LLM4Sec
Workshop on the use of Large Language Models for Cybersecurity

Capital One

AutoBnB-RAG: Enhancing Multi-Agent Incident Response with Retrieval-Augmented Generation

Zefang Liu (Capital One, Georgia Institute of Technology)
Arman Anwar (Georgia Institute of Technology)

Paper ID: S14202

Georgia Tech

Introduction

Why Incident Response Is Hard Today?

- Cyber threats are fast, multistage, and constantly evolving
- Traditional incident response is human driven and slow under pressure
- LLM based agents show promising reasoning and collaboration abilities
- But without access to external knowledge, they may still hallucinate or miss critical context
- Opportunity: retrieve real knowledge in real time to improve accuracy and speed

Image source: https://swimlane.com/blog/incident-response/

Simulation Framework

Backdoors & Breaches (B&B) as the Foundation:

- Cooperative tabletop game for realistic incident response training by Black Hills Information Security
- Goal: defenders uncover 4 hidden attack stages within 10 turns
- Four attack categories: Initial Compromise → Pivot & Escalate → C2 / Exfiltration → Persistence
- 50+ attack cards + 12 procedure cards used for detection and investigation
- Each turn: team selects one procedure → 20-sided dice roll → success (11+) or failure
- Four “established” procedures receive a +3 modifier to represent real world maturity
- If all attack stages are revealed within 10 turns → team wins

