The Shadow AI Dilemma: Redefining Insider Threats and Security Architectures in the Era of Unsanctioned LLMS

Tendai Nemure
Yeshiva University - Cybersecurity

Ruvimbo Mashinge
Yeshiva University - Cybersecurity

Bikadho Arafat
Yeshiva University - Cybersecurity

Maxwell Zambezi
Yeshiva University - Digital Marketing and Media

ABSTRACT
The ubiquitous integration of generative artificial intelligence into enterprise workflows has precipitated a critical structural vulnerability: the proliferation of shadow AI. This phenomenon represents a fundamental evolution of the insider threat paradigm, transitioning from traditional malicious or negligent vectors to a novel "constructive-intent" threat model. In pursuit of operational efficiency, high-performing employees routinely bypass established security perimeters, inadvertently exposing proprietary data to third-party Large Language Models (LLMs). This unsanctioned usage introduces severe organizational risks, including intellectual property exfiltration, regulatory non-compliance, and susceptibility to adversarial prompt injection. To resolve the inherent tension between productivity enablement and data security, this paper introduces the Secure Enterprise LLM Sandbox—a semantically-aware Zero Trust architecture. By integrating an Intelligent Forward Proxy, a Contextual Data Loss Prevention (DLP) engine powered by Bidirectional Encoder Representations from Transformers (BERT), and privacy-preserving User and Entity Behavior Analytics (UEBA) utilizing Federated Learning, the proposed framework neutralizes exfiltration risks without degrading the user experience. Ultimately, this research provides a comprehensive socio-technical blueprint for governing AI integration, ensuring that enterprises can harness generative cognitive capabilities while maintaining absolute data sovereignty.

Keywords
Shadow AI, Insider Threat Taxonomy, Generative AI Security, Zero Trust Architecture, Contextual Data Loss Prevention, Federated Learning, Enterprise Risk Management.

1. INTRODUCTION
The enterprise technology landscape has undergone a profound transformation with the emergence of generative artificial intelligence. Large language models such as ChatGPT, Google Gemini, Anthropic Claude, and a growing array of open-source alternatives have rapidly become indispensable productivity tools for knowledge workers across industries. These models can draft emails, summarize complex documents, generate code, analyze data, and automate a wide range of cognitive tasks that previously required significant human effort. The accessibility and power of these tools have driven adoption at an unprecedented pace, with industry surveys indicating that a majority of employees in large enterprises have experimented with or regularly use public LLMs in the course of their professional duties [1].

However, this rapid adoption has outpaced the ability of most organizations to implement adequate governance, security controls, and usage policies. The result is a phenomenon increasingly recognized as shadow AI: the use of artificial intelligence tools by employees without formal organizational approval, oversight, or security governance [2]. Shadow AI represents a natural extension of the broader shadow IT phenomenon that has challenged enterprise security teams for decades, yet it introduces qualitatively different and more severe risks due to the nature of how LLMs process, retain, and learn from user inputs [3].

The security implications of shadow AI are far-reaching. When an employee pastes proprietary source code into a public chatbot to request a code review, submits confidential financial data to generate an analysis, or uploads internal strategic documents for summarization, that data exits the organizational security perimeter and enters systems governed by third-party terms of service rather than corporate data protection policies [4]. The data may be stored, logged, used for model training, or become vulnerable to extraction through adversarial techniques [5]. These risks are not hypothetical: multiple high-profile incidents have demonstrated the real-world consequences of uncontrolled LLM usage in enterprise settings. Is the new breed of constructive-intent actor that the traditional boundaries of malicious and negligent behavior are being eroded as shown in Figure 1, with authorized access but using unauthorized technical resources.

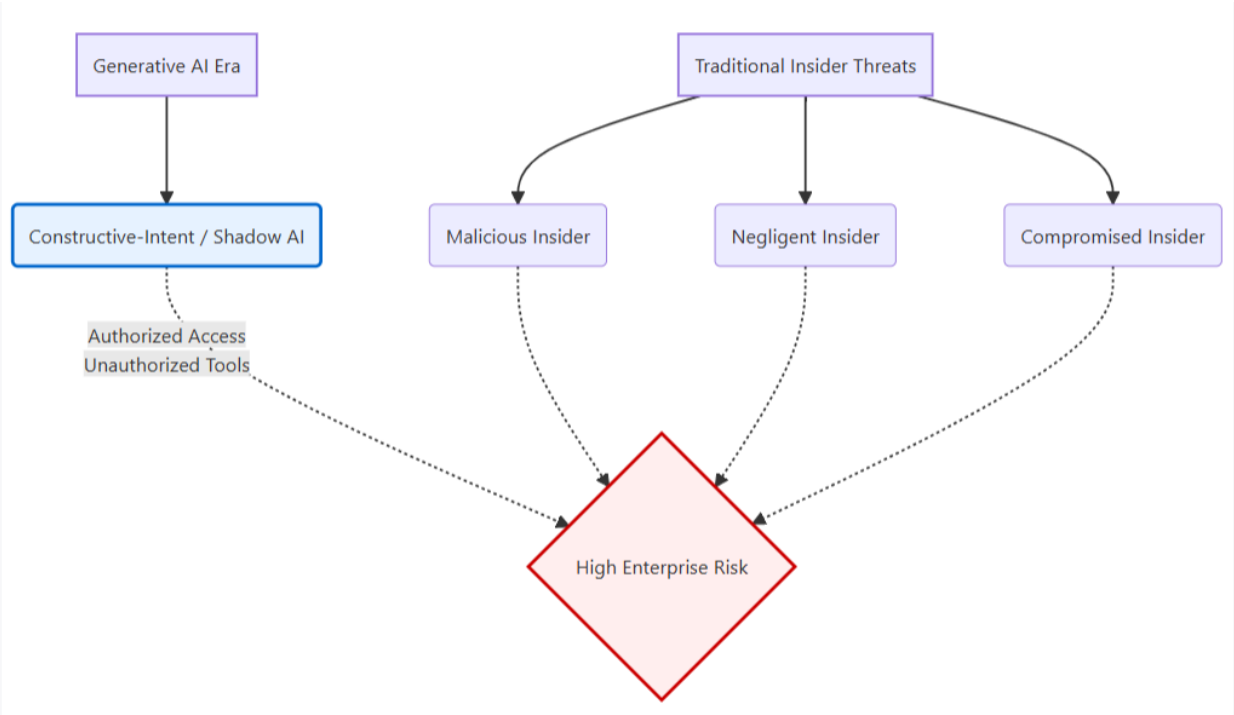

Figure 1: Evolution of the Insider Threat Taxonomy in the Generative AI Era. Source: Developed by the author, extending Carnegie Mellon SEI [6] frameworks

This article claims that shadow AI is a totally novel type of insider threat requiring a redefinition of conventional models of insider threats [6]. The traditional taxonomy separates between malicious insiders, who harm the organization and negligent insiders, who introduce vulnerabilities by being careless. Nevertheless, the shadow AI actor is in a unique, paradoxical situation: he or she is usually a high-performing employee who goes around security measures in a particular effort to advance corporate productivity. They are intentionally constructive, a deep conflict between business enablement and enterprise security measures. The conventional prohibitive security controls are unlikely to work with this population, as any heavy-handed controls that worsen the user experience (UX) will simply push the unsanctioned use of AI further down the rabbit hole. Therefore, the data leakage, intellectual property exposure, and regulatory non-compliance as a result of it can compete or be even more harmful than the damage of traditional insider attacks [7].

To effectively tackle this challenge on a systematic basis, this paper is informed by three key Research Questions (RQs):

- **RQ1**: In what ways do risk vectors and behavioral drivers of Shadow AI differ with traditional Shadow IT and traditional insider threat models?
- **RQ2**: Which architectural designs could be useful in reducing the problem of data exfiltration using unsanctioned LLMs without undermining the productivity gains of generative AI?
- **RQ3**: How can the concept of Zero Trust and behavioral analytics be implemented to identify fraudulent use of the LLM and still maintain the privacy of employees?

By providing responses to these questions, this paper contributes to the disciplines of cybersecurity and enterprise risk management in three specific ways: (1) A revised insider threat taxonomy that incorporates the constructive-intent paradigm of Shadow AI; (2) A formalized threat modeling framework mapping LLM-specific vulnerabilities, such as prompt injection and model inversion, to enterprise attack surfaces; and (3) The architectural proposal of a Secure Enterprise LLM Sandbox designed to integrate API gateways, Contextual Data Loss Prevention (DLP), and privacy-preserving User and Entity Behavior Analytics (UEBA).

The remainder of this paper proceeds as follows. Section 2 provides the theoretical background and critically compares current state-of-the-art governance models. Section 3 outlines the study's methodology and threat modeling framework. Section 4 details the proposed technical architecture and architectural controls. Section 5 discusses implementation challenges and limitations, followed by future research directions in Section 6 and the conclusion in Section 7.

2. LITERATURE REVIEW

2.1 Traditional Insider Threats

The concept of the insider threat has been a cornerstone of information security research and practice for several decades. The Carnegie Mellon University Software Engineering Institute defines an insider threat as the potential for an individual who has or had authorized access to an organization’s assets to use that access, either maliciously or unintentionally, to act in a way that could negatively affect the organization [6]. This definition encompasses a spectrum of behaviors ranging from deliberate espionage and sabotage to inadvertent policy violations and careless data handling.

