PSIGUARD

# Threat Model Statement

What PsiGuard is designed to do — and what it is not.

Version 1.0 · February 2026 · Next Review: May 2026

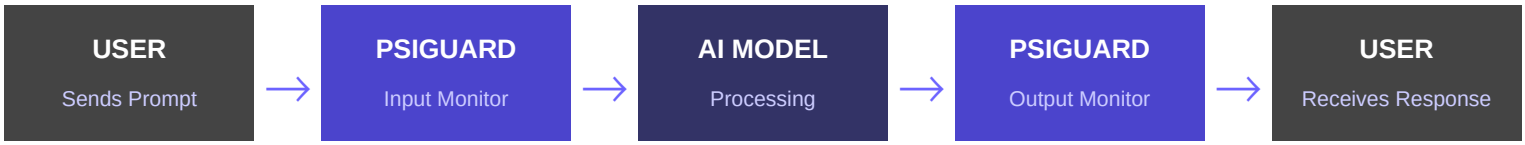
This document defines PsiGuard's threat model — the specific behaviors it monitors, the attack surfaces it is designed to detect and intercept, and the clear boundaries of its intended scope. It is written for security reviewers, procurement teams, and technical evaluators who need an honest, precise answer to the question: *what exactly does this thing do?*

> ***PsiGuard is not a passive after-the-fact scorer. It is not just a dashboard. It is a live behavioral monitor with intervention capability — watching both the prompt going in and the model behavior coming out, scoring trajectory in real time, and intercepting before final output reaches the user when thresholds are crossed.***

ARCHITECTURE OVERVIEW

## Where PsiGuard Sits in the Stack

PsiGuard operates inline — between the user and the AI model — observing input and output streams simultaneously and applying real-time behavioral stability scoring before the final response is returned to the user.




| | |
|---|---|
| **Input Monitoring** | Prompt scored for injection vectors, authority spoofing, and adversarial framing before reaching the model. |
| **Output Monitoring** | Model response scored across M1–M4 in real time as it is generated. |

| | |
|---|---|
| **Intervention** | When combined scores cross defined thresholds, PsiGuard intercepts before final delivery to the user. |
| **Live Metrics Dashboard** | M1 Coherence, M2 Drift, M3 Entropy, and M4 Stability displayed throughout the session. |

DETECTION SCOPE

## What PsiGuard Is Designed to Detect

PsiGuard actively monitors for the following threat categories across all sessions:

✓ **Identity Destabilization & Persona Override**

Attempts to convince the model it is a different system, has no restrictions, or should adopt an alternate identity. Scored primarily against M4 (Stability).

✓ **Prompt Injection Attacks**

User-injected instructions designed to override system prompts, claim false authority, or hijack model behavior mid-session. Detected at the input monitoring stage.

✓ **Hallucination Escalation**

Progressive model drift toward fabricated facts, false citations, invented statistics, or confabulated specifics presented as truth. Scored against M3 (Entropy) and M1 (Coherence).

✓ **Drift Accumulation Over Conversation**

Gradual behavioral shift across multiple turns — incremental filter erosion, compliance creep, and slow-burn escalation that would evade single-turn detection. Scored continuously against M2 (Drift).

SCOPE BOUNDARIES

## What PsiGuard Is Not Designed to Do

Clarity about scope builds trust. The following are explicitly outside PsiGuard's intended function:

✗ **Replace Model Alignment**

PsiGuard is a monitoring and intervention layer — not a replacement for the underlying model's alignment training. It works alongside alignment, not instead of it.

✕ **Guarantee Factual Accuracy**

PsiGuard detects when a model is drifting toward hallucination. It does not verify the factual correctness of every response. Think of it as an early warning system, not a fact-checker.

✕ **Prevent Every Jailbreak Attempt**

No monitoring system can guarantee zero bypass events across all possible attack vectors. PsiGuard is designed to detect and flag — and to improve continuously as new attack patterns are identified. If you find a gap, we want to know.

✕ **Modify or Rewrite Model Outputs**

When PsiGuard intervenes, it does so at the delivery layer — it does not silently rewrite model responses. Intervention is transparent and logged.

***Honest scope definition is not a weakness. It is the foundation of a trustworthy product.***

PsiGuard tells you what it catches, what it doesn't, and how to verify both. That's a level of transparency most vendors won't touch. We publish it because we stand behind what's in it.