CHIMERA: HARNESSING MULTI-AGENT LLMS FOR AUTOMATIC INSIDER THREAT SIMULATION

PREPRINT VERSION

Jiongchi Yu
Singapore Management University
Singapore, Singapore
jcyu.2022@smu.edu.sg

Xiaofei Xie
Singapore Management University
Singapore, Singapore
xfxie@smu.edu.sg

Qiang Hu*
Tianjin University
Tianjin, China
qianghu@tju.edu.cn

Yuhan Ma
Tianjin University
Tianjin, China
mayuhan@tju.edu.cn

Ziming Zhao
Zhejiang University
Hangzhou, China
zhaoziming@zju.edu.cn

ABSTRACT

Insider threat, which can lead to unacceptable losses, is a widespread and significant security concern, making its detection essential. Although recent machine-learning-based insider-threat detection (ITD) techniques have shown encouraging results, their progress is limited by a persistent bottleneck involving the scarcity of high-quality data. The paradox is that enterprise internal data is highly sensitive and typically inaccessible, while public datasets are either limited in real-world coverage or, in the case of synthetic data, lack rich semantic information and realistic behavioral patterns. As a result, there is a crucial need for the construction of real-world insider threat datasets.

To address this challenge, we propose *Chimera*, the first large language model (LLM)-based multi-agent framework to automatically simulate both benign and malicious insider activities, as well as collect logs across diverse enterprise environments. Based on analysis of organizational composition and structural characteristics of the organization, *Chimera* customizes each LLM agent to represent an individual employee by detailed role modeling and couples with modules of group meetings, pairwise interactions, and self-organized scheduling. In this way, *Chimera* can accurately reflect the complexities of real-world enterprise operations. The current version of *Chimera* consists of 15 distinct types of manually abstracted insider attacks, such as intellectual property theft and system sabotage. Using *Chimera*, we simulate the benign and attack activities across three typical data-sensitive organizational scenarios, including technology company, finance corporation, and medical institution, and generate a new dataset named *ChimeraLog* to facilitate the development of machine learning-based ITD methods.

To evaluate the quality and authenticity of *ChimeraLog*, we conduct comprehensive human studies and quantitative analyses. The results demonstrate both the diversity and realism of the dataset. Further expert analysis highlights the presence of realistic threat patterns as well as explainable activity traces. In addition, we evaluate the effectiveness of existing insider threat detection methods on *ChimeraLog*. The average F1-score achieved is 0.83, which is notably lower than the score of 0.99 observed on the baseline dataset CERT, thereby illustrating the greater difficulty posed by *ChimeraLog* for threat detection tasks.

Keywords Internal Threat Detection · Multi-Agent Systems · Large Language Model

*Corresponding Author.

1 Introduction

Insider threats, defined as security incidents that originate within an organization, have emerged as a critical concern in modern enterprise environments. Recent studies indicate that more than 50% of organizations have experienced insider incidents, with 29% incurring remediation costs exceeding $1 million Systems [2025], Gurucul [2024]. Compounding this challenge, insider attacks are exceptionally difficult to detect as they are perpetrated by trusted individuals possessing legitimate access to organizational systems. These malicious activities span various forms, including horizontal propagation such as lateral movement within internal networks, unauthorized data exfiltration, IT sabotage, fraud, and espionage, as well as vertical escalation such as privilege escalation PurpleSec [2025]. Prominent cases that illustrate these threats include incidents involving Edward Snowden's data exfiltration at the NSA Harding [2014], the sabotage of IT systems at Tesla by an employee in 2018, and the Capital One data breach caused by insider exploitation of privilege-escalation vulnerabilities Agnihotri and Bhattacharya [2024], Novaes Neto et al. [2020]. Traditional security defenses struggle to detect insider threats due to their legitimate access privileges and the subtlety of malicious activities Schoenherr and Thomson [2020], Al-Mhiqani et al. [2020]. Inadequate security infrastructure further exacerbates the insider threat landscape. Recent reports Gurucul [2024], Six [2025] revealed that only 36% of organizations have implemented comprehensive monitoring systems to proactively detect insider activity, highlighting significant gaps in current insider threat defenses.

To address insider threats, multiple insider threat detection (ITD) methods have been proposed to analyze internal activity logs, such as user login information and employee email communications, within organizational systems He et al. [2024a], Hong et al. [2022], Kotb et al. [2025]. Among these methods, rule-based solutions typically demand extensive manual effort to define detection rules and to collect reference data tailored to the systems under protection. Such approaches tend to be ad hoc, restricted for internal use because of privacy considerations, and vulnerable to high false alarm rates when the underlying systems evolve Gurucul [2024], Erola et al. [2022]. Consequently, frequent manual updates are necessary, making rule-based methods costly and difficult to generalize.

On the other hand, although machine-learning-based ITD methods have demonstrated promising detection performance Janjua et al. [2020], Hong et al. [2022], Kotb et al. [2025], these approaches require large-scale, high-quality, and precisely labeled datasets to train models capable of differentiating subtle threat behaviors from legitimate activities. However, obtaining adequate datasets for effective model training remains a significant practical challenge.

Specifically, the challenges associated with the lack of high-quality ITD datasets are fourfold:

- **Privacy Constraints.** Insider activities typically involve sensitive and proprietary organizational data, making it difficult to share such information externally for analysis.
- **Unrealistic Data.** Publicly available ITD datasets, such as the CERT insider threat dataset Glasser and Lindauer [2013], are generally synthetic and lack semantic richness. Both benign organizational behaviors and threat scenarios are artificially designed rather than derived from authentic interactions within real organizations Yuan and Wu [2021], Lindauer et al. [2014]. These synthetic datasets also commonly omit critical system-level log modalities found in practical environments, such as network traffic and system call logs, limiting their utility and realism for practical scenarios.
- **High Cost.** Due to the sheer volume and complexity of internal activity logs, collecting and labeling datasets from real-world environments can be prohibitively costly. This cost increases substantially as systems evolve rapidly, as commonly observed in modern software development practices, and as organizations scale and their user base expands. For instance, large enterprises routinely generate millions of log entries per day, leading to exponential increases in labeling and maintenance expenses as the organization grows Gurucul [2024], Yu et al. [2025]. Although datasets such as TWOS (The Wolf of SUTD) Harilal et al. [2017] attempt to address these concerns by collecting data from human activities within realistic, yet controlled environments, their small scale, consisting of only six student teams, and narrow scenario coverage significantly restrict their broader applicability.
- **Lack of Adaptability.** The frequent updating of enterprise systems results in significant distribution shifts in log data Han et al. [2023], Yu et al. [2025], which can adversely affect the performance of ITD models trained on outdated datasets. Additionally, insider threat scenarios represented in existing datasets often appear ad hoc, applicable only within specific system configurations. Therefore, datasets must be regularly updated and maintained, which increases cost and operational complexity.

Motivation. The aforementioned challenges significantly restrict the preparation of high-quality datasets for ITD, consequently hindering the research and development of effective ITD methods. In the absence of realistic and representative data, ITD approaches frequently experience high false-positive rates and exhibit limited generalization ability when deployed across diverse operational environments. Therefore, there is an urgent need for novel automated approaches capable of generating representative, high-fidelity insider threat datasets without incurring excessive costs or compromising privacy.

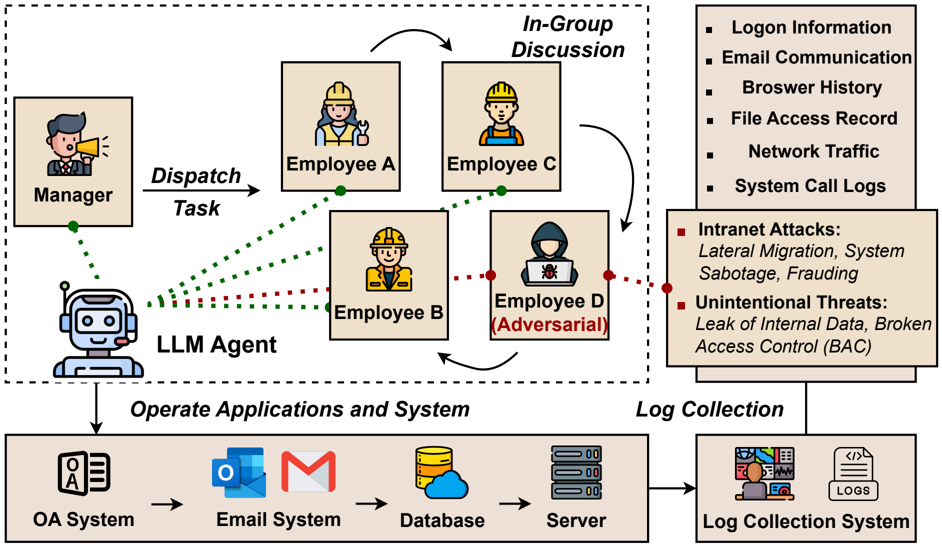


Figure 1: Automation of insider threat simulation.

To effectively support practical security practices, an automated ITD data generation framework, as illustrated in Figure 1, should ideally incorporate the following key attributes: (1) flexible scenario simulations tailored to domain-specific software and communication protocols; (2) realistic simulation of benign user activities that accurately represent genuine organizational behavior; (3) adaptive threat simulation capable of replicating diverse internal attack behaviors originating from insider threats; and (4) comprehensive log collection mechanisms capable of producing instrumented and labeled logs, thus minimizing reliance on manual annotation efforts.

Having witnessed the success of LLM and LLM-based agents, which can be utilized to simulate human behaviors across a variety of domains, such as software development He et al. [2024b], Jin et al. [2024], event planning Hong et al. [2023], Park et al. [2023], and society modeling Lin et al., Xie et al. [2024]. In this work, we propose *Chimera*, the first LLM-based multi-agent framework specifically designed for generalized insider threat simulation. *Chimera* constructs adaptive enterprise societies to realistically model insider threat scenarios and collect high-fidelity log data in a fully automated and flexible manner. At its core, *Chimera* utilizes LLM-based agents to simulate enterprise activities, where each agent represents a distinct organizational member endowed with an individual working role, a personality, and a defined set of responsibilities within the company.

To achieve realistic simulations of both benign user behaviors and insider threat activities, *Chimera* incorporates specialized features tailored specifically for insider scenario simulation. These include: (1) a multi-stage task specification workflow that allows agents to organize and prioritize daily responsibilities; (2) reflective memory capabilities enabling agents to recall and reference past daily activities, thereby maintaining consistent behaviors over time; and (3) unrestricted and context-rich communication among agents, facilitating realistic interactions and coordination.

