13th INTERNATIONAL SYMPOSIUM ON
DIGITAL FORENSICS AND SECURITY
WENTWORTH INSTITUTE OF TECHNOLOGY
BOSTON, MA, USA
APRIL 24 – 25, 2025

AutoBnB: Multi-Agent Incident Response with Large Language Models

Zefang Liu
Georgia Institute of Technology

Georgia Tech

Motivation

Why is this important?

- Cyberattacks are growing in complexity and frequency.
- Effective incident response (IR) is essential but difficult to scale with human-only teams.

Challenges in Traditional IR:

- Requires coordination across multiple roles under time pressure.
- Human bottlenecks in decision-making and expertise.

Opportunity:

- Large Language Models (LLMs) are strong in reasoning, communication, and decision support.
- Can we leverage LLMs as autonomous agents to enhance IR collaboration?

Georgia Tech

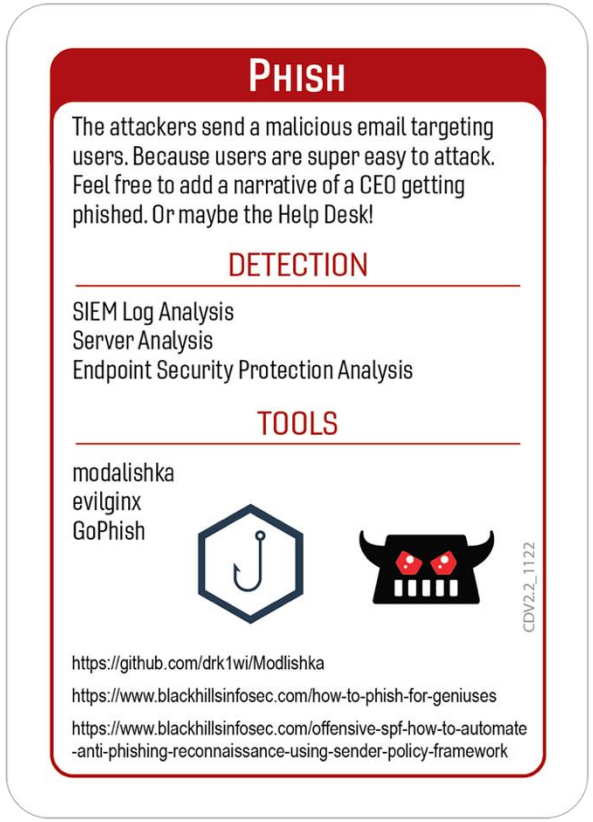
Backdoors & Breaches

What is Backdoors & Breaches (B&B)?

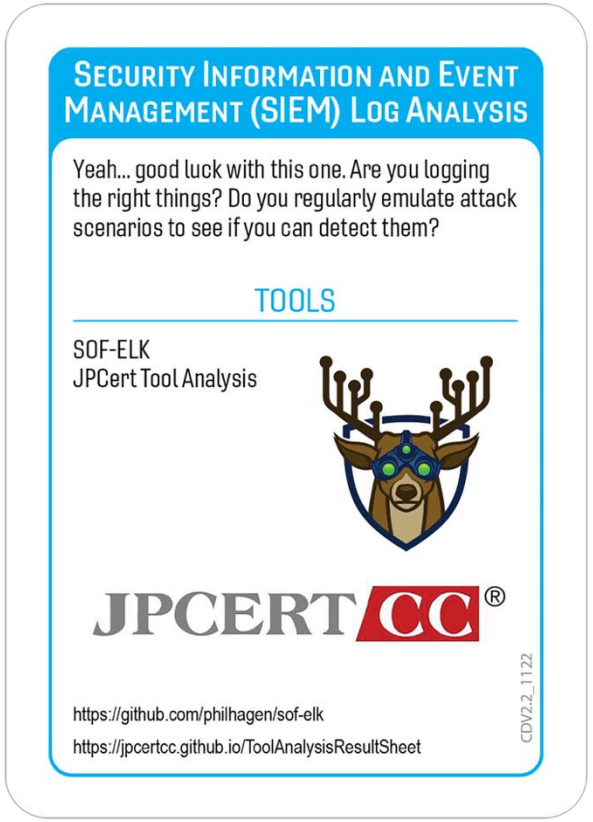
- A tabletop card game simulating real-world cybersecurity incident response.
- Developed by Black Hills Information Security and Active Countermeasures for training and education.

