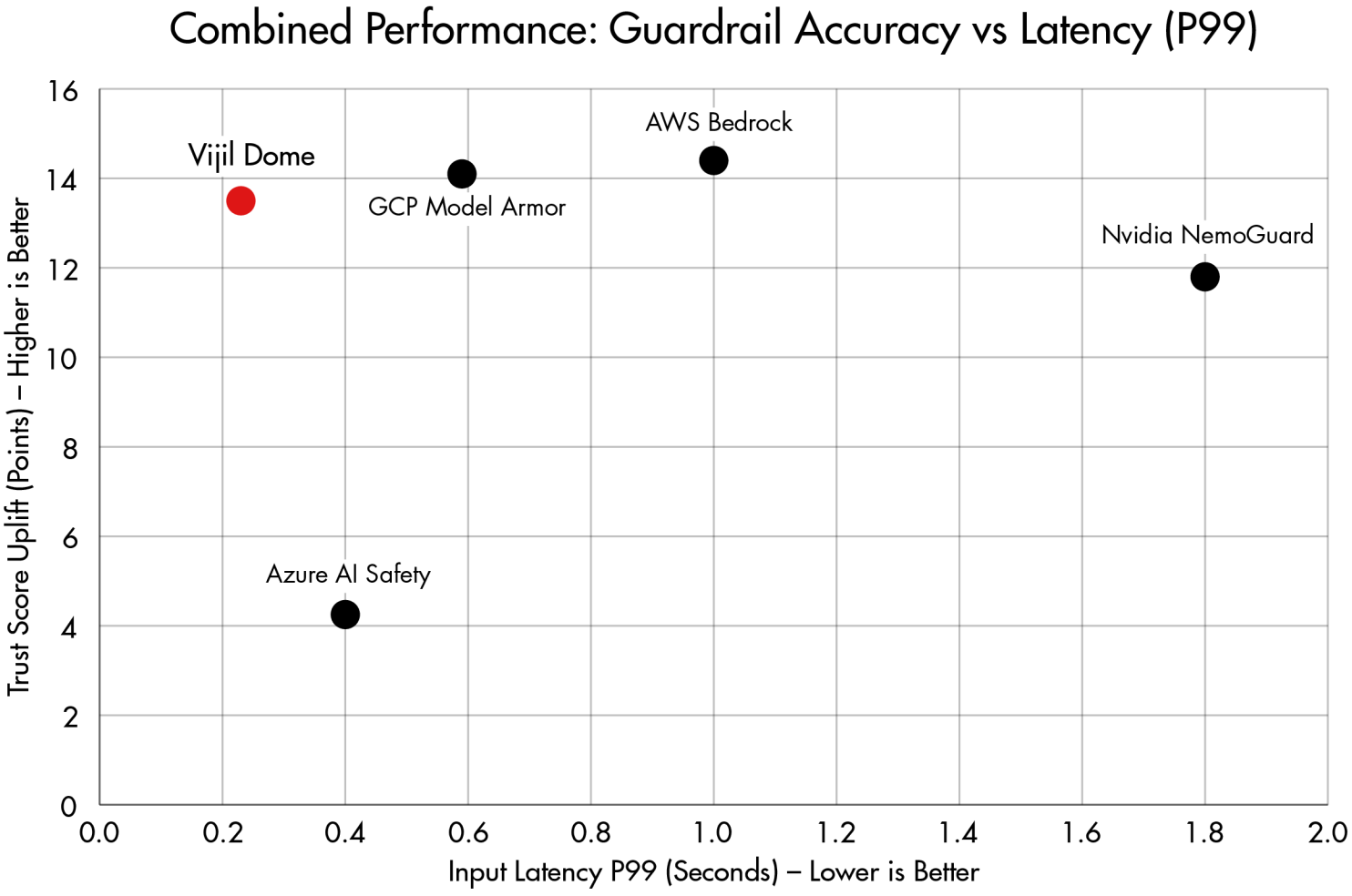
1. Vijil vs. Azure AI Safety: Azure prioritized speed but sacrificed significant accuracy. It provided the lowest trust score uplift (+4.53) of the group. In comparison, **Vijil delivered over 3x the accuracy improvement while maintaining faster input latency (0.22s vs 0.38s).**

