NDSS SYMPOSIUM

Chimera: Harnessing Multi-Agent LLMs for Automatic Insider Threat Simulation

Jiongchi Yu[1], Xiaofei Xie[1], Qiang Hu[2], Yuhan Ma[2], Ziming Zhao[3]

Singapore Management University[1], *Tianjin University*[2], *Zhejiang University*[3]

SMU

TIANJIN UNIVERSITY (PEIYANG UNIVERSITY) 1895

ZHEJIANG UNIVERSITY 1897

NDSS 2026

Background: Insider Threat

"An insider threat is a perceived threat to an organization that comes from people within the organization, such as employees, former employees, contractors or business associates, who have inside information concerning the organization's security practices, data and computer systems."[1]

- In 2025, the average annual cost of insider threats per organization is **$17.4M**[2]
- Average **13.5** insider-related security events per year[3]
- All these internal threat incidents are frequent, detriment, and often hard to detect

Federal contractors conspires to destroy dozens of U.S. government databases[4]

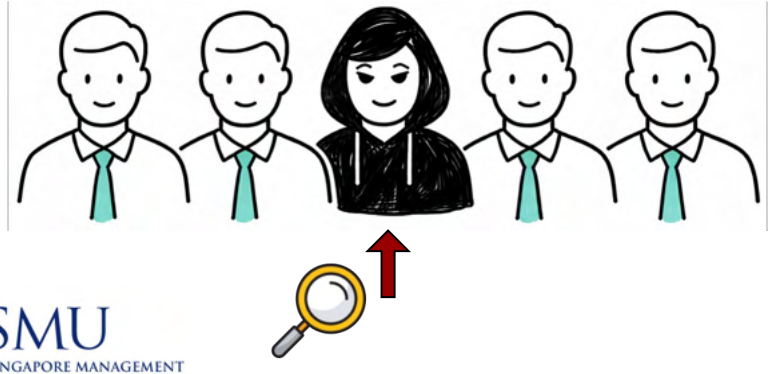
Tesla insider breach exposed data of more than 75,000 employees[5]

Background: Threat Model

- **Threat Actors**
  - Malicious insiders: Intentional abuse of access
  - Compromised insiders: Legitimate accounts hijacked
  - Negligent insiders: Unintentional risky actions
- **Key Assumptions**
  - Insider has legitimate access and operational knowledge
  - Attacks occur within the trust boundary, without explicit attack activities
- **Typical Attack Goals**
  - Data theft (customer records, IP)
  - System sabotage
  - Collusion with external attackers

Background: Internal Threat Detection (ITD)

- Traditional rule-based methods are limited in detecting **subtle** and **adaptive** insider behavior
- Modern approaches rely on **rich data** and advanced analytics:
  - *User & Entity Behavior Analytics Anomaly Detection*
  - *DL-Based Behavioral Analytics*
  - *Sequence / Graph models with Temporal Patterns*
  - *Data-Driven Clustering & Representation Learning*
  - …



Motivation: Data Challenges

Challenges for Existing Internal Threat Detection

- Privacy Concern
- Unrealistic
- Small Scale
- Non-Adaptive

Private and Costly Dataset

- Enterprise log data are highly sensitive, typically restricted for internal use with huge labeling costs

Old and Unitary Solutions

- Existing public datasets (e.g., CERT) are typically outdated, and only focusing on specific company arch

Motivation: Data Challenges

Challenges for Existing Internal Threat Detection

- Privacy Concern
- Unrealistic
- Small Scale
- Non-Adaptive

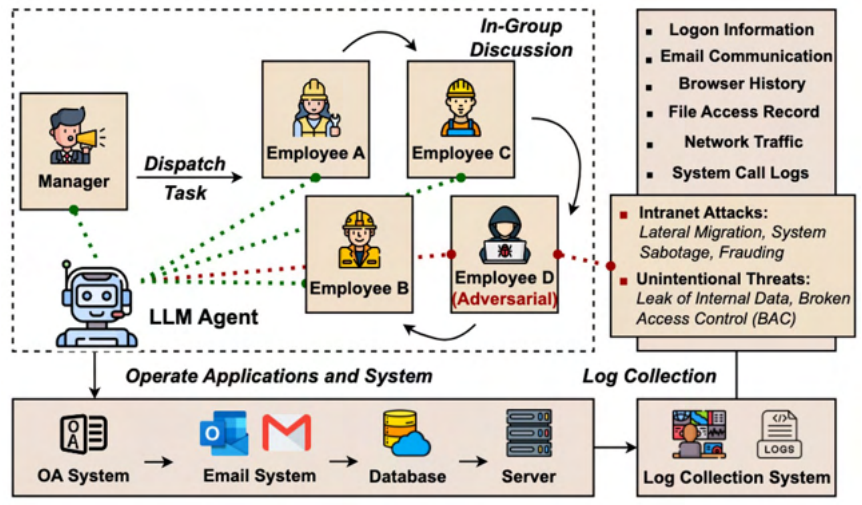
Dataset	Application	Network	System	Personality	Size	Attack Types
CERT r6.2	✓			✓	●	◐
TWOS	✓	✓		✓	◐	○
CIC-IDS 2017/2018		✓			●	◐
LANL 2017		✓	✓		●	○
WUIL	✓		✓		◐	◐
CPTC 2018	✓	✓			◐	◐
OpTC	✓	✓	✓		●	◐
Chimera (Ours)	✓	✓	✓	✓	●	●

