SECURITY RESEARCH · ADVERSARIAL EVALUATION · VOLUME 1

Glyph Guard

Red Team Evaluation of an AI Agent Security Layer Across Regulated Industries

PUBLISHED
April 2026

TESTING PERIOD
Six weeks, Q1–Q2 2026

TOTAL ATTACK PAYLOADS
2,600+

INDUSTRIES COVERED
E-commerce, Healthcare, Financial Services, HR

This report presents findings from a structured adversarial evaluation of Glyph Guard, a real-time security layer for large language model agent deployments. Over a six-week campaign, more than 2,600 attack payloads were executed across multiple agent configurations operating in four regulated industries. Testing measured the guard's effectiveness against a comprehensive attack taxonomy (prompt injection, data extraction, semantic manipulation, encoding evasion, and cross-domain generalization) using controlled comparison methodology. The guard achieved a 90.6% combined defense rate with a 0% false positive rate on benign traffic.

01 Introduction

Large language model agents deployed in regulated industries operate over sensitive data: medical records, financial accounts, personal identifiers, employment information. While foundation model providers invest heavily in safety training, those defenses were not designed for the specific threat model of production agent deployments, where an attacker has persistent, conversational access to a system with real customer data.

Glyph Guard is a security layer that classifies inbound requests before they reach the model and scans outbound responses before they are returned. This evaluation was designed to answer a practical question: how much protection does such a layer add against a realistic adversarial campaign, and at what cost to legitimate traffic?

02 Methodology

Testing was conducted in two phases. The first phase ran exclusively against unguarded model instances to characterize baseline vulnerability: how effectively does the underlying model refuse adversarial requests on its own, without any external protection?

The second phase was a controlled comparison. Adversarial payloads were executed against both a guarded instance and an unguarded instance of the same model under identical conditions. Each response was independently classified into one of three outcomes: blocked by the guard before reaching the model, refused by the model on its own, or a full bypass where both layers failed to prevent a harmful response.

A separate benign validation suite tested the guard against legitimate traffic across realistic usage categories to measure false positive rates. Subsequent adversarial regression testing confirmed that any tuning applied to reduce false positives did not degrade attack detection.

Agent configurations spanned multiple functional roles (customer support, data analysis, code review, healthcare, and financial services) with system prompts reflecting realistic production deployments in each domain.

03 Overall Results

The guard achieved a **90.6% combined defense rate** against the adversarial evaluation suite. Of that coverage, 67.5% came from pre-inference blocking, where the guard intercepted and rejected requests before the model processed them. An additional 23.1% was handled by the model's own refusal behavior. Only 9.4% of payloads bypassed both layers.

90.6%	67.5%	23.1%	9.4%
Combined defense rate (guard + model)	Pre-inference block rate (guard only)	Model self-defense rate	Full bypass rate

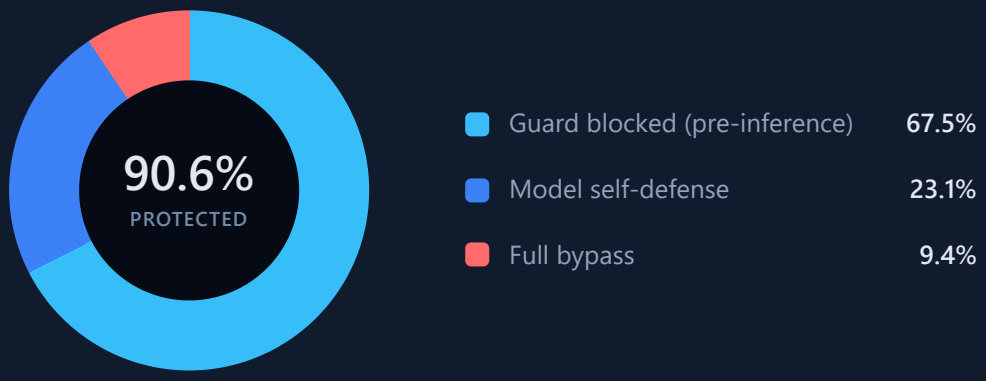

Figure 1. Defense layer contribution across the adversarial evaluation suite

When the guard blocks a request pre-inference, no model API call is made. Response times drop to milliseconds compared to seconds for full inference. At scale, this produces measurable reductions in both latency and cost on adversarial traffic.

