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

Zefang Liu (Capital One, Georgia Tech), *Arman Anwar* (Georgia Tech)
Contact: liuzefang@gatech.edu

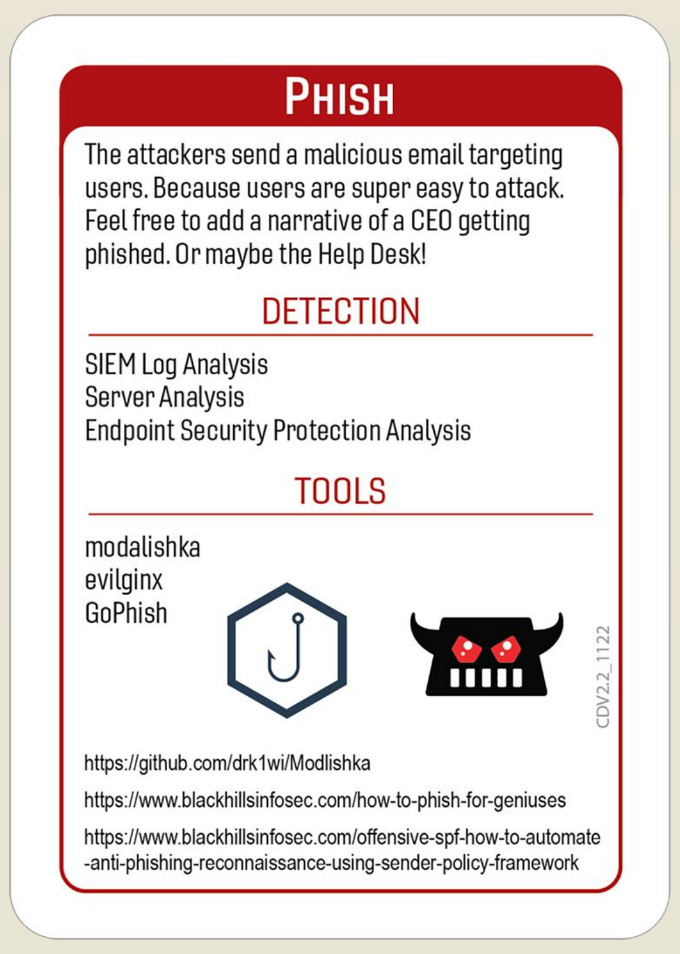
ICDM 2025
LLM4Sec

Introduction

- Modern **cyber threats** are complex and multi-stage, demanding rapid detection and coordinated response.
- Effective **Incident Response (IR)** relies on timely decisions, teamwork, and adaptive reasoning.
- **Large Language Models (LLMs)** can simulate and support IR teams but often lack access to external knowledge.
- **AutoBnB-RAG** extends the AutoBnB framework with **Retrieval-Augmented Generation (RAG)**, enabling agents to fetch and apply external cybersecurity information during collaborative investigations.

Simulation Framework

- The simulation is built on **Backdoors & Breaches (B&B)**, a tabletop game for training **cyber incident response** teams.
- In B&B, defenders uncover a hidden **four-stage attack path** by selecting investigative procedure cards and rolling a **20-sided die**; rolls of **11 or higher** reveal an attack card.
- The four stages are **Initial Compromise**, **Pivot and Escalate**, **Command and Control (C2) and Exfiltration**, and **Persistence**.
- Each team has **10 turns** to reveal all stages and win the game; otherwise, the incident remains unresolved.
- **AutoBnB-RAG** digitizes this process with **LLM agents** replacing human players.
- The system automates game logic, card handling, and dice rolls to ensure consistent, repeatable evaluations.