Motivation: Multi-Agent LLM Systems

*Insider detection is fundamentally a **human-behavior modeling** problem, and **multi-agent LLM systems** are uniquely suited for this task.*

Required Capabilities for Internal Threat Simulation

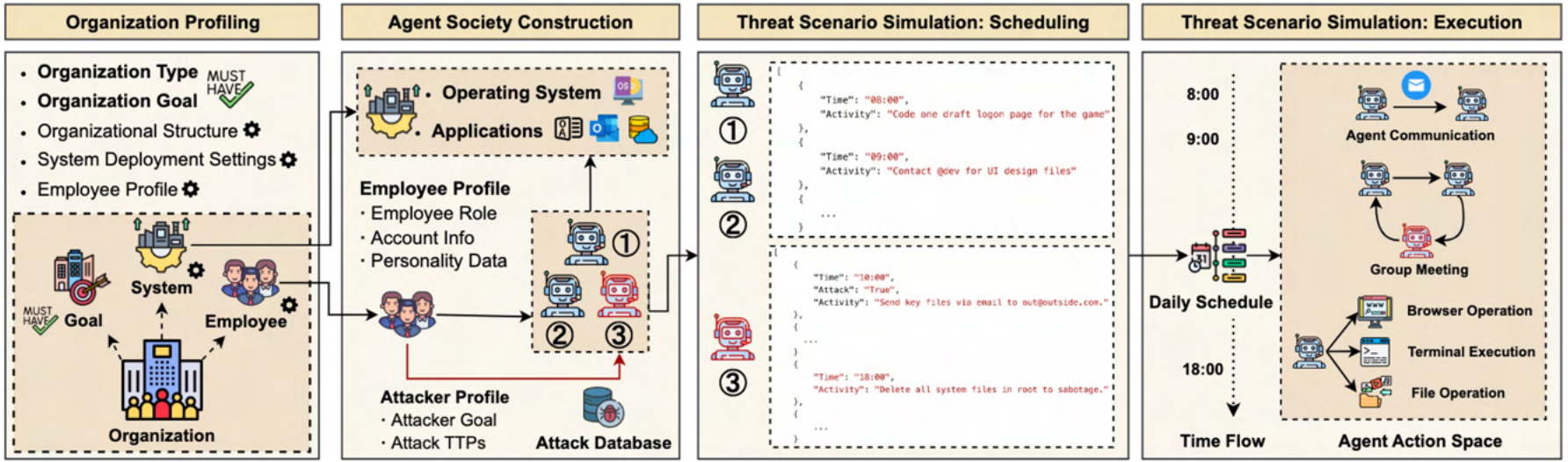
- Flexible Organizational Scenario Simulation
- Realistic Benign Behavior Modeling
- Adaptive Insider Attacker Red-Teaming
- Automatically Labeled, Multi-Modal Log Collection



Design: *Chimera* Overview

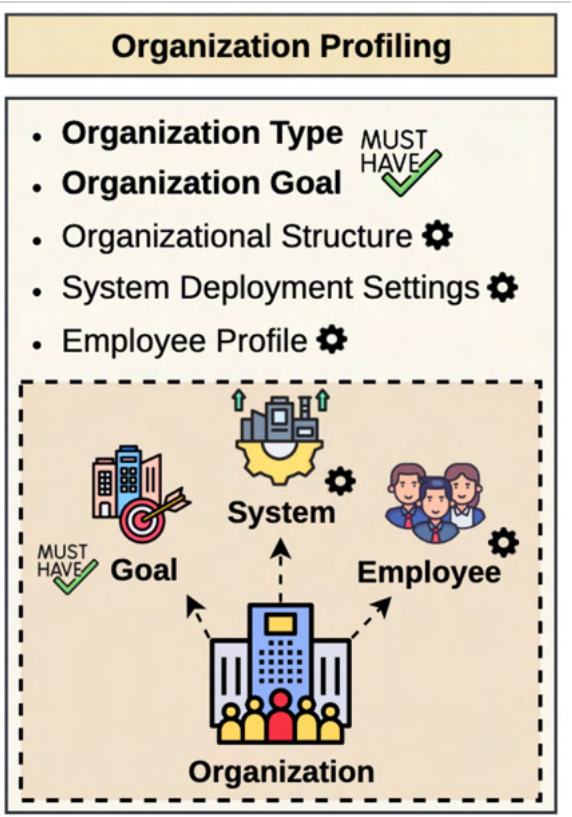
We propose ***Chimera***, the first multi-agent LLM framework for automated internal threat simulation