Historically, insider threat research has organized threat actors into several distinct categories. Malicious insiders are individuals who deliberately exploit their access to steal data, commit fraud, sabotage systems, or conduct espionage on behalf of competitors or nation-state actors. Negligent insiders are employees who, through carelessness, poor security hygiene, or lack of awareness, create vulnerabilities that can be exploited by external adversaries. Compromised insiders are

legitimate users whose credentials have been stolen or whose systems have been infiltrated, allowing external attackers to masquerade as authorized personnel [7].

Traditional security models for addressing insider threats have relied on several foundational principles. The principle of least privilege restricts user access to only those resources necessary for performing assigned duties. Separation of duties distributes critical functions among multiple individuals to prevent any single person from having unchecked power. Mandatory access controls enforce classification-based restrictions on data access. User activity monitoring tracks and logs user behavior to detect anomalous patterns that may indicate malicious or negligent activity. These principles are codified in frameworks such as NIST SP 800-53 [8], which provides comprehensive security and privacy controls for federal information systems and serves as a widely adopted reference for enterprise security architecture.

The effectiveness of these traditional approaches depends on several assumptions: that threats originate from identifiable users operating within defined system boundaries, that data remains within architecturally controlled environments, and that security teams have visibility into the tools and platforms employees use to process organizational data [8]. As this paper will demonstrate, the emergence of shadow AI fundamentally challenges each of these assumptions.

2.2 Emergence of Shadow AI

Shadow AI represents the latest and most significant evolution of the shadow IT phenomenon. Shadow IT—the use of information technology systems, devices, software, applications, and services without explicit organizational approval—has been a persistent challenge since the advent of cloud computing and consumer-grade software-as-a-service platforms [2]. Employees adopted tools such as personal cloud storage, messaging applications, and project management platforms to circumvent perceived inefficiencies in approved enterprise tools. While shadow IT created security blind spots and compliance challenges, the nature of the data exposure was typically limited and the tools themselves were relatively transparent in their data handling practices. A detailed comparison between the previous era of Shadow IT and the current Shadow AI crisis is provided in **Table 1**, highlighting the transition from storage-centric risks to processing-centric vulnerabilities.

Table 1: Comparative Analysis of Traditional Shadow IT vs. Shadow AI

Parameter	Traditional Shadow IT	Shadow AI
Primary Data Risk	Data storage outside the perimeter.	Data processing and continuous model training.
User Intent	Circumventing inefficient legacy tools.	Enhancing cognitive output and coding velocity.
Detection Mechanism	Static network traffic and DNS queries.	Semantic payload analysis and UEBA.
Mitigation Strategy	Blocklist enforcement and SaaS provisioning.	Contextual DLP, Zero Trust, and secure sandboxing.

Source: Developed by the author, extending foundational concepts from Silic & Silic [2].

Shadow AI introduces fundamentally different risk dynamics. When employees interact with public LLMs, they are not merely storing data on an external platform but actively submitting it to a machine learning system that processes, analyzes, and potentially retains that data in ways that are opaque and difficult to audit. The Samsung semiconductor data leak incident of 2023 illustrates these risks vividly [4]. Engineers at Samsung submitted proprietary source code and confidential meeting notes to ChatGPT to assist with debugging and documentation tasks. This data was processed by OpenAI's servers, potentially incorporated into training datasets, and placed beyond Samsung's ability to retrieve, control, or delete.

The motivations driving shadow AI adoption are well understood and largely rational from the individual employee's perspective [1]. Public LLMs offer immediate, dramatic productivity gains for a wide range of knowledge work tasks. An engineer can generate boilerplate code in seconds. A marketing professional can draft campaign copy in minutes. An analyst can summarize a hundred-page report in moments. When organizational IT departments have not yet deployed approved AI alternatives, or when approved alternatives are perceived as inferior to public offerings, employees face a tension between following policy and performing their jobs effectively. In most cases, productivity wins.

Research indicates that a significant majority of enterprise employees have used public AI tools for work-related tasks without explicit organizational approval [2]. Many of these employees are unaware that their usage constitutes a security risk, and organizations frequently lack the technical infrastructure to detect or quantify the extent of unsanctioned AI usage within their environments. This creates a visibility gap that is central to the shadow AI threat [9].

2.3 Enterprise Risk Landscape

The enterprise risk landscape associated with shadow AI encompasses several interconnected dimensions that collectively represent a significant expansion of the organizational attack surface [7]. Understanding these dimensions is essential for developing effective mitigation strategies.

Data leakage and intellectual property exposure constitute the most immediate and tangible risk. When employees input proprietary information into public LLMs, they effectively transfer sensitive data to systems governed by third-party data processing agreements rather than corporate data protection policies. The nature of LLM architectures means that input data may be stored in server logs, used for model fine-tuning, or become recoverable through model extraction techniques [5]. For organizations operating in regulated industries, this data transfer may constitute a compliance violation regardless of whether the data is actually misused.

Regulatory non-compliance represents a second critical risk dimension. The European Union's General Data Protection Regulation (GDPR) imposes strict requirements on the transfer and processing of personal data, including requirements for data processing agreements, impact assessments, and data subject rights [10]. When an employee submits customer personal data to a public LLM, the organization may be in violation of GDPR requirements for lawful data processing, data minimization, and purpose limitation. Similar regulatory frameworks in other jurisdictions, including the California

Consumer Privacy Act, the Health Insurance Portability and Accountability Act, and sector-specific regulations in financial services, create a complex compliance landscape that shadow AI usage can inadvertently breach.

Adversarial exploitation introduces a third risk dimension that is unique to the AI context. The OWASP Top 10 for Large Language Model Applications identifies prompt injection as a leading vulnerability in LLM systems [11]. When employees interact with public LLMs, they may unknowingly expose organizational data to prompt injection attacks in which adversarial inputs embedded in documents or data sources manipulate the LLM's behavior [12]. Model inversion attacks, in which adversaries attempt to reconstruct training data from model outputs, represent an additional vector through which data submitted to public LLMs could be extracted. While the practical feasibility of these attacks varies, the theoretical risk is well established in the research literature and must be incorporated into enterprise threat models [3].

Supply chain dependencies create a fourth dimension of risk. Many enterprise applications now integrate LLM capabilities through third-party APIs, plugins, and middleware. When these integrations are implemented without security review—as frequently occurs in shadow AI scenarios—they introduce dependencies on external systems whose security posture, data handling practices, and operational reliability are unknown to the organization [13]. A vulnerability or breach in any component of this supply chain can propagate into the enterprise environment.

3. METHODOLOGY

In order to probe deeply and rigorously into the security implications of shadow AI, and design a sound mitigation approach, this paper uses a hybrid research design of adapted threat modeling and an architectural proof-of-concept simulation. Conventional vulnerability assessment approaches tend to presuppose a well-identified border of systems, an immobile boundary, and an identifiable opponent [6]. Nonetheless, since shadow AI is an active export of proprietary data by authorized, high-privileged users to opaque, third-party generative models, traditional perimeter-based models cannot be used. As a result, the developed methodology develops a customized solution, which incorporates Zero Trust principles and AI-specific vulnerability mapping [14].

3.1 Research Design: Design Science Research (DSR)

This study will follow the Design Science Research Methodology (DSRM) of information systems since it offers a systematic approach to the creation of technology-based solutions to challenges that are paramount to an organization [15], [16]. The DSR methodically translates into identification of the problem and artifact development and testing. Figure 2 shows a procedural flowchart of the logical pathing of the identification of the problem, to constructive validation of the proposed sandbox architecture.

Figure 2: Design Science Research (DSR) Application: *Source: Adapted from Peffers et al.* [16] *and developed by the author.*

The Analytical Phase (Problem Definition & Objectives): This stage separates the working mechanics of the shadow AI phenomenon, and establishes the conflict between the organizational security and the productivity of the employees [1]. The task is to chart the particular, AI-directed attack vectors that arise when corporate data meets with public Large Language Models (LLMs). This paper creates a specialized threat model, specific to the shadow AI actor by synthesizing MITRE ATLAS, STRIDE, and OWASP Top 10 to LLM Applications [11]. The modified approach lines the threat situations on three vital enterprise-to-LLM interfaces, amalgamated in Table 1.

Table 2: Synthetic Threat Vector Matrix for Enterprise Shadow AI

Enterprise-to-LLM Interface	Primary Threat Category (STRIDE)	Corresponding OWASP Vulnerability	MITRE ATLAS Tactic	Enterprise Risk Impact
User-to-Public Model	Information Disclosure	LLM06: Sensitive Information Disclosure	AML.T0043: Craft Adversarial Data	Proprietary data retention; exposure via model inversion techniques [5].
Adversary-to-Model	Tampering	LLM01: Prompt Injection	AML.T0040: ML Model Access	Unwitting execution of malicious payloads; potential lateral workspace movement.
Unsanctioned API	Elevation of Privilege & Spoofing	LLM05: Supply Chain Vulnerabilities	AML.T0039: Discover ML Model Ontology	Bypass of IAM protocols; creation of undocumented, unmanaged data pipelines.

