ICML
International Conference
On Machine Learning

Benchmarking LLM-Assisted Blue Teaming via Standardized Threat Hunting

Yuqiao Meng [1] Luoxi Tang [1] Feiyang Yu [2] Xi Li [3] Guanhua Yan [1] Ping Yang [1] Zhaohan Xi [1]

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

DUKE UNIVERSITY 1838

THE UNIVERSITY OF ALABAMA AT BIRMINGHAM 1969

ICML
International Conference
On Machine Learning

Background: The Growing Cyber Threat Crisis

- The scale and sophistication of cyber threats are outpacing manual defense capabilities.
  - **Vulnerability Surge**: 38% increase in CVEs reported in 2024 (11,000+ more than 2023).
  - **Defender Pressure**: Blue teams must identify, analyze, and respond in near real-time.
  - **LLM Potential**: Recent advances show promise in isolated tasks (malware analysis, fuzzing).

ICML
International Conference
On Machine Learning

Problem: Fragmented Research Limitations

Isolated Research	Most LLM security studies focus on isolated tasks, ignoring the chained nature of real-world threat hunting.
Task Dependencies	Hunting requires understanding patterns, behaviors, and mitigation in sequence—outputs of one stage drive the next.
Hallucination Risk	Open-ended reasoning often leads to inconsistencies and unreliability in high-stakes defensive operations.

Introducing CYBERTEAM

Unprecedented Scope: A large-scale repository of 450,000+ samples from 23 authoritative vulnerability and intelligence sources.

Standardized Workflow: Models realistic blue team practices by capturing task dependencies from attribution to response.

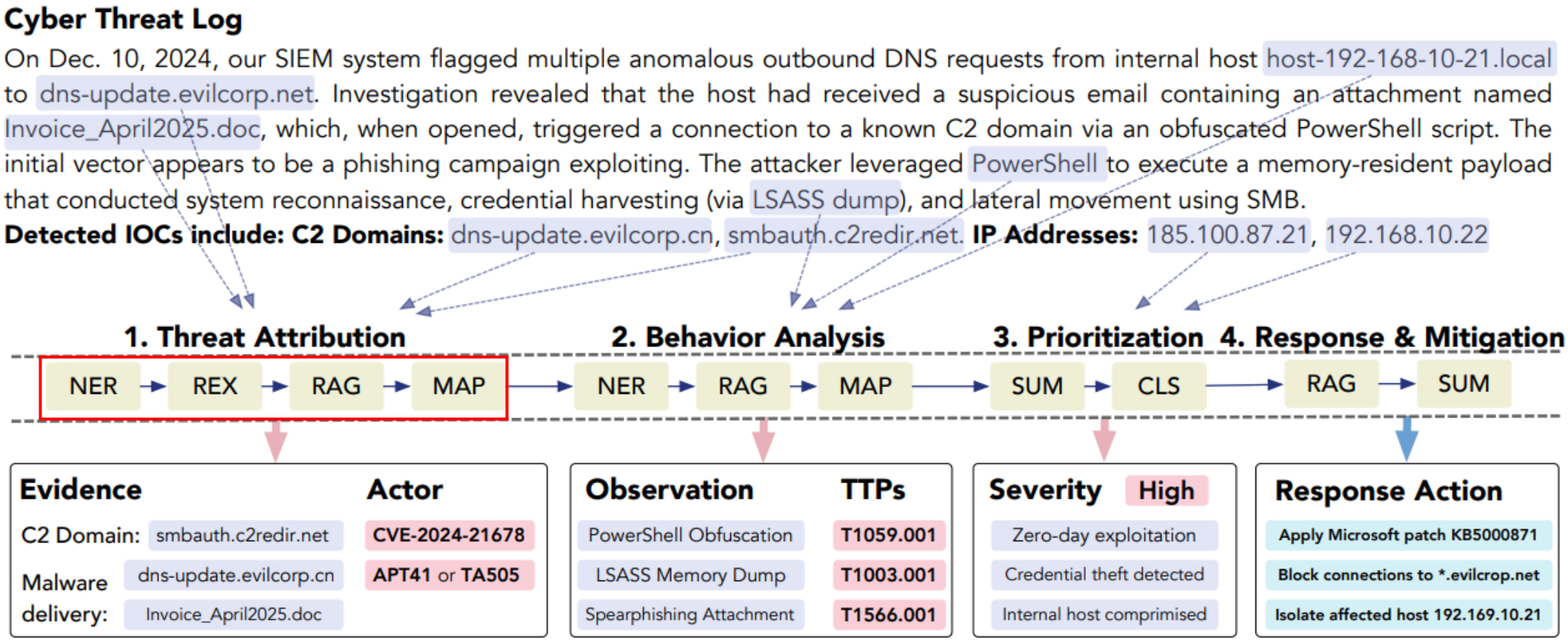
Modularized Reasoning: Integrates 30 tasks with 9 functional modules (NER, RAG, etc.) to balance flexibility and reliability.