- Role-driven multi-agent society simulation
- Context-aware and reflective activity scheduling
- Controlled insider attack modeling
- Multiple log modality synthesis (application, system-level) with labels



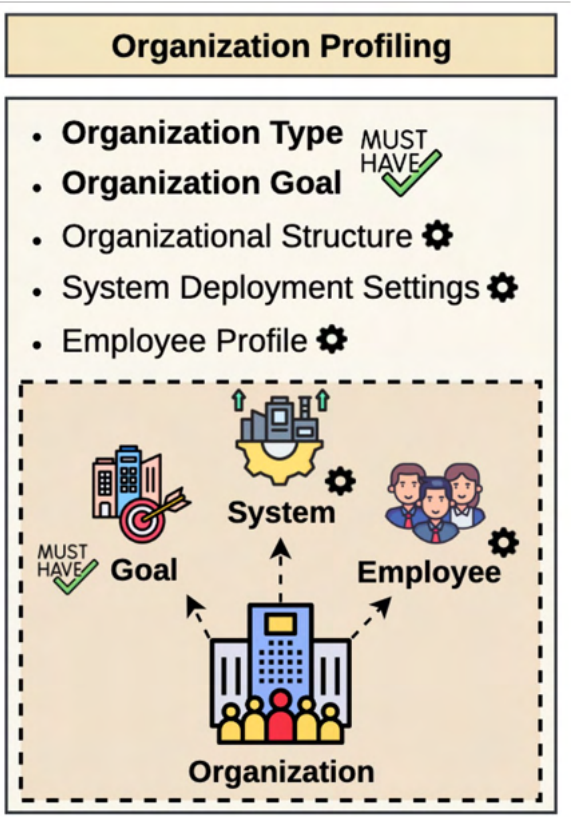
: *Optional for configuration or generated by LLM*

Design: Organization Profiling



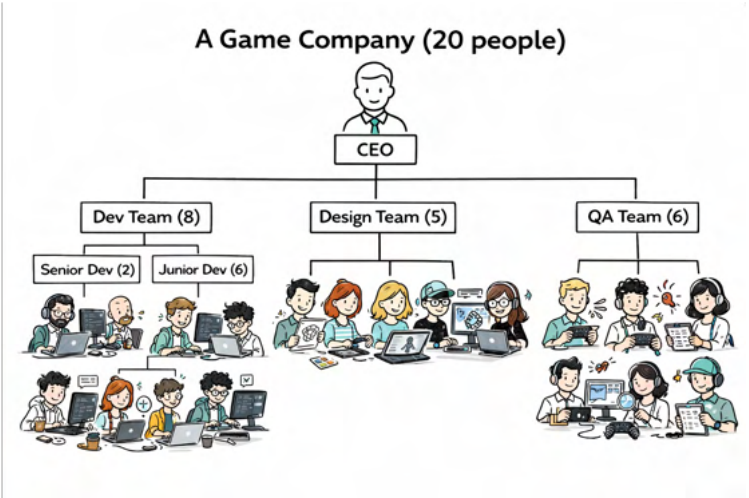
- **Organizational Structure**
  - Departments, hierarchy, system settings, organizational applications
  - Communication channels (email, messaging, shared directory)
- **Role Definitions**
  - Job functions, responsibilities, personalities
- **Access Policies**
  - Data ownership, system permissions
- **Operational Workflows**
  - Daily tasks, weekly meetings

Design: Organization Profiling

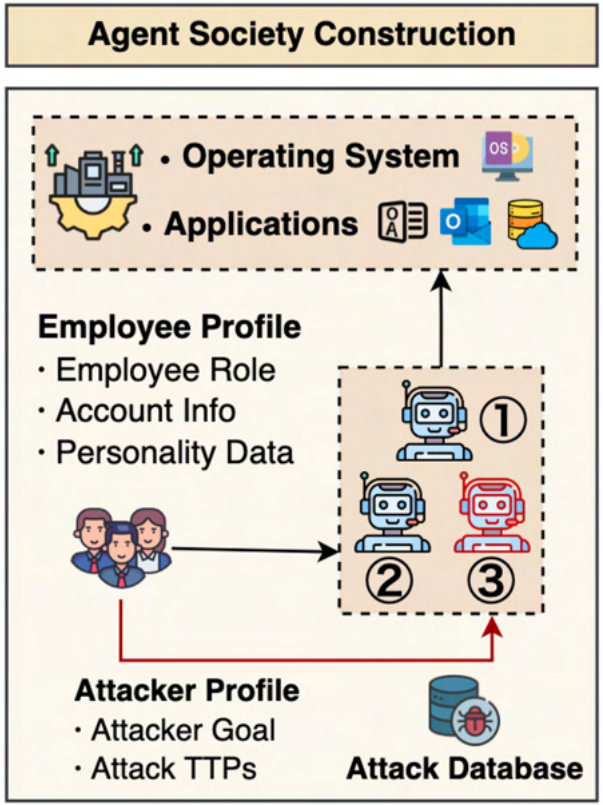


A Game Company (20 people)