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

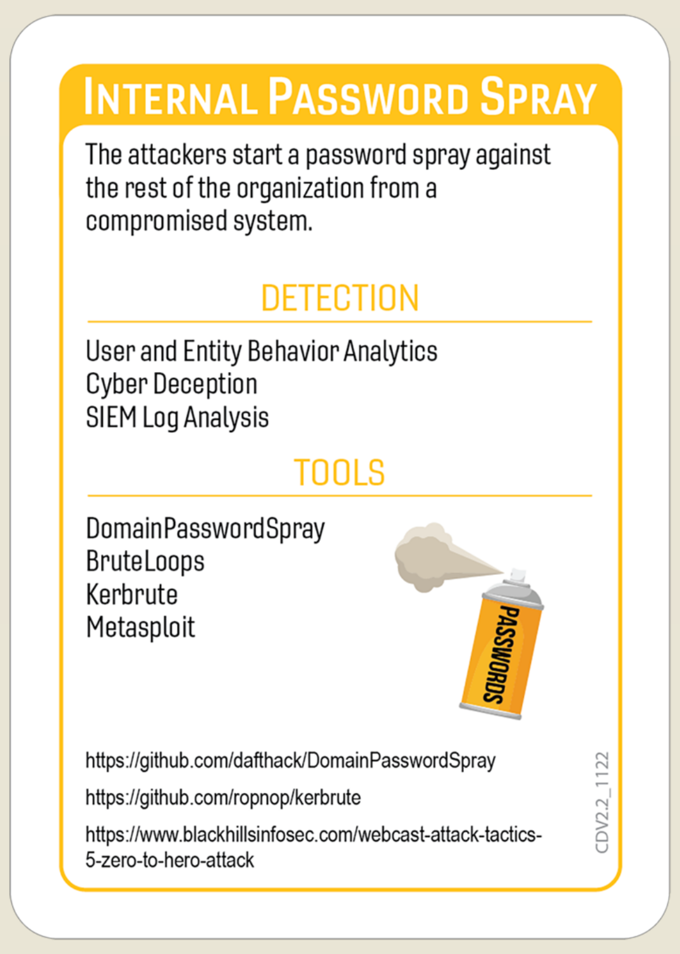
TOOLS

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

Initial Compromise



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

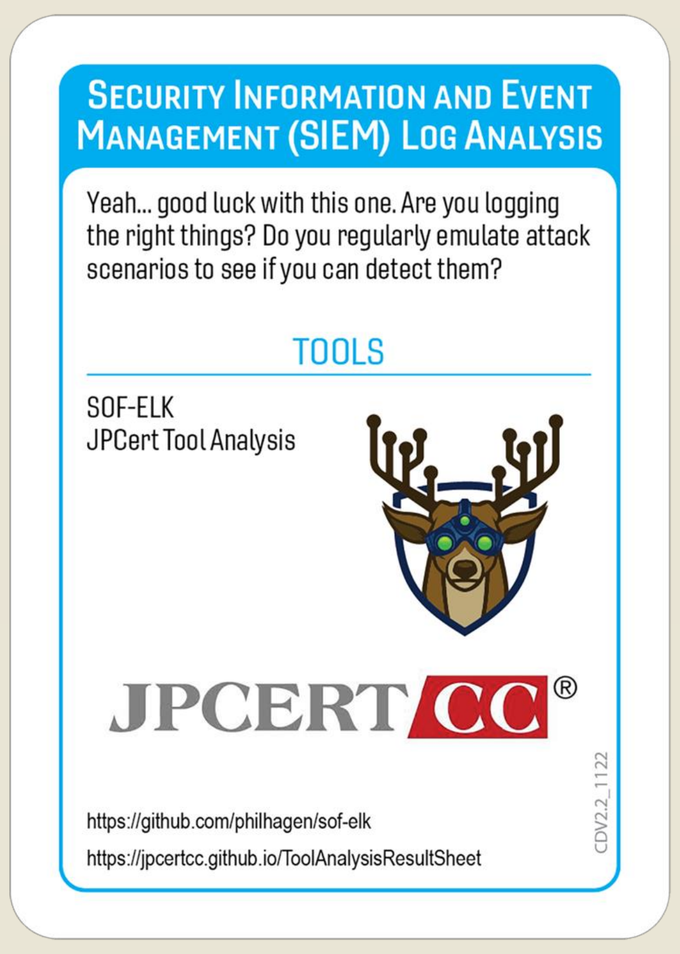
TOOLS

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

Pivot and Escalate



SECURITY INFORMATION AND EVENT MANAGEMENT (SIEM) LOG ANALYSIS

Yeah... good luck with this one. Are you logging the right things? Do you regularly emulate attack scenarios to see if you can detect them?