Actionable Intelligence: First benchmark focused on transforming raw logs into structured, procedural defense actions.

ICML
International Conference
On Machine Learning

Comprehensive Task Lifecycle

A CYBERTEAM threat hunting example equipped with operational modules.



ICML
International Conference
On Machine Learning

Threat Hunting Tasks

1. Threat Attribution
2. Behavior Analysis
3. Prioritization
4. Response & Mitigation

Task	Analytical Target	Standard Operation	#Data	Metric
	Threat Attribution			
Malware Identification	Malware delivery or toolset	NER, SUM + Reasoning	15,742	F1
Signature Matching	Techniques from known threat groups	NER, SIM + Reasoning	5,166	F1
Temporal Pattern Matching	Known work schedules	REX + Reasoning	4,203	Sim
Affiliation Linking	Source organizations	NER, MAP + Reasoning	17,583	F1
Geographic Analysis	Geographic or cultural indicators	NER, SIM + Reasoning	6,164	F1
Victimology Profiling	Targeted victims or attacker motives	NER, REX + Reasoning	18,612	F1
Infrastructure Extraction	Domains, IPs, URLs, or file hashes	NER, REX, SUM + Reasoning	24,129	F1
Actor Identification	The threat group or actor (e.g., APT28)	NER, RAG, MAP + Reasoning	17,823	F1
Campaign Correlation	Threat campaigns or incidents	NER, MAP + Reasoning	27,762	F1
	Behavior Analysis			
File System Activity Detection	Suspicious file creation, deletion, or access	SPA, NER, SUM + Reasoning	4,653	Sim
Network Behavior Profiling	Patterns of external communication (e.g., C2)	SPA, NER, SUM + Reasoning	2,617	Sim
Credential Access Detection	Theft or misuse of credentials	SPA, NER, SUM + Reasoning	2,492	Sim
Execution Context Analysis	Execution behaviors by user or process	SPA, NER, SUM + Reasoning	23,888	Sim
Command & Script Analysis	Suspicious commands or scripts	SPA, NER, SUM + Reasoning	20,232	F1
Privilege Escalation Inference	Privilege escalation attempts	SPA, NER, SUM + Reasoning	15,953	Sim
Evasion Behavior Detection	Evasion or obfuscation techniques	SPA, NER, SUM + Reasoning	8,973	Sim
Event Sequence Reconstruction	Timeline of attack-related events	SUM + Reasoning	23,265	Sim
TTP Extraction	Tactics, techniques, and procedures	RAG, MAP + Reasoning	28,292	F1
	Prioritization			
Attack Vector Classification	Exploitation vectors (e.g., network, local, physical)	SUM, CLS + Reasoning	17,448	Acc
Attack Complexity Classification	Level of hurdles required to carry out the attack	SUM, CLS + Reasoning	17,116	Acc
Privileges Requirement Detection	Level of access privileges an attacker needs	SUM, CLS + Reasoning	18,030	Acc
User Interaction Categorization	If exploitation requires user participation	SUM, CLS + Reasoning	17,075	Acc
Attack Scope Detection	If the vulnerability affects one/multiple components	SUM, CLS + Reasoning	18,570	Acc
Impact Level Classification	Consequences on confidentiality, integrity, and availability	SUM, CLS + Reasoning	18,736	Acc
Severity Scoring	A numerical score indicating the overall attack severity	SUM, MATH + Reasoning	11,507	Dist
	Response & Mitigation			
Playbook Recommendation	Relevant response actions based on threat type	RAG, SUM + Reasoning	10,718	Hit
Security Control Adjustment	Firewall rules, EDR settings, or group policies	RAG, SUM + Reasoning	9,929	Sim
Patch Code Generation	Code snippets to patch the vulnerability	RAG, SUM + Reasoning	11,341	Pass
Patch Tool Suggestion	Security tools or utilities	RAG, SUM + Reasoning	9,763	Hit
Advisory Correlation	Security advisories or best practices	RAG, SUM + Reasoning	24,511	Hit

