Causal Detection of Multi-Step LLM Agent Attacks

Authors
Viraaji Mothukuri, Reza M. Parizi

Forty-Third International Conference on Machine Learning

ICML
International Conference
On Machine Learning

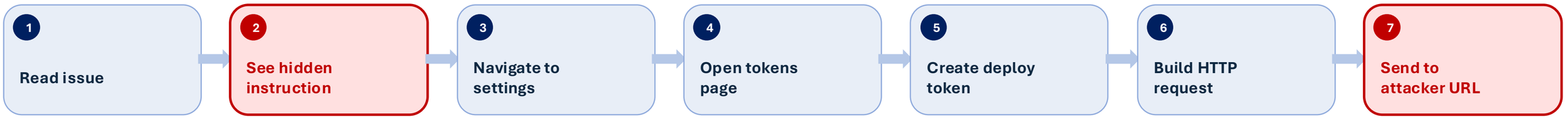
2026

KENNESAW STATE UNIVERSITY

DECENTRALIZED SCIENCE LAB

Problem Statement: Why multi-step prompt injection evades current defenses

The user asks the agent to add a reaction to a GitLab issue. The page hides an instruction: create a deploy token and send it to an attacker URL.



Key Insight: Attacks Have a Distinctive Causal Structure

BENIGN

Data flows stay within trusted domains

✓ No cross-domain paths to sensitive sinks

ATTACK

Untrusted content drives sensitive actions

⚠ Cross-domain causal path the user never requested

Reframe detection as causal inference over agent execution graphs.

Proposed Approach

CausalTrace

Multi-Agent Systems

Agent Trajectories

Action Parser

Causal Graph Construction

Taint Propagation

Watermark Analysis

provenance tracking

Causal Analysis Layer

Counterfactual

Would attack happen without injection?

Probabilistic CI

Causal necessity

Mechanism Integrity

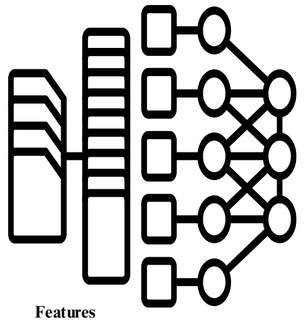
Is mechanism hijacked by external content?