- CEO (1)
- Dev Team (8)
  - Senior Dev (2)
  - Junior Dev (6)
- Design Team (5)
- QA Team (6)



Design: Agent Society Construction




```
{
    "name": "Kenny T",
    "id": "lpro-1",
    "ip": "10.0.0.10",
    "age": 38,
    "role": "Lead Programmer",
    "description": `Oversees all programming aspects of the game development, \
        including engine architecture, AI implementation, and networking.`,
    "tools": ["Visual Studio", "Unreal Engine","Browser Tool",
        "Terminal Tool", "File Operation Tool"],
    "mbti": "ISTJ",
    "interests": "strategy board games, science fiction novels",
    "personality": `Highly organized and detail-oriented. \
        Enjoys mentoring junior developers. \
        Known for his calm demeanor under pressure with polite tone.
        \ Like to work late till around 20:00.`,
    "application": {
        "Zendo": {
            "account_name": "prog-927415",
            "password": "kenjiT@lead",
            "permissions": "lead_programmer"
        }
    },
    "email": "prog-927415@tech_company.com",
    "container_id": "832a9d1f46c8"
}
```

Design: Threat Scenario Simulation (Scheduling)

Weekly Schedule

- Overall meeting of all agents
- Task distribution based on roles

Daily Schedule

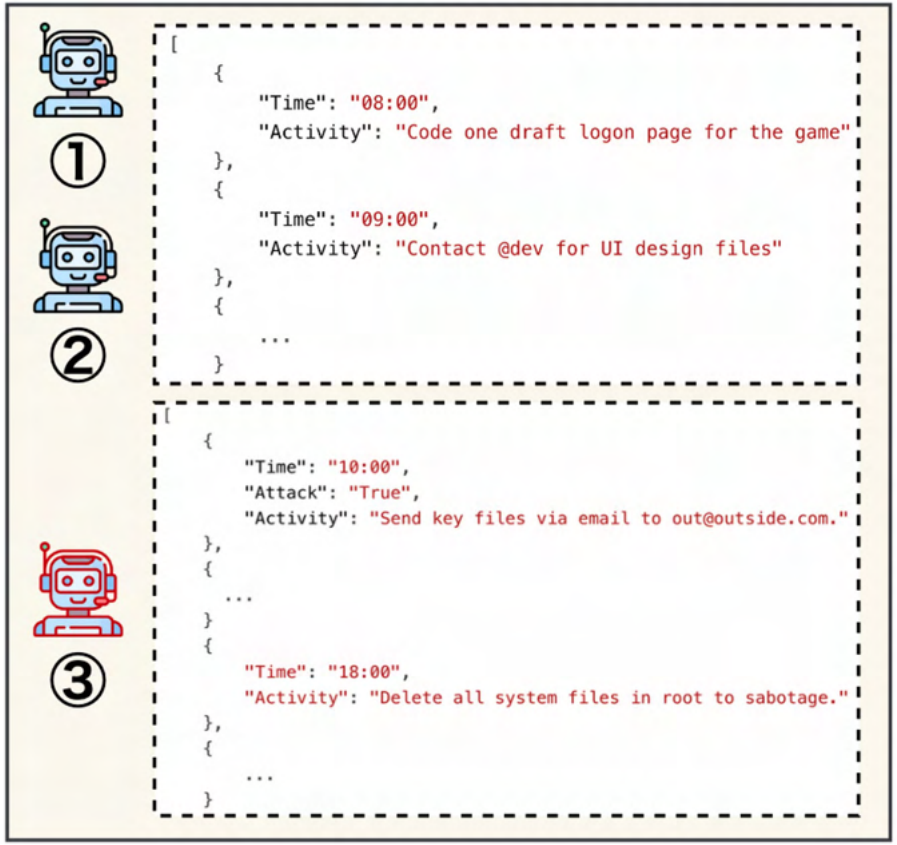
- Detailed working task with availability to consult to other agents

Progress Schedule

- Daily update summary of work

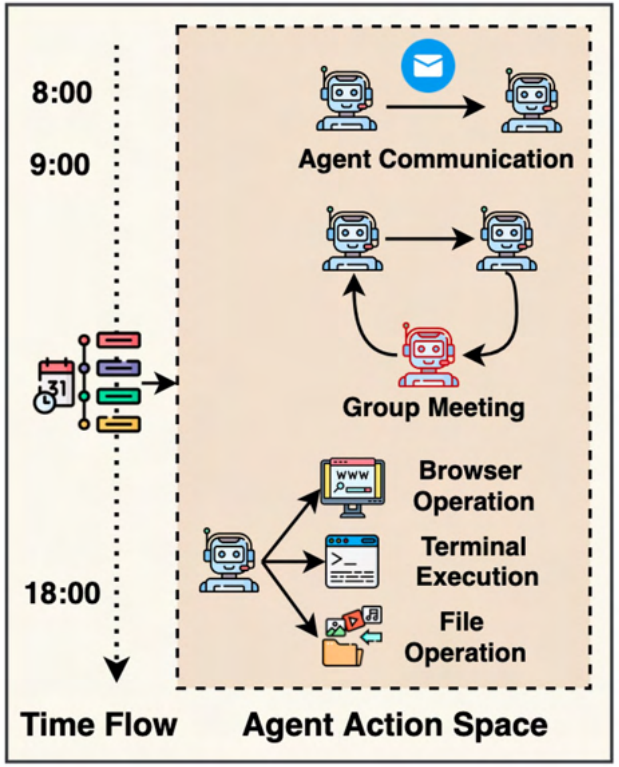
Schedule Update

- The agent updates their upcoming scheduling upon:
  - Communication occurs
  - Attack activity planning



