

WiP: COG-IMMUNE

Toward a Cognitive Immune System for Large Language Models

Sylvester Kaczmarek
Imperial College London
research@sylvesterkaczmarek.com

Core claim

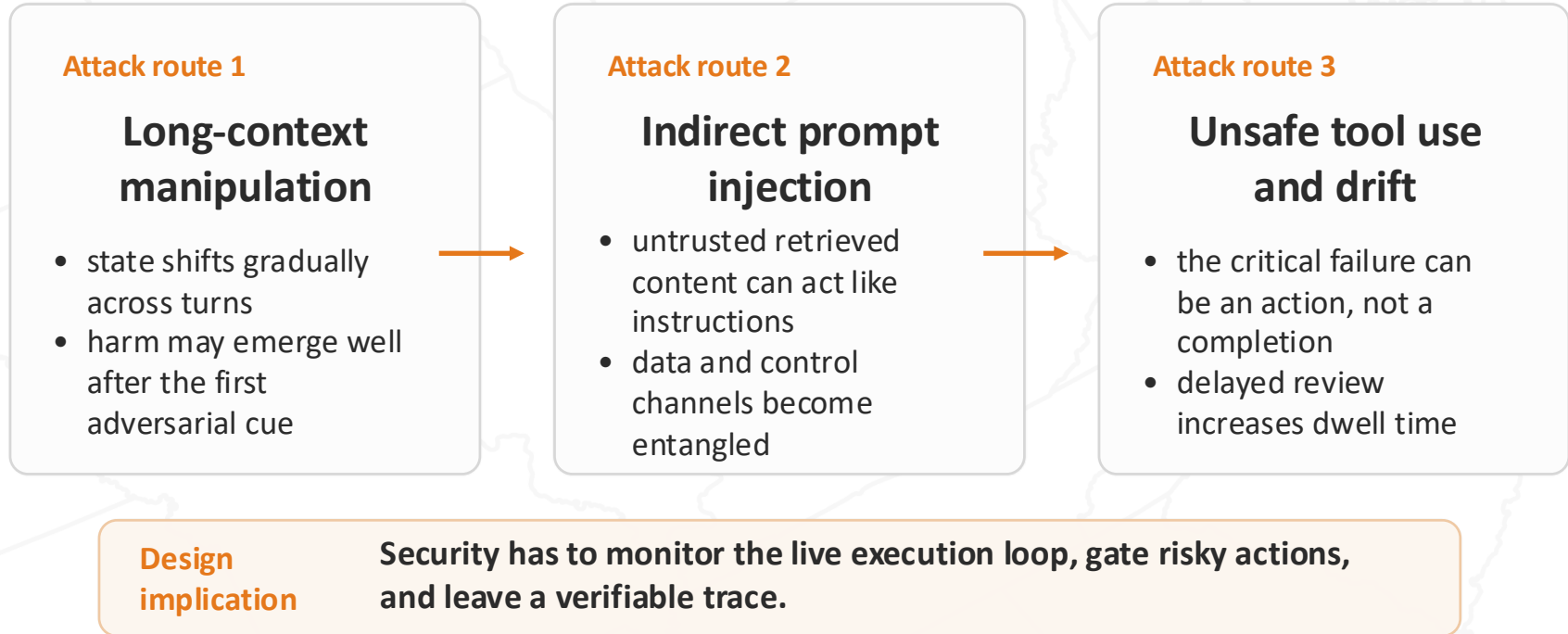
Tool-using LLM systems need continuous risk sensing, graded containment, and compact signed evidence.