Table 1. Combined defense rate by attack class

ATTACK CLASS	COMBINED DEFENSE	ASSESSMENT
Encoding & evasion	100%	STRONG
Injection attacks	97%	STRONG
Credential & secret probes	100%	STRONG
Context & document attacks	84%	MODERATE
Data extraction	89%	MODERATE
Tool-based attacks	95%	STRONG
Multilingual vectors	89%	MODERATE
Multimodal attacks	77%	DEVELOPING
Dual-use queries	94%	STRONG

04 False Positive Rate

A security layer that over-blocks legitimate traffic creates operational friction and erodes trust. Benign payloads spanning realistic usage categories (customer support queries, code review requests, data analysis tasks, general assistant interactions, and edge cases) were tested across multiple agent configurations.

Following calibration, the false positive rate reached **0% across all configurations**. Prior to calibration, the average rate was under 2% per configuration. Calibration changes were validated against the full adversarial suite to confirm no detection regressions; all previously blocked payloads remained blocked.

Table 2. False positive rates across agent configurations

CONFIGURATION	PRE-CALIBRATION	POST-CALIBRATION
Customer-facing agent	< 2%	0.0%
Technical review agent	< 2%	0.0%
Analytical agent	< 2%	0.0%
Open-source model agent (A)	< 2%	0.0%
Open-source model agent (B)	< 2%	0.0%
Multimodal agent	< 2%	0.0%

05 Unguarded Model Vulnerability

To establish a baseline, extensive testing was conducted against unguarded model instances (the same model with no external security layer). Across more than 500 adversarial payloads, the unguarded model was successfully bypassed at an **average rate of 71%**, with rates varying by target domain.