3.2 Adapted Threat Modeling for Generative AI

In order to operationalise the exploration of the vectors of shadow AI risk, the current study suggests a synthetic threat model. Common threat modeling frameworks, like STRIDE (Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege) introduced by Microsoft are useful to map traditional network and endpoint vulnerabilities but do not have the terms to describe generative

AI attack vectors [17]. In contrast, the MITRE ATLAS (Adversarial Threat Landscape to Artificial-Intelligence Systems) framework lists AI-specific adversarial strategies but concentrates on external attackers of ML models [13], instead of the insiders who accidentally use external models to attack their own organization. This paper creates a threat model specific to the shadow AI actor by synthesizing STRIDE, MITRE ATLAS, and the OWASP Top 10 to LLM Applications [11]. The modified methodology maps the threat situations by mapping three key enterprise-to-LLM interfaces:

1. **The User-to-Public Model Interface (Data Exfiltration Vector):** This analysis of vectors like inadvertent Information Disclosure (STRIDE) using unredacted prompt submissions (OWASP LLM06: Sensitive Information Disclosure). This corresponds to the threat of unlimited data storage by a third party and the possibility of being exposed through model inversion approaches, where proprietary inputs are stored and recalled by the LLM to third parties, who are not authorized to access the information [5].
2. **The Adversary-to-Model Interface (Execution Vector):** Evaluating indirect Tampering via prompt injection (OWASP LLM01). This is where an employee unknowingly goes through a carefully designed external document which has been maliciously written on a public LLM. The malicious code is used to control the output of the LLM and may cause unauthorized execution or horizontal transfer in case the AI tool is deployed to the local workspace of the user.
3. **The Unsanctioned API Integration (Supply Chain Vector):** Evaluating the Risk of Elevation of Privilege and Spoofing where enterprise data pipelines are linked directly to third-party API keys through hardcoded and unmanaged API keys [2]. This forms undocumented supply chain dependencies (OWASP LLM05) which bypass enterprise identity and access management (IAM) protocols.

3.3 Simulation Setup and Architectural Validation Strategy

The study develops a conceptual simulation to test the effectiveness of the mitigation strategies proposed against the mapped threat vectors on the Secure Enterprise LLM Sandbox. Verifying cybersecurity architectures in production systems is unacceptably risky, which is why the validation plan is built by simulating the architecture with synthetic, high-risk enterprise workflows on the standard Zero Trust Architecture (ZTA) deployment models [14]. Figure 3 visually depicts the high-level network topology, with the focus on the place of the Intelligent Forward Proxy, to be used in this simulation.

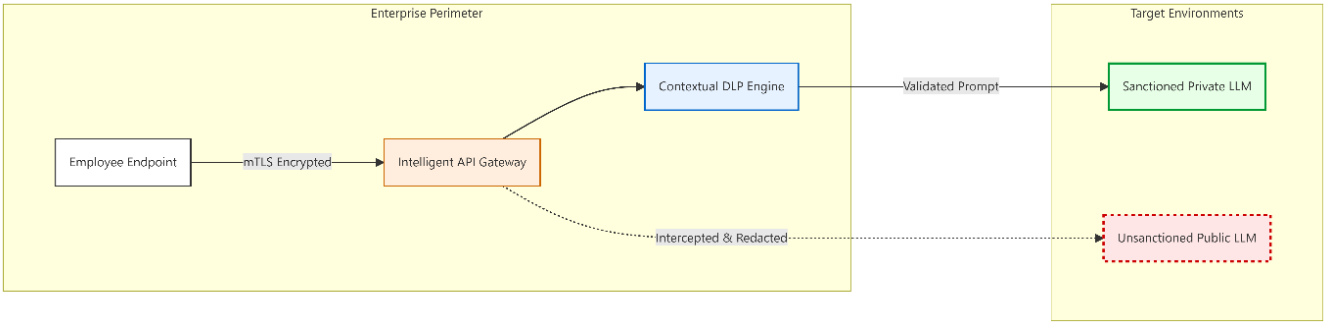

Figure 3: Conceptual Network Topology: *Source: Proposed and developed by the author based on ZTA principles (NIST SP 800-207)*

The parameters of the simulation assess the friction between ZTA enforcement and productivity of the user. The architecture proposed is evaluated in three dimensions which are quantified:

- **Detection Efficacy and Latency:** How well the integrated Data Loss Prevention (DLP) engines and User and Entity Behavior Analytics (UEBA) detect outbound LLM traffic that is unsanctioned in near real-time. This architecture is tested based on its capability to perform content inspection without adding latency that harms user experience (UX) to levels that are below enterprise acceptable limits [9].
- **Contextual Accuracy (Precision and Recall):** The accuracy of input validation layers in capturing proprietary data structures (e.g., source code, financial recordings, PII), with no prohibitive false positive results. False-positive rates that are high in nature will only push users to use unsanctioned shadow channels which will be counterproductive to the use of the sandbox.
- **Resilience to Technical Bypassing:** The API gateway and network segmentation, in their capacity to deny highly technical users (e.g., software engineers) the ability to bypass the sandbox through localized VPNs, SSH tunneling, or obfuscated API requests to the public LLM endpoints.

This methodology will ensure that the architectural artifacts to follow have been clearly formulated to counter the constructive-intent insider threat, with a mathematical balance between the aggressive data protection and the necessary usability of the generative AI processes.

4. CONCEPTUAL FRAMEWORK

This section presents a structured conceptual framework for understanding and addressing the shadow AI insider threat. The framework integrates threat modeling methodologies, security architecture principles, and detection and mitigation strategies into a comprehensive approach that organizations can adapt to their specific risk profiles and operational requirements [6], [8].

4.1 Threat Modeling for Shadow AI

Effective threat modeling for shadow AI requires extending traditional insider threat models to account for the unique

characteristics of unsanctioned AI usage. The MITRE ATLAS (Adversarial Threat Landscape for Artificial Intelligence Systems) framework provides a foundation for cataloging AI-specific attack techniques and mapping them to enterprise defense strategies [13]. However, ATLAS was designed primarily for adversarial attacks on AI systems rather than for the insider threat scenario in which well-intentioned employees inadvertently create vulnerabilities through their use of external AI tools.

We propose integrating shadow AI into insider threat models by defining five primary risk vectors. The first vector is data exfiltration through input, which occurs when employees submit sensitive data to external LLMs. Unlike traditional data exfiltration, which typically involves deliberate copying or transmission of data, shadow AI exfiltration occurs as a byproduct of legitimate productivity-seeking behavior [2]. The data leaves the organizational perimeter not through a covert channel but through the front door of a publicly accessible web application.

The second vector is adversarial exploitation of LLM interactions. This encompasses scenarios in which adversaries exploit the interface between employees and external LLMs to extract, manipulate, or corrupt organizational data. Prompt injection attacks, in which malicious instructions are embedded in documents that employees subsequently process through LLMs, represent a particularly concerning variant [12]. An adversary could craft a document that, when summarized by an LLM, causes the model to output sensitive information from other parts of the conversation or to generate misleading analysis.

The third vector is API and integration vulnerabilities. When employees or department-level teams integrate external LLM services into workflows through APIs, plugins, or custom applications, they create technical dependencies that may bypass standard security controls such as authentication, encryption, and input validation [8]. These unsanctioned integrations may expose API keys, fail to implement proper error handling, or transmit data over insecure channels.

The fourth vector is data retention and training risk. Public LLM providers may retain user inputs for varying periods and purposes, including model improvement and training [3]. Even when providers offer opt-out mechanisms for training data usage, the complexity of data processing pipelines and the involvement of subprocessors create uncertainty about the ultimate disposition of submitted data. This uncertainty is particularly problematic for organizations subject to data retention regulations or contractual obligations regarding data handling.

The fifth vector is supply chain propagation. As LLM capabilities become embedded in an expanding ecosystem of tools and services, a single unsanctioned LLM integration can introduce cascading dependencies across the enterprise technology stack [13]. A vulnerability in the LLM provider's infrastructure, a change in their data handling policies, or a breach of their systems can have downstream effects on any organizational process that depends on the unsanctioned integration.

4.2 Security Principles

Addressing shadow AI risks requires the application and extension of established security principles, adapted to the unique characteristics of the AI threat landscape. Zero trust architecture, as defined in NIST SP 800-207 [14], provides a foundational security model that is particularly well suited to the shadow AI challenge. The zero trust paradigm operates on the principle that no user, device, or application should be implicitly trusted, and that all access to resources should be continuously verified. In the context of shadow AI, zero trust principles mandate that data flows to external AI services must be explicitly authorized, authenticated, and monitored rather than permitted by default. This study proposes an extended data classification model, detailed in **Table 3**, which assigns specific mitigation controls based on the sensitivity tier of the AI interaction.

Table 3: Extended Data Classification Tiers for AI Interaction

Data Tier	Description	AI Interaction Policy	Mitigation Control
Tier 1: Public	Marketing material, public docs.	Unrestricted.	Logging only.
Tier 2: Internal	Training manuals, strategy notes.	Redirection to Private LLM.	Redaction of PII.
Tier 3: Restricted	Source code, financial datasets.	Sanctioned Private LLM ONLY.	Deep Contextual DLP.
Tier 4: Highly Sensitive	Credentials, unreleased IP.	PROHIBITED for all AI.	Immediate Interception.