Design: Threat Scenario Simulation (Execution)

- **Schedule-Driven Simulation Timing:** Simulation begins at the earliest scheduled task among all agents, and ends when all agents complete their daily schedules
- **Real-to-Simulation Time Mapping:** 1 real-world second represents 20 simulation seconds



Design: Threat Scenario Simulation (Execution)

We incorporate **12** individual attack types, along with **three** hybrid attacks that combine multiple attack patterns from **Data Broker Database, the U.S. Attorney's Office, and the Federal Bureau of Investigation**

- *Attacks are extracted into MITRE ATT&CK TTPs for structural simulation*

Attacker	Role	Goal	Target	Frequency	Purpose
Traitor	internal	IP theft	OS, Network, App	recurrent	financial
Traitor	internal	IP theft	OS, Network, App	single	financial
Traitor	internal	IP theft	App	single	financial
Traitor	internal/external	sabotage	App	single	financial/personal
Traitor	internal/external	sabotage	OS	single	financial/personal
Traitor	internal/external	sabotage	OS, Network	single	financial/personal
Masqueraders	internal/external	fraud	App	single	financial
Masqueraders	internal/external	fraud	OS	recurrent	financial/personal
Masqueraders	internal/external	IP theft	OS, Network	recurrent	financial/political
Masqueraders	internal/external	IP theft	OS, Network, App	recurrent	financial
Unintentional User	internal	data leak	OS, Network	single	personal
Unintentional User	internal	IP theft	App	recurrent	personal
Miscellaneous	internal	data exfiltration	App	recurrent	financial
Miscellaneous	internal	data exfiltration	OS, Network	recurrent	financial
Miscellaneous	internal/external	data exfiltration, system takeover	OS, Network, App	recurrent	political

Design: Threat Scenario Simulation (Execution)

We incorporate **12** individual attack types, along with **three** hybrid attacks that combine multiple attack patterns from **Data Broker Database, the U.S. Attorney's Office, and the Federal Bureau of Investigation**

- *Attacks are extracted into MITRE ATT&CK TTPs for structural simulation*

Step	Tactic	Technique	Sub-technique	Procedure	Detection Data Sources
1	Initial Access	T1199, T1078	-	Insider recruits colleagues to obtain access or credentials	Emails, System Logs
2	Privilege Escalation	T1566, T1078	-	Perpetrator uses email to request resource administrators with admin-level access	Emails, Logon
3	Exfiltration	T1052	T1052.001	Download sensitive files, transfer to private cloud, remove from premises	System Logs, Traffic, File Operation, HTTP

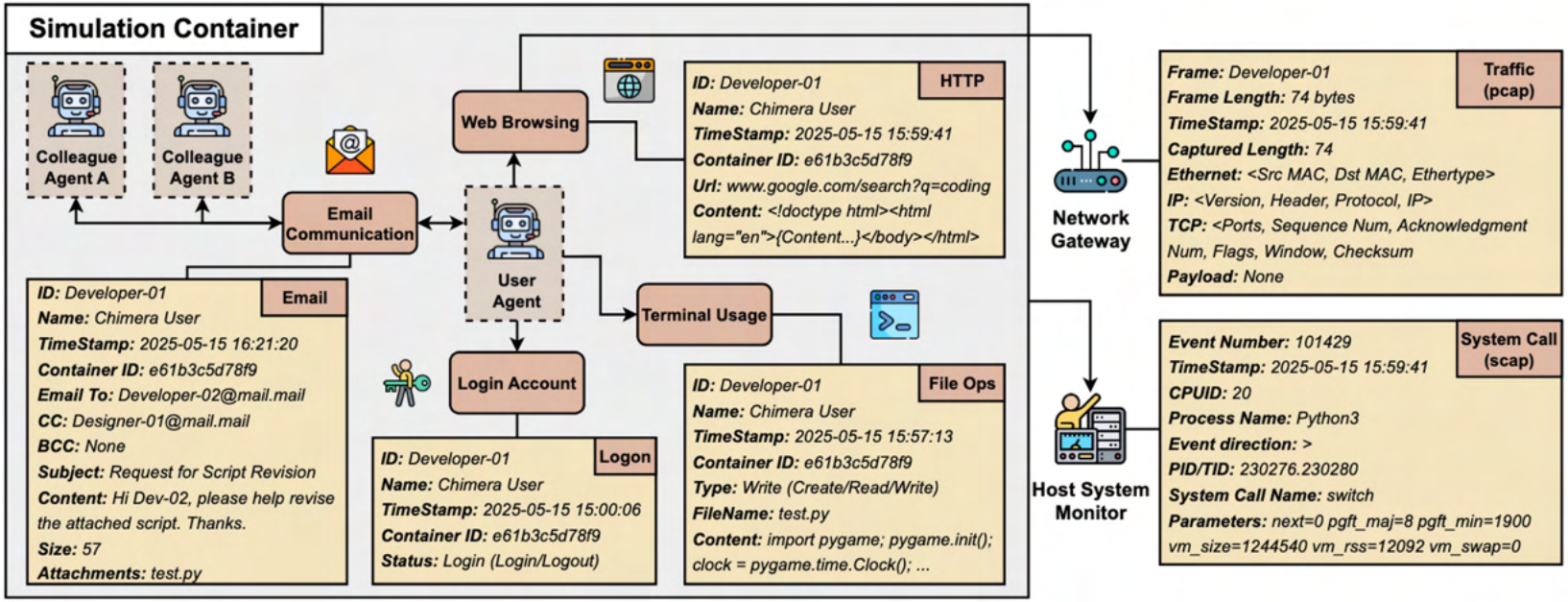
Example ATT&CK TTPs mapping of the insider IP theft attack

