Large Language Models (LLMs) and Generative AI in Cybersecurity and Privacy: A Survey of Dual-Use Risks, AI-Generated Malware, Explainability, and Defensive Strategies

Kiarash Ahi, Saeed Valizadeh
Virelya Intelligence Research Labs
ahi@virelya.org
kiarash.ahi@uconn.edu

***Abstract*— Large Language Models (LLMs) and generative AI (GenAI) systems, such as ChatGPT, Claude, Gemini, LLaMA, Copilot, Stable Diffusion by OpenAI, Anthropic, Google, Meta, Microsoft, Stability AI, respectively, are revolutionizing cybersecurity, enabling both automated defense and sophisticated attacks. These technologies power real-time threat detection, phishing defense, secure code generation, and vulnerability exploitation at unprecedented scales. LLM-generated malware alone is projected to account for 50% of detected threats in 2025, up from just 2% in 2021, emphasizing the need for next-generation security frameworks.**

This paper presents a comprehensive survey of the beneficial and malicious applications of LLMs in cybersecurity, including zero-day detection, DevSecOps, federated learning, synthetic content analysis, and explainable AI (XAI). Drawing on a review of over 70 academic papers, industry reports, and technical documents, this work synthesizes insights from real-world case studies across platforms like Google Play Protect, Microsoft Defender, Amazon Web Services (AWS), Apple's App Store, OpenAI Plugin Stores, Hugging Face Spaces, and GitHub, alongside emerging initiatives like the SAFE Framework and AI-driven anomaly detection.

We conclude with practical recommendations for responsible and transparent LLM deployment, including model watermarking, adversarial defense, and cross-industry collaboration—setting a new benchmark for rigorous, holistic cybersecurity research at the intersection of AI and threat defense—and offering a roadmap for secure, scalable LLM systems that serves as a critical reference for researchers, engineers, and security leaders navigating the complex challenges of AI-driven cybersecurity.

***Keywords*, Large Language Models (LLMs), Generative AI, Cybersecurity, Dual-Use AI, AI-Driven Malware, Explainable AI (XAI), Zero-Day Detection, Federated Learning, Platform Integrity, ChatGPT, Claude, Gemini, LLaMA, Copilot, Stable Diffusion, OpenAI, Anthropic, Google, Meta, Microsoft, Stability AI, Amazon Web Services (AWS), Apple App Store, OpenAI Plugin Stores, Microsoft Defender, Google Play Protect, GitHub, AI Governance, Deepfakes, Synthetic Content, AI Security, Threat Detection, Anomaly Detection**

I. Introduction

The rapid evolution of artificial intelligence has placed Large Language Models (LLMs) and generative AI at the forefront of software innovation and cybersecurity transformation. Originally developed to enhance natural language understanding, LLMs such as GPT-4, PaLM, and Gemini are now widely adopted across industries to automate code generation, accelerate development workflows, and enable intelligent decision-making [1]–[3]. However, this widespread adoption has created a double-edged sword: LLMs empower defenders—especially platform administrators like Google Play, Apple App Store, and other enterprise app platforms—to perform static code scanning, automate threat detection, and improve code quality in real time. Yet simultaneously, those same models are exploited by attackers to generate malware, obfuscate code, and discover vulnerabilities at scale. This duality introduces complex security and governance challenges, underscoring the urgent need for systematic analysis, responsible deployment, and robust defensive frameworks [4].

This paper presents a comprehensive survey of both the risks and opportunities associated with LLMs in cybersecurity. We explore their dual-use nature, recent industry and academic advances, and how both defenders and adversaries leverage these models for tasks such as code generation, malware design, zero-day detection, and DevSecOps, supported by architectural comparisons, benchmark studies, and