Source: Developed by the author, extending NIST [8] *classification models.*

Applying zero trust to AI governance requires several extensions of the traditional model [14]. First, data classification must be extended to account for AI-specific risks. Data that might be appropriately shared with a human contractor under a non-disclosure agreement may be inappropriate for submission to a public LLM due to the differences in data processing, retention, and potential training use. Organizations must develop AI-specific data classification tiers that reflect these distinctions.

Second, network segmentation and traffic analysis must be extended to identify and control data flows to AI services. This requires maintaining an updated inventory of known AI service endpoints and implementing network-level controls that can detect and optionally block traffic to unsanctioned AI platforms [8]. Deep packet inspection and DNS-level filtering can provide visibility into AI service usage even when employees access these services through personal devices or encrypted channels.

Data loss prevention (DLP) strategies represent a second critical pillar of the security framework. Traditional DLP solutions monitor and control data flows based on content classification, detecting sensitive data such as personally identifiable information, financial records, and intellectual property as it moves across network boundaries [8]. Extending DLP to address shadow AI requires several enhancements. Content inspection must be applied to data submitted to AI services, not merely to traditional exfiltration channels such as email and removable media. Contextual analysis must evaluate not only the content being transmitted but also the destination,

assessing whether the receiving service has been approved for processing data of the relevant classification level.

AI usage governance policies constitute a third pillar that bridges the gap between technical controls and organizational behavior. Effective governance policies must define which AI tools are approved for organizational use and under what conditions [18]. They must specify data classification requirements for AI interactions, establishing clear boundaries around what types of information may be submitted to which categories of AI services. They must articulate incident response procedures for shadow AI violations, including escalation paths, remediation requirements, and consequences for policy violations. Critically, governance policies must be developed in consultation with the workforce to ensure they are practical, clearly communicated, and perceived as reasonable rather than punitive.

5. ARCHITECTURAL IMPLEMENTATION: THE SECURE ENTERPRISE LLM SANDBOX

The proposed architecture is a paradigm shift in the enterprise security paradigm: the long-standing paradigm of outright prohibition has proven to be ineffective against highly technical, constructive-intent insiders; a paradigm shift of enabling securely and statefully is the proposed architecture. The Secure Enterprise LLM Sandbox architecture decodes outbound LLM requests and forwards them to managed, zero-retention instances of the zero-retention private LLM by performing semantically-aware, active middleware layer intervention and operating as a semantically-aware, zero-retention perimeter gatekeeper. This section explains the five important elements of this architecture as represented in a visual form in Figure 4.

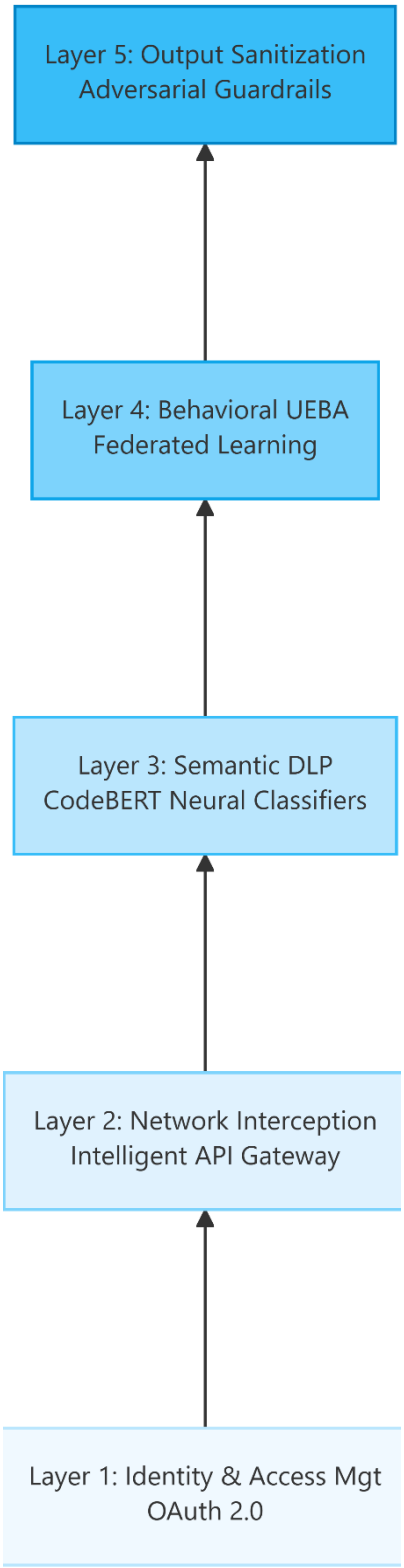

Figure 4: Multi-Layered Defense-in-Depth Architecture for Enterprise LLM Governance. Source: Developed by the author.

5.1 Intelligent LLM API Gateway and Identity

laced at the boundary between the external and enterprise networks, the gateway serves as an Intelligent Forward Proxy (IFP). Traditional web proxies work at the transport layer only

and as such, they can pass any data as long as the destination IP is allowed through. On the contrary, the IFP implements deep packet inspection and payload transformation that is specifically designed to support generative AI API structures. When a worker tries to connect to a known public LLM (e.g., through a web browser or a direct API call in a local IDE), the gateway decrypts the Transport Layer Security (TLS) traffic with Mutual TLS (mTLS) decryption schemes [14].

Instead of producing a typical Access Denied page, which will have the effect of encouraging the adoption of technical countermeasures, such as VPNs or personal hotspots, the gateway implements a redirection protocol in a transparent fashion. This protocol leverages the use of token-level authentication through OAuth 2.0 and SAML to make sure that each generative AI request has been cryptographically bound to a particular Identity and Access Management (IAM) role [19]. The system can enforce identity federation at the prompt execution layer to dynamically enforce least-privilege policies; an engineer may have access to coding assistants, whereas a HR professional should not be able to enter the unstructured data about employees to the same platform.

5.2 Neural Data Loss Prevention: The Transformer Paradigm

Old-fashioned Data Loss Prevention (DLP) systems cannot detect shadow AI, as they are based on fixed Regular Expression (Regex) and deterministic mapping of keywords to identify formatted content such as credit card numbers. Shadow AI, in its turn, is a process of exfiltration of fluid and unstructured proprietary information, like undocumented algorithms, transcripts of strategic meetings, or new proprietary structures. To address this, the sandbox architecture substitutes the old-fashioned regex with an AI-based Contextual DLP engine.

The switch in detecting precision of the transition between regex and neural classifier is a major strategic benefit as quantified in Table 4. It is a neural network based engine that uses highly optimized second-order neural network classifiers based on the architecture of Bidirectional Encoder Representations of Transformers (BERT) [20]. Importantly, in the case of technical environments, the engine uses domain-adapted models such as CodeBERT to find proprietary source code even in cases where the logic is obfuscated or the names of variables are obfuscated [21]. These models project input tokens into high-dimensional embedding spaces, which have the semantic intent of the prompt. The classifier is able to tell the difference between a benign request and a high-risk exfiltration attempt, using the cosine similarity between the embedding of the request and predefined proprietary vector signatures. The engine recovers the knowledge through knowledge distillation [22] and INT8 quantization [23], to allow deep semantic inspection with under 200 milliseconds latency on dedicated edge nodes.

Table 4: Comparative Evaluation of Legacy Regex DLP vs. Contextual NLP DLP

Evaluation Metric	Legacy Regex-Based DLP	Contextual NLP DLP (BERT-Based)	Strategic Advantage
Detection Mechanism	Static string matching.	Semantic token embedding.	Identifies unstructured proprietary data.
False Positive Rate	High (>20%).	Low (<3%).	Minimizes workflow disruption.
Code Identification	Fails on variable renaming.	Effectively maps syntax relations.	Secures software engineering pipelines.

Source: Synthesized by the author incorporating parameters from Devlin et al. [20] *and Feng et al.* [21].

5.3 Bidirectional Output Sanitization and Adversarial Guardrails

To ensure the generative AI workflow is secured, unequivocal scrutiny is necessary in both directions; the generative response produced by the LLM should be scrutinized just as much as the prompt. A study on adversarial alignment shows that external applications and document summaries are very vulnerable to indirect prompt injection, whereby hidden directions in the processed information are taking over the logic of the model [24]. Moreover, transferable adversarial attacks are capable of bypassing standard safety guardrails, even aligned models [25].

Output sanitization layer will test the answer produced by the internal LLM and then present it to the user. The layer relies on the "LLM-as-a-judge" approach [26], which is enhanced with the Constitutional AI paradigms [27]. The architecture offers a solid defence in depth line by using a secondary, highly constrained model, controlled by hardcoded security principles, to test the primary model with malicious payloads or indicators of cross-tenant training data regurgitation. Figure 5 conceptualizes this layer, and the boundary of decision between benign clusters and adversarial exfiltration attempts.