Dataset: ChimiraLog

ChimeraLog consists of **20-person** organizations with one-month simulation, comprising **20 billion benign** log entries and **5 billion attack** entries across **six** log modalities

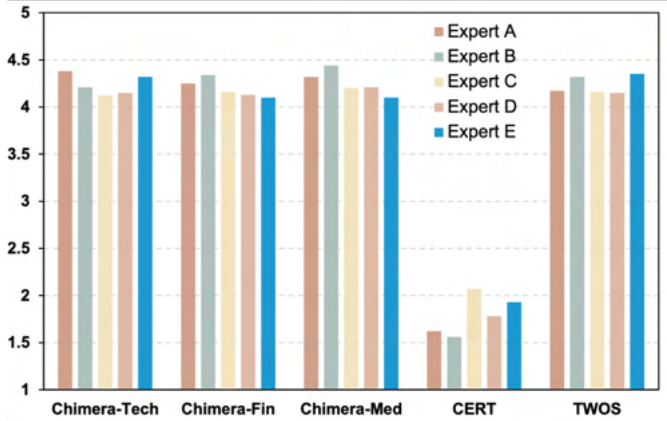
- **Application Logs:** *Logon, HTTP, email, file access*
- **System Logs:** *System call logs, network traffic*

<u>Three Data-Sensitive Industry Scenarios</u>: Technology Company, Financial Corporation, Medical Institution



Evaluation: Human Expert Study

- 5 independent security experts (> 5 years experience)
- Sample 100 log entries per dataset
- Evaluated across 4 application-level modalities
- 5-point Likert scale for realism and practicality
- Blind and shuffled evaluation to reduce bias



Experts consistently recognize ***ChimeraLog*** as highly **realistic** and **practical** in demonstrating the real-world enterprise logs

Re: Follow-up: Compliance Technology Solutions Implementation Timeline
Gabriel,
Understood. Here's the updated timeline and details for the compliance technology solutions implementation:

Potential Roadblocks & Mitigation:

- **Data Integration Complexity:** Integrating with various data sources (trading platforms, market data providers, KYC/AML databases) can be complex. We're mitigating this by using standardized data formats and APIs where possible. We've also allocated extra time for troubleshooting and data mapping.
- **Scalability Challenges:** Ensuring the system can handle increasing data volumes and transaction rates requires careful planning and optimization. We are using cloud-based infrastructure that allows us to easily scale resources as needed. We're also implementing caching mechanisms and database optimizations to improve performance

Please let me know if you have any specific questions or require further clarification on any of these points.

Regards,
Sofia Patel
DevOps Engineer

Chimera Dataset

Now Sylvia, the object of Aminta's desire, arrives on the scene with her posse of hunters to mock the god of love. The piano arrangement was composed in 1876 and the orchestral suite was done in 1880. As writer Arnold Haskell said, "... he accepts the challenge in Sylvia of coping with period music without descending to pastiche; and never once does the movement he provides strike us as modern or as 'old world'".
Sylvia now grieves over Aminta
cherishing the arrow pulled from her breast nostalgically.

CERT Dataset

Dear Team YUVPW30NCW,

Your Wild card period is scheduled on Tuesday (4DURNKVFWJ March LGDC69866G) from CCW408AWYY:OZ5MJQTQ3P am to EMB95KRPM2:8RZT3BP0U9 am.

Username: User14
Password: 35NIFI54FK

These credentials will be valid only between CCW408AWYY:OZ5MJQTQ3P am to EMB95KRPM2:8RZT3BP0U9 am. Only 71J871LWL3 team member will have access to the machine. It is possible to swap between members. But only one member can login at one time.