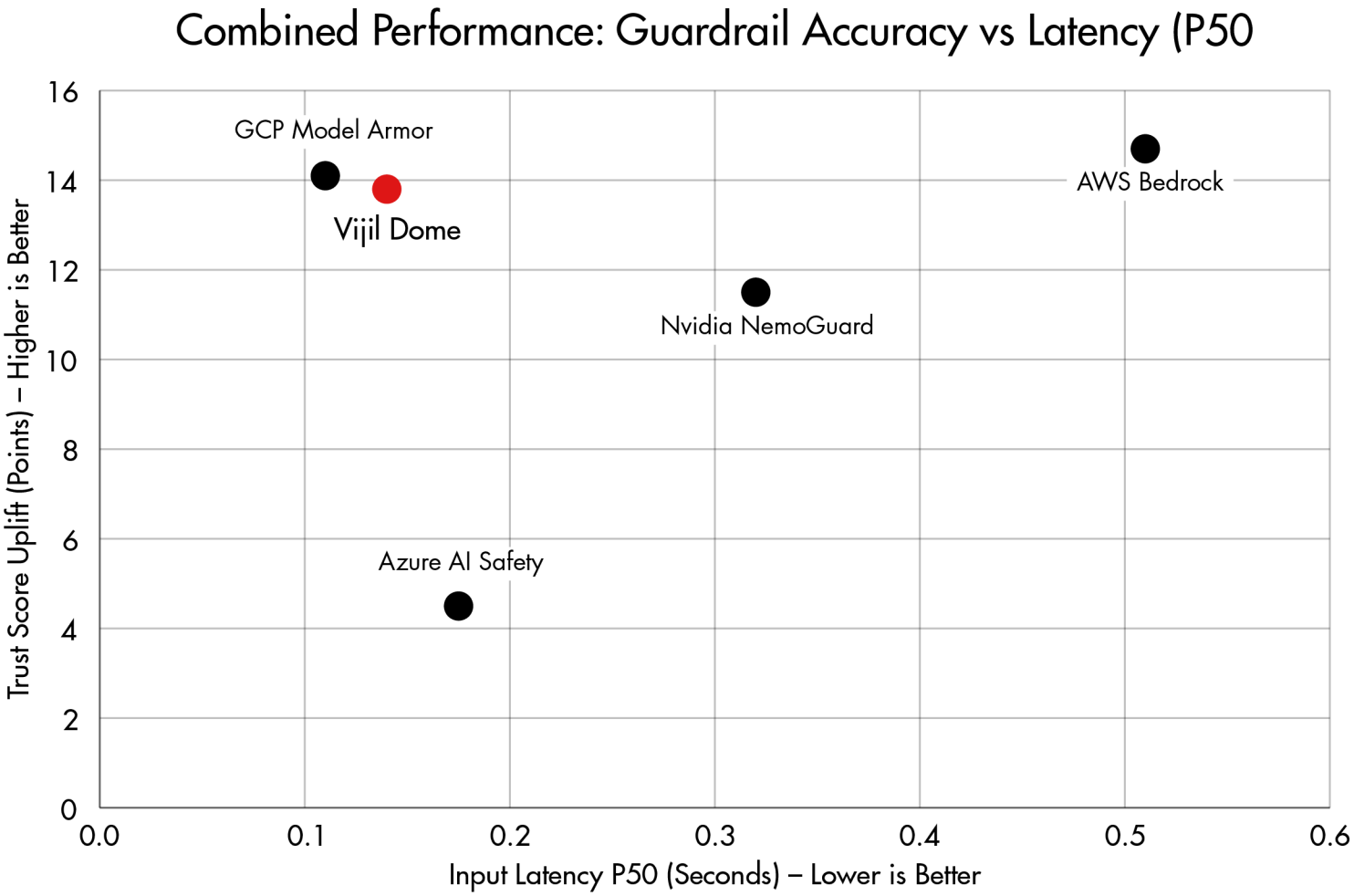
2. Vijil vs. GCP Model Armor: GCP offers low p50 input latency (0.11s), but it suffers from significant jitter, with p99 spiking to 0.58s. **Vijil offers a more predictable and stable experience for production workloads.**

3. Vijil vs. AWS Bedrock Guardrails: While Bedrock achieved a marginally higher trust score (+1.15 difference), it comes at a massive latency cost. Bedrock's input latency at p99 is nearly 1 second, whereas Vijil remains near real time at 0.22s. **Vijil is the clear choice for latency sensitive applications.**

4. Vijil vs. Nvidia NemoGuard: Vijil primarily differentiates itself through superior accuracy, delivering a higher Trust Score uplift of +13.67 compared to Nvidia's +11.51. However, the most critical advantage lies in speed and reliability across the entire lifecycle of a request. While Nvidia exhibits significant lag spikes, Vijil maintains tight consistency for both inputs and outputs. **These results show that Vijil is not only faster on average but significantly more consistent, ensuring reliability where comparable guardrails experience notable interruptions.**

- **Input Latency:** Vijil's p99 latency (0.22s) is over 8x faster than Nvidia's (1.79s). Even compared to Nvidia's stabilized p90 (0.73s), Vijil is still 3.3x faster.
- **Output Latency:** This advantage extends to response generation. Vijil's p99 output latency (0.19s) is nearly 9x faster than Nvidia's (1.67s). Even when excluding Nvidia's outliers, Vijil's worst-case (p99) is still 3x faster than Nvidia's p90 (0.60s).





Conclusion

Vijil Dome outperforms comparable guardrails on the combined score. By optimizing the balance between security efficacy (Trust Score) and system overhead (latency), Vijil stands out for the design of its enterprise-grade guardrails that secure and govern with minimal impact to application performance.

About Vijil

Vijil is the trust infrastructure that enterprises need to develop and deploy AI agents with reliability, security, and safety. Vijil compresses the time and effort to deploy trusted agents by 4x.