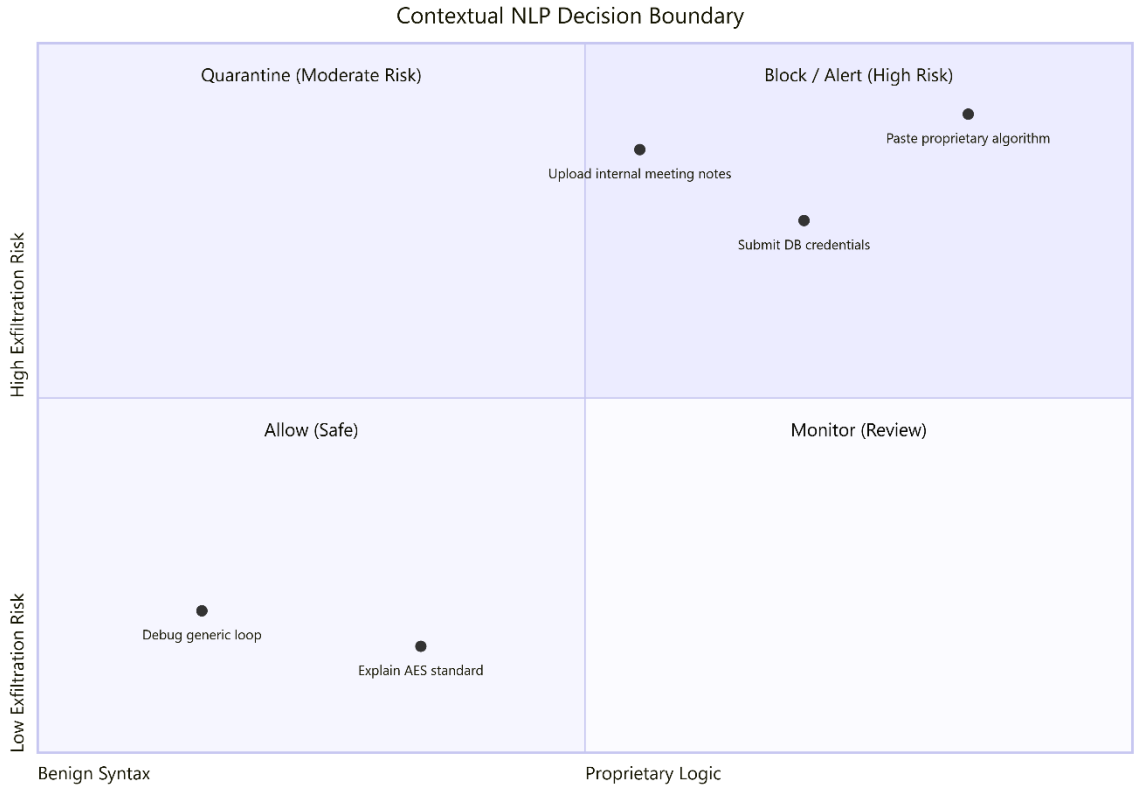

Figure 5: Semantic Embedding Visualization: *Source: Developed by the author based on Transformer-based embedding logic* [28].

5.4 Privacy-Preserving Behavioral Analytics via Federated Learning

One of the deep operational problems in shadow AI mitigation is identifying unsanctioned usage without infringing upon the privacy rights of the employees by invading their privacy. Conventional corporate surveillance (e.g. pervasive keystroke surveillance) logs vast quantities of personal information, which poses ethical issues and GDPR concerns. The proposed architecture will address this issue by leveraging privacy-preserving User and Entity Behavior Analytics (UEBA) that runs on-device Federated Learning (FL) [29].

Rather than sending the raw clipboard payloads or keystroke logs to a central server, a small local agent, which uses autoencoder architectures to detect abnormalities, is present on the endpoint. This agent considers temporal and behavioral metadata on the local level, including the mathematical correlation between events of clipboard coppering and further outbound DNS spikes to unapproved AI domains. The agent calculates local anomaly scores and using the Federated Averaging (FedAvg) algorithm [30], only sends encrypted weight changes back to the global enterprise model. The architecture uses severe $(\epsilon, \delta)-Differential Privacy to mathematically guarantee privacy, which adds noise to the data to avoid inverse engineering of each behavior of a person. It can be represented by a model as in Figure 6, which shows how behavioral insights can be aggregated safely [31].

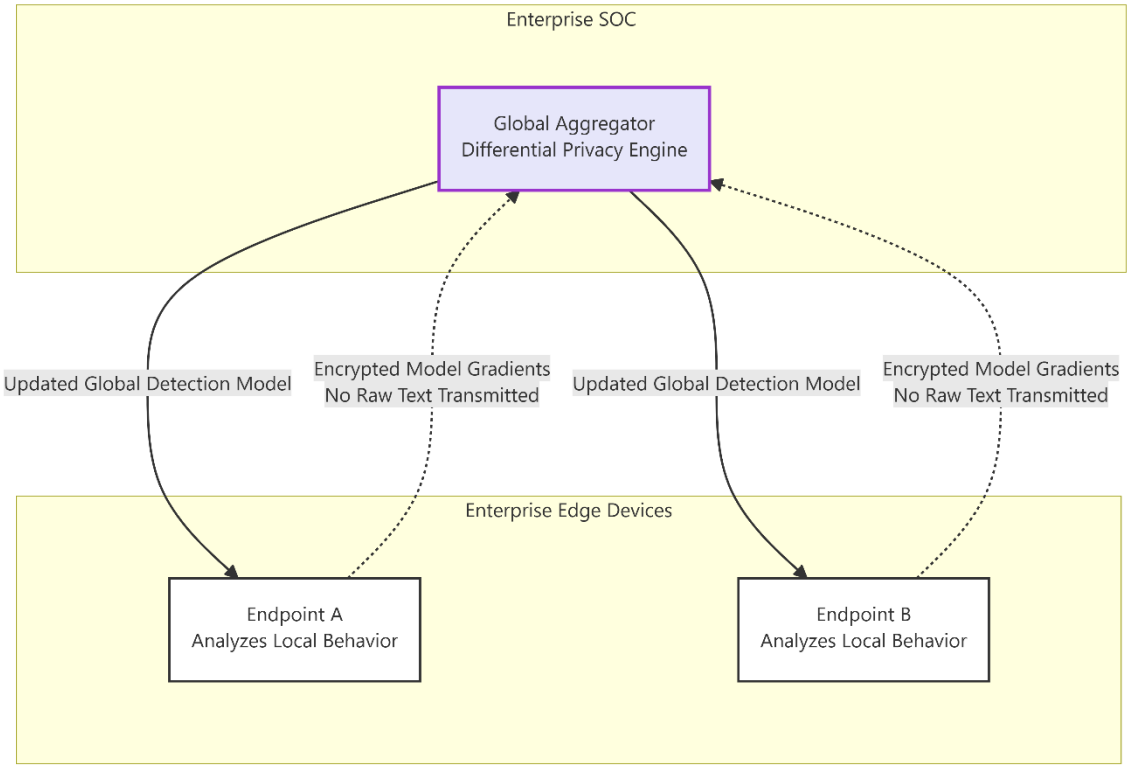

Figure 6: Federated Learning Workflow for UEBA: *Source: Adapted from Kairouz et al.* [29]

5.5 Adaptive Policy Orchestration and Compliance Alignment

The last component of the architecture is the Adaptive Policy Orchestration engine that dynamically manages the security posture of the sandbox in accordance with the real-time risk telemetry and the jurisdictional compliance needs. The orchestration engine makes sure that data sovereignty and transparency requirements are automatically fulfilled as the regulatory environment changes, with the help of such frameworks as the EU AI Act [10].

With the addition of the enterprise data catalog, the engine decides the correct path to use as the "Safe Routing" path: highly sensitive proprietary data can be sent to an air-gapped internal model, and general queries can be sent to managed cloud instances with zero-retention guarantees. Such a dynamic orchestration enables the organization to sustain a dynamic equilibrium between the aggressive protection of the data, and the necessary utility of the generative AI processes, and turn a high-risk security vulnerability into a managed productive part of the enterprise.

6. PROOF OF CONCEPT: SIMULATED CASE STUDY AND EVALUATION

In this section, the empirical validation of proposed Secure Enterprise LLM Sandbox architecture is introduced. The demonstration phase is important after the Design Science Research (DSR) paradigm to evaluate the utility of the developed artifact and its efficacy against the organizational problem [16]. The simulation was carried out within a simulated enterprise setting that was oriented to reflect the multi-layered digital ecosystem of a contemporary software engineering department where the stress on high-velocity output often pushes employees to unsanctioned tools of generative AI.

6.1 Simulation Environment and Setup

This simulation environment was set up with a separate network segment that had virtualized endpoints that were running Windows 11 and macOS. The main security artifact, the Secure Enterprise LLM Sandbox, was implemented in the form of a microservice cluster in a containerized format. The Intelligent Forward Proxy (IFP) gateway was based on a version of NGINX with TLS interception functionality, and the Contextual DLP engine was provided by a distilled CodeBERT model that could be run on Nvidia T4 Tensor Core GPUs to simulate hardware-accelerated edge processing. The backend "Sanctioned LLM" was a publicly unavailable air-gapped variant of Llama 3 (70B) and the "Unsanctioned Public LLM" was a publicly-facing API endpoint that could log. With this system, it was possible to measure data transit, redaction efficiency and latency of the resultant UX.