Through a carefully structured workflow, given a scenario, *Chimera* either automatically generates organizational structures, employee roles, and individual personalities or accepts user-provided input. Each agent then independently plans and executes semantically meaningful and contextually coherent daily activities, such as attending group meetings, emailing colleagues, and engaging in document or code creation. To realistically simulate insider attacks, *Chimera* integrates dedicated attacker agents following penetration testing paradigms. These malicious agents simulate insider attackers who launch threats while simultaneously maintaining their routine work duties to avoid detection. All attacks are guided by abstracted attack specifications, enabling malicious agents to adaptively apply specific attack techniques relevant to the simulated organizational context and scenarios.

We deploy *Chimera* in three representative, data-sensitive enterprise scenarios, namely, technology companies, financial corporations, and medical institutions, to realistically simulate the activities of a 20-person organization. In our simulation, we consider 15 insider threat scenarios from diverse perspectives, adaptively applying them across each enterprise environment. Over a one-month simulation period, we collect a dataset named *ChimeraLog*, comprising approximately 20 billion normal log entries and 5 billion attack log entries. *ChimeraLog* includes six distinct log modalities. Four modalities pertain to application-level logs, specifically login information, email communications, web browsing history, and file operation records. Two modalities cover system-level logs, namely network traffic and system logs. To the best of our knowledge, *ChimeraLog* represents the largest and most diverse ITD dataset available to date, in terms of dataset size, variety of log modalities, and breadth of attack scenarios.

We first perform a comprehensive evaluation comprising quantitative analyses and human studies to assess the quality of *ChimeraLog*. This evaluation rigorously assesses the dataset quality in terms of realism and the effectiveness of *Chimera* in simulating authentic insider threat scenarios. Additionally, we conduct benchmarking analyses on existing ITD methods to evaluate their detection performance, investigating both the challenge presented by *ChimeraLog* and the generalization capability of these ITD approaches across diverse scenarios and distributional shifts. We observe that all existing ITD methods suffer significantly under distribution shifts, highlighting the necessity for automated simulation frameworks like *Chimera*. Meanwhile, we find that *ChimeraLog* is more challenging for ITD, and models trained on *ChimeraLog* exhibit stronger generalization than existing datasets.

Contributions. We summarize our contributions as follows:

- We design and develop a novel LLM-based multi-agent framework, named *Chimera*, to simulate the user behavior of enterprise employees and insider threats. *Chimera* supports diverse enterprise scenarios and roles, enabling the generation of realistic and diverse log events without the need for manual behavior scripting.
- Based on *Chimera*, we construct a new dataset named *ChimeraLog* that significantly outperforms existing datasets in both realism and complexity. *ChimeraLog* covers 15 distinct insider attack scenarios derived from real-world cases, encompassing six log modalities and contains 25 billion log entries, representing 160 hours of recording with fine-grained labels.
- We conduct an extensive evaluation to measure the quality of *ChimeraLog* and the effectiveness of ITD methods on *ChimeraLog*. The results of human studies show that *ChimeraLog* is consistently rated as more realistic and plausible, which is similar to the real-world dataset *TWOS*.
- We share promising insights gained from the deployment of *Chimera* in real-world environments. Our findings highlight the potential of leveraging LLM-based multi-agent frameworks for fully automated applications in software engineering and security domains.

2 Background and Related Works

In this section, we first introduce the threat model of ITD and then summarize the background and related works with respect to existing approaches for ITD and applications of LLM-based MAS.

2.1 Threat Model

System Model. We consider the simulation of the enterprise environment where employees operate within defined organizational roles (e.g., developers, analysts, administrators). In *Chimera*, each employee is represented by an LLM-powered agent, instantiated within an isolated virtual environment (i.e., a Docker container) and subject to role-based access control (RBAC) policies. This design ensures that agents' behaviors are both individualized and governed by realistic organizational boundaries. For example, developers are prevented from accessing sensitive financial or medical data unless explicitly authorized. To enable comprehensive activity monitoring, *Chimera* captures logs from six complementary modalities. Four correspond to application-layer behaviors, including login records, email communications, web browsing history, and file operations. The remaining two focus on system-level activities: network traffic (captured via packet inspection) and system calls (traced at the container level). These multi-modal logs provide a rich context for analyzing both normal workflows and malicious behaviors. Additionally, Chimera simulates essential organizational dynamics such as daily task scheduling, collaborative meetings, and peer-to-peer interactions among agents. These features are critical to preserving temporal and semantic consistency in the generated data, thereby improving the realism of both benign and adversarial behaviors.

Attacker Capabilities and Assumptions. As shown in Table 1, *Chimera* assumes the presence of insider threats arising from three attacker archetypes, each reflecting a distinct level of intent, access, and potential damage. These threat actors are explicitly modeled within the simulation framework to emulate realistic and context-aware adversarial behaviors.

- **Malicious insiders** are legitimate employees who intentionally abuse their authorized access to harm the organization. Their objectives may include intellectual property theft, financial fraud, or operational sabotage. They often leverage their contextual knowledge of internal workflows and privileges to escalate access, exfiltrate data covertly, or manipulate systems while maintaining the appearance of normal behavior.

- **Masqueraders** are external adversaries who gain unauthorized access by compromising internal credentials, commonly via phishing or credential leakage. Although they lack long-term familiarity with the environment, they can impersonate employees to steal sensitive information, disrupt services, or act as covert conduits for third-party infiltration. Their success often depends on mimicking typical usage patterns to evade detection mechanisms.
- **Unintentional insiders** are well-meaning employees whose negligent actions, such as misconfigurations, weak passwords, or falling victim to phishing, lead to unintended security breaches. While lacking malicious intent, they may still cause significant damage, including inadvertent data exposure or the unintentional escalation of adversarial access.

Moreover, in real-world cases U.S. Department of Justice [2025], Wilkens et al. [2021], most insider attack incidents involve multiple combined attack stages, and certain attacks may involve long incubation periods during which malicious activities remain undetected, effectively catching system maintainers off guard. Therefore, we include three hybrid attacks based on real-world impactful insider threat cases Tavani and Grodzinsky [2014], DepartmentofJustice [2003], U.S. Department of Justice [2020] for analysis. In this way, we aim to challenge ITD systems in realistic and complex ways. For example, a single scenario might involve an insider who gradually escalates privileges through exploitation, and later, as an administrator, exfiltrates sensitive data. Such behavior would be reflected in activity logs that appear consistent with normal administrative actions, except for subtle anomalies, thus requiring advanced detection methods to identify threats.

These attacker models are incorporated into the simulation of *Chimera* via adversarial agent roles. Each type introduces varying levels of stealth, knowledge, and behavioral deviation, thereby enabling a comprehensive evaluation of insider threat detection methods under realistic operational assumptions.

Trust Relationships. Our threat model assumes that the underlying infrastructure, such as the host operating systems, containerization platform, and logging mechanisms, is fully trusted and uncompromised. These components serve solely as a reliable substrate for executing simulation logic and capturing behavioral traces without interference or tampering. In contrast, all LLM agents representing employees are considered untrusted, including those modeling malicious or colluding insiders. These agents may attempt to exploit system permissions, circumvent access controls, or inject adversarial behavior into collaborative workflows. Notably, even agents simulating benign employees may act as inadvertent threat vectors by processing adversarial content, such as phishing emails or misleading peer communications. To preserve the realism and integrity of enterprise operations, *Chimera* enforces configurable RBAC policies that constrain each agent's access to organizational resources.

2.2 Insider Threat Detection

Insider threats typically refer to security risks that originate within the trusted boundary of an organization Homoliak et al. [2019]. Insiders have intimate knowledge of the organization's systems and normal procedures, which can enable them to carry out malicious activities in ways that are difficult to distinguish from routine actions. Previous work Homoliak et al. [2019], Mazzarolo and Jurcut [2019] categorizes insider threats into broad types based on the perpetrator's relationship to the organization and intent. For example, insiders can be divided into masqueraders and traitors. A masquerader is an outside actor or unauthorized user who manages to gain insider credentials and impersonate a legitimate user, while a traitor is one who abuses his or her privileges to perform malicious acts. In our work, we consider both types of attackers but include unintentional insiders, who inadvertently cause harm without malicious intent.

Table 1: Comparison of ITD Datasets.

Dataset	Application Logs	Network Traffic	System Logs	Personality	Log Size	Attack Types
CERT r6.2 Glasser and Lindauer [2013]	✓			✓	●	◐
TWOS Harilal et al. [2017]	✓	✓		✓	○	○
CIC-IDS 2017/2018 Sharafaldin et al. [2018]		✓			●	◐
LANL 2017 ken [2017]		✓	✓		●	○
WUIL Camina et al. [2014]	✓		✓		◐	◐
CPTC 2018 cpt [2018]	✓	✓			◐	◐
OpTC opt [2020]	✓	✓	✓		●	◐
Chimera (Ours)	✓	✓	✓	✓	●	●

Existing ITD Datasets. Due to privacy and legal constraints, internal enterprise activity data is rarely released publicly. Consequently, most existing ITD datasets are either synthetically generated Glasser and Lindauer [2013], Sharafaldin et al. [2018] or collected through controlled environments with limited realism Harilal et al. [2017]. Table 1 compares representative ITD datasets across several key dimensions. Specifically, *CERT* Glasser and Lindauer [2013] is one of the most widely used synthetic ITD datasets. It includes application-layer logs such as logon records, file accesses, and email communications, generated for over 100 simulated users over several months. However, the log data lacks semantic information, and its behaviors are rule-based and repetitive, which limits realism. *TWOS* Harilal et al. [2017] captures human-generated activity from a controlled five-day red-team/blue-team competition involving university students. While its logs exhibit greater behavioral authenticity, TWOS is constrained in both scope and scale. *CIC-IDS 2017/2018* Sharafaldin et al. [2018] focuses on external intrusion detection rather than insider threats. It simulates network traffic using predefined benign profiles and inserts attacks such as DDoS, brute-force, and botnets. However, it lacks application-level logs and any notion of user roles or personalities. *LANL 2017* ken [2017] provides a rich set of real internal authentication and network flow logs over 58 days. Despite its scale and realism, it only covers identity management and system-level behaviors, with no application semantics or attack annotations. *WUIL* Camina et al. [2014] (Windows User Interaction Logs) consists of real Windows GUI usage traces (e.g., mouse events, file access, registry changes). It reflects fine-grained user activity but lacks labeled threat behaviors. *CPTC 2018* cpt [2018] captures red-team/blue-team activity from a collegiate penetration testing competition. While it contains real attacker actions and offers some application and network logs, the data is noisy, fragmented, and lacks continuity of benign behavioral patterns. *OpTC* opt [2020] is a DARPA-backed dataset collected from simulated real-world enterprise environments with embedded red team attacks. It includes comprehensive logs across multiple modalities and timeframes, but still lacks semantic realism in user behaviors.

