

CyberProfiler

In-Model Defence For Agentic Systems

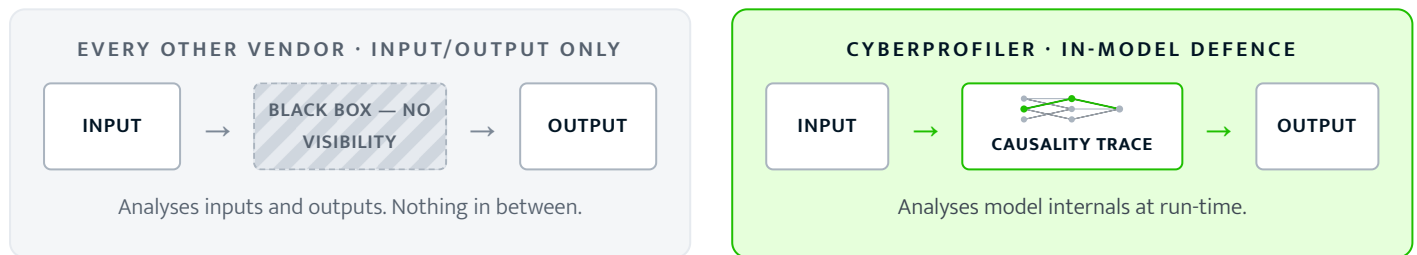
Introducing **In-Model Defence** a new class of prompt injection defence.

WHY IT MATTERS

- **99%+ Accuracy**
Identifies 'malicious intent' doesn't rely on phrasing or input text.
- **Does Away With Patching**
Update frequency doesn't need to chase attackers' new patterns.
- **Model Agnostic & Fast**
A guardrail for closed or open LLM models, completes in under 75ms

A NEW APPROACH TO PROMPT INJECTION DEFENCE

Every existing approach (classifiers, LLM-as-judge, embeddings) competes on the same ground, at the input/output boundary, reading what goes in and out of a model but oblivious to what happens in between. Attackers know this, so they simply evolve techniques to pass the filter. CyberProfiler is the first product in a new category 'In-Model Defence', accessing and analysing model internals at run-time.



A new layer of visibility, not a better filter

Instrumented into the model's actual reasoning path. CyberProfiler catches what traditional input/output tools cannot see.

Profiles intent, not vocabulary

Catches existing and evolving zero-day attacks. Updates per attack class (24 covered today), far less often than per-phrasing filter patches.

Rock solid audit evidence

A Causality Trace is a 'system of record for inference'; what audit, incident investigation and regulatory reporting need.

DEPLOYMENT TIERS

TIER A · OUT-OF-THE-BOX

90%+ accuracy

Trained on 24 classes of prompt injection and operates, out-the-box as an input guardrail against a reference model.

<75ms latency

TIER B · YOUR DATA

95%+ accuracy

Tuned on your traffic patterns and edge cases to close the domain gap, operates as an input guardrail.

<75ms latency

TIER C · YOUR DATA & MODEL

99%+ accuracy

Instrumented directly into your model's reasoning path and trained on your production model and data specification.

<35ms latency

DEFENCE & SECURITY

HEALTHCARE

CRITICAL NATIONAL INFRASTRUCTURE

FINANCIAL SERVICES

Download the SDK, protection running immediately.

Technology proven with Cisco, National Cyber Security Centre (NCSC is part of GCHQ, IEEE research award, AMLUCS 2026. Patent-pending Causality Trace technology.