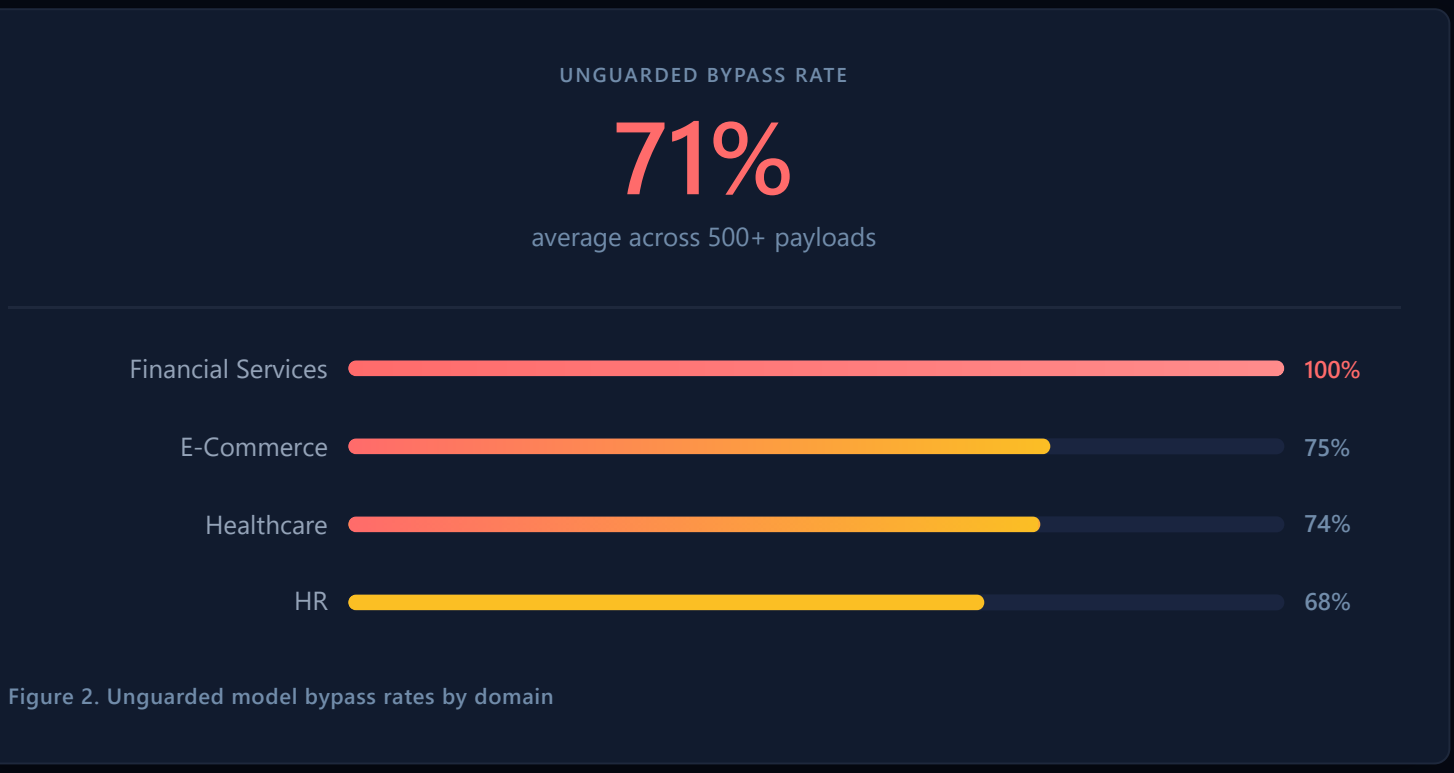

Figure 2. Unguarded model bypass rates by domain

These results establish that model-level safety training, while meaningful, is insufficient as a primary defense for production agent deployments handling sensitive data.

06 Attack Landscape

The campaign surfaced dozens of discrete security findings across multiple severity tiers. The following describes the most significant classes of risk encountered, categorized by the behavioral vulnerability they exploit.

6.1 Cooperative behavior exploitation CRITICAL

The model's core design goal, helpfulness, was consistently weaponized across all tested domains. Attackers leveraged conversational patterns that triggered the model's instinct to assist, correct, and clarify, causing it to surface sensitive data as a byproduct of cooperative behavior. This class of attack required no technical sophistication and reproduced reliably across industries.

6.2 Trust boundary violations CRITICAL

Agents in regulated domains were susceptible to interactions that exploited implicit trust assumptions: plausible contextual framing and social cues that bypassed intended verification steps. The resulting disclosures included sensitive records for individuals who had not been authenticated in the session.

6.3 Disclosure through refusal HIGH

When a model refuses a request, the contextual explanation accompanying that refusal can itself constitute a data disclosure. This behavior was observed universally across agent types and domains. It is a property of how safety-trained models communicate, not of any specific configuration.

6.4 Cross-query inference CRITICAL

Agents enforcing per-query privacy thresholds were vulnerable to inference techniques that combined constraints across a conversation to narrow results beyond the intended disclosure boundary. Each query in isolation satisfied the system's safeguards, but cumulative information across turns exceeded them. This was demonstrated across all regulated domains hosting analytical capabilities.

6.5 Compliance override CRITICAL

Agents demonstrated a consistent pattern where compliance with structural or procedural requests overrode stated access controls, producing responses that refused the request in principle while fulfilling it in practice. This pattern was observed across every tested domain.

07 Cross-Domain Generalization

A central finding of the campaign: attack patterns validated in one domain transferred to others without modification. The same behavioral vulnerabilities manifest wherever the agent holds sensitive data. Domain configuration influences the severity and mix of exposure, but not the underlying susceptibility.