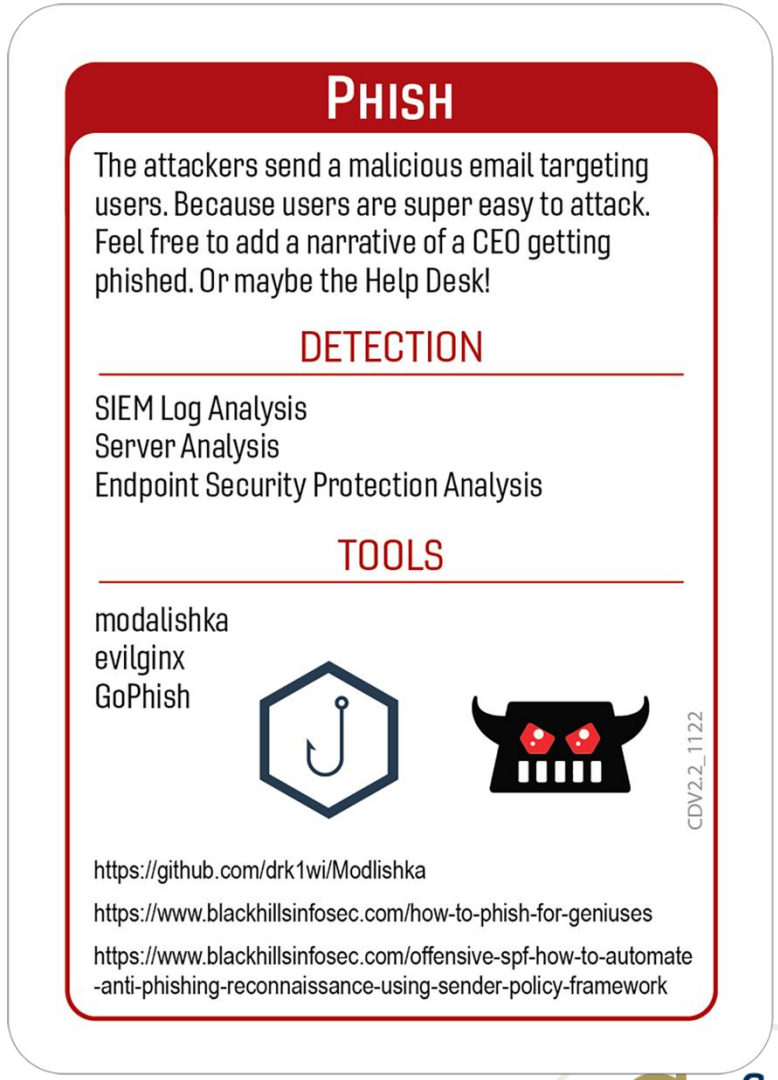


Image source: https://www.blackhillsinfosec.com/tools/backdoorsandbreaches/

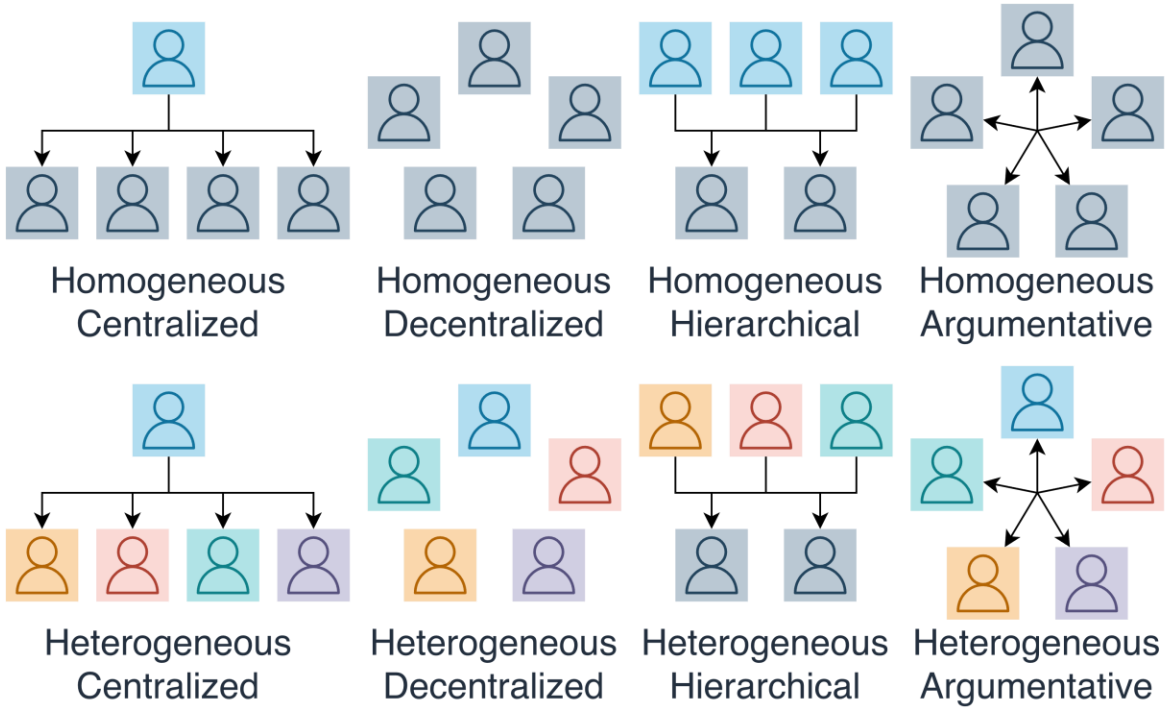
Team Structures

LLM-Based Simulation Setup:

- Human players in B&B are fully replaced by LLM agents
- Environment automatically handles rules, card selection, and dice logic
- One agent acts as the incident captain
- Five defender agents communicate and decide collaboratively

Team Structure Variants:

- Centralized vs. decentralized coordination
- Homogeneous vs. heterogeneous expertise roles
- Hierarchical experience levels
- Also explore argumentative teams that actively challenge each other