TOOLS

SOF-ELK
JPCert Tool Analysis

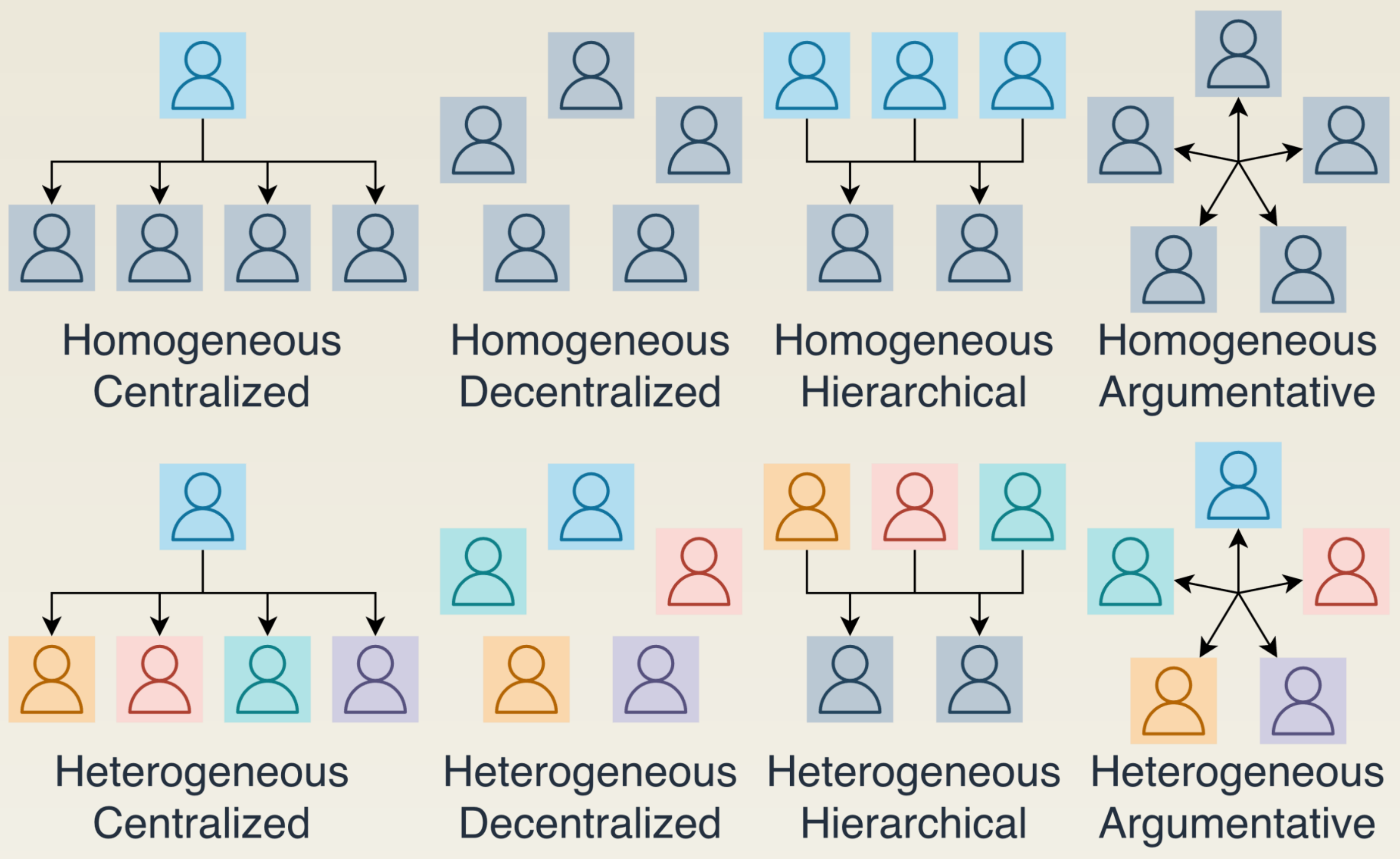
https://github.com/philhagen/sof-elk
https://jpcertcc.github.io/ToolAnalysisResultSheet