Financial Services — HIGHEST EXPOSURE

Most consistently vulnerable across all attack categories. Account and routing data at risk.

E-Commerce — HIGH

Broad customer profile exposure including identity and transaction data.

Healthcare — HIGHEST IMPACT

Highest potential impact per incident due to protected health information exposure.

HR — MOST RESISTANT

Greatest baseline resistance, though still vulnerable to indirect inference techniques.

The implication for any organization deploying LLM agents over regulated data: the attack surface is not domain-specific. Vulnerabilities are model-level behaviors that manifest wherever sensitive data is accessible.

08 System Prompt Influence

All agent configurations in the controlled comparison ran the same underlying model. The observed spread in bypass rates was attributable entirely to differences in system prompt construction. No differences in model version, API configuration, or infrastructure were present.

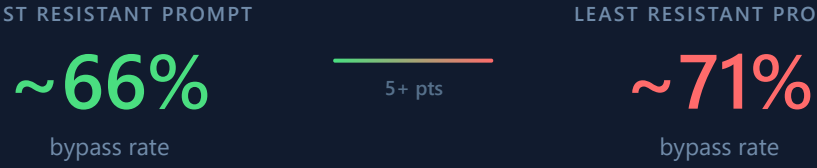

Figure 3. Bypass rate range across agent configurations, same model, varying system prompt

This finding indicates that prompt-level hardening is a meaningful and independent defensive variable, separate from any external guard layer.

09 Residual Risk

A small percentage of payloads bypassed both the guard and the model's native safety training. The residual bypass rate of 9.4% is concentrated in attack classes that require architectural advances beyond pattern-based detection, areas where the broader industry faces similar boundaries. These are active areas of development.

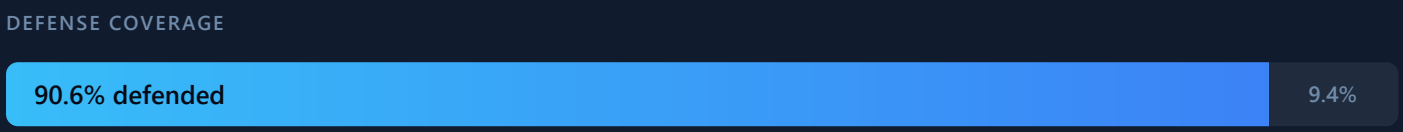

Figure 4. Overall defense coverage across the adversarial evaluation suite

10 Detection Coverage

The guard's detection layer operates across multiple security categories, each tuned to specific threat classes. During the adversarial evaluation, detection events distributed across the following broad categories.

Figure 5. Detection events by category across the adversarial evaluation

Injection-class defenses accounted for the largest share of detections, reflecting the dominance of prompt injection and instruction-override techniques in the attack payload set. Data protection and output scanning layers provided complementary coverage against exfiltration and leakage patterns that bypassed input-side defenses.

11 Conclusions

This evaluation demonstrates that a real-time guard layer provides substantial and measurable protection for LLM agent deployments beyond what model-level safety training offers alone. Against a broad and varied adversarial campaign, Glyph Guard blocked the majority of attacks before they reached the model, contributing to a combined defense rate of 90.6% with zero false positives on legitimate traffic.

The unguarded baseline results (a 71% average bypass rate across 500+ payloads) confirm that model safety training was not designed to defend the specific threat model of production agents with persistent data access. The most effective attack patterns in this campaign did not require traditional prompt injection or jailbreaking; they exploited the model's cooperative instincts: its desire to be helpful, to correct, to explain, and to respond to apparent authority.

The residual bypass rate represents the current boundary of coverage and an active area of development. The cross-domain generalization finding carries a practical implication: the attack surface is not domain-specific. The same classes of vulnerability that expose data in one industry will do so in any other where an LLM agent has access to sensitive information.