ITD Methods. Existing work has been proposed using different machine learning-based methods to detect insider threats in large organizations. Early approaches applied traditional classifiers on engineered features from system logs. For example, Support Vector Machine (SVM) Janjua et al. [2020] classifiers have been trained on user activity statistics to distinguish benign from malicious profiles. These methods can achieve very high accuracy on balanced datasets, and they benefit from interpretability and strong performance with limited data.

Deep neural networks have also been adopted. Convolutional Neural Networks (CNNs) Hong et al. [2022] have been applied by converting user behavior sequences into “image-like” matrices and using convolution and pooling to extract spatiotemporal features. Graph-based models, such as Graph Convolutional Networks (GCNs) Hong et al. [2022], treat an organization’s user interactions as a graph, propagating each user’s profile and activity features along edges to capture relational and community structure. In recent years, hybrid deep models have emerged. For example, Deep Synthesis-based Insider Intrusion Detection (DS-IID) model Kotb et al. [2025] uses deep feature synthesis to automatically generate rich user profiles from event logs, and then applies a binary deep learning classifier to detect insiders.

Recently, the introduction of large language models (LLMs) has brought new opportunities for log-based insider threat detection. Audit-LLM Song et al. [2024] proposes a multi-agent collaboration framework, where different LLM agents are assigned specialized roles to jointly analyze large-scale organizational logs. Similarly, LogGPT Qi et al. [2023] leverages the powerful language modeling capability of LLMs to process log sequences as natural language, demonstrating strong performance in anomaly detection tasks under zero-shot settings. RedChronos Li et al. [2025] introduces a production-level LLM-based log analysis system, incorporating query-aware voting and semantic-expansion algorithms to enhance detection accuracy and automation in real-world SOC operations.

2.3 LLM-Based Multi-Agent Systems

Multi-agent systems have emerged as a transformative paradigm in both cybersecurity and software engineering, offering decentralized, collaborative, and adaptive solutions to complex challenges. Comprising multiple autonomous agents that interact within shared simulated environments, MAS enable scalable and resilient architectures capable of handling tasks that are difficult for monolithic systems to manage effectively. Established frameworks such as Camel Li et al. [2023], AutoGen Wu et al. [2023], and MetaGPT Hong et al. [2023] exemplify this trend. More recent open-source and simulation-focused systems like AgentSims Lin et al. provide GUI-driven sandboxes for evaluating LLM agents in custom social and planning tasks, while Alympics Mao et al. [2023] applies game-theoretic settings to probe strategic decision-making by LLM agents. Specialized simulation platforms such as BotSim Qiao et al. [2025] model malicious social botnets using LLM-powered bots to emulate coordinated misinformation campaigns, and social-network simulation systems with LLM-empowered agents explore emergent social behavior in networked settings. MAS can also realistically simulate human user behavior for evaluation and modeling Wang et al. [2025]. In medical domains, MedSentry Chen et al. [2025] explores safety vulnerabilities in LLM-based multi-agent architectures under adversarial prompts. However, no prior work has applied MAS in simulating enterprise insider threats.

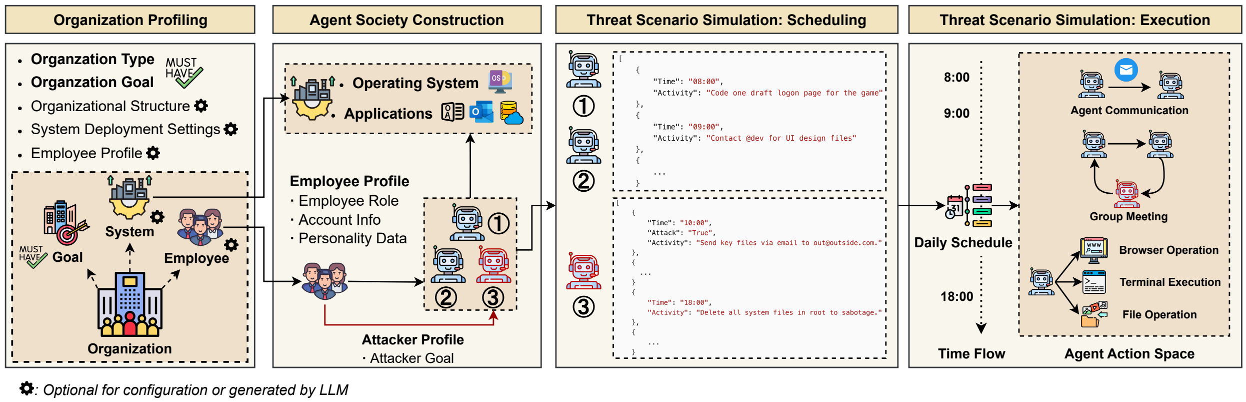


Figure 2: The workflow of *Chimera* for automated insider threat simulation.

3 Design of Chimera

3.1 Overview

The purpose of *Chimera* is to use LLM-based multi-agents to simulate the operations of an organization, such as the activities of the employees of a software development company, where each agent models an employee. To achieve this, as illustrated in Figure 2, *Chimera* operates through three key phases: **Organization Profiling**, **Agent Society Construction**, and **Threat Scenario Simulation**.

Algorithm 1 outlines the detailed procedure for the threat simulation of *Chimera* and corresponding log collection. Specifically, ❶ Given the type (e.g., game company) and the overall goal (e.g., develop a game software) of the organization, *Chimera* can accept the user provided profile of the organization as the framework input or automatically generate this profile (Line 1-2), including the application systems deployed, the number of employees, and the roles assigned to each employee in the organization under simulation (OUS). ❷ Next, *Chimera* constructs an agent society (Line 4-12) in which each agent is personalized according to the roles and characteristics of OUS employees, including adversarial insiders with specific attack objectives. ❸ Finally, *Chimera* simulates organizational activities, including the insider threat scenarios. For each simulated day, benign agents generate detailed schedules based on their daily objectives (Lines 14–15). Meanwhile, attacker agents adjust their schedules to account for malicious objectives (Lines 16–17). After schedule generation, all agents collaborate toward organizational goals, such as software development, while attacker agents conduct covert adversarial actions alongside their legitimate tasks. Throughout the simulation, comprehensive logs are collected for all agent actions, capturing both application-level activities (such as email communication and login information) and system-level events (network traffic and system logs).

3.2 Organization Profiling

To realistically simulate organizational operations, the first step involves acquiring both the *organizational structure* and *system settings*. The *organizational structure* encompasses the number of employees, the primary business objective (e.g., a game company developing a new game, a financial corporation preparing a hedge-fund project), and the roles of each employee. The organizational goal directly affects application-level logging behaviors and shapes the semantic content of log entries. *System settings* are equally important and include the operating systems used by the organization, as well as applications deployed on organizational servers (e.g., Office Automation (OA) systems, and domain-specific software, such as internal Git services). Differences in system configurations, even software versioning, can substantially alter the resulting system-level logs Dragoi et al. [2022], Yu et al. [2025]. Thus, faithful simulation requires adherence to the actual system settings of the target OUS.

Since large corporations often deploy legacy software for compatibility, which may keep the potential vulnerabilities with fixes, such system configuration details are not always available (e.g., due to privacy concerns). Therefore, *Chimera* supports two modes: 1) automatically synthesizes organizational profiles by *Chimera* itself, and 2) accepts and employs users' pre-defined organizational profiles. To ensure realistic simulations, *Chimera* can construct organizational settings grounded in established models of multi-agent coordination following previous works Hong et al. [2023] and scale the number of employees to align with the intended organizational goals.

Algorithm 1: Workflow of CHIMERA

Input: *OrgProfile*: (i) (Mandatory) organization goal and system setting $\mathcal{G}$, (ii) simulation date $\mathcal{D}$, (iii) employee profiles $\mathcal{E}$, (iv) attack profile $\mathcal{P}$

Output: Application / system-level logs $\mathcal{L}$.

```
// Phase 1: Organisation Profiling
if IsMissing(OrgProfile) then
    OrgProfile ← GenerateOrgProfile_LLM(𝒢);
SetupSystemEnv(OrgProfile);
// Phase 2: Agent Society Construction
A ← ∅ ;                                              // Initialize agents
;
foreach employee e ∈ OrgProfile.employees do
    a ← CreateAgentBundle(e);
    AssignProfile(a, ℰ);
    EquipTools(a, {terminal, browser, file-ops});
    if e ∈ 𝒫 then
        AssignAttackObjective(a, 𝒫));
    a.e ← e;
    A ← A ∪ {a};
// Phase 3: Threat Scenario Simulation
for day = 1 to D do                                  // day-level time loop
    foreach agent a ∈ 𝒜 do
        GenerateDailySchedule(a, day);
        if a.e ∈ 𝒫 then
            UpdateAttackSchedule(a, day, 𝒫));
    foreach timeslot t on day do
        foreach agent a ∈ 𝒜 parallel do
            ExecuteTasksOrAttack(a, t);
        UpdateSchedulesAfterComms(𝒜, t);
// Phase 4: Unified Logging
ℒ ← CollectLogs(𝒫);
return ⟨ℒ, 𝒫⟩;
```

3.3 Agent Society Construction

With the organizational config and environment established, *Chimera* proceeds to generate an LLM agent society, which aims to represent each employee in the organization with an LLM agent. Building on prior works Li et al. [2023], Yang et al. [2024], we employ multiple LLM agents per employee and form an “agent bundle" for greater behavioral realism. Each agent bundle comprises multiple functional LLM agents equipped with specialized tools, including a *user* agent (responsible for task distribution and planning), and an *assistant* agent (responsible for providing solutions and using tools such as a terminal execution tool, browser search tool, and file operation tool). These tools are pre-defined with fine-grained prompts from the reference of existing open-sourced tools (e.g., Camel Li et al. [2023]) to ensure robust, accurate tool usage, enabling LLM agents to interact with the system environment much like real employees in industrial scenarios. *Chimera* supplies each agent bundle with key employee information extracted from the configuration file provided by the user (e.g., name, role) and system identifiers retrieved through real-world systems (e.g., container ID, IP address), assembling a unified employee profile as exampled in the Figure 3. To simulate malicious behavior, *Chimera* identifies certain agents as insiders who will execute threats in the system. These adversary agents are assigned specific objectives and plan their attacks while still performing routine duties for a more challenging yet realistic setting in the real-world scenarios.

3.4 Threat Scenario Simulation

Once the system environment and agent society are prepared, *Chimera* simulates realistic organizational operations. To achieve this, each employee agent has a well-defined action space, such as participating in group meetings for planning, creating daily task schedules, and executing tasks to achieve organizational goals. By default, the access-control model of *Chimera* restricts employees from accessing others’ data unless explicitly permitted, echoing practical permission

```
{
    "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"
}
```


Figure 3: Example of employee profile.

boundaries (e.g., as in the *CERT* dataset). Each employee is provisioned with a dedicated Docker container, ensuring individualized system environments.