CDV2.2_1122

Procedure

Team Structures

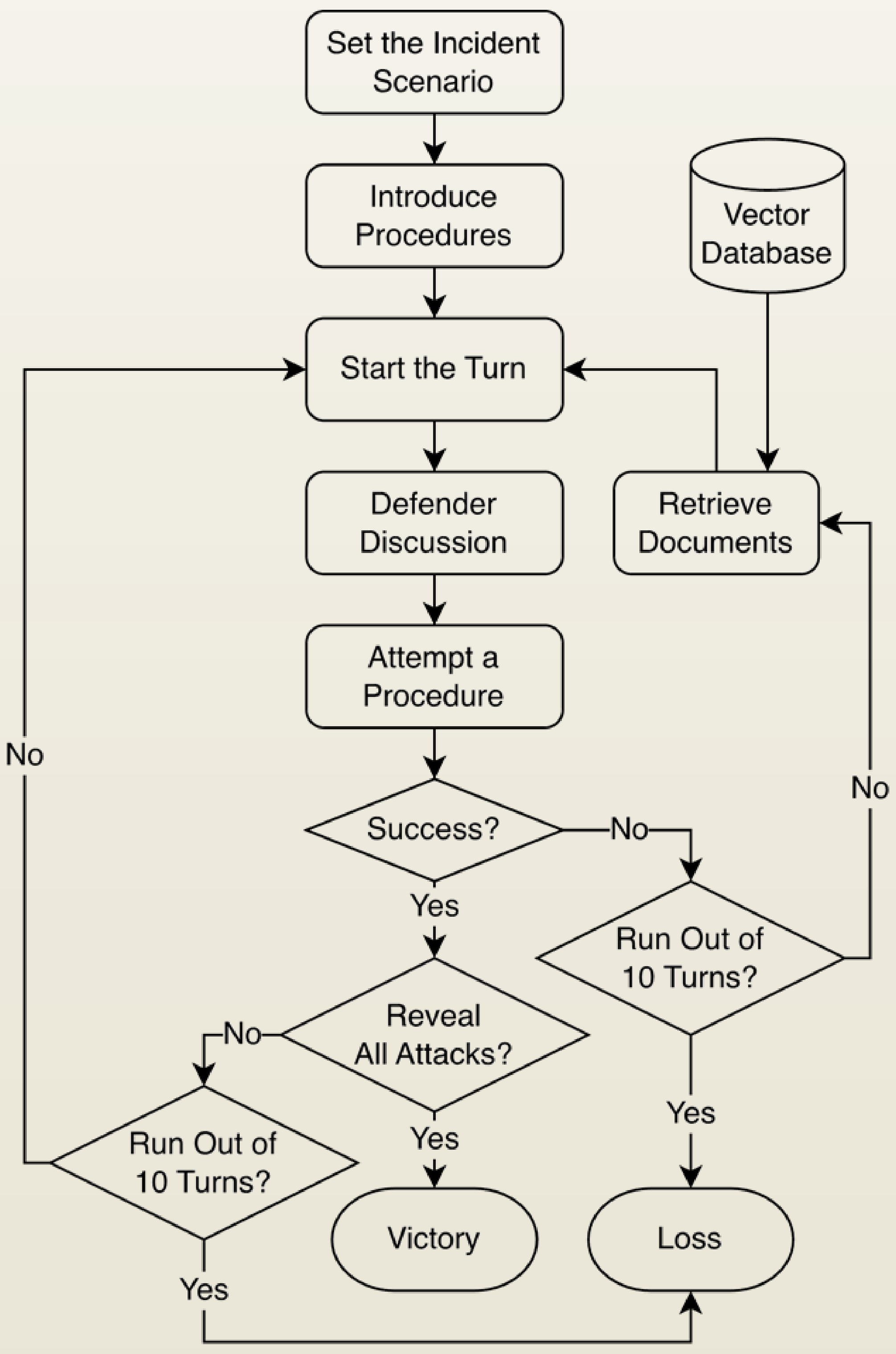
- **Eight organizational models** were tested to examine how coordination style affects IR.
- Structures included **centralized**, **decentralized**, **hierarchical**, and **argumentative** variants, each in **homogeneous** or **heterogeneous** form.
- **Homogeneous teams** shared the same role expertise, while **heterogeneous teams** combined defenders with different specialties.