ICML
International Conference
On Machine Learning

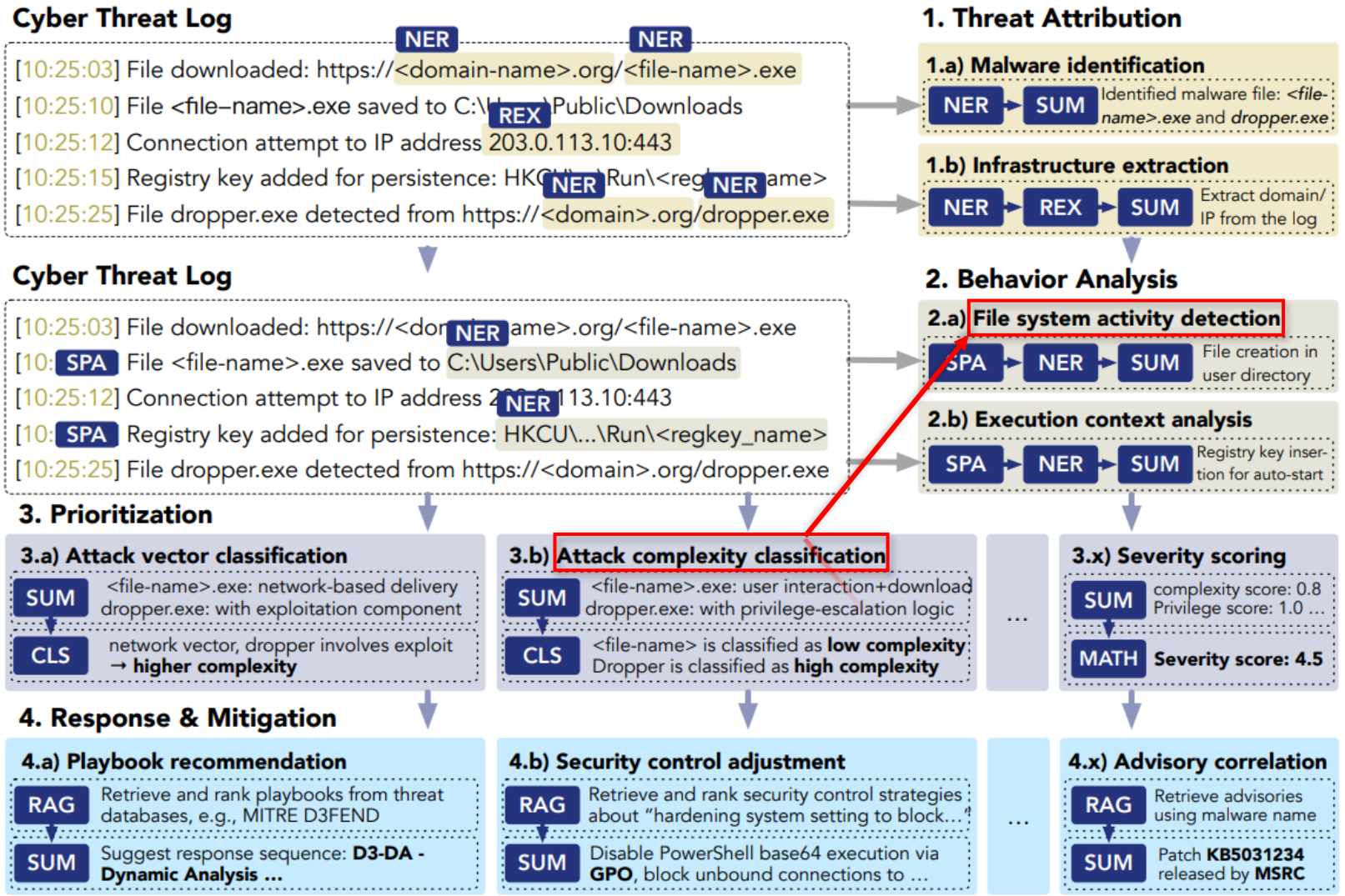
Data Sources

Benchmark	Focus	#Data	#Task	#Source	Coverage	Unique Feature
CWE-Bench-Java (Li et al., 2025)	Java vulnerability	120	4	1	Four CWE classes	Large-scale Java codes
CTIBench (Alam et al., 2024)	Cyber Threat Intelligence	2,500	3	6	CVE, CWE, CVSS, ATT&CK	Multi-choice questions (MCQ)
SevenLLM-Bench (Ji et al., 2024)	Report understanding	91,401	28	N/A	Bilingual instruction corpus	Synthetic Data, MCQ, QA
SWE-Bench (Jimenez et al., 2024)	Software bug fixing	2,294	12	1	GitHub issues	Python repository
CYBERTEAM (Ours)	Blue team threat hunting	452,293	30	23	Threat-hunting lifecycle (3.1)	Open Generation, Standardized Reasoning Env

Vulnerability DBs (e.g., NVD, MITRE) Structural Ground Truth, Provides deterministic standard ontologies (CVE, CWE, CAPEC) for model grounding.
Intel Platforms (e.g., VirusTotal, AlienVault) Real-World Live Telemetry, Incorporate active, multi-vendor Indicators of Compromise (IoCs) and behavioral logs.
Industry Reports (e.g., Mandiant, Unit 42) Expert-Level Threat Narratives, Semi-structured reports that test LLMs' advanced text comprehension and long-context reasoning.

ICML International Conference On Machine Learning

Standardized Threat Hunting with Operational Modules



ICML
International Conference
On Machine Learning

Experimental Research Questions

RQ1: Strategy: How effective is **standardization** compared to popular open-ended reasoning strategies (ICL, CoT, ToT)?

RQ2: Capabilities: Can current LLMs accurately solve **individual** threat-hunting tasks across the full lifecycle?

RQ3: Robustness: How do models handle **noisy inputs** (token-level vs. semantic noise) in realistic environments?