Georgia Tech

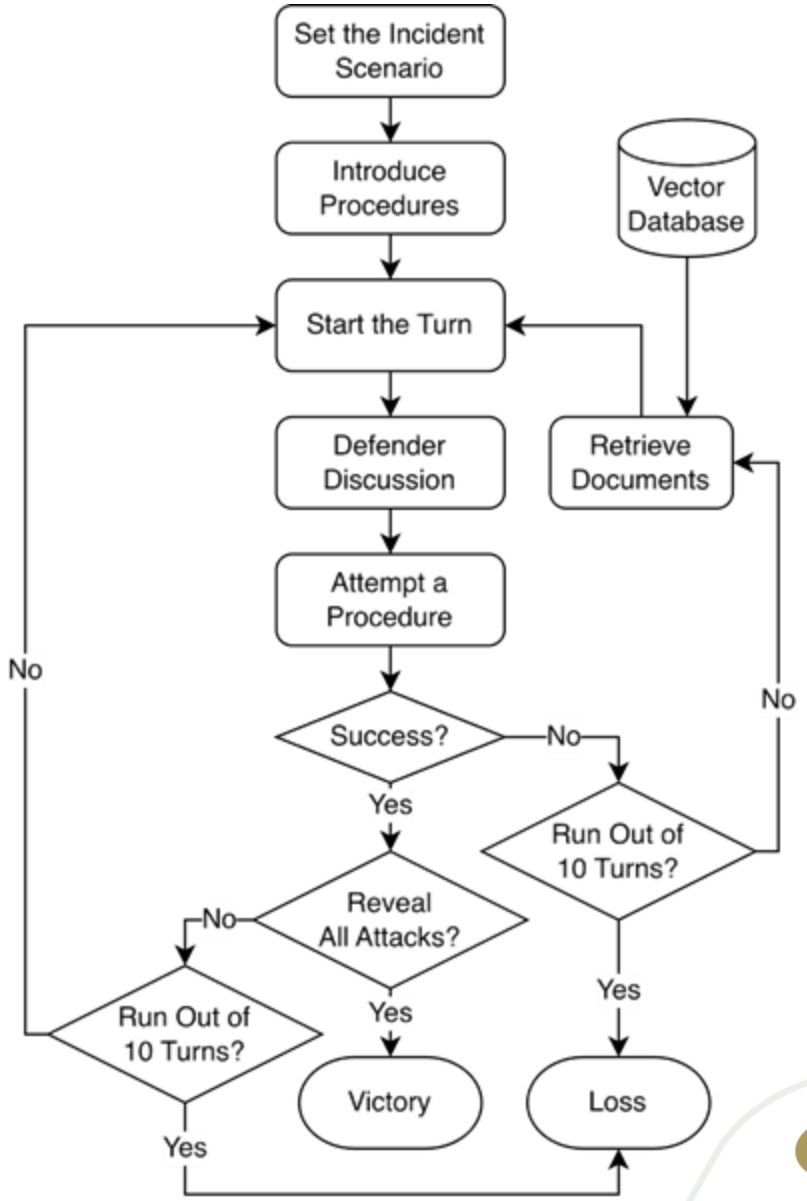
Retrieval-Augmented Generation

Why Retrieval Matters:

- LLMs can reason well but may hallucinate without real-world knowledge
- External knowledge is often required during incident investigation

Our Integration:

- Retrieval happens after a failed investigation step
- A dedicated retrieval agent pulls relevant knowledge
- Information is returned quietly and used by the team in the next turn



Georgia Tech

External Knowledge Sources

RAG-Wiki (Webpage Collection):

- 125 curated cybersecurity webpages
- Sources include Wikipedia, MITRE ATT&CK, Microsoft Learn, and security blogs
- Covers technical concepts, attack techniques, and defensive strategies

RAG-News (Synthetic Incident Reports):

- 100 realistic narrative-style incident simulations
- Generated to reflect real multistage attack investigations
- Helps agents learn from past breach-style scenario reasoning

Webpages Collection for RAG-Wiki

Source Category	Count	Percentage
Wikipedia	67	53.6%
MITRE ATT&CK	9	7.2%
Microsoft Learn / Support	6	4.8%
CISA / Government	3	2.4%
Cybersecurity Blogs / Vendors	27	21.6%
Other	13	10.4%
Total	**125**	**100%**

Experimental Setup

- Implemented using the AutoGen framework with GPT-4o
- Each simulation includes 1 incident captain + 5 defender agents
- 8 different team structures evaluated (coordination and expertise vary)
- Each structure tested across 30 independent runs
- Compared retrieval settings:
  - No Retrieval (baseline)
  - RAG-Wiki
  - RAG-News

Experimental Results

Key Observation:

- Retrieval consistently improves performance across all team structures

Notable Trends:

- Largest gains seen in centralized and hierarchical teams
- RAG-News often outperforms RAG-Wiki
- Argumentative teams also show improvement, but smaller than centralized teams
- No team structure performs best without retrieval

Win Rates (%) and Performance Gains

Team	Base	RAG-Wiki	RAG-News
Homo. Cen.	20.0	**50.0** (+30.0)	60.0 (+40.0)
Hetero. Cen.	30.0	43.3 (+13.3)	63.3 (+33.3)
Homo. Decen.	33.3	40.0 (+6.7)	43.3 (+10.0)
Hetero. Decen.	26.7	**50.0** (+23.3)	50.0 (+23.3)
Homo. Hier.	23.3	40.0 (+16.7)	43.3 (+20.0)
Hetero. Hier.	30.0	36.7 (+6.7)	**70.0** (+40.0)
Homo. Arg.	23.3	43.3 (+20.0)	46.7 (+23.4)
Hetero. Arg.	30.0	46.7 (+16.7)	53.3 (+23.3)

Ablation Studies

What We Tested:

- Effect of number of retrieved documents (Top 1, Top 3, Top 5)
- Effect of retrieval chunk size (1k vs 5k characters)

Key Findings:

- Performance remains stable across different Top k values
- Larger chunks are generally more helpful because more context is preserved
- Retrieval is robust and does not require precise fine tuning

Win Rates (%) for Varying Numbers of Retrieved Documents

Setting	Top-1	Top-3	Top-5
RAG-Wiki	46.7	**50.0**	46.7
RAG-News	60.0	60.0	**63.3**

Win Rates (%) for Different Document Chunk Sizes

Setting	1k Chars	5k Chars
RAG-Wiki	33.3	**50.0**
RAG-News	**63.3**	60.0

Credential Stuffing on The North Face

Turn	Procedure	Roll	Modifier	Success	Revealed Incident	Retrieval
1	User and Entity Behavior Analytics	10	+3	Yes	Internal Password Spray	No
2	SIEM Log Analysis	12	+3	Yes	-	Yes
3	Server Analysis	19	+0	Yes	Credential Stuffing	No
4	Network Threat Hunting - Zeek/RITA Analysis	17	+0	Yes	HTTPS as Exfil	No
5	Endpoint Security Protection Analysis	10	+0	No	-	Yes
6	Endpoint Analysis	5	+0	No	-	Yes
7	Endpoint Security Protection Analysis	4	+0	No	-	Yes
8	Endpoint Analysis	20	+0	Yes	New User Added	No

Roundcube Exploit at Cock.li

Turn	Procedure	Roll	Modifier	Success	Revealed Incident	Retrieval
1	Endpoint Security Protection Analysis	2	+3	No	-	Yes
2	SIEM Log Analysis	6	+0	No	-	Yes
3	Network Threat Hunting - Zeek/RITA Analysis	4	+0	No	-	Yes
4	Server Analysis	12	+0	Yes	Web Server Compromise	No
5	User and Entity Behavior Analytics	8	+0	No	-	Yes
6	Endpoint Analysis	13	+0	Yes	Local Privilege Escalation	No
7	Network Threat Hunting - Zeek/RITA Analysis	19	+0	Yes	HTTP as Exfil	No
8	Endpoint Security Protection Analysis	1	+3	No	-	Yes
9	Endpoint Analysis	7	+0	No	-	Yes
10	Endpoint Security Protection Analysis	14	+3	Yes	Registry Keys for Persistence	No

Supply Chain Attack on Gluestack

Turn	Procedure	Roll	Modifier	Success	Revealed Incident	Retrieval
1	SIEM Log Analysis	9	+3	Yes	Weaponizing Active Directory	No
2	Endpoint Analysis	2	+3	No	-	Yes
3	Endpoint Security Protection Analysis	17	+0	Yes	Malware Injection Into Client Software	No
4	Network Threat Hunting - Zeek/RITA Analysis	11	+0	Yes	Supply Chain Attack	No
5	Firewall Log Review	8	+0	No	-	Yes
6	Network Threat Hunting - Zeek/RITA Analysis	12	+0	Yes	Gmail, Tumblr, Salesforce, Twitter as C2	No

Conclusion

Key Takeaways:

- LLM agents can realistically simulate incident response teams
- Retrieval augmentation clearly improves performance
- RAG-News provides strong benefits through narrative context than RAG-Wiki
- Team structure influences effectiveness, with centralized and hierarchical teams benefiting the most

Broader Insight:

- Reasoning alone is not enough, while informed reasoning is essential for incident responses.

Thank You!

Georgia Tech

Backdoors & Breaches Cards

PHISH

The attackers send a malicious email targeting users. Because users are super easy to attack. Feel free to add a narrative of a CEO getting phished. Or maybe the Help Desk!

DETECTION

SIEM Log Analysis
Server Analysis
Endpoint Security Protection Analysis

TOOLS

modalishka
evilginx
GoPhish

https://github.com/drk1wi/Modlishka

https://www.blackhillsinfosec.com/how-to-phish-for-geniuses

https://www.blackhillsinfosec.com/offensive-spf-how-to-automate-anti-phishing-reconnaissance-using-sender-policy-framework

CDV2.2_1122

INTERNAL PASSWORD SPRAY

The attackers start a password spray against the rest of the organization from a compromised system.

DETECTION

User and Entity Behavior Analytics
Cyber Deception
SIEM Log Analysis

TOOLS

DomainPasswordSpray
BruteLoops
Kerbrute
Metasploit

https://github.com/dafthack/DomainPasswordSpray

https://github.com/ropnop/kerbrute

https://www.blackhillsinfosec.com/webcast-attack-tactics-5-zero-to-hero-attack

CDV2.2_1122

HTTP AS EXFIL

The attackers use HTTP as an exfil method. This is usually used in conjunction with some type of stego. For example, VSAgent uses base64 encoded __VIEWSTATE as an exfil field.