Homogeneous Centralized
Homogeneous Decentralized
Homogeneous Hierarchical
Homogeneous Argumentative
Heterogeneous Centralized
Heterogeneous Decentralized
Heterogeneous Hierarchical
Heterogeneous Argumentative

Experimental Setup

- Simulations run with **GPT-4o** via **AutoGen** (temperature 0.7).
- Each simulation includes **1 incident captain** and **5 defenders**.
- Each of the 8 team structures evaluated over **30 runs**.
- **Top 3 documents** are retrieved using **LangChain** and **Chroma** (OpenAI embeddings).

Retrieval



- **Retrieval** is triggered after failed procedures to simulate real-world reference consultation.
- A **retrieval agent** returns relevant context to support the team's reasoning.
- **Two knowledge sources:**
  - **RAG-Wiki:** 125 curated webpages (Wikipedia, MITRE ATT&CK, Microsoft Learn, etc.)
  - **RAG-News:** 100 synthetic narrative incident reports

Experimental Results

- **Retrieval improved success rates** across all 8 team structures in simulated incident response.
- **Argumentative team configurations**, newly introduced in this study, also benefited from retrieval integration.
- **Retrieval-augmented teams** showed clearer investigative progress and fewer repeated failed actions compared to non-retrieval runs.

Win Rates (%) and Performance Gains

Team	Base	RAG-Wiki	RAG-News
Homogeneous Centralized	20.0	**50.0** (+30.0)	60.0 (+40.0)
Heterogeneous Centralized	30.0	43.3 (+13.3)	63.3 (+33.3)
Homogeneous Decentralized	33.3	40.0 (+6.7)	43.3 (+10.0)
Heterogeneous Decentralized	26.7	**50.0** (+23.3)	50.0 (+23.3)
Homogeneous Hierarchical	23.3	40.0 (+16.7)	43.3 (+20.0)
Heterogeneous Hierarchical	30.0	36.7 (+6.7)	**70.0** (+40.0)
Homogeneous Argumentative	23.3	43.3 (+20.0)	46.7 (+23.4)
Heterogeneous Argumentative	30.0	46.7 (+16.7)	53.3 (+23.3)

Ablation Studies

- Retrieval performance remained stable across different **top-k values** and improved with larger **chunk sizes**.

Numbers of Documents

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

Document Chunk Sizes

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

Conclusion

- **AutoBnB-RAG** integrates retrieval-augmented generation into multi-agent incident response simulations.
- Retrieval grounding improves **decision quality**, **coordination**, and **overall success rates** across diverse team structures.
- The framework demonstrates how **external knowledge** can enhance **reasoning**, **adaptability**, and **realism** in LLM-driven cybersecurity research.

References

Wu, Qingyun, et al. "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation." *ICLR 2024 Workshop on Large Language Model Agents*.
Liu, Zefang. "Multi-Agent Collaboration in Incident Response with Large Language Models." *AAAI 2025 Workshop on Multi-Agent AI in the Real World*.
Liu, Zefang. "AutoBnB: Multi-Agent Incident Response with Large Language Models." *2025 13th International Symposium on Digital Forensics and Security (ISDFS)*. IEEE, 2025.

SCAN ME

Paper ID:S14202