For benign employees, on each day, they will create a daily schedule following the example format shown in 2. Then, employees will perform tasks based on the schedule. When there is communication between employees, the schedule after the current time will be updated in response to the discussion results. During each simulation date, benign agents generate daily schedules and execute tasks as scheduled. Besides, if agents communicate with each other, future schedules are dynamically updated based on interaction outcomes. The simulation continues until all tasks have been completed. Similarly, for adversarial employees, *Chimera* seamlessly embeds attack activities within their regular daily routines, allowing malicious actions to be concealed among legitimate work tasks. Each attack scenario is instantiated according to real-world adversarial tactics, techniques, and procedures (see Section 2.2 for details). After analyzing the target employee's overall schedule, *Chimera* identifies the most opportune timing for attack execution within the simulation period. The agent's schedule is updated to reflect the specific attack objectives and intent, ensuring that simulated attacks are contextually realistic. Throughout the simulation, all attack logs are automatically labeled, as malicious activities are pre-defined and scheduled by attacker agents with explicit adversarial intent. This labeling process ensures both the convenience and the high quality of the collected log data, supporting downstream research and evaluation tasks.

3.5 Log Collection

High-quality logs are crucial to developing machine learning-based ITD methods. However, unified logging across diverse system configurations is challenging, as different organizations adopt varied software stacks and log formats. *Chimera* addresses this by deploying each agent within a Docker container, capturing network traffic and system logs at the container level. This approach ensures that all activities attributable to a specific agent are comprehensively recorded, faithfully representing the employee's behavior in the OUS. Unlike existing ITD data sets, *Chimera* offers fine-grained employee action tracking, which can be mapped to each employee action during the simulation. More importantly, *Chimera* performs the LLM agent instrumentation, which lets the agent log all collaboration information, as well as the tool call traces for further in-depth analysis. Note that traditional dataset collection approaches rely on a finite set of pre-defined data sources, often failing to satisfy the in-distribution requirements of ITD tasks. In contrast, *Chimera* is highly adaptable and can support all enterprise systems with the proper configuration. All required software is deployed with the assistance of LLM agents, enabling broad applicability and minimal manual intervention.

4 ChimeraLog Dataset

4.1 Dataset Overview

To evaluate the effectiveness of the *Chimera* framework in simulating insider threats in realistic organizational settings, we construct a new dataset, *ChimeraLog*, by deploying *Chimera* in three typical data-sensitive enterprise scenarios: technology companies, financial corporations, and medical institutions. In each scenario, a cohort of 20 employee agents is simulated continuously for one month. We choose a month-long period to allow complex collaborative workflows that require more time than a few days, while keeping the overall simulation within practical time and cost constraints. Nonetheless, *Chimera* remains fully configurable to support arbitrary durations. Across the three scenarios, the agents enact 15 real-world insider attacks, encompassing both benign and malicious behaviors.

ChimeraLog is collected under three foundation models, including OpenAI GPT-4o, Google Gemini-2.0-Flash, and DeepSeek V3. The dataset spans application-level logs, including login events, email communications, web browsing history, and file operations, as well as system-level logs, capturing network traffic (in Packet Capture format Wikipedia [2025] and considers multiple pathing Zhao et al. [2025]) and system logs (System Capture format Sysdig [2025]).

In total, *ChimeraLog* comprises approximately 2.0 billion application-level events, including 0.2 billion logons, 0.6 billion email records, 0.8 billion web histories, and 0.4 billion file operations, alongside 4.5 billion network packets and 18.2 billion system log entries, representing over 160 hours of agent activity. By contrast, widely used public datasets like *CERT* and *TWOS* include only application-level logs and lack the system-level modalities and scale provided by *ChimeraLog*.

4.2 Simulation Scenarios

For data-sensitive organizations, insider threat detection is critical, as internal information compromise has severe consequences. We select three representative data-sensitive organization types, assigning a domain-specific task to each for dataset collection:

- **Technology companies** are particularly susceptible to insider threats due to the rapid pace of innovation and the intensive use of information technology Wright [2025]. Common insider attack patterns include intellectual property (IP) theft (e.g., stealing source files) and IT sabotage (e.g., developers spam threat emails). In our simulation, the organizational goal is to develop a game software, i.e., a third-person shooter game.

- **Financial corporations** are high-value targets due to their management of substantial assets, and insider threats can result in significant financial losses Randazzo et al. [2004]. Typical insider scenarios include fraud or data exfiltration, such as an analyst misusing credentials to steal proprietary trading algorithms, or a broker executing unauthorized trades for personal gain. For this scenario, the corporation's goal is to design a market-neutral statistical arbitrage fund.

- **Medical institutions** manage sensitive patient information, making insider attacks especially damaging Metomic [2025]. Typical attacks involve illicit access to electronic health records, selling protected health information, or sabotaging medical systems. In our simulation, the institution was tasked with completing electronic health record (EHR) collection and conducting seasonal influenza trend analysis.

4.3 Insider Threats

Insider threats are highly diverse, exhibiting countless variations, as insiders possess significantly broader action spaces compared to traditional vulnerability exploitation scenarios. In *ChimeraLog*, we adopt the unified categorization framework proposed in previous work Homoliak et al. [2019], which synthesizes existing taxonomies using the methodology of 'who, what, where, when, why, and how' (5W1H) Yang et al. [2011]. Given the complexity and scale of contemporary insider threats, our goal is to capture as many attack types as possible to support comprehensive simulation and evaluation.

To this end, we explore existing public databases that report real-world insider incidents, including the Data Broker Database Privacy Rights Clearinghouse [2025], the U.S. Attorney's Office United States Department of Justice [2025], and the Federal Bureau of Investigation Federal Bureau of Investigation [2025]. In total, we incorporate 12 individual attack types, along with three hybrid attacks that combine multiple attack patterns. All hybrid attack scenarios are selected based on real-world cases documented by the U.S. Department of Justice (DOJ). A summary and detailed description of each attack considered are provided in Table 2.

Table 2: Attacks considered in *ChimeraLog*. All the attacks are summarized from the existing 5W1H taxonomy Homoliak et al. [2019].

Attacker	Role	Goal	Target	Frequency	Purpose
Traitor	internal	theft of intellectual property	OS, Network, App	recurrent	financial
Traitor	internal	theft of intellectual property	OS, Network, App	single	financial
Traitor	internal	theft of intellectual property	App	single	financial
Traitor	internal/external	sabotage	App	single	financial/personal
Traitor	internal/external	sabotage	OS	single	financial/personal
Traitor	internal/external	sabotage	OS	single	financial/personal
Masqueraders	internal/external	fraud	App	single	financial
Masqueraders	internal/external	fraud	Data	recurrent	financal/personal
Masqueraders	internal/external	theft of intellectual property	OS, Network	recurrent	financial/political
Masqueraders	internal/external	theft of intellectual property	OS, Network, App	recurrent	financial
Unintentional User	internal	data leak	OS, Network	single	personal
Unintentional User	internal	theft of intellectual property	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	App, Network, OS	recurrent	political

5 Evaluation

5.1 Overview

Based on *ChimeraLog*, we conduct a comprehensive evaluation to assess the dataset quality and the effectiveness of existing ITD methods in detecting insider threats in *ChimeraLog*. Our evaluation consists of the following steps:

- **Quality Evaluation.** We conduct a human study with security experts to compare *ChimeraLog* against two established insider threat datasets (i.e., *CERT* and *TWOS*) in terms of realism.
- **ITD Evaluation.** We benchmark four representative machine learning based ITD methods (SVM, CNN, GCN, and DS-IID) using *ChimeraLog*. We further evaluate the cross-dataset generalization (e.g., training on the *ChimeraLog* scenarios A vs. testing on *ChimeraLog* scenario B or *CERT*) to investigate how well models trained on one data distribution can detect threats in another.

5.2 Evaluation Setup

Dataset. Our evaluation covers three datasets, including our constructed *ChimeraLog*, *CERT* insider threat dataset (v6.2), and *TWOS*. For human study, we follow previous research Yang et al. [2022] and randomly sample 100 log entries from each dataset as the study subjects. For the ITD evaluation, for each dataset, we separate it into training, validation, and test sets. To ensure fair evaluation and prevent overfitting, we set aside 10% of each dataset as a test set (never seen during training). The remaining 90% is split into training (80%) and validation (20%) sets. We maintain the original class proportions (normal vs. malicious instances) in all splits. For cross-dataset experiments, we similarly use 10% of the target dataset as test data and train on the entirety of the source dataset's training split. All datasets are preprocessed to a common feature format so that the baseline models can be applied uniformly. In particular, we extract user-day behavioral features following prior work Homoliak et al. [2019] for *CERT* and *ChimeraLog*, which includes aggregating log characteristics for all the log events per user per day.

ITD Models. Four machine learning based ITD methods are considered in the evaluation. Specifically, SVM Janjua et al. [2020] adopts a kernel-based binary classifier that separates normal and malicious user-day profiles using a Radial Basis Function (RBF) kernel. It constructs a hyperplane in high-dimensional feature space to maximize the margin between attack and non-attack instances. Temporal CNN Hong et al. [2022] adopts a convolutional architecture processing fixed-length sequences of user behavior vectors for all days. Stacks convolutional layers with ReLU activations and max-pooling to extract local temporal patterns, followed by fully connected layers for classification. GCN Hong et al. [2022] applies graph convolutions to propagate node features across neighborhoods, followed by node-level classification. Recent research Kotb et al. [2025] has also adopted deep learning based methods in detecting insider threat log behaviors. We choose Deep Synthesis Insider Intrusion Detection (DS-IID) hkotb [2025], which combines the LSTM-based model and the autoencoder to jointly learn the ability to reconstruct normal logs and the attack log classification result. We follow the publicly available implementation Wayne-on-the-road [2025], hkotb [2025] to build models and set parameters following the corresponding papers.

Foundation LLM and LLM Agent Frameworks. *Chimera* relies on LLMs as cognitive engines to simulate human-like organizational behaviors and agentic operations. Importantly, *Chimera* is designed as a flexible multi-agent paradigm for insider threat simulation. This flexibility allows it to be adapted to multiple LLM agent frameworks and foundation models. In our experiments, we employ three models, including Google Gemini-2.0-Flash, OpenAI GPT-4o, and DeepSeek V3.

We implement *Chimera* based on famous multi-agent frameworks Camel Li et al. [2023] and owl Hu et al. [2025]. The multi-agent frameworks provide role-playing communicative architecture that supports multiple agents with individual roles and coordinated interactions. Since the design of *Chimera* is based on these standardized multi-agent platforms, migration to other agent frameworks such as AutoGen Wu et al. [2024] or custom MCP-enabled systems requires only minimal engineering effort.