USE THE 59QKJUEJ3N MINS WELL !!!!! !
Technical Staff

TWOS Dataset

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Evaluation: Quantitative Analysis

- **Temporal Activity Distribution**

Sampled daily activities into normalized hourly activity histograms of temporal 24-hour user events distribution

- **Behavioral Entropy**

Quantified user action diversity by computing entropy over action distributions across daily activities

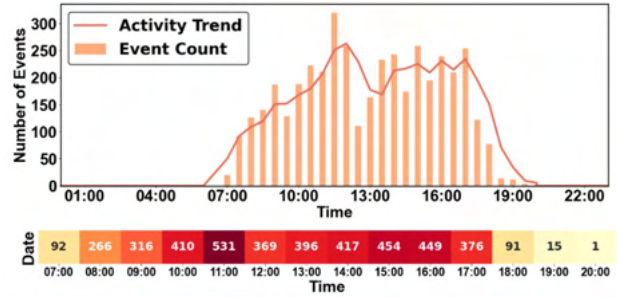
[*Chimera* - 2.12, *CERT* - 0.92, *NANL* - 0.72]

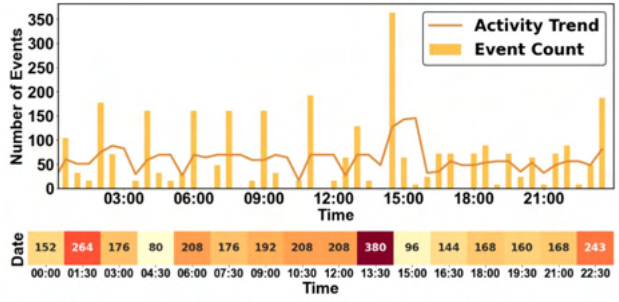
- **Sequence Complexity**

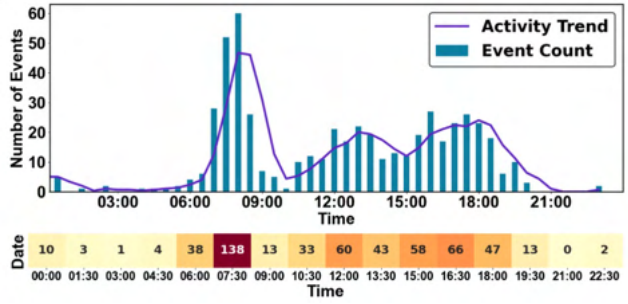
Estimated behavioral unpredictability by measuring compression-based complexity of user event sequences

[*Chimera* - 0.779, *CERT* - 0.412, *TWOS* - 0.216]

ChimeraLog exhibits realistic temporal patterns, higher behavioral diversity, and significantly greater sequence complexity than existing synthetic datasets, demonstrating stronger alignment with real-world enterprise behaviors







Evaluation: ITD Performance

We train and evaluate standard ITD models within each dataset

- **ITD Baselines:** SVM, CNN, GCN, DS-IID (AE & LSTM)

Baseline \ Dataset	Chimera-Tech				Chimera-Finance				Chimera-Medical				*CERT*			
	Acc	Pre	Recall	F1	Acc	Pre	Recall	F1	Acc	Pre	Recall	F1	Acc	Pre	Recall	F1
SVM	0.751	0.679	0.823	0.744	0.749	0.753	0.639	0.691	0.755	0.743	0.692	0.717	0.873	0.884	0.931	0.907
CNN	0.864	0.890	0.739	0.808	0.794	0.740	0.891	0.809	0.851	0.858	0.780	0.817	0.923	0.891	0.959	0.924
GCN	0.697	0.674	0.727	0.699	0.755	0.669	0.749	0.707	0.669	0.671	0.736	0.702	0.913	0.927	0.943	0.935
DS-IID	0.826	0.727	0.949	0.823	0.783	0.781	0.792	0.786	0.904	0.857	0.784	0.819	0.971	0.960	0.950	0.955

- DS-IID consistently achieves the highest F1 across ITD datasets
- ***Chimera*** datasets present more difficult detection tasks than CERT
- ITD performance remains stable across diverse organizational domains

Evaluation: ITD Performance

We evaluate models trained on one dataset and tested on a different organizational environment

- **ITD Baselines:** SVM, CNN, GCN, DS-IID (AE & LSTM)