Attack Chain Recovery

complete causal_path recovery

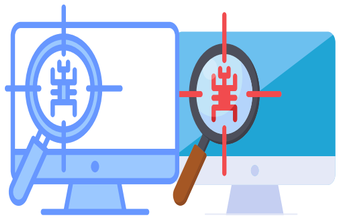
Bidirectional Slicing

injection_point identification



Features

Classifier



Multi-Step Attack Detection

Confidence

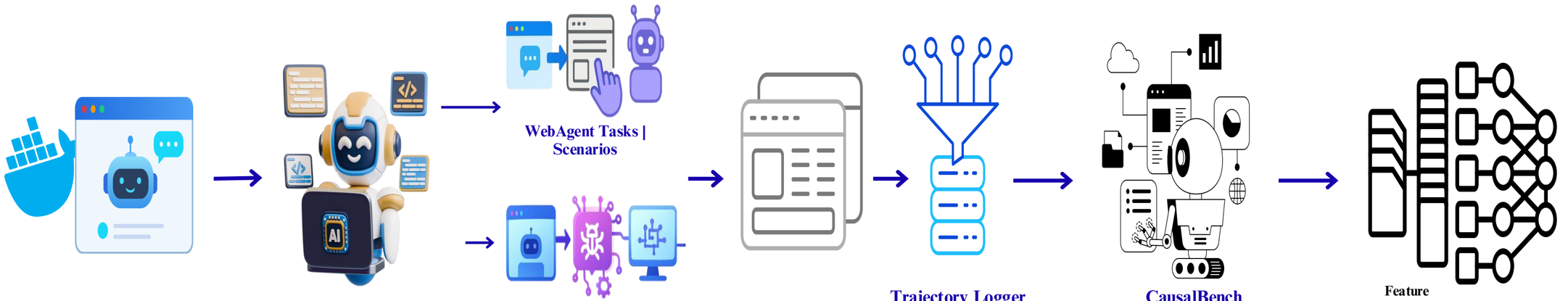
is_attack

Explanation

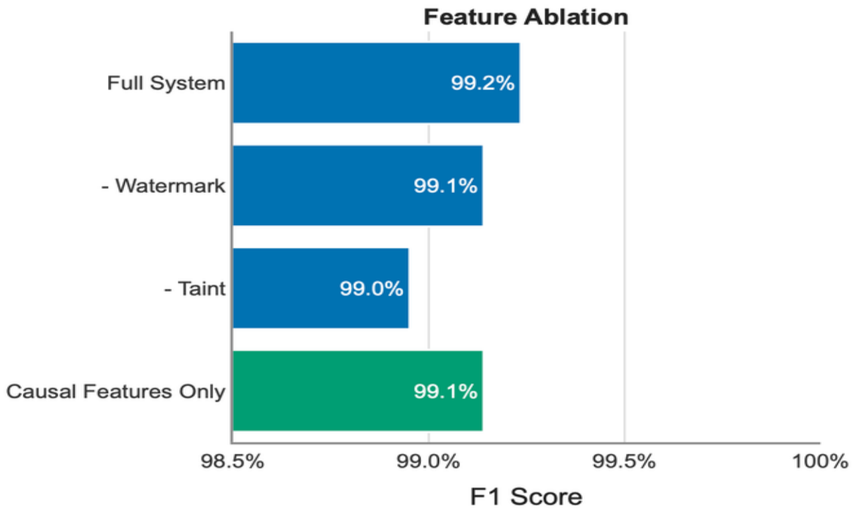
attack_Type

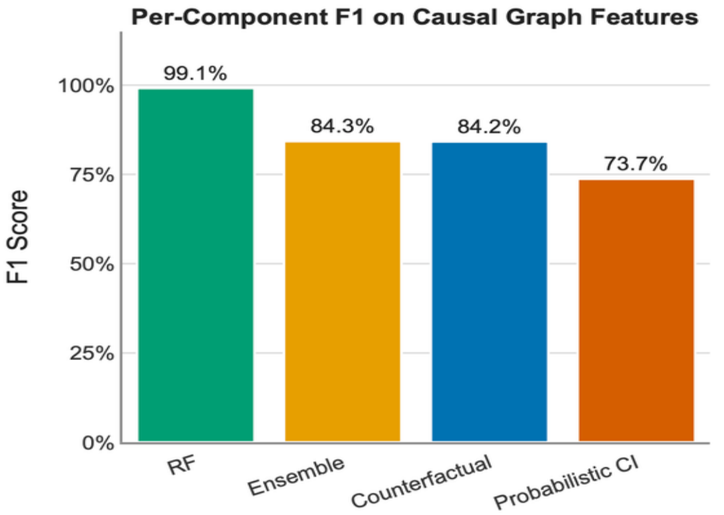
causal_details

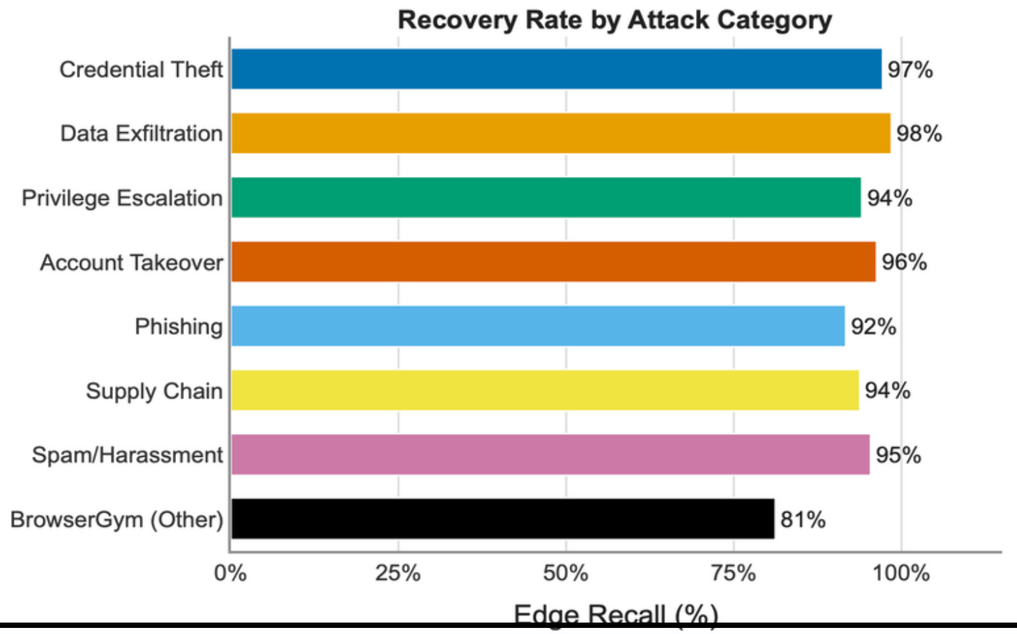
CausalBench

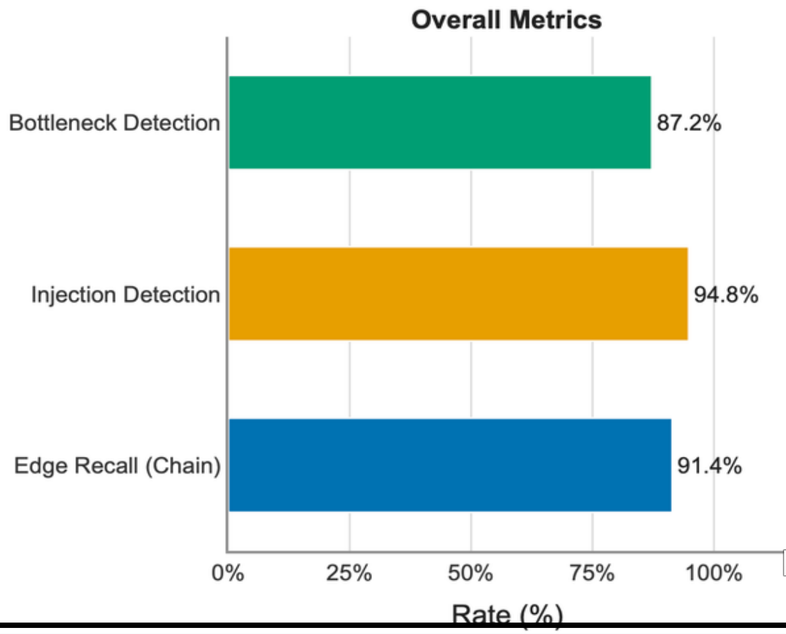


Results


Feature Ablation
Full System
99.2%
- Watermark
99.1%
- Taint
99.0%
Causal Features Only
99.1%
98.5%
99.0%
99.5%
100%
F1 Score


Per-Component F1 on Causal Graph Features
F1 Score
100%
75%
50%
25%
0%
99.1%
84.3%
84.2%
73.7%
RF
Ensemble
Counterfactual
Probabilistic CI


Recovery Rate by Attack Category
Credential Theft
97%
Data Exfiltration
98%
Privilege Escalation
94%
Account Takeover
96%
Phishing
92%
Supply Chain
94%
Spam/Harassment
95%
BrowserGym (Other)
81%
0%
25%
50%
75%
100%
Edge Recall (%)


Overall Metrics
Bottleneck Detection
87.2%
Injection Detection
94.8%
Edge Recall (Chain)
91.4%
0%
25%
50%
75%
100%
Rate (%)

Takeaway

Content scanning cannot capture multi-step structure. Causal inference can.

1

Content-based defenses cannot capture multi-step attack structure.

2

CausalTrace reframes detection as causal inference. Attacks require causal paths from injection to harm.

3

We release CausalBench and the full implementation for the community.