5.3 Evaluation Metrics.

For ITD baseline evaluation, to mitigate the effect of randomness introduced by the model training process and improve the reliability of the results, we repeat all experiments five times and report the average results. We evaluate the performance of ITD methods using three standard metrics: *Precision*, *Recall*, and *F1-Score*. Specifically, *Precision* denotes the proportion of true attacks among all instances classified as attacks. *Recall* measures the proportion of actual attacks that are correctly identified by the model. The *F1-Score* is the harmonic mean of *Precision* and *Recall*, providing a balanced measure of the classifier's effectiveness. A higher F1-score indicates stronger overall performance. We define a True Positive (TP) as an attack log that is correctly labeled as an attack, while a False Positive (FP) refers to a normal log that is incorrectly labeled as an attack. The formulas for each metric are summarized as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

6 Evaluation Result

6.1 Quality of *ChimeraLog* Dataset

To better fit the real-world applications in deploying ITD methods, given the data-centric tasks' essence, we aim to assess the authenticity of the *ChimeraLog* against the existing *CERT* and *TWOS* datasets. We first conduct a human expert study of the quality of *ChimeraLog* in terms of the realism and practicality of the dataset, then we conduct quantitative analysis regarding the event frequency.

Human Study. We aim to evaluate the realism and practical utility of our collected dataset, specifically examining whether the logs accurately reflect real-world activities and how likely these scenarios are to occur in actual enterprise environments. Given that *ChimeraLog* contains over 20 billion log entries, it is infeasible to manually inspect every entry. To address this, we follow established research methodologies Yang et al. [2022], Neyman [1992] and apply stratified sampling, selecting 100 log entries (e.g., exampled email collected in Figure 6) from each dataset (*ChimeraLog*, *CERT*, and *TWOS*) for human assessment.

Our sampling strategy employs stratification across three dimensions (i.e., dataset, log modality, and behavior). For each dataset, we maintain a fixed total of 100 sampled entries. To maximize interpretability for human experts, we restrict our analysis to four application-level log modalities (i.e., Logon, Email, Web history, and File operations), while excluding system call and network flow records, as these are not human-readable and would hinder consistent evaluation within practical time constraints. Within each dataset, we also balance the sample to include an equal proportion of benign and attack entries (i.e., 50 benign and 50 attack logs per dataset).

$$n_{select}^{(c)} \;=\; n^{(c)} \cdot \frac{N_m^{(c)}\, S_m^{(c)}}{\sum_{j \in \mathcal{M}} N_j^{(c)}\, S_j^{(c)}},\; m \in \mathcal{M},\; c \in \{\mathrm{B}, \mathrm{A}\}.$$

Sample allocation follows Neyman allocation Neyman [1992] principles, whereby the number of selected entries for each stratum is determined based on within-stratum variability. Specifically, M denotes the total number of log modalities, $N_m^{(c)}$ represents the total number of logs in modality m for class c, and $n^{(c)}$ is the target sample size for each class $c \in \{\mathrm{Benign}, \mathrm{Attack}\}$, which in our study is 50. Since the within-stratum standard deviation $S_m^{(c)}$ is not available a priori, we revert to proportional allocation based on stratum size, rounding $n_{select}^{(c)}$ to the nearest integer for each modality and class.

We invited five independent experts, each with at least five years of experience in security and artificial intelligence, from esteemed universities and leading security companies. All of the experts possess deep familiarity with insider threat scenarios within large corporations. Each expert individually evaluated the 100 sampled log snippets for each dataset. The experts were presented with the same set of log entries, which were shuffled to prevent bias. For each log entry, the experts rated its realism and practicality using a five-point Likert scale, drawing on their professional expertise. The evaluation questions included prompts such as, *"The timestamps and event frequency align with a typical workday rhythm"* and *"Overall, I would be inclined to believe that this log segment was captured in a real production environment."* A rating of 1 indicated that the log content was not realistic, while a rating of 5 indicated strong agreement that the log patterns closely reflected real-world activities.

We calculate the average ratings given to each of the questions in the dataset per participant, and present the results in Figure 4. The x-axis distinguishes each dataset we used for evaluation, where we separate the three scenarios of *ChimeraLog* into three rating candidates, and the y-axis shows the average ratings. The results presented in Figure 4 demonstrate that all three organizational scenarios simulated in *ChimeraLog* received expert recognition for their high degree of realism, comparable to the real-world *TWOS* dataset. Specifically, the five participating experts awarded an average realism score of 4.20 to *ChimeraLog*, which is only marginally lower than the 4.25 average score assigned to *TWOS*. This suggests that experts perceive the logs in both datasets as highly natural and realistic. In contrast, the *CERT* dataset received consistently low scores, with an average of 1.78, reflecting experts' views that its logs lack realism. The primary criticism was that *CERT* focuses primarily on system graph construction and populates logs with randomly generated, semantically impoverished content. We quantify inter-rater agreement on the 5-point realism scale using Krippendorff's alpha Krippendorff [2004] (α) to demonstrate the consistency and significance of expert ratings. We report both the point estimate and bootstrap 95% confidence intervals for α. Our analysis indicates high inter-rater reliability, with $\alpha = 0.87$, and the corresponding confidence interval supporting the robustness of these results. This high level of agreement demonstrates that expert ratings are highly consistent, thereby confirming the reliability of the evaluation.

As illustrated by the example email in Figure 6, *Chimera* produces comprehensive and logically consistent content, clearly surpassing both *TWOS* and especially *CERT* in semantic richness and naturalness, which are consistent with expert feedback. While the *CERT* dataset is widely used in research, it consists mostly of meaningless word combinations that fail to convey any real semantics or actionable information. Although the *TWOS* dataset logs are sourced from actual human communications, privacy concerns and practical limitations necessitate masking key information, resulting in brief and often incomplete exchanges.

Quantitative Analysis. We further conduct a quantitative analysis of the three datasets, focusing on the realism of benign activities and the richness of attack-related activities. As shown in Figure 5, the TWOS dataset, although

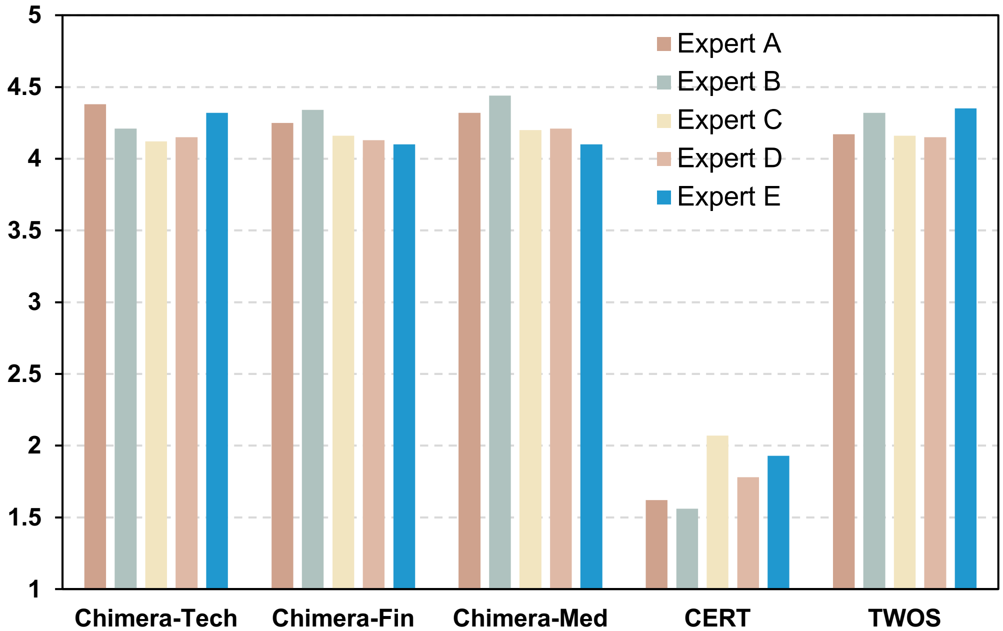


Figure 4: Results of realism study by human experts. The y-axis corresponds to the average ratings (1 refers to very unrealistic; 5 refers to very realistic). The x-axis represents each dataset.

collected from real human participants, is derived from a competition setting. This leads to activity patterns that span the full 24-hour day, which do not realistically reflect typical working hours in real-world companies. While this dataset offers a variety of actions, identifying threats within it requires more detailed action modeling and semantic context. In contrast, the CERT dataset is specifically designed to replicate user activity patterns. Although its log content lacks meaningful semantic information, the overall activity patterns appear normal in comparison to real-world settings.

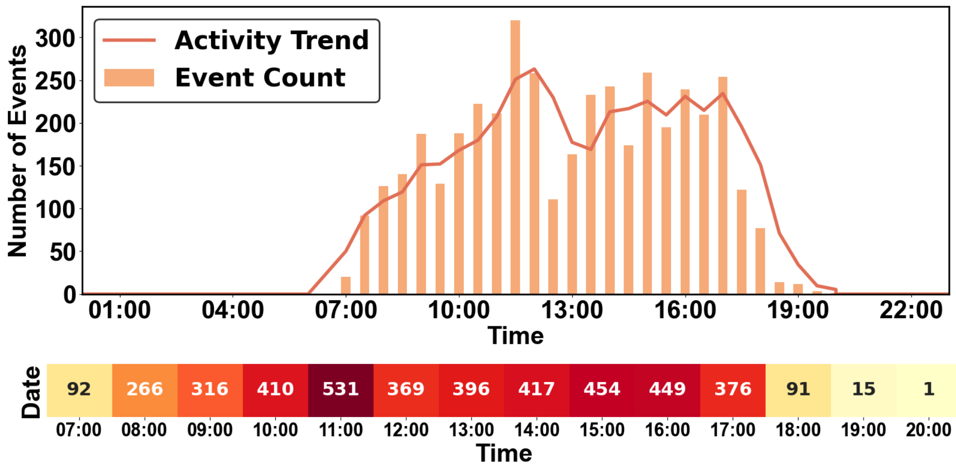
Findings 1: *ChimeraLog* demonstrates similar authenticity comparable to the real-world *TWOS* dataset, which is a significant improvement over *CERT*, whose logs lack meaningful semantic content. At the same time, *ChimeraLog* retains the advantages of *CERT* in terms of realistic activity patterns, which are not present in *TWOS*. This combination highlights the practical potential of *ChimeraLog* for use in the ITD domain.