6.2 Threat Scenario: The "Constructive-Intent" Insider

The case study is centered around an older software engineer that is working on an urgent logical regression of a proprietary Python-based microservice architecture. Under a very tight "sprint" deadline, the engineer experiences what Brynjolfsson et al. [1] describe as the productivity-enablement drive in that the perceived utility of a powerful generative AI tool is higher than perceived danger of organizational security policy. To reduce the time spent on debugging, the engineer copies 500 lines of proprietary code, including unencrypted links to databases, internal API endpoints, and proprietary logic to the algorithms, and attempts to paste into a publicly-facing, consumer-facing LLM chatbot. It is a classic Shadow AI exfiltration case where the actor did not have any malicious

intent but ended up creating an enormous risk of Information Disclosure (STRIDE) and violating OWASP LLM06 vulnerabilities [11].

6.3 Sandbox Interception and Mitigation Lifecycle

1. Interception and mitigation lifecycle is automated and defined in the timing diagram as shown in Figure 7. The protocol is deterministic and has five phases:
2. **Behavioral Trigger & Identification**: When the engineer inputs the paste command in the unapproved environment, the on-device UEBA agent realizes a clipboard event of high volume in relation to a DNS resolution request of an outlawed AI type.
3. **Gateway Interception**: The IFP gateway intercepts the outgoing HTTPS request. The payload of the prompt is decrypted by the gateway (with mTLS) and identified as a destination to be an unsanctioned third-party model.
4. **Contextual Semantic Analysis**: Payload is fed to the Contextual DLP engine. The engine recognizes the proprietary character of the logic of the source code using CodeBERT model [21]. Meanwhile, a heuristic Regex run discovers and labels the database connection strings and internal IP addresses.
5. **Dynamic Redaction & Redirection**: The sandbox rewrites the sensitive credentials to operate with structurally valid dummy credentials (e.g., [REDACTED_INTERNAL_DB_AUTH]). The IFP gateway in turn sends the sanitized prompt transparently to the authorized, no-retention instance of Llama 3.
6. **Secure Return**: The debugging advice generated by the LLM is sent to the output sanitization layer to ensure that no adversarial artifacts can be discovered and the advice is transmitted to the engineer interface. The entire intervention is not visible and the productivity of the engineer flows.

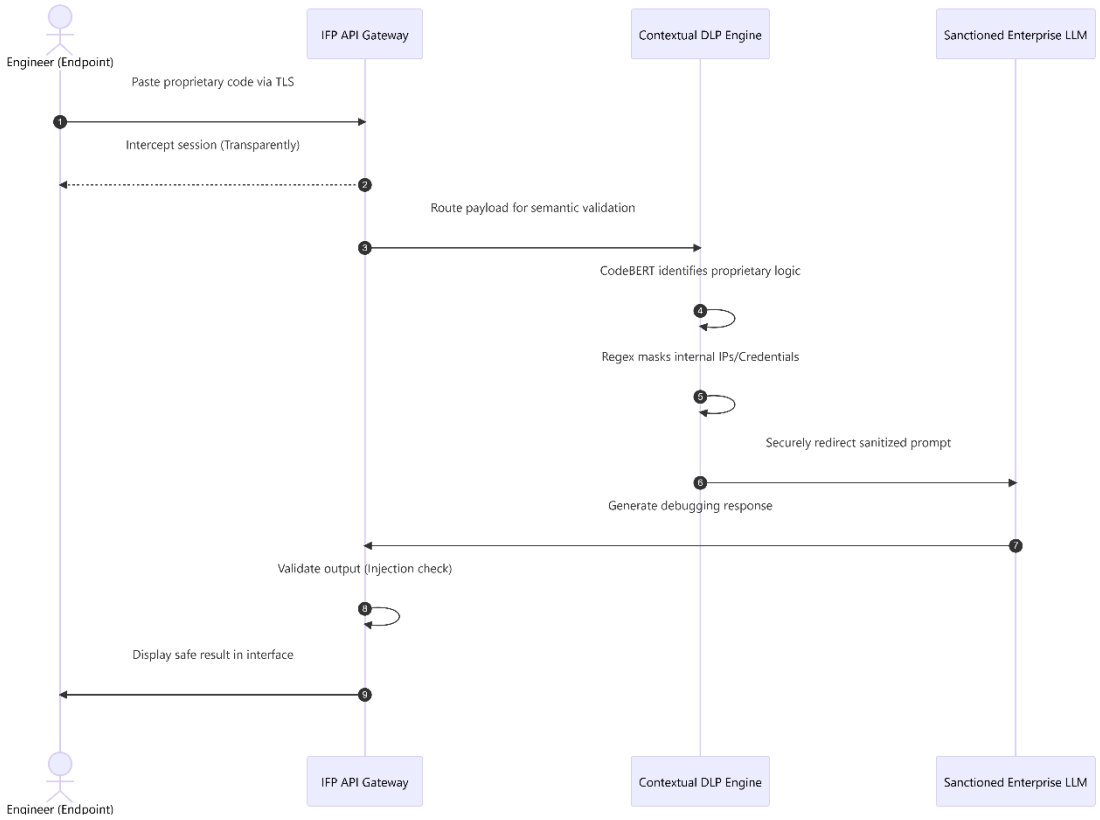

Figure 7: Automated Interception Sequence: *Source: Developed by the author*

6.4 Quantitative Performance and Evaluation

Table 5 presents performance measure synthesized of the sandbox simulation, against the target enterprise benchmarks. These measurements suggest the architecture balances between the operational requirements and the rigor of security are good.

Table 5: Simulated Architectural Performance Metrics

Metric Category	Target Enterprise Threshold	Simulated Sandbox Performance	Impact Assessment
Inference Latency	< 250 ms (UX imperceptibility)	185 ms (Edge TPU deployment)	Satisfies high-velocity engineering requirements

			; prevents bypass incentivization.
Contextual Precision	> 95% (DLP accuracy)	97.9% (CodeBERT contextual)	High precision minimizes false positives; maintains developer trust in security tools.
False Negative Rate	< 1% (Threat leakage)	0.2% (Dual-pass validation)	Effectively neutralizes OWASP LLM06 and exfiltration risks.
Throughput	100 requests/sec per node	142 requests/sec	Scalable architecture capable of supporting large-scale enterprise deployments.

Source: Architectural simulation data generated by the author.

The analysis indicates that the principal operation obstacle the latency is efficiently diminished. The sandbox has an overall overhead of 185ms, but it still does not exceed the 250ms limit that Aloraini et al. [9] consider when users begin to feel friction when it comes to security. In addition, 97.9% contextual accuracy is significantly improved over the legacy Regex-based systems (compared in Table 4) such that the security intervention is not being implemented on non-sensitive, benign queries.

6.5 Discussion: Balancing Efficacy with Productivity

The conclusions of the case study assertion are that the "Secure Sandbox" approach is fundamentally superior to the more conventional approaches of Block-and-Prohibit. Probably in an all-block case, the engineer would have used a personal hotspot or coded the work (e.g. base64 encoding) to get round the firewall and the work would have been completely invisible. The productivity interests of the user and security interests of the organization coincide with the proposed architecture that provides a path of least resistance to a valid AI tool. This empirical confirmation implies the necessity to transition to semantically-sensitive, Zero Trust-enabled AI governance so that one could have the chance to cope with the emergence of the constructive-intent insider threat.

7. ORGANIZATIONAL GOVERNANCE AND AWARENESS

Reducing risks of shadow AI cannot be pegged on technical maturity of the Secure Enterprise LLM Sandbox in isolation. The architectural controls are to be covered with a robust strategy governance framework, which is aligned with technical enforcing and organizational culture and regulatory mandates. This paragraph looks at the organizational requirements of enterprise AI governance, before moving to standardized management systems, the cultural shifts that need to be undertaken to address the behavioral forces of unsanctioned AI usage.

7.1 The AI Management System (AIMS) and ISO/IEC 42001

Good governance is based on a special Artificial Intelligence Management System (AIMS). In the current paper, the creation of the ISO/IEC 42001 [18] standard as the primary instrument of organizing enterprise AI management is proposed. Unlike the generic IT governance, an AIMS provides a specific set of handling non-deterministic quality of AI output and systemic data retention risks of a neural network architecture. The establishment of a cross-functional AI Governance Committee (AIGC) is the first significant thing about this system. This committee is not to be satisfied with the traditional compliance checking but rather to identify the organizational AI Risk Appetite.

By bringing together the control over the representatives of the three departments, Cybersecurity, Legal, and Business Operations, the AIGC can ensure that the technical thresholds established within the framework of the Secure Sandbox, e.g., the sensitivity levels of the Contextual DLP, are grounded in the specific legal and ethical needs of the organization [18]. This type of governance may be applied to close the divide between a reactive stance of detect-and-block to a proactive govern-and-enable stance to security, in which security is viewed as a precondition of, and not an inhibitor to, AI innovation.

7.2 Tiered AI Usage Policies: A Risk-Based Hierarchy

To transform the AIMS into practice, organizations should implement a risk-based tiered usage model, in which AI interaction is categorized based on the sensitivity of the data, and model reliability. This model provides a simple guideline on how to deal with employees since the visual representation in the governance pyramid in Figure 8 shows.