DETECTION

Network Threat Hunting - Zeek/RITA Analysis
Firewall Log Review

TOOLS

Metasploit Reverse HTTP Payloads
C2 Matrix

https://www.thec2matrix.com/

CDV2.2_1122

MALICIOUS SERVICE

The attackers add a service that starts every time the system starts.

DETECTION

Endpoint Security Protection Analysis
Memory Analysis
Endpoint Analysis

TOOLS

Meterpreter Persistence Modules
msconfig.exe
SILENTTRINITY
Sysinternals:
- autoruns.exe

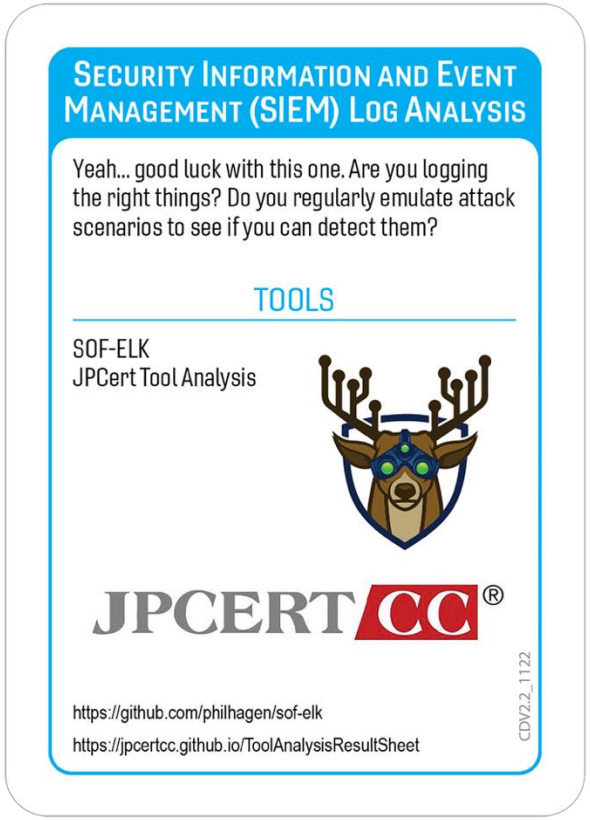
https://github.com/byt3bl33d3r/SILENTTRINITY

https://learn.microsoft.com/en-us/sysinternals/

CDV2.2_1122

Image source: https://www.blackhillsinfosec.com/tools/backdoorsandbreaches/

Backdoors & Breaches Cards



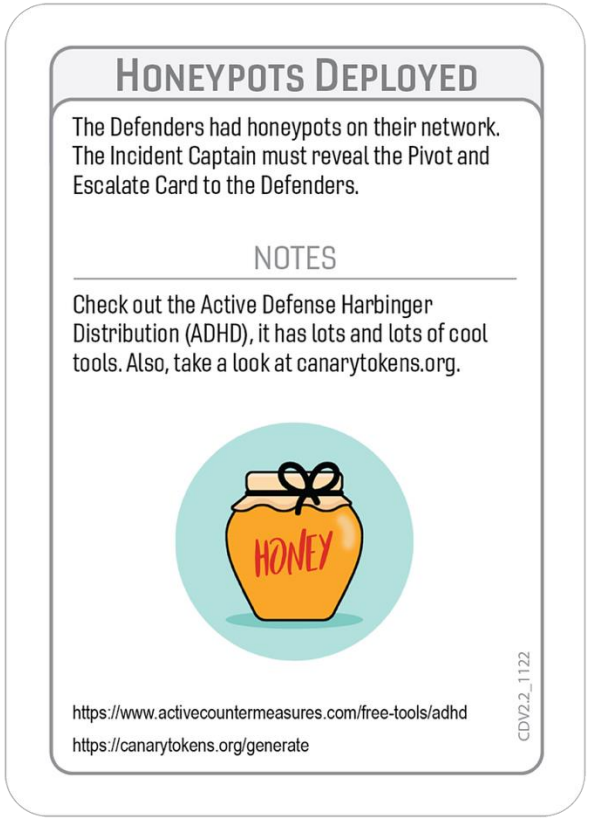


Image source: https://www.blackhillsinfosec.com/tools/backdoorsandbreaches/