6.2 Effecitvness of Existing ITD in *ChimeraLog*

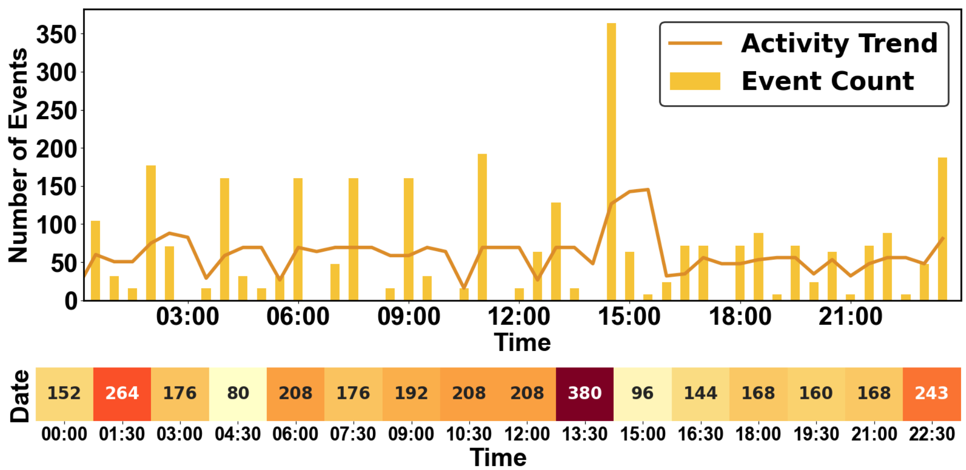
Table 3: Evaluation results of ITD models in different scenarios of *Chimera* and *CERT* dataset. The best (worst) results on each dataset are highlighted with a green (pink) background.

Dataset / Baseline	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

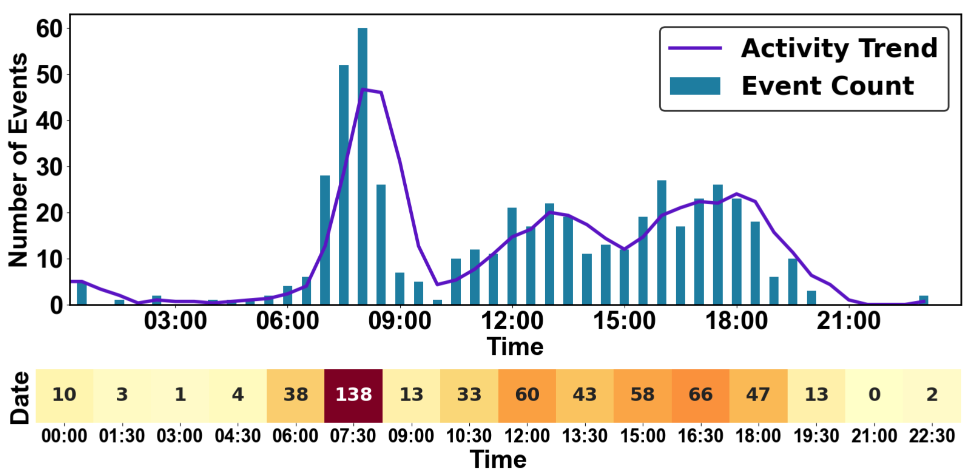
We evaluate the effectiveness of existing log-based anomaly detection (LAD) methods on our *Chimera*-generated dataset. The average performance of each method is summarized in Table 3. Overall, existing ITD methods can identify most insider threats, but their performance varies substantially. While they achieve strong results on the *CERT* dataset, their performance fluctuates by as much as 20% on *ChimeraLog*, highlighting the increased complexity and challenge presented by our dataset compared to the purely simulated *CERT* logs. Notably, Chimera-Finance emerges as the most challenging scenario, suggesting that internal threat behaviors within financial institutions are more deeply concealed than in other contexts. This finding underscores the urgent need for the development of more robust ITD methods tailored to real-world environments.



(a) Daily benign employee activity of *ChimeraLog*.



(b) Daily benign employee activity of *TWOS*.



(c) Daily benign employee activity of *CERT*.

Figure 5: Comparison of benign employee activities among the ITD datasets.

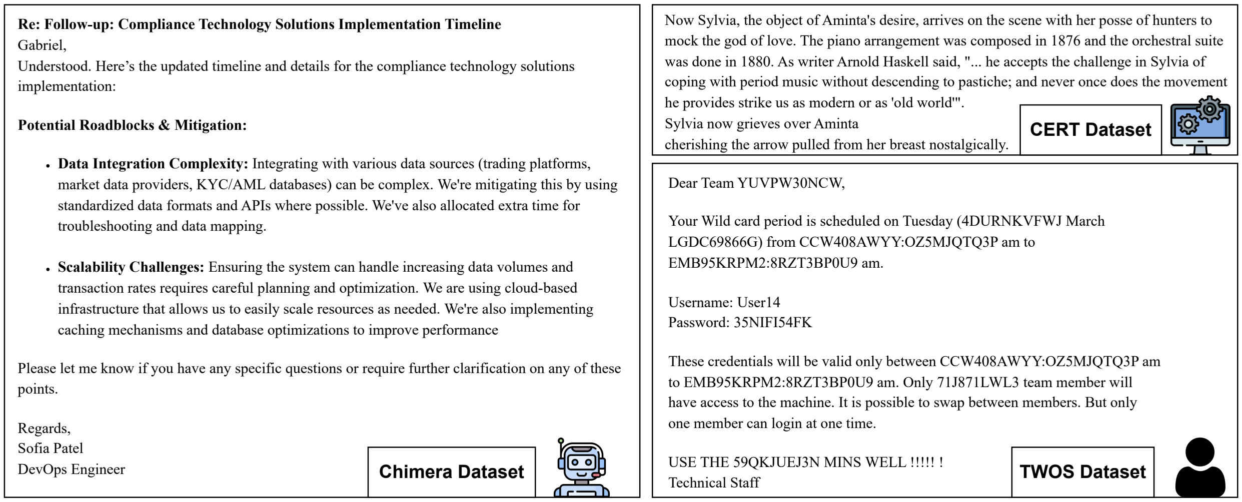


Figure 6: Example of the email communication data in three datasets.

Furthermore, we also examine how well existing ITD methods generalize when faced with distributional shifts. To this end, we train ITD models on one dataset (e.g., Chimera-Tech) and evaluate them on another (e.g., *CERT*). The detailed results are presented in Table 4.

The findings are compelling. The first conclusion is that the distribution shift substantially undermines model performance. For instance, in the DS-IID evaluation, the F1-score drops by 49.1% when comparing performance from the in-distribution scenario (on *CERT*) to the out-of-distribution scenario (Chimera-Finance). This dramatic performance decline underscores the critical need to develop methods that are robust to distribution shifts within the ITD domain. Comparing datasets further highlights that models trained on *CERT* exhibit poor generalization to *ChimeraLog*, which typically results in 100% false positive rates, thereby underscoring the limitations of systems trained solely on synthetic data when deployed in realistic settings.

Findings 2: *ChimeraLog* is more challenging for ITD models compared with *CERT*, and models trained on the *ChimeraLog* demonstrate better generalization. While ITD methods are effective in detecting insider threats, all methods experience significant performance declines when faced with distribution shifts.

Table 4: Cross-Domain Evaluation of ITD models in *Chimera*.

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	$0.468^{\downarrow 0.487}$

6.3 Performance under Different Foundation Models

We further investigate how different foundation models affect the effectiveness of *Chimera*. Concretely, we employ three models, including Gemini, GPT, and DeepSeek, to simulate a technology company and collect logs. Then, we quantify the quality of collected logs.

Table 5 summarizes the statistical information of the dataset collected from different foundation models. Despite receiving the same prompt inputs, the simulation behaviors varied across foundation models. Notably, DeepSeekoperated agents frequently extend work activities late into the night, sometimes until midnight, for tasks such as iterative document refinement. This divergence appears to result from non-terminating logical loops, as agents attempt to converge on content without bifocal agreement, despite prompt-provided constraints. Such behavior is not observed in simulations with OpenAI or Gemini.

Figure 7 illustrates the distribution of activities over different foundation models. OpenAI's foundation model facilitates the most diverse and communicative log outputs, producing higher volumes of events and richer insider interactions compared to Gemini and DeepSeek. In terms of task execution, *Chimera* performs well on atomized, agent-specific tasks, such as searching for API references or emailing colleagues, achieving nearly a 100% success rate. Among the three, Gemini exhibits the highest failure rate on initial attempts, with 22.5% of atomized tasks unresolved on the first try. Error analysis reveals that approximately 85% of failures are due to incorrect tool usage (e.g., misformatted parameters or hallucinated function calls like browser_tool or search_google). Additional operational issues included API rate limits and occasional foundation model timeout errors.

Table 5: Average statistics of simulated dataset from different foundation models.

Model	Daily Start	Daily End	Event	Insider Communication	Task Failure Rate	Attack Step
OpenAI	7:41 AM	19:48 PM	6627	2829	985	30
Gemini	6:58 AM	19:01 PM	4522	349	1017	25
DeepSeek	7:43 AM	22:47 PM	4735	390	872	26

To investigate the usability of datasets generated by *Chimera* with different foundation models, we use them to build machine learning based ITD models and evaluate their effectiveness. Results shown in Table 6 indicate that logs from all three foundation models are effective for training detection models, yielding performance consistent with our prior evaluation findings.

Table 6: Evaluation results of ITD models in the dataset collected from different foundation models.

Dataset / Baseline	ChimeraLog-OpenAI				ChimeraLog-Gemini				ChimeraLog-DeepSeek				CERT			
	Acc	Pre	Recall	F1	Acc	Pre	Recall	F1	Acc	Pre	Recall	F1	Acc	Pre	Recall	F1
SVM	0.758	0.68	0.831	0.748	0.792	0.647	0.781	0.708	0.798	0.724	0.85	0.782	0.916	0.890	0.937	0.913
CNN	0.859	0.889	0.748	0.812	0.813	0.864	0.735	0.794	0.825	0.893	0.766	0.825	0.921	0.900	0.937	0.918
GCN	0.693	0.669	0.735	0.700	0.731	0.709	0.755	0.731	0.736	0.704	0.75	0.726	0.918	0.894	0.937	0.915
DS-IID	0.834	0.732	0.948	0.826	0.801	0.724	0.902	0.803	0.863	0.695	0.948	0.802	0.973	0.790	0.860	0.824

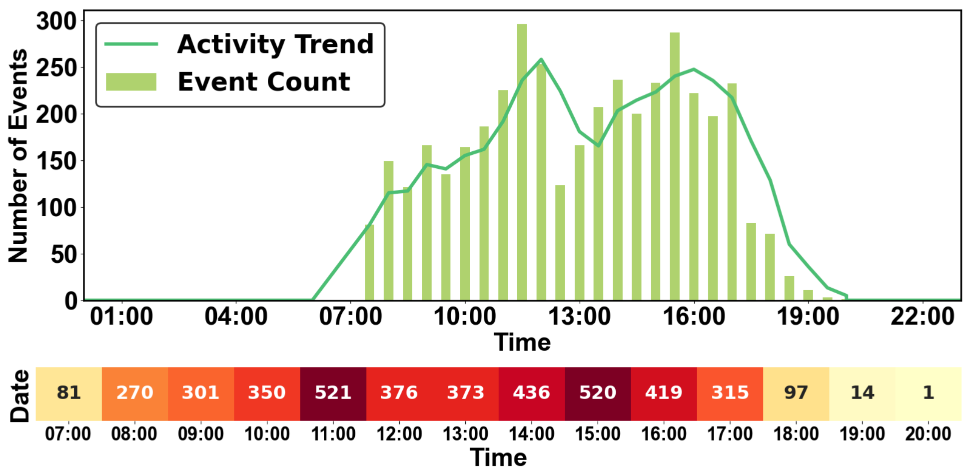
Findings 3: The choice of deployed foundation model affects the quality of generated data by *Chimera*, where GPT-4o serves as the best model. It is a trade-off to consider the data quality and budget.