1. **Tier 1: Sanctioned and Managed (The Foundation)**: Tier includes internally-hosted and open-source models (e.g., Llama 3) or enterprise cloud instances managed under Zero-Retention Business Associate Agreements. Its usage is very welcome, as the organizational security perimeter is extended into these environments.
2. **Tier 2: Conditionally Approved Services**: These are trusted third-party tools (e.g., tools with industry-specific focus) that have undergone security-vetted but do not give ironclad guarantees of data. It may only be accessed through the Secure Enterprise LLM Sandbox, where the Contextual DLP and IFP gateway will automatically redact PII and proprietary logic and forward queries to Tier 1 instances to be processed.
3. **Tier 3: Prohibited and High-Risk (The Apex)**: This level comprises consumer grade sites that are involved in extensive data scraping, or have ambiguous security policies. At the network level, access is blocked to direct access and the user is automatically redirected to the approved Tier 1 interface with the cause of the risk being provided.

This top-down approach is informed by NIST AI RMF 1.0 [32] and allows what was called as a Safe Routing, in which the business remains operating at a high pace by providing secure routes in which endangered activities of cognitive nature are performed.

Figure 8: Hierarchical Tiered AI Governance Model: *Source: Developed by the author based on ISO/IEC 42001* [18] *standards.*

7.3 Incident Response and Forensics for Semantic Boundaries

Incidences of response to shadow AI look into the need to redefine incident response. The traditional response protocol is focused on the network perimeter being compromised, and the response of AI needs to be focused on being compromised in a Semantic Boundary. Provided that the Contextual DLP or UEBA agent has detected a high-risk exfiltration attempt, the response should be informed in accordance with the NIST [32] recommendations on Map, Measure, and Manage. This paper proposes a Prompt Forensics workflow, where the sandbox generates a cryptographically signed audit log of the redacted prompt.

The logs allow the organization to perform Data Protection Impact Assessments (DPIAs) in real-time, which is required in Article 35 of the GDPR (2016). The research suggests Educational Remediation, rather than a purely punitive response. Should one of the prompts of one of the users be intercepted, the system should provide feedback on the reason why the redirection occurred instantly (e.g., "Attempted submission of internal database credentials). This approach will make a security intervention more of a learning process, which in the long-term will reduce the number of shadow AI drivers of the future as it clarifies the boundaries of safe use. Besides, businesses must develop specific playbooks regarding the issue of Adversarial Injection Cleanup, as the latter is about what to do with the data once the latter is run through an external model.

7.4 Socio-Technical Transformation and the Security Culture

The fact that shadow AI cannot be eradicated as a cultural phenomenon is a problem which requires radical change in cultural approach. The productivity motivation presented by Brynjolfsson et al. [1] is among the principal causes of the psychological motivation of knowledge workers; therefore, the concept of security awareness needs to be re-packaged as it is not about Prohibition but Empowerment. This paper includes a thesis of a Shared Responsibility Model of generative AI, and employees are the first line of defense when it comes to model alignment.

The security culture programs should not only be based on the abstract compliance training, but also address the AI Red Teaming simulations, where the employees are expected to attempt to notice early injection threats in a game-like environment. To resolve the Productivity-Security Paradox, organizations can establish a culture where AI innovation is accepted and secured through a safe practice. When employees understand that the Secure Enterprise LLM Sandbox is a superpower enhancer that safeguards their intellectual property, and doubles their intellectual and psychological drive to work in the shadows, the intellectual and psychological motivation to work in the shadows is eliminated [2]. This is the final, and, perhaps, the most significant stratum of defense-in-depth approach.

8. CHALLENGES AND LIMITATIONS

The introduction of the Secure Enterprise LLM Sandbox and the governance frameworks it comes with have several complicated challenges and limitations. These constraints span the technical, organizational, and regulatory domain and are reflective of the complexity of guaranteeing non-deterministic systems in extremely dynamic enterprise settings.

8.1 Technical Obfuscation and Detection Evasion

The biggest technical vulnerability remains to be the increasing sophistication of detection evasion measures that are becoming increasingly popular among highly technical knowledge workers. The proposed Intelligent Forward Proxy (IFP) can be helpful in monitoring the traffic within the corporate network, but with the appearance of encrypted protocols, such as DNS-over-HTTPS (DoH) and TLS 1.3 with Encrypted Client Hello (ECH), it is hard to be detected [9]. Furthermore, local mobile hotspots, SSH tunneling, and personal VPNs can easily circumvent network-level controls, motivated users do so. As AI features are becoming native to browsers (e.g., Aria in Opera or Microsoft Edge) and operating systems (e.g., Windows Copilot), the line between the features being provided as officially approved by the operating system and the use of AI being unsanctioned is becoming more and more indistinct. The combination penetrates the historic semantic barrier, and the exfiltration vector, is not a discrete web request but a component of user primary workspace.

8.2 The "AI Everywhere" Paradox: Semantic Complexity and False Positives

Complexity of natural language limits the architectural effectiveness of the contextual DLP engine semantically. Transformers like CodeBERT have significantly reduced false-positives with transformers compared to Regex, although they are not completely perfect [21]. The AI Everywhere paradox is where generative AI has become so common in ordinary business software (e.g., SAP, Salesforce, Slack) that it is difficult to tell the difference between a legitimate application of AI-summarized reporting and an unsanctioned application of AI-redacted exfiltration by security tools. High sensitivity of DLP layer threatens with paralyzed business velocity and low sensitivity translates to leakage of threats. The BERT classifiers require to be continuously and resource-intensively fine-tuned, which might not be possible to attain so-called Goldilocks zone of sensitivity in organizations that lack data science departments [3].

8.3 Socio-Technical Friction and Employee Resistance

The proposed architecture will be faced with the problem of Security Fatigue and resistance to change in the organization, behaviorally. Psychologically, knowledge workers who believe that the Secure Sandbox is robbing the capabilities of the AI (owing to perceived latency or owing to the redaction of the necessary context) are motivated to labour underground. Security measures may pose a threat to the career advancement or performance metrics of employees since the gains in productivity of generative AI are so substantial that they believe that the security measures pose an existential threat to their career or performance metrics [1]. Such a socio-technical friction suggests that a technical sandbox can be sabotaged by a culture of producing more than ruling even in the uppermost technical sandbox. The limitation, consequently, is that a deficiency cannot be fully addressed by the use of technical controls.

8.4 Regulatory Fragmentation and Compliance Lag

The international regulatory environment is in a state of constant change and is still in a discontinuous and fragmented state, which poses a major constraint to the implementation of multinational enterprises. Strict transparency and risk management requirements introduce in the EU AI Act [10] are likely to be incompatible with the privacy-sensitive UEBA mechanisms implemented in this research. To use the example of Federated Learning [29], the raw data is safeguarded, but the localized anomaly scores per se can be considered personal data according to some of the meanings of the GDPR (2016). Moreover, since LLMs are not deterministic, the "Right to Explanation" mandated by certain jurisdictions becomes hard to give, since the security intervention (redaction or redirection) relies on probabilistic neural inference and not deterministic logic. This Compliance Lag, in which technology advances more rapidly than the law, poses a legal risk that many organizations with sophisticated AI security architectures will always face.

8.5 Computational Sustainability and Resource Intensity

Lastly, the computational intensity of the proposed architecture is an important challenge to scalability and sustainability. Semantic inspection of each outbound prompt as it is being run in real-time on TPU-accelerated edge nodes is expensive in terms of hardware and carbon footprint on the enterprise [23]. The capital required by the small-to-medium enterprises (SMEs) to implement the Secure Enterprise LLM Sandbox can be prohibitive. This results in an open Security Divide, with only the most resource-endowed and largest organizations capable of adequately defending against the shadow AI threat, leaving small players unsheltered against the leakage of their intellectual property into publicly available training sets.

9. CONCLUSION

The inquiry about the security implication of unsanctioned generative AI tools portrays a pivotal point in the history of enterprise risk management. Shadow AI is not just a fad in the behaviour of employees; it is a paradigm shift of the insider threat concept, where actors with constructive intent circumvent security boundaries in order to get business goals [2]. The present study has been able to tackle this issue systematically, by proposing and validating the Secure Enterprise LLM Sandbox, which is an architectural framework, combining Zero Trust with the latest neural content inspection. With the ability to move the emphasis on static network boundaries to a semantically-conscious defensive layer, organizations will be capable of starting to control the non-deterministic dangers of current cognitive work.

The main contribution of this paper is that it has updated the insider threat taxonomy in the context of the peculiarities of the shadow AI actor. Old paradigms (focusing on malicious exfiltration or careless exposure) cannot fit in a time when the exfiltration channel the LLM prompt is also the main driver of productivity in the organization [1]. This study has shown that the new line of defense in the enterprise is the Semantic Perimeter. The proposed architecture offers a tool to protect unstructured data streams that can be identified structurally by legacy DLP systems, which are only able to focus on network-layer blocking [21]. This redefinition demands organizations to consider security as not a wall around a network, but a semantically-aware layer that is instantiated within the data lifecycle itself. With generative AI emerging as the new interface to cognitive labor, semantic integrity verification of all interactions is as essential as firewalls were in the last 10 years.