Card Types:

- **Attack Cards:** Represent stages of a cyberattack (i.e., Initial Compromise, Pivot and Escalate, Command & Control (C2) and Exfiltration, Persistence).
- **Procedure Cards:** Detection methods (e.g., log analysis, threat hunting).
- **Inject Cards:** Random disruptive events (excluded).
- **Consultant Cards:** Special aids (also excluded).



Attack Card



Procedure Card

Backdoors & Breaches card images used for research only.
Original content © Black Hills Information Security.

Backdoors & Breaches

Goal:

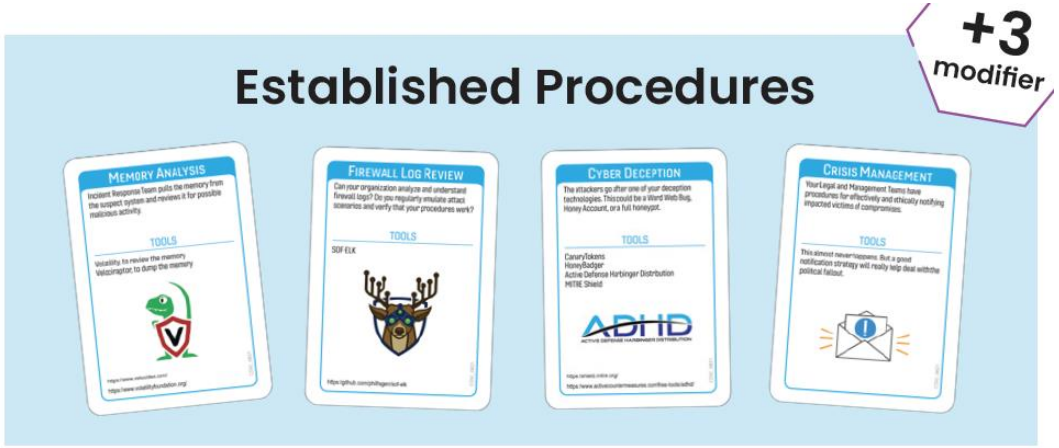
- Defenders work to reveal **four hidden attack cards** (one per attack phase) within **10 turns**.

Roles:

- **Incident Captain:** Runs the game and sets the scenario.
- **Defenders:** Choose and apply procedures to identify attacks.

How it plays:

- Each turn, defenders select a **procedure** and roll a 20-sided die.
- Success (11+) may uncover a related hidden threat.
- Team coordination and strategy are critical.



Backdoors & Breaches card images used for research only.
Original content © Black Hills Information Security.

Methodology

Simulation Framework:

- Implemented using the **AutoGen** multi-agent framework.
- Agents powered by **GPT-4o** simulate incident response roles.
- Each game includes **1 incident captain** and **5 defender agents**.
- Agents interact via a shared group chat.
- Decisions are made autonomously based on role knowledge and communication.

Team Structures Explored:

- **Centralized:** One leader coordinates decisions.
- **Decentralized:** No leader; agents decide by consensus.
- **Hybrid:** Experienced agents support beginners.
- Each structure tested in **homogeneous** (generalist) and **heterogeneous** (specialist or mixed-experience) variants.

