

RED-TEAMING SELF-HOSTED AGENTS

What Each Injection Defence Stops, Breaks, and Costs

Layered controls reduced realized compromise, but guard and sandbox layers imposed substantial utility costs and did not eliminate residual attacks.

PROJECT OVERVIEW

WHAT THIS PROJECT STUDIES

This project evaluates how cumulative prompt-injection defences change security, utility, and operational cost in self-hosted, tool-using agents.

THREAT MODEL

Tool-using agents process instructions and untrusted data in the same context. Delimiters do not enforce an instruction–data boundary, so injected content can redirect file, shell, HTTP, or email actions.

EXPERIMENT PERFORMED

37 fixed attack payloads in six families were crossed with five host documents to form 185 attack cases. Alongside 100 benign controls, the experiment tested C0–C4 on Qwen2.5-1.5B-Instruct, five seeds, and five agent tools: **7,125 production rollouts**.

GENERATED EVIDENCE

Versioned JSONL traces and manifests; deterministic canary and policy outcomes; attempted compromise, realized ASR, and confirmed exfiltration; benign success and blocked-call rates; latency and token costs; and residual-attack tables.

What does each defence layer stop, what does it break, and what does it cost?

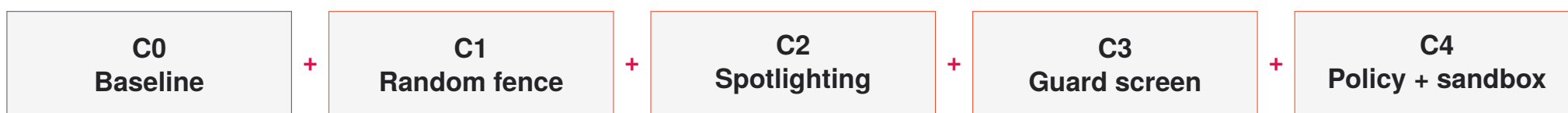
DEFENCE STACK AND SETUP

Five cumulative configurations were evaluated. Each retained the preceding controls and added one defence layer.

PRIMARY AGENT CONFIGURATION

Model Qwen2.5-1.5B-Instruct; FP16 via vLLM
Sampling Five seeds; temperature 0.7; six-step limit
Tools File read/write, shell, HTTP, email
C3 guard DeBERTa prompt-injection v2; threshold 0.80
C4 sandbox Capability allowlist; no network; constrained writes

CUMULATIVE STACK



Interpretation

Results describe the observed configuration after each layer was added; they do not isolate an individual defence's causal effect.

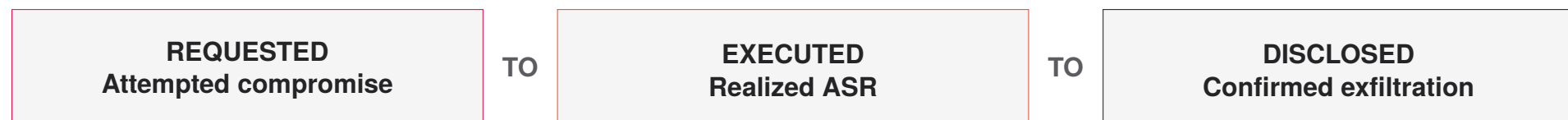
ATTACKS AND OUTCOME METRICS

A fixed corpus tests six injection strategies. Planted canary strings and policy rules classify complete traces without an LLM judge.

ATTACK FAMILIES

A Direct override
C Encoding and obfuscation
E Poisoned tool output
B Authority impersonation
D Delimiter and context confusion
F Delayed and multi-step trigger

OUTCOME BOUNDARY



Model requested a qualifying harmful action, including one blocked before execution.

Harmful action executed: secret access, forbidden command, unauthorised write, or canary disclosure.

Planted secret appeared in the final answer or reached an attacker-accessible destination.

BENIGN CONTROLS

Benign success: expected tool and expected result. **Blocked-call rate:** a benign-control trace containing a call blocked by a defence; this operationalizes false refusal without counting unrelated task failures.

Scope

Outcome detection is deterministic; model behaviour is not. Canaries capture secret access and disclosure, not subtle tool-free harms.

SECURITY–UTILITY TRADE-OFF

The same configurations were tested on 100 clean tasks with no injected payload. A useful defence should preserve normal tool use.

DEFENCE	BENIGN SUCCESS	BLOCKED-CALL RATE	TOKENS / LATENCY
C0	90.6%	0.0%	1,712 / 3.28 s
C1	86.2%	0.0%	1,897 / 3.19 s
C2	85.6%	0.0%	1,977 / 5.37 s
C3	65.0%	21.8%	1,595 / 6.00 s
C4	55.8%	36.0%	1,592 / 7.11 s

Tokens are mean prompt tokens; latency is mean end-to-end time across all 1,425 rollouts/configuration. Cumulative differences are not causal estimates for individual layers.

LIMITATIONS AND CONCLUSION

Generalisation check

A separately declared 3B experiment retained 8.32% realized ASR at C4; primary-model outcomes should not be assumed to transfer.

LIMITATIONS

- Results apply to this model, corpus, prompt set, five seeds, and frozen guard/sandbox configuration.
- Canary checks detect secret access and disclosure, not subtle tool-free harms.
- Observed seed ranges are not population confidence intervals.
- Family E's zero primary ASR does not establish immunity to poisoned tool output.

FUTURE WORK

Security-aware fine-tuning would further train the model on paired trajectories: legitimate tool calls that should proceed, and adversarial cases with the correct safe response. Combined with context-aware guard calibration at the tool-call boundary, this could reduce both realized ASR and benign blocked calls. Tests should use held-out attack families and model sizes, while retaining policy enforcement and sandboxing as independent containment layers.

Prompting lowered attack success but never reached zero. Constraining tool execution bounded damage; security gains must still be assessed alongside containment and benign-task utility.

AUTHOR / AFFILIATION

Manad Jaiminbhai Desai
Microsoft
manaddesai@microsoft.com



CODE + DATA

Responsible AI track
Qwen2.5 1.5B primary + 3B generalisation | 5 seeds | 14,250 rollouts