7 Discussion

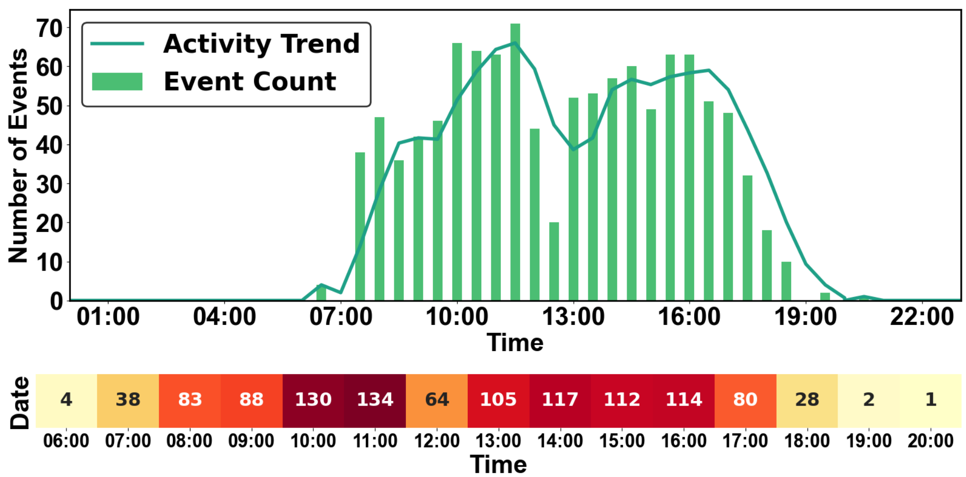
7.1 Implications

Based on our evaluation results and findings, we identify several key implications that could serve as promising guidance for future research.

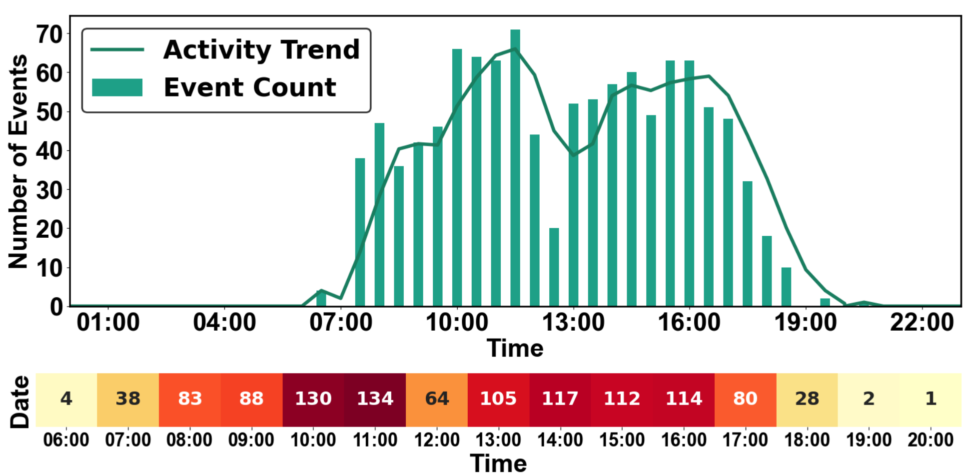
- **Promising usage of multi-agents.** Many works Li et al. [2023], Wu et al. [2023] employ LLM-based multi-agents to solve concrete tasks, such as software development Hong et al. [2023]. In addition, one great use of multi-agents should be simulating scenarios and constructing datasets that are difficult to collect manually, similar to *Chimera*.
- **Distribution shift matters.** Distribution shifts have a significant impact on the effectiveness of ITD methods. As such, users of ITD systems must carefully consider the effects of distribution shifts and ensure timely maintenance of ITD models when such shifts are likely to occur. Furthermore, there is an urgent need to develop new ITD methods that are both robust and effective in detecting challenging insider threat behaviors.
- **Towards fully automation with advanced framework** The success of *Chimera* stems from LLM capabilities, yet it still falls short of realistic and intelligent simulation. Building threat environments and attack workflows remains manual. We urgently need advanced LLM-powered automation, such as threat-environment generation, personality-aware agent simulation, and autonomous penetration testing, to reduce manual work and enable smarter, scalable simulations.



(a) Daily benign employee activity based on OpenAI.



(b) Daily benign employee activity based on Gemini.



(c) Daily benign employee activity based on DeepSeek.

Figure 7: Comparison of benign employee activities produced by *Chimera* with different foundation models.

7.2 Threat to Validity

The *insider threats* are the implementations of models and baselines. During our evaluation, we unify the result-reporting logic of existing baselines to ensure fair baseline comparison. We prioritize using publicly available implementations for the baselines. Another concern comes from the generation of the organization profiles from *Chimera* might not be realistic. To mitigate this concern, human experts are involved to ensure that *ChimeraLog* consists of the proper data.

The *external threats* come from the considered internal attacks, ITD methods, and the base LLM used in *Chimera*. Due to the complexity and huge search space, we cannot cover them all. In addressing this, we follow existing research in attack taxonomy and collect 15 of the most representative attacks, which include 12 summarized orthogonality identical insider threats as well as three hybrid attacks that come from famous real-world threat breaches. Moreover, the current version of *ChimeraLog* supports only three scenarios. We only choose the most severe and common data-sensitive scenarios for collection and make the *Chimera* highly configurable for convenient future extension. For the concern of the base LLM used in *Chimera*, we also implement a highly configurable and adaptable component to support other foundation models and agent frameworks.

7.3 Future Work

- **Hierarchical enterprise structures:** We plan to extend *Chimera* to simulate complex organizations (e.g. departments, branches, subsidiaries) with cross-team interactions. It includes cross-organizational cooperation, which is more realistic for real-world scenarios. Besides, the cross-organization scenarios will also introduce more attack surfaces for the internal attackers.
- **Human-in-the-loop simulations:** Future versions of *Chimera* will support incorporating feedback from security experts and let users participate in the company as attackers. For example, humans might review generated activity sequences or refine agent goals, enabling the system to self-correct and produce more nuanced behavior.
- **Longer timelines:** Building on recommendations from the literature, we will scale up simulations to cover months or years. This will capture evolving behavior trends (such as user churn, seasonal workloads, or progressive breaches) and test models against concept drift. Longer logs will help ensure that detection methods remain effective over time.

7.4 Ethics and Societal Impact

- **Scope of insider threat** *Chimera* is explicitly designed to simulate insider threats within an enterprise security context (e.g. intellectual-property theft, sabotage, or unauthorized access). For the hybrid attacks we considered, our focus is solely on malicious insiders seeking personal or corporate gain within organizational boundaries. We do not take a side on the whistleblowers and politically motivated disclosures, but only study the abnormal activities from the perspective of the organization.
- **Privacy and legality:** All data in *ChimeraLog* are synthetic (e.g., user employee names and IP). No personal information from the real world is used. This avoids privacy, legal, or ethical issues that arise. We aim to simulate the result to be as realistic and match the provided configuration settings as closely as possible. We do not ensure that the configuration generated by the LLM model does not leak personal data Chu et al. [2024]. So we adopt the slight vocabulary mutation of the generated result to circumvent the usage of the leaked data. We ensure that generated logs contain no actual personal identifiers or confidential information.

8 Conclusion

In this paper, we introduced a multi-agent-based framework *Chimera*, which is designed for simulating internal corporate activities in enterprise environments. Based on *Chimera*, we constructed a new dataset *ChimeraLog* that contains diverse internal operation logs to support evaluating insider threat detection methods. Human studies demonstrated that *ChimeraLog* is as realistic as real-world insider threat datasets. Experiments on four ITD methods showed that *ChimeraLog* is more challenging than existing datasets, and distribution shifts posed significant concerns for ITD methods. Based on these findings, we summarized several implications from different perspectives. We believe this work can, to some extent, facilitate future research in enhancing enterprise security.

References

DTEX Systems. 2025 Cost of Insider Threats Global Report. `https://ponemon.dtexsystems.com/`, 2025.

Gurucul. 2024 insider threat report. `https://gurucul.com/2024-insider-threat-report/`, 2024.

PurpleSec. Privilege Escalation Attacks: Types, Examples, And Prevention. https://purplesec.us/learn/privilege-escalation-attacks/, 2025.

Luke Harding. *The Snowden files: The inside story of the world's most wanted man*. Guardian Faber Publishing, 2014.

Arpita Agnihotri and Saurabh Bhattacharya. *Tesla: Knowledge Sharing or Knowledge Protection*. SAGE Publications: SAGE Business Cases Originals, 2024.

Nelson Novaes Neto, Stuart Madnick, Anchises Moraes G de Paula, and Natasha Malara Borges. A case study of the capital one data breach. *Stuart E. and Moraes G. de Paula, Anchises and Malara Borges, Natasha, A Case Study of the Capital One Data Breach (January 1, 2020)*, 2020.

Jordan Richard Schoenherr and Robert Thomson. Insider threat detection: A solution in search of a problem. In *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, pages 1–7. IEEE, 2020.

Mohammed Nasser Al-Mhiqani, Rabiah Ahmad, Z Zainal Abidin, Warusia Yassin, Aslinda Hassan, Karrar Hameed Abdulkareem, Nabeel Salih Ali, and Zahri Yunos. A review of insider threat detection: Classification, machine learning techniques, datasets, open challenges, and recommendations. *Applied Sciences*, 10(15):5208, 2020.

Signpost Six. Insider risk trend report 2025. https://offers.signpostsix.com/insider-risk-trend-report-2025/, 2025.

Daojing He, Xin Lv, Xueqian Xu, Sammy Chan, and Kim-Kwang Raymond Choo. Double-layer detection of internal threat in enterprise systems based on deep learning. *IEEE Transactions on Information Forensics and Security*, 2024a.