ICML
International Conference
On Machine Learning

RQ1: Standardization vs. Open-Ended

Method		Cybersecurity Agent			Industry-Leading LLM							
		LY	DH	SL	GK	G5	QW	GM	CD	L3.1	L4	GA
Playbook Recommend												
Open-ended	ICL5	42.3$_{\pm 1.9}$	54.2$_{\pm 2.4}$	54.7$_{\pm 2.3}$	65.8$_{\pm 2.6}$	74.4$_{\pm 2.8}$	52.8$_{\pm 2.0}$	79.4$_{\pm 1.8}$	63.7$_{\pm 2.1}$	65.8$_{\pm 2.6}$	55.8$_{\pm 2.0}$	54.9$_{\pm 2.1}$
	ICL10	44.1$_{\pm 1.8}$	52.5$_{\pm 2.6}$	55.3$_{\pm 2.7}$	66.5$_{\pm 2.7}$	75.8$_{\pm 2.7}$	53.6$_{\pm 2.1}$	80.2$_{\pm 1.9}$	64.9$_{\pm 2.2}$	66.4$_{\pm 2.7}$	56.4$_{\pm 2.1}$	55.5$_{\pm 2.2}$
	CoT	51.6$_{\pm 2.2}$	50.6$_{\pm 2.4}$	50.5$_{\pm 2.2}$	79.6$_{\pm 1.8}$	90.5$_{\pm 1.7}$	67.5$_{\pm 2.5}$	80.1$_{\pm 1.6}$	81.4$_{\pm 1.7}$	77.3$_{\pm 2.0}$	67.3$_{\pm 2.3}$	66.4$_{\pm 2.2}$
	ToT	48.1$_{\pm 2.4}$	53.3$_{\pm 2.6}$	54.3$_{\pm 2.5}$	76.5$_{\pm 2.1}$	86.4$_{\pm 1.8}$	71.4$_{\pm 2.4}$	83.5$_{\pm 1.7}$	77.2$_{\pm 2.0}$	82.1$_{\pm 1.8}$	72.1$_{\pm 2.1}$	71.2$_{\pm 2.2}$
Standardized (Ours)		**67.2**$_{\pm 1.7}$	**58.4**$_{\pm 2.1}$	**66.8**$_{\pm 1.9}$	**85.9**$_{\pm 1.4}$	**92.7**$_{\pm 1.3}$	**79.3**$_{\pm 2.0}$	**91.8**$_{\pm 1.3}$	**89.3**$_{\pm 1.6}$	**89.7**$_{\pm 1.5}$	**79.7**$_{\pm 1.9}$	**78.8**$_{\pm 2.0}$
Security Control Adjust												
Open-ended	ICL5	51.5$_{\pm 2.4}$	66.3$_{\pm 2.3}$	43.9$_{\pm 2.7}$	63.1$_{\pm 2.5}$	71.6$_{\pm 0.8}$	50.6$_{\pm 2.2}$	65.8$_{\pm 2.1}$	79.2$_{\pm 1.9}$	61.5$_{\pm 2.4}$	51.5$_{\pm 2.0}$	50.6$_{\pm 1.1}$
	ICL10	53.2$_{\pm 2.5}$	68.4$_{\pm 2.4}$	45.6$_{\pm 2.8}$	64.0$_{\pm 2.1}$	73.1$_{\pm 2.7}$	51.2$_{\pm 0.1}$	66.4$_{\pm 2.2}$	80.1$_{\pm 1.8}$	62.3$_{\pm 2.4}$	52.3$_{\pm 2.1}$	51.4$_{\pm 2.1}$
	CoT	60.3$_{\pm 2.0}$	70.5$_{\pm 2.1}$	68.4$_{\pm 1.9}$	71.6$_{\pm 1.9}$	81.5$_{\pm 1.7}$	59.8$_{\pm 2.4}$	79.2$_{\pm 1.7}$	77.2$_{\pm 1.9}$	77.9$_{\pm 1.8}$	67.9$_{\pm 1.4}$	63.0$_{\pm 2.1}$