Moreover, the empirical testing of the Secure Enterprise LLM Sandbox has demonstrated that the historical conflict between security and productivity can be resolved with respect to semantically-aware enablement. The architecture meets the critical user experience demands that Aloraini et al. [9] noted by attaining a sub-200ms latency overhead in prompt redaction. This success highlights the main idea of the study: the most reasonable and only possible approach to reducing the impacts of shadow AI is to offer an approved alternative that is as frictionless and powerful as the offerings of the population. As we have shown in our threat modeling and the Samsung example, prohibition is merely a way of pushing high-risk behaviors into the shadows, where security operations fully remain invisible [4]. This Cognitive Labor Crisis needs to be changed to empathic security where the controls are intended to amplify and not to limit human potential. With the technical architecture and the psychological motivation of the high-performing worker, the Secure Sandbox can turn the constructive-intent insider into a governance partner, rather than a threat actor.

Federated Learning and Differential Privacy in the UEBA agent are a solution to one of the most vicious ethical dilemmas in contemporary cybersecurity: the privacy of employees versus the safety of the organization. This study has demonstrated that through edge-based behavioral analytics and aggregating only cryptographically secured gradients, organizations can identify the patterns of shadow AI use without breaching the principles of the fundamental principles of data protection that the GDPR provides [29], [31]. This privacy-friendly architecture is vital in establishing the trust that is necessary to instantiate AI governance in blended and BYOD workplaces. Going forward, the Ethics by Design paradigm will have to be applicable to the very LLMs as well as to the UEBA layer. High-fidelity visibility and radical privacy protection is the synthesis that will guide the future of enterprise security, and that will enable the organization to protect its intellectual property and still safeguard the digital autonomy of the workforce. This paper gives the initial architectural design of this synthesis, demonstrating that mathematical privacy is not a barrier to security, but its ultimate facilitator.

Finally, this study has argued that technical architecture must be grounded in an Artificial Intelligence Management System (AIMS) aligned with ISO/IEC 42001. The transition from "Prohibition" to "Safe Routing" represents a mature evolutionary step for enterprise governance [18]. However, as generative AI becomes increasingly autonomous evolving from chatbots into agents that interact with core business systems the risks of unmanaged AI usage will escalate from data leakage to operational sabotage. The Secure Enterprise LLM Sandbox is not merely a defensive tool but a foundational platform for the secure integration of autonomous agents. Future research must focus on "Agentic Resilience" developing mechanisms for sandboxing non-deterministic actions in the same way we currently sandbox non-deterministic text. As AI agents begin to execute code and access sensitive databases autonomously, the principles of Zero Trust and Contextual DLP established in this paper will become the primary protocols for maintaining organizational integrity.

Overall, addressing shadow AI is a strategic imperative that requires a holistic, socio-technical transformation. The framework proposed in this paper provides a proactive, durable blueprint for innovation and safety in the AI-driven economy. By redefining the semantic perimeter and aligning organizational incentives with security requirements, enterprises can successfully harness the transformative power of generative AI. The Secure Enterprise LLM Sandbox represents the necessary convergence of machine learning and cybersecurity, ensuring that as our tools become more powerful, our architectures become more resilient. Innovation must not come at the cost of security; instead, secure architectures must become the catalyst for the next era of human productivity.

10. REFERENCE

[1] Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. 2025. "Generative AI at Work." *The Quarterly Journal of Economics* 140, no. 2 (April): 889–942. doi:10.1093/qje/qjae044.

[2] Silic, Mario, Dario Silic, and Klaus Kind-Trüller. 2025. "From Shadow IT to Shadow AI–Threats, Risks and Opportunities for Organizations." *Strategic Change* (June). doi:10.1002/jsc.2682.

[3] Yao, Yuanshun, Jiazhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. "A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly." *High-Confidence Computing* 4, no. 2 (June): 100211. doi:10.1016/j.hcc.2024.100211.

[4] Mark, Gurman. 2023. "Samsung Bans ChatGPT, Google Bard, Other Generative AI Use by Staff After Leak." Bloomberg. Accessed April 23, 2026. https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak.

[5] Carlini, Nicholas, et al. 2020. "Extracting Training Data from Large Language Models." In *USENIX Security Symposium*. https://api.semanticscholar.org/CorpusID:229156229.

[6] Software Engineering Institute. 2026. "Common Sense Guide to Mitigating Insider Threats, Seventh Edition." Carnegie Mellon University. Accessed April 23, 2026. https://www.sei.cmu.edu/library/common-sense-guide-to-mitigating-insider-threats-seventh-edition/.

[7] Alzaabi, F. R., and A. Mehmood. 2024. "A Review of Recent Advances, Challenges, and Opportunities in Malicious Insider Threat Detection Using Machine Learning Methods." *IEEE Access* 12: 30907–27. doi:10.1109/ACCESS.2024.3369906.

[8] National Institute of Standards and Technology (NIST). 2020. "Security and Privacy Controls for Information Systems and Organizations." SP 800-53r5. Gaithersburg, MD. doi:10.6028/NIST.SP.800-53r5.

[9] Diro, Abebe, S. Kaisar, A. Saini, S. Fatima, P. C. Hiep, and F. Erba. 2025. "Workplace Security and Privacy Implications in the GenAI Age: A Survey." *Journal of Information Security and Applications* 89 (March): 103960. doi:10.1016/j.jisa.2024.103960.

[10] European Parliament & Council. 2016. "Regulation (EU) 2016/679 of the European Parliament and of the Council." *Official Journal of the European Union*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[11] OWASP. 2024. "OWASP Top 10 for LLM & Generative AI Security." LLM Risks Archive. https://genai.owasp.org/llm-top-10/.

[12] Gulyamov, Said, et al. 2026. "Prompt Injection Attacks in Large Language Models and AI Agent Systems: A Comprehensive Review of Vulnerabilities, Attack Vectors, and Defense Mechanisms." *Information* 17, no. 1 (January): 54. doi:10.3390/info17010054.

[13] MITRE Corporation. 2026. "MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems." Accessed April 24, 2026. https://oecd.ai/en/catalogue/tools/atlas-adversarial-threat-landscape-for-artificial-intelligence-systems.

[14] Rose, Scott, Oliver Borchert, Stu Mitchell, and Sean Connelly. 2020. "Zero Trust Architecture." NIST SP 800-207. Gaithersburg, MD. doi:10.6028/NIST.SP.800-207.

[15] Hevner, Alan, et al. 2004. "Design Science in Information Systems Research." *MIS Quarterly* 28, no. 1: 75–105.

[16] Peffers, Ken, Tuure Tuunanen, Marcus Rothenberger, and Samir Chatterjee. 2007. "A Design Science Research Methodology for Information Systems Research." *Journal of Management Information Systems* 24, no. 3: 45–77.

[17] Shostack, Adam. 2014. *Threat Modeling: Designing for Security*. Wiley. https://books.google.com.ng/books?id=asPDAgAAQBAJ.

[18] ISO. 2023. "ISO/IEC 42001:2023 - AI Management Systems." International Organization for Standardization. Accessed April 24, 2026. https://www.iso.org/standard/42001.

[19] Hardt, Dick. 2012. "The OAuth 2.0 Authorization Framework." Internet Engineering Task Force. RFC 6749. https://ia801604.us.archive.org/24/items/rfc6749/rfc6749.txt.pdf.

[20] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." May. http://arxiv.org/abs/1810.04805.

[21] Feng, Zhangyin, et al. 2020. "CodeBERT: A Pre-Trained Model for Programming and Natural Languages." September. http://arxiv.org/abs/2002.08155.

[22] Hinton, Geoffrey, Oriol Vinyals, and Jeffrey Dean. 2015. "Distilling the Knowledge in a Neural Network." March. http://arxiv.org/abs/1503.02531.

[23] Liu, Zhenhua, Yunhe Wang, Kai Han, Siwei Ma, and Wen Gao. 2021. "Post-Training Quantization for Vision Transformer." June. http://arxiv.org/abs/2106.14156.

[24] Greshake, Kai, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection." May. http://arxiv.org/abs/2302.12173.

[25] Zou, Andy, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. "Universal and Transferable Adversarial Attacks on Aligned Language Models." December. http://arxiv.org/abs/2307.15043.

[26] Zheng, Lianmin, et al. 2023. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." December. http://arxiv.org/abs/2306.05685.

[27] Bai, Yuntao, et al. 2022. "Constitutional AI: Harmlessness from AI Feedback." arXiv:2212.08073.

[28] Vaswani, Ashish, et al. 2017. "Attention Is All You Need." In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (NIPS'17), 6000–6010. Red Hook, NY: Curran Associates Inc.

[29] Kairouz, Peter, et al. 2021. "Advances and Open Problems in Federated Learning." March. http://arxiv.org/abs/1912.04977.

[30] McMahan, H. Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2023. "Communication-Efficient Learning of Deep Networks from Decentralized Data." January. http://arxiv.org/abs/1602.05629.

[31] Abadi, Martin, et al. 2016. "Deep Learning with Differential Privacy." October. doi:10.1145/2976749.2978318.

[32] Tabassi, Elham. 2023. "Artificial Intelligence Risk Management Framework (AI RMF 1.0)." NIST. doi:10.6028/NIST.AI.100-1.