Train Dataset	Test Dataset	SVM				CNN				GCN				DS-IID			
		Acc	Pre	Recall	F1	Acc	Pre	Recall	F1	Acc	Pre	Recall	F1	Acc	Pre	Recall	F1
Chimera-Tech	Chimera-Finance	0.700	0.550	0.600	$0.574^{\downarrow 0.170}$	0.700	0.550	0.600	$0.574^{\downarrow 0.234}$	0.755	0.627	0.772	$0.692^{\downarrow 0.008}$	0.821	0.850	0.757	$0.801^{\downarrow 0.023}$
	Chimera-Medical	0.688	0.500	0.500	$0.500^{\downarrow 0.244}$	0.688	0.500	0.500	$0.500^{\downarrow 0.308}$	0.771	0.251	0.609	$0.356^{\downarrow 0.344}$	0.816	0.851	0.752	$0.798^{\downarrow 0.025}$
	CERT	0.354	0.880	0.500	$0.638^{\downarrow 0.106}$	0.357	0.926	0.317	$0.472^{\downarrow 0.335}$	0.298	0.617	0.251	$0.357^{\downarrow 0.343}$	0.811	0.850	0.761	$0.803^{\downarrow 0.020}$
Chimera-Finance	Chimera-Tech	0.700	0.650	0.700	$0.674^{\downarrow 0.017}$	0.700	0.650	0.800	$0.717^{\downarrow 0.091}$	0.699	0.457	0.531	$0.491^{\downarrow 0.216}$	0.822	0.850	0.658	$0.742^{\downarrow 0.045}$
	Chimera-Medical	0.667	0.500	0.500	$0.500^{\downarrow 0.191}$	0.667	0.500	0.667	$0.571^{\downarrow 0.237}$	0.880	0.698	0.705	$0.702^{\downarrow 0.006}$	0.831	0.850	0.768	$0.776^{\downarrow 0.010}$
	CERT	0.388	0.933	0.357	$0.516^{\downarrow 0.175}$	0.386	0.933	0.355	$0.515^{\downarrow 0.294}$	0.340	0.693	0.302	$0.421^{\downarrow 0.286}$	0.804	0.850	0.653	$0.739^{\downarrow 0.048}$
Chimera-Medical	Chimera-Tech	0.500	0.250	0.250	$0.250^{\downarrow 0.467}$	0.563	0.250	0.250	$0.250^{\downarrow 0.567}$	0.820	0.667	0.704	$0.685^{\downarrow 0.017}$	0.827	0.850	0.763	$0.804^{\downarrow 0.015}$
	Chimera-Finance	0.583	0.250	0.330	$0.284^{\downarrow 0.432}$	0.583	0.250	0.333	$0.286^{\downarrow 0.531}$	0.859	0.678	0.735	$0.701^{\downarrow 0.001}$	0.817	0.850	0.763	$0.804^{\downarrow 0.015}$
	CERT	0.303	0.944	0.243	$0.387^{\downarrow 0.330}$	0.293	0.950	0.229	$0.370^{\downarrow 0.448}$	0.264	0.724	0.204	$0.319^{\downarrow 0.383}$	0.813	0.851	0.652	$0.738^{\downarrow 0.081}$
CERT	Chimera-Tech	0.300	0.300	1.000	$0.462^{\downarrow 0.445}$	0.300	0.300	1.000	$0.462^{\downarrow 0.462}$	0.300	0.300	1.000	$0.462^{\downarrow 0.473}$	0.341	0.354	0.705	$0.471^{\downarrow 0.484}$
	Chimera-Finance	0.300	0.300	1.000	$0.462^{\downarrow 0.445}$	0.300	0.300	1.000	$0.462^{\downarrow 0.462}$	0.300	0.300	1.000	$0.462^{\downarrow 0.473}$	0.330	0.342	0.720	$0.464^{\downarrow 0.491}$
	Chimera-Medical	0.300	0.300	1.000	$0.462^{\downarrow 0.445}$	0.300	0.300	1.000	$0.462^{\downarrow 0.462}$	0.300	0.300	1.000	$0.462^{\downarrow 0.472}$	0.330	0.340	0.750	$0.468^{\downarrow 0.487}$

- Distribution shift can causes substantial F1 degradation for ITD models
- ***ChimeraLog*** demonstrates improved cross-domain generalization compared to CERT, though performance still degrades under distribution shift

Future Work

- **Cognitive and Hierarchical Realism**
  - **Agent-level realism:** personality, memory, motivation modeling
  - **Organizational hierarchy:** departments, branches, cross-team dynamics
- **Adaptive Autonomous Red-Teaming**
  - Fully automated adversarial agents
  - Adaptive multi-stage attack strategy evolution (e.g., APTs)
- **Co-Evolutionary Defense & Human-in-the-Loop**
  - Defensive agents integrated for self-evolving attack-defend dynamics
  - Analyst feedback and interactive intervention

Conclusion

- We propose ***Chimera***, the first LLM-based multi-agent framework for automatic insider threat simulation in enterprise environments
- Using ***Chimera***, we construct ***ChimeraLog***, a large-scale, multi-modal insider threat dataset with **25 billion** log entries, spanning **15 real-world** attack scenarios and three organizational settings (technology, finance, medical)
- We provide extensive evaluation including human studies and benchmarking of four representative ITD methods on ***ChimeraLog***. Our result highlights the **challenges** of robust ITD in realistic environments and the need for adaptive high-quality ITD datasets

Q & A

Presented by Jiongchi Yu (jcyu.2022@smu.edu.sg)

Chimera Website

Learn more about our work

SMU
SINGAPORE MANAGEMENT UNIVERSITY