	ToT	66.7$_{\pm 1.9}$	72.1$_{\pm 2.0}$	61.6$_{\pm 2.2}$	77.2$_{\pm 1.7}$	86.9$_{\pm 0.9}$	66.3$_{\pm 2.3}$	73.6$_{\pm 2.1}$	73.1$_{\pm 2.0}$	72.8$_{\pm 0.3}$	62.8$_{\pm 2.3}$	61.9$_{\pm 2.2}$
Standardized (Ours)		**74.2**$_{\pm 1.6}$	**77.6**$_{\pm 1.5}$	**80.1**$_{\pm 1.7}$	**83.4**$_{\pm 1.5}$	**91.0**$_{\pm 1.1}$	**74.7**$_{\pm 1.8}$	**88.5**$_{\pm 1.5}$	**86.5**$_{\pm 0.7}$	**86.4**$_{\pm 1.7}$	**76.4**$_{\pm 1.8}$	**75.5**$_{\pm 0.9}$
Patch Code Generation												
Open-ended	ICL5	10.8$_{\pm 1.0}$	49.8$_{\pm 3.1}$	29.2$_{\pm 2.7}$	57.5$_{\pm 2.7}$	59.7$_{\pm 1.5}$	39.3$_{\pm 2.9}$	63.7$_{\pm 2.2}$	47.5$_{\pm 2.5}$	49.2$_{\pm 0.6}$	39.2$_{\pm 2.9}$	38.3$_{\pm 3.8}$
	ICL10	12.6$_{\pm 1.6}$	51.2$_{\pm 3.0}$	31.5$_{\pm 2.8}$	59.1$_{\pm 2.6}$	60.4$_{\pm 1.4}$	40.1$_{\pm 1.8}$	64.9$_{\pm 0.3}$	48.6$_{\pm 0.4}$	50.1$_{\pm 1.1}$	40.1$_{\pm 1.7}$	39.2$_{\pm 2.7}$
	CoT	24.5$_{\pm 2.2}$	54.7$_{\pm 2.4}$	55.1$_{\pm 2.1}$	59.7$_{\pm 2.3}$	77.6$_{\pm 1.8}$	54.7$_{\pm 2.3}$	65.3$_{\pm 2.0}$	66.3$_{\pm 2.0}$	67.4$_{\pm 1.8}$	57.4$_{\pm 2.1}$	51.5$_{\pm 2.2}$
	ToT	25.3$_{\pm 2.1}$	50.9$_{\pm 2.5}$	58.3$_{\pm 2.0}$	63.1$_{\pm 2.2}$	73.8$_{\pm 1.9}$	50.2$_{\pm 2.5}$	69.8$_{\pm 1.9}$	61.4$_{\pm 2.1}$	62.9$_{\pm 2.0}$	52.9$_{\pm 2.3}$	52.2$_{\pm 2.3}$
Standardized (Ours)		**29.7**$_{\pm 1.9}$	**63.4**$_{\pm 1.8}$	**60.2**$_{\pm 2.0}$	**73.8**$_{\pm 1.6}$	**88.7**$_{\pm 1.2}$	**65.4**$_{\pm 2.0}$	**82.6**$_{\pm 1.4}$	**79.2**$_{\pm 1.6}$	**80.6**$_{\pm 1.5}$	**70.6**$_{\pm 1.8}$	**69.7**$_{\pm 1.9}$
Patch Tool Suggestion												
Open-ended	ICL5	48.2$_{\pm 2.6}$	65.2$_{\pm 2.3}$	61.5$_{\pm 2.5}$	70.2$_{\pm 2.1}$	80.7$_{\pm 1.8}$	59.2$_{\pm 2.4}$	74.1$_{\pm 2.1}$	68.5$_{\pm 2.3}$	70.3$_{\pm 2.3}$	60.3$_{\pm 2.4}$	59.4$_{\pm 2.6}$