Wei Hong, Jiao Yin, Mingshan You, Hua Wang, Jinli Cao, Jianxin Li, and Ming Liu. Graph intelligence enhanced bi-channel insider threat detection. In *International Conference on Network and System Security*, pages 86–102. Springer, 2022.

Hazem M Kotb, Tarek Gaber, Salem AlJanah, Hossam M Zawbaa, and Mohammed Alkhathami. A novel deep synthesis-based insider intrusion detection (ds-iid) model for malicious insiders and ai-generated threats. *Scientific Reports*, 15(1):207, 2025.

Gurucul. Insider threat detection tools requirements. https://gurucul.com/blog/insider-threat-detection-tools-requirements/, 2024. Accessed: 2025-05-31.

Arnau Erola, Ioannis Agrafiotis, Michael Goldsmith, and Sadie Creese. Insider-threat detection: Lessons from deploying the citd tool in three multinational organisations. *Journal of Information Security and Applications*, 67:103167, 2022.

Faisal Janjua, Asif Masood, Haider Abbas, and Imran Rashid. Handling insider threat through supervised machine learning techniques. *Procedia Computer Science*, 177:64–71, 2020.

Joshua Glasser and Brian Lindauer. Bridging the gap: A pragmatic approach to generating insider threat data. In *2013 IEEE Security and Privacy Workshops*, pages 98–104. IEEE, 2013.

Shuhan Yuan and Xintao Wu. Deep learning for insider threat detection: Review, challenges and opportunities. *Computers & Security*, 104:102221, 2021.

Brian Lindauer, Joshua Glasser, Mitch Rosen, Kurt C Wallnau, and L ExactData. Generating test data for insider threat detectors. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, 5(2):80–94, 2014.

Jiongchi Yu, Xiaofei Xie, Qiang Hu, Bowen Zhang, Ziming Zhao, Yun Lin, Lei Ma, Ruitao Feng, and Frank Liau. Cashift: Benchmarking log-based cloud attack detection under normality shift. *arXiv preprint arXiv:2504.09115*, 2025.

Athul Harilal, Flavio Toffalini, John Castellanos, Juan Guarnizo, Ivan Homoliak, and Martín Ochoa. Twos: A dataset of malicious insider threat behavior based on a gamified competition. In *Proceedings of the 2017 international workshop on managing insider security threats*, pages 45–56, 2017.

Dongqi Han, Zhiliang Wang, Wenqi Chen, Kai Wang, Rui Yu, Su Wang, Han Zhang, Zhihua Wang, Minghui Jin, Jiahai Yang, et al. Anomaly detection in the open world: Normality shift detection, explanation, and adaptation. In *NDSS*, 2023.

Junda He, Christoph Treude, and David Lo. Llm-based multi-agent systems for software engineering: Literature review, vision and the road ahead. *ACM Transactions on Software Engineering and Methodology*, 2024b.

Haolin Jin, Linghan Huang, Haipeng Cai, Jun Yan, Bo Li, and Huaming Chen. From llms to llm-based agents for software engineering: A survey of current, challenges and future. *arXiv preprint arXiv:2408.02479*, 2024.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 3(4):6, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. Agentsims: An open-source sandbox for large language model evaluation, 2023. *URL https://arxiv. org/abs/2308.04026*.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, et al. Can large language model agents simulate human trust behavior? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

U.S. Department of Justice. U.s. government employee arrested for attempting to provide classified information to foreign government, 5 2025. URL `https://www.justice.gov/opa/pr/us-government-employee-arrested-attempting-provide-classified-information-foreign-government`. Accessed: 2025-05-31.

Florian Wilkens, Felix Ortmann, Steffen Haas, Matthias Vallentin, and Mathias Fischer. Multi-stage attack detection via kill chain state machines. In *Proceedings of the 3rd Workshop on Cyber-Security Arms Race*, pages 13–24, 2021.

Herman T Tavani and Frances S Grodzinsky. Trust, betrayal, and whistle-blowing: Reflections on the edward snowden case. *ACM SIGCAS Computers and Society*, 44(3):8–13, 2014.

US DepartmentofJustice. A review of the fbi's performance in deterring, detecting, and investigating the espionage activities of robert philip hanssen. 2003.

U.S. Department of Justice. Russian national arrested for conspiracy to introduce malware into a nevada company's computer network. `https://www.justice.gov/archives/opa/pr/russian-national-arrested-conspiracy-introduce-malware-nevada-companys-computer-network`, August 2020. Accessed: 2025-05-31.

Ivan Homoliak, Flavio Toffalini, Juan Guarnizo, Yuval Elovici, and Martín Ochoa. Insight into insiders and it: A survey of insider threat taxonomies, analysis, modeling, and countermeasures. *ACM Computing Surveys (CSUR)*, 52(2): 1–40, 2019.

Guerrino Mazzarolo and Anca Delia Jurcut. Insider threats in cyber security: The enemy within the gates. *arXiv preprint arXiv:1911.09575*, 2019.

Iman Sharafaldin, Arash Habibi Lashkari, Ali A Ghorbani, et al. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, 1(2018):108–116, 2018.

Comprehensive, multi-source cyber-security events. `https://csr.lanl.gov/data/cyber1/`, 2017. [Online; accessed 31-May-2025].

J Benito Camina, Carlos Hernández-Gracidas, Raúl Monroy, and Luis Trejo. The windows-users and-intruder simulations logs dataset (wuil): An experimental framework for masquerade detection mechanisms. *Expert Systems with Applications*, 41(3):919–930, 2014.

Global collegiate penetration testing competition (cptc). `https://cp.tc/`, 2018. [Online; accessed 31-May-2025].

Operationally transparent cyber (optc) data release. `https://github.com/FiveDirections/OpTC-data`, 2020. [Online; accessed 31-May-2025].

Chengyu Song, Linru Ma, Jianming Zheng, Jinzhi Liao, Hongyu Kuang, and Lin Yang. Audit-llm: Multi-agent collaboration for log-based insider threat detection. *arXiv preprint arXiv:2408.08902*, 2024.

Jiaxing Qi, Shaohan Huang, Zhongzhi Luan, Shu Yang, Carol Fung, Hailong Yang, Depei Qian, Jing Shang, Zhiwen Xiao, and Zhihui Wu. Loggpt: Exploring chatgpt for log-based anomaly detection. In *2023 IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, pages 273–280. IEEE, 2023.

Chenyu Li, Zhengjia Zhu, Jiyan He, and Xiu Zhang. Redchronos: A large language model-based log analysis system for insider threat detection in enterprises. *arXiv preprint arXiv:2503.02702*, 2025.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36: 51991–52008, 2023.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Tao Ge, and Furu Wei. Alympics: Llm agents meet game theory–exploring strategic decision-making with ai agents. *arXiv preprint arXiv:2311.03220*, 2023.

Boyu Qiao, Kun Li, Wei Zhou, Shilong Li, Qianqian Lu, and Songlin Hu. Botsim: Llm-powered malicious social botnet simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14377–14385, 2025.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, et al. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37, 2025.

Kai Chen, Taihang Zhen, Hewei Wang, Kailai Liu, Xinfeng Li, Jing Huo, Tianpei Yang, Jinfeng Xu, Wei Dong, and Yang Gao. Medsentry: Understanding and mitigating safety risks in medical llm multi-agent systems. *arXiv preprint arXiv:2505.20824*, 2025.

Marius Dragoi, Elena Burceanu, Emanuela Haller, Andrei Manolache, and Florin Brad. Anoshift: A distribution shift benchmark for unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 35: 32854–32867, 2022.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. Oasis: Open agents social interaction simulations on one million agents. *arXiv preprint arXiv:2411.11581*, 2024.

Wikipedia. Packet capture appliance. `https://en.wikipedia.org/wiki/Packet_capture_appliance`, 2025. [Online; accessed 31-May-2025].

Ziming Zhao, Zhaoxuan Li, Xiaofei Xie, Zhipeng Liu, Tingting Li, Jiongchi Yu, Fan Zhang, and Binbin Chen. Verify all traffic: Towards zero-trust in-network intrusion detection against multipath routing. *IEEE Journal on Selected Areas in Communications*, 2025.

Sysdig. System capture. `https://sysdig.com/learn-cloud-native/what-is-an-scap-file/`, 2025.

Verrion Wright. Top 10 security concerns facing high-growth tech today. `https://bigid.com/blog/top-10-security-concerns-facing-high-growth-tech-today/`, 2025. Accessed: 2025-05-31.

Marisa Reddy Randazzo, Michelle Keeney, Eileen Kowalski, Dawn Cappelli, and Andrew Moore. *Insider threat study: Illicit cyber activity in the banking and finance sector*. United States Secret Service, 2004.

Metomic. Healthcare and insider threats: Securing patient data from within. `https://www.metomic.io/resource-centre/healthcare-and-insider-threats`, 2025. Accessed: 2025-05-31.

Liu Yang, Zhigang Hu, Jun Long, and Tao Guo. 5w1h-based conceptual modeling framework for domain ontology and its application on stpo. In *2011 Seventh International Conference on Semantics, Knowledge and Grids*, pages 203–206. IEEE, 2011.

Privacy Rights Clearinghouse. Data broker database. `https://privacyrights.org/data-brokers`, 2025. Accessed: 2025-05-31.

United States Department of Justice. Offices of the united states attorneys. `https://www.justice.gov/usao`, 2025. Accessed: 2025-05-31.

Federal Bureau of Investigation. Federal bureau of investigation. `https://www.fbi.gov`, 2025. Accessed: 2025-05-31.

Zhou Yang, Jieke Shi, Junda He, and David Lo. Natural attack for pre-trained models of code. In *Proceedings of the 44th International Conference on Software Engineering*, pages 1482–1493, 2022.

hkotb. Ds-iid. `https://github.com/hkotb/insider-threats`, 2025. Accessed: 2025-05-31.

Wayne-on-the-road. B-citd. `https://github.com/Wayne-on-the-road/B-CITD`, 2025. Accessed: 2025-05-31.

Mengkang Hu, Yuhang Zhou, Wendong Fan, Yuzhou Nie, Bowei Xia, Tao Sun, Ziyu Ye, Zhaoxuan Jin, Yingru Li, Zeyu Zhang, Yifeng Wang, Qianshuo Ye, Ping Luo, and Guohao Li. Owl: Optimized workforce learning for general multi-agent assistance in real-world task automation, 2025. URL `https://github.com/camel-ai/owl`.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

Jerzy Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. In *Breakthroughs in statistics: Methodology and distribution*, pages 123–150. Springer, 1992.

Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433, 2004.

Junjie Chu, Zeyang Sha, Michael Backes, and Yang Zhang. Reconstruct your previous conversations! comprehensively investigating privacy leakage risks in conversations with gpt models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL https://arxiv.org/abs/2402.02987.