Sensing • Containment • Evidence

Why this matters

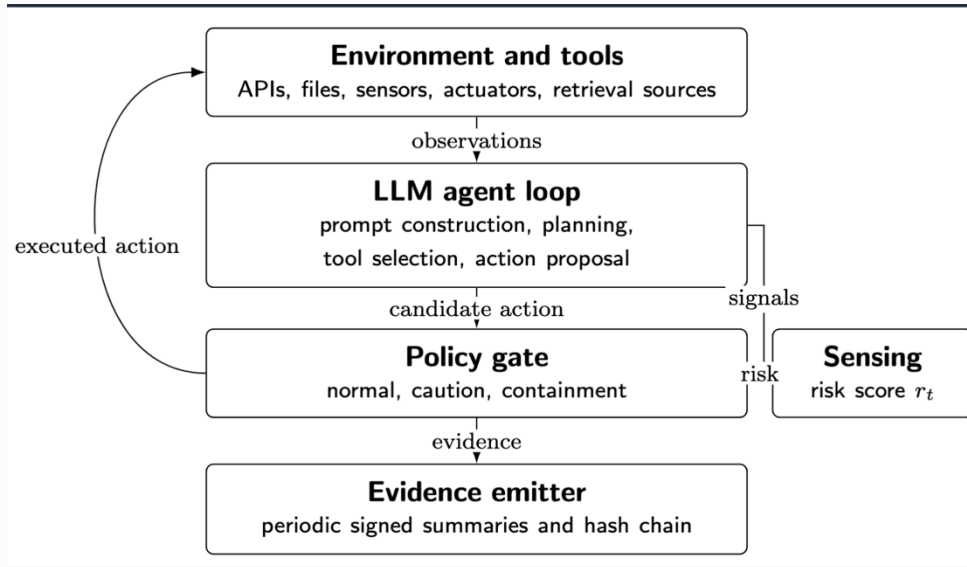
- Perimeter filters do not address attacks that unfold across turns, tools, and changing state.
- The object of control is the system's actions, not only the text it produces.
- Delayed oversight calls for low-bandwidth telemetry that remains verifiable after the fact.

Why perimeter defenses fail in agentic systems



COG-IMMUNE architecture

Closed-loop assurance around the agent execution loop



Three functions: continuous sensing over execution-time signals; graded containment as risk rises; low-bandwidth signed evidence for later verification

Policy gate and signed evidence protocol

Risk-gated action selection

The gate narrows the allowable action set as risk rises.

Normal	full tool set subject to normal access control
Caution	read-only tools, bounded parameters, rate limits, added confirmations
Containment	minimal safe tools only; no writes or privilege transitions without approval

Escalation uses thresholds and hysteresis to reduce oscillation.

Evidence object

Each window emits a compact signed summary that supports remote audit.

state digest | mode | counter | sig
hash chain to previous window →

Operational benefits

- missing windows are detectable under packet loss
- approval can be verified outside the model context
- review can focus on flagged windows instead of full traces

Evaluation design and current empirical progress

Attack classes

- multi-turn jailbreaks and long-context manipulation
- indirect prompt injection through retrieved content
- tool misuse, privilege escalation attempts, and exfiltration
- slow drift across long episodes
- degraded oversight with delayed approvals and packet loss

Primary metrics: time-to-detect, time-to-contain, unsafe-action rate, false positives, and evidence overhead

Current status

Controlled harness • **feature weights** • **ablations** • **static-caution baseline**

The updated manuscript now includes early controlled-harness experiments and ablation-oriented analysis. The next step is to quantify detection, containment, and overhead trade-offs across text-only and autonomy-style loops.

Baselines: perimeter filters, external attack classifiers, and runtime gates without evidence objects

Designed for comparison against both text-only agents and delayed-oversight autonomy loops.

Takeaways and discussion

Key takeaways

- agentic LLM security should be framed as a closed-loop assurance problem over actions and state
- graded containment can reduce harm while preserving useful capability under uncertainty
- compact signed evidence is valuable when review is delayed and bandwidth is constrained
- the architecture is meant to be testable, comparative, and deployable in delayed-oversight settings

Feedback for discussion

- most realistic evasion routes against the sensing layer
- best benchmark harnesses for delayed-oversight evaluation
- deployment assumptions for signed approvals and evidence review

Contact

Sylvester Kaczmarek
Imperial College London
research@sylvesterkaczmarek.com