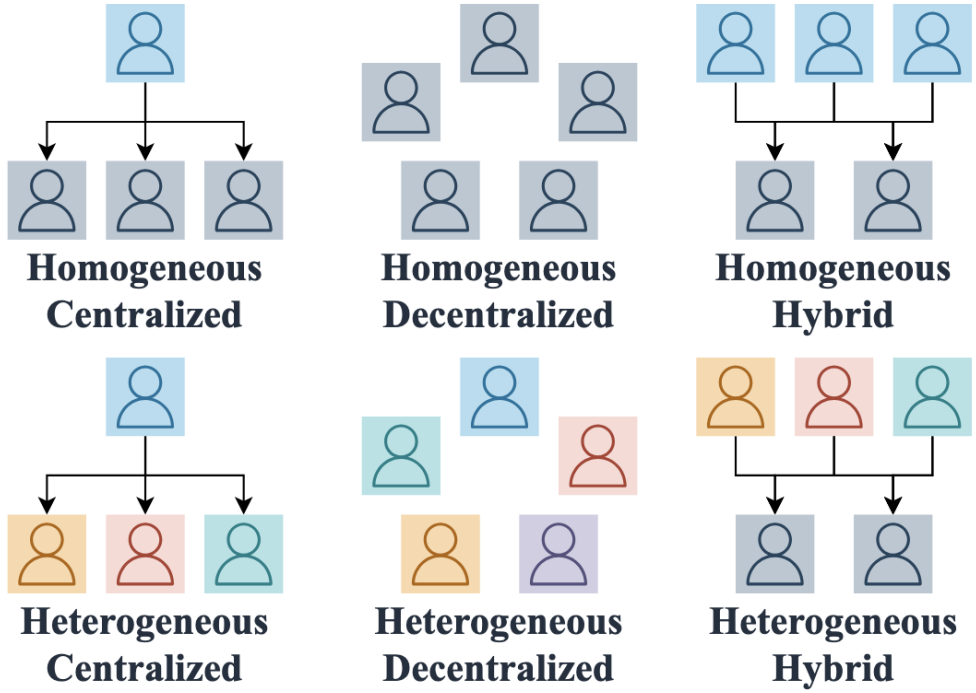


Fig. 1: Visualization of defender team structures used in the *Backdoors & Breaches* incident response simulation.

Georgia Tech

Experiment Setup & Example

Simulation Parameters:

- Total of **150 games** conducted.
- Each team configuration was tested in **25 simulations**.

Agent Configuration:

- All agents powered by **GPT-4o**, temperature set to 1.
- Predefined role prompts assigned to guide behavior.

Roles per Simulation:

- **1 Incident Captain**: Initializes the scenario, enforces rules, facilitates turns.
- **5 Defender Agents**: Roles vary by team structure.

Consistency Controls:

- Same pool of attack and procedure cards across all simulations for each seed.

TABLE I: Turn-by-turn game trajectory from a simulation using the homogeneous centralized team structure.

Turn	Procedure	Base Roll	Modifier	Success	Revealed Incident
1	Network Threat Hunting - Zeek/RITA Analysis	13	+3	Yes	Windows BITS
2	SIEM Log Analysis	15	+3	No	-
3	Server Analysis	16	+3	No	-
4	Firewall Log Review	15	+3	No	-
5	Endpoint Analysis	16	+0	Yes	New Service Creation/Modification
6	Memory Analysis	13	+0	Yes	Malicious Driver
7	User and Entity Behavior Analytics	16	+0	Yes	Insider Threat

Experiment Results

Key Findings:

- **Heterogeneous Hybrid** and **Homogeneous Decentralized** teams had the highest win rates (36%).
- **Centralized teams** (both homogeneous and heterogeneous) showed lower success (24%–32%).
- Decentralized teams benefited from more balanced decision-making.
- Hybrid teams gained from knowledge diversity and mentorship dynamics.

Takeaway:

- Combining **expertise diversity** with **collaborative decision-making** tends to produce better outcomes under uncertainty.

TABLE II: Performance summary across team structures.

Team	Games	Victory	Loss	Win (%)
Homo. Centralized	25	6	19	24.0
Hetero. Centralized	25	8	17	32.0
Homo. Decentralized	25	9	16	36.0
Hetero. Decentralized	25	6	19	24.0
Homo. Hybrid	25	7	18	28.0
Hetero. Hybrid	25	9	16	36.0

Georgia Tech

Ablation Studies

Impact of Team Size (Homogeneous Decentralized Teams):

- **Best performance:** Teams of **3 or 4** members.
- **Larger teams** (5–6) showed a decline in performance.
- Likely due to increased coordination overhead.

Impact of Team Composition (Homogeneous Hybrid Teams):

- **Best performance:** 2 experts + 3 beginners.
- Adding more experts did not always improve results.
- Too many experts may reduce diversity in perspectives or cause role overlap.

Takeaway:

- Effective collaboration benefits from **compact teams** and a **balanced skill mix**.

TABLE V: Impact of team size on performance using the homogeneous decentralized structure.

Team Size	Games	Victory	Loss	Win (%)
3	25	11	14	44.0
4	25	11	14	44.0
5	25	9	16	36.0
6	25	7	18	28.0

TABLE VI: Impact of team members on performance using the homogeneous hybrid structure.