	ICL10	49.1$_{\pm 2.5}$	64.7$_{\pm 2.4}$	63.1$_{\pm 2.6}$	71.0$_{\pm 2.1}$	81.9$_{\pm 1.8}$	60.3$_{\pm 2.4}$	74.9$_{\pm 2.0}$	69.8$_{\pm 2.2}$	71.4$_{\pm 2.2}$	61.4$_{\pm 2.4}$	60.5$_{\pm 2.5}$
	CoT	53.6$_{\pm 2.2}$	70.1$_{\pm 2.1}$	77.2$_{\pm 1.8}$	80.5$_{\pm 1.6}$	91.4$_{\pm 1.4}$	70.3$_{\pm 2.0}$	81.7$_{\pm 1.6}$	79.1$_{\pm 1.8}$	79.6$_{\pm 1.9}$	69.6$_{\pm 2.0}$	68.7$_{\pm 2.1}$
	ToT	56.5$_{\pm 2.0}$	71.8$_{\pm 2.0}$	68.1$_{\pm 2.2}$	77.1$_{\pm 1.8}$	87.6$_{\pm 1.6}$	74.5$_{\pm 1.9}$	86.3$_{\pm 1.5}$	83.7$_{\pm 1.2}$	84.2$_{\pm 1.6}$	74.2$_{\pm 1.8}$	67.3$_{\pm 2.2}$
Standardized (Ours)		**69.1**$_{\pm 1.8}$	**76.5**$_{\pm 1.7}$	**77.7**$_{\pm 1.7}$	**88.7**$_{\pm 1.3}$	**98.2**$_{\pm 1.1}$	**83.6**$_{\pm 1.6}$	**93.2**$_{\pm 1.3}$	**91.2**$_{\pm 1.5}$	**92.1**$_{\pm 1.4}$	**82.1**$_{\pm 1.6}$	**81.2**$_{\pm 1.7}$
Advisory Correlation												
Open-ended	ICL5	21.7$_{\pm 3.8}$	57.5$_{\pm 2.6}$	63.8$_{\pm 2.5}$	66.0$_{\pm 2.3}$	68.5$_{\pm 2.3}$	48.5$_{\pm 3.3}$	62.4$_{\pm 2.6}$	56.8$_{\pm 2.7}$	58.7$_{\pm 2.6}$	48.7$_{\pm 3.1}$	47.8$_{\pm 3.2}$
	ICL10	22.9$_{\pm 3.7}$	59.1$_{\pm 2.6}$	64.7$_{\pm 2.5}$	67.2$_{\pm 2.3}$	69.4$_{\pm 2.2}$	49.2$_{\pm 3.1}$	63.2$_{\pm 2.5}$	58.1$_{\pm 2.6}$	59.5$_{\pm 1.2}$	49.5$_{\pm 3.1}$	48.6$_{\pm 3.2}$
	CoT	49.5$_{\pm 2.3}$	71.4$_{\pm 2.0}$	69.5$_{\pm 2.1}$	68.5$_{\pm 2.1}$	81.8$_{\pm 1.7}$	61.7$_{\pm 2.4}$	77.5$_{\pm 1.8}$	76.2$_{\pm 1.9}$	76.3$_{\pm 0.9}$	66.3$_{\pm 2.1}$	65.4$_{\pm 1.1}$
	ToT	46.8$_{\pm 2.3}$	73.2$_{\pm 1.9}$	67.2$_{\pm 2.2}$	72.1$_{\pm 1.9}$	85.5$_{\pm 1.6}$	64.8$_{\pm 2.2}$	73.1$_{\pm 1.9}$	72.5$_{\pm 2.0}$	71.8$_{\pm 2.1}$	61.8$_{\pm 2.3}$	60.9$_{\pm 2.3}$
Standardized (Ours)		**73.4**$_{\pm 1.7}$	**78.8**$_{\pm 1.5}$	**77.1**$_{\pm 1.6}$	**81.6**$_{\pm 1.4}$	**93.6**$_{\pm 1.2}$	**76.5**$_{\pm 1.8}$	**86.9**$_{\pm 1.4}$	**84.5**$_{\pm 1.6}$	**84.9**$_{\pm 1.5}$	**74.9**$_{\pm 1.7}$	**74.0**$_{\pm 1.7}$