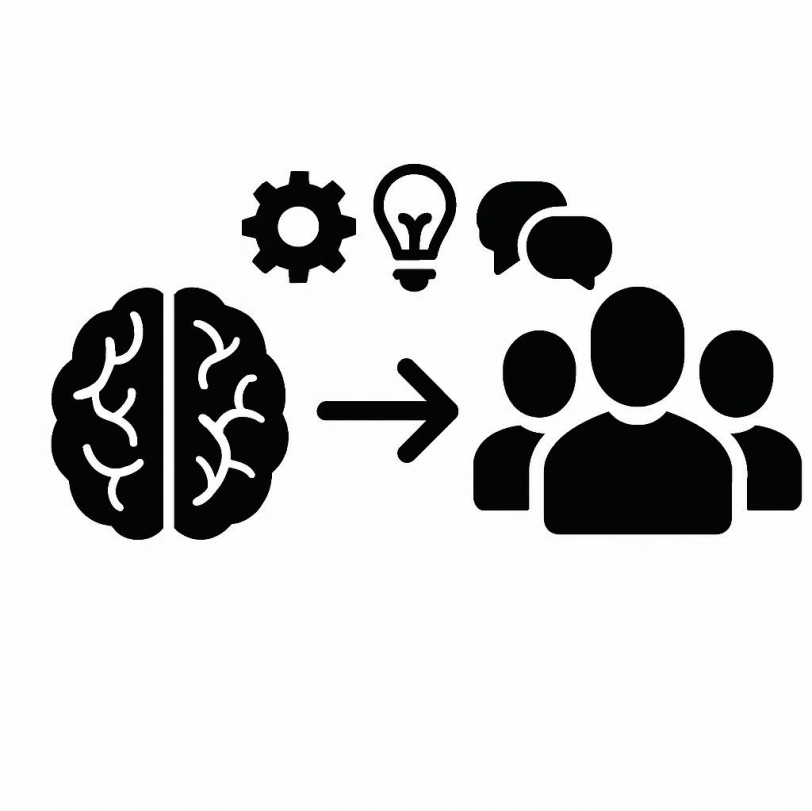
Experts	Beginners	Games	Victory	Loss	Win (%)
0	5	25	7	18	28.0
1	4	25	6	19	24.0
2	3	25	10	15	40.0
3	2	25	7	18	28.0
5	0	25	8	17	32.0

Georgia Tech

Conclusion

Key Takeaways:

- LLM-based agents can simulate effective collaboration in incident response.
- **Team structure matters:** decentralized and hybrid teams performed best.
- **Diversity and manageable size** support better coordination and outcomes.



Future Directions:

- Integrate real-world data and incident scenarios.
- Improve long-term memory and learning across simulations.
- Explore dynamic team structures and threat environments.

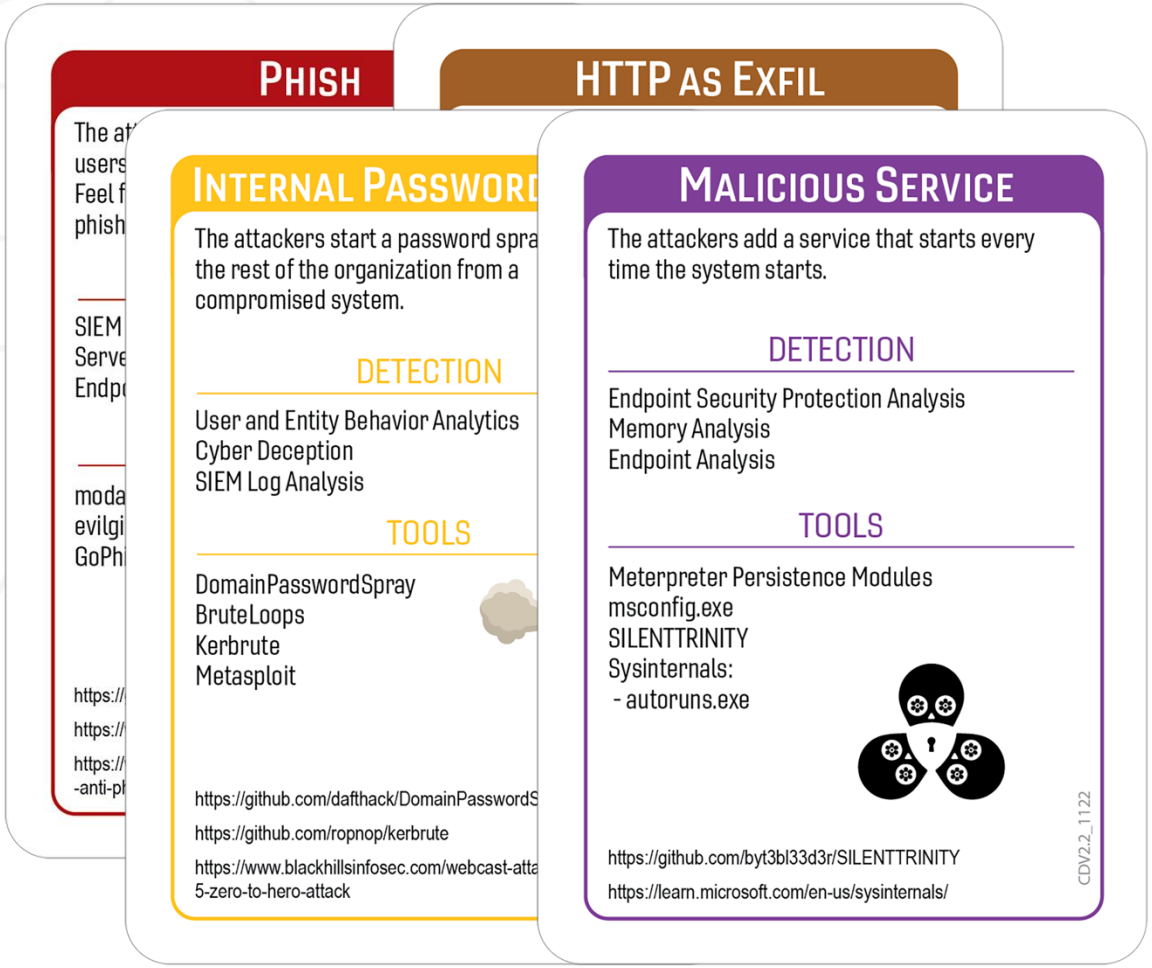
Thank You!

B&B Cards

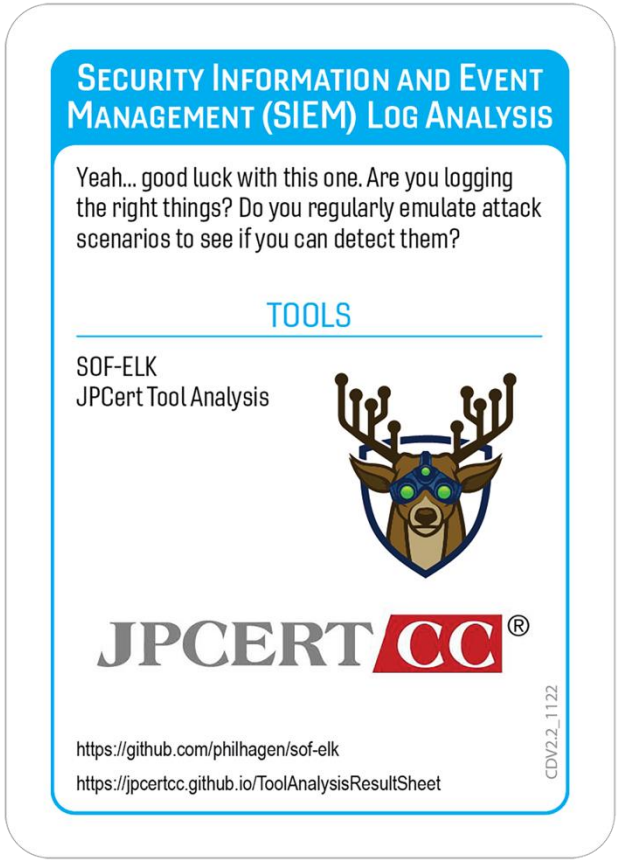
GitHub Repo

Georgia Tech

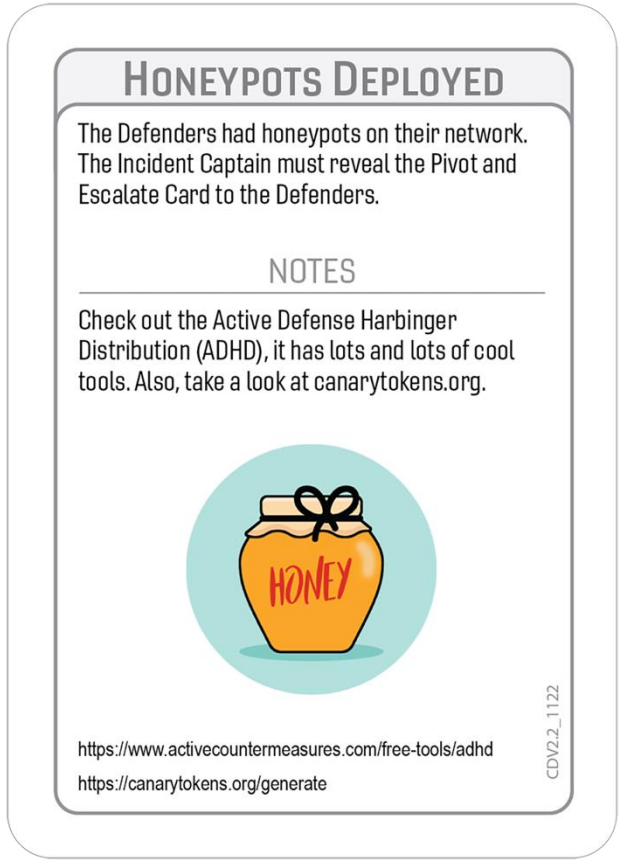
Backdoors & Breaches Cards



Attack Cards



Procedure Cards



Inject Cards

Backdoors & Breaches card images used for academic research under fair use. Original content © Black Hills Information Security.

Team Structure Details

Homogeneous Centralized Structure

- 1 team leader
- 4 generalist team members

Heterogeneous Centralized Structure

- 1 team leader
- 1 endpoint security expert
- 1 network traffic analysis expert
- 1 log and behavioral analysis expert
- 1 deception and containment expert

Homogeneous Hybrid Structure

- 3 generalist experts
- 2 beginners

Homogeneous Decentralized Structure

- 5 generalist team members

Heterogeneous Decentralized Structure

- 1 endpoint security expert
- 1 network traffic analysis expert
- 1 log and behavioral analysis expert
- 1 deception and containment expert
- 1 incident response expert

Heterogeneous Hybrid Structure

- 1 endpoint security expert
- 1 network traffic analysis expert
- 1 log and behavioral analysis expert
- 2 beginners

Georgia Tech

Agent Prompt Design

Incident Captain Prompt:

- Initializes the Backdoors & Breaches scenario.
- Selects hidden attack cards and available procedures.
- Enforces game rules, manages turns, and announces outcomes.
- Does **not** reveal solutions or provide strategic advice.

Defender Agent Prompts:

- Each agent is assigned a role-specific prompt that defines:
- Area of expertise (e.g., endpoint security, network analysis).
- Preferred procedures based on that domain.
- Communication style (collaborative, analytical, assertive, etc.).