Standardized workflows significantly outperform open-ended reasoning by decomposing complexity and reducing error propagation. (Data shown for GPT-5.1 on Playbook Recommendation)

ICML
International Conference
On Machine Learning

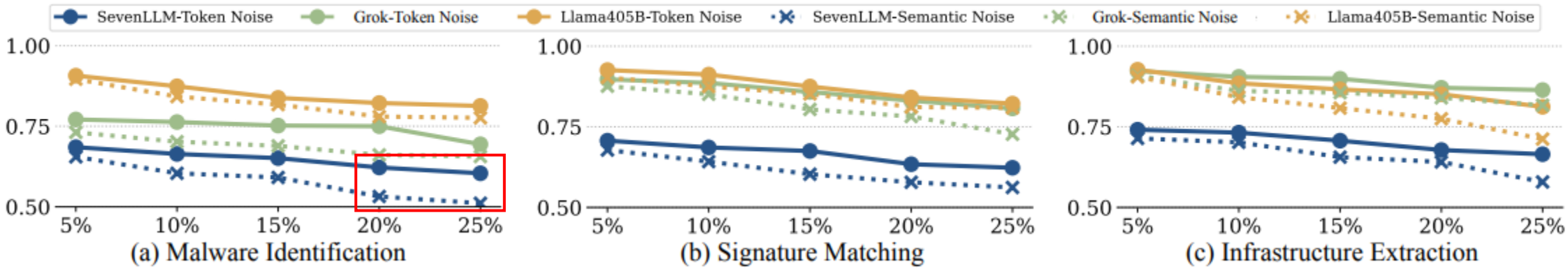
RQ2: Performance by Task Type

Analytical Category	Representative Task	Standard Operation	G5 Perf (%)
Threat Attribution	Actor Identification	NER + RAG + MAP	94.1
Behavior Analysis	Event Reconstruction	SUM + Reasoning	88.6
Prioritization	Severity Scoring	SUM + MATH	86.5
Response	Patch Code Gen	RAG + SUM	88.7

Insight: Models maintain consistent logic even across the longest task chains (up to 30 stages).

ICML
International Conference
On Machine Learning

RQ3: Robustness against Noise



Key Findings on Noise

Models show high resilience to token-level noise (character swaps) but struggle with **semantic-level noise** (paraphrased misleading context).

Perturbing just 10% of semantic context leads to a much sharper performance decline compared to 10% character-level noise.

Thank You !
Q&A

BINGHAMTON
UNIVERSITY
STATE UNIVERSITY OF NEW YORK