Examples of Defender Roles:

- **Generalist:** Balanced knowledge across all phases; uses broad procedures.
- **Endpoint Security Expert:** Prioritizes host-based investigations.
- **Network Traffic Analyst:** Focuses on packet flow and firewall activity.
- **Behavioral Log Analyst:** Looks for anomalies in user behavior and log events.
- **Deception and Containment Expert:** Suggests mitigation strategies and countermeasures.
- **Beginner:** Limited knowledge, asks questions, relies on others for guidance.

Georgia Tech

Experiment Results

TABLE III: Reveal rates of frequently appearing attack cards.

Attack Card	Appearances	Reveal (%)
Exfiltration Over Physical Medium	48	64.6
Credential Stuffing	42	66.7
Domain Fronting as C2	36	77.8
Access Token Manipulation	30	73.3
Local Privilege Escalation	24	75.0
HTTP as Exfil	24	79.2
Application Shimming	24	37.5
Windows Service Recovery Actions	24	45.8
Windows BITS	24	95.8
External Cloud Access	24	45.8
Insider Threat	24	37.5
Bring Your Own (Exploited) Device	18	83.3
Internal Password Spray	18	88.9
Supply Chain Attack	12	75.0
Malware Injection Into Client Software	12	41.7

TABLE IV: Procedure effectiveness by usage and success rate.

Procedure Name	Usage	Success	Rate (%)
SIEM Log Analysis	188	51	27.1
Network Threat Hunting	185	86	46.5
User and Entity Behavior Analytics	160	58	36.2
Memory Analysis	144	8	5.6
Endpoint Security Protection Analysis	138	62	44.9
Endpoint Analysis	129	70	54.3
Firewall Log Review	128	46	35.9
Server Analysis	90	9	10.0
Cyber Deception	80	23	28.7
Physical Security Review	48	20	41.7
Isolation	39	0	0.0
Crisis Management	7	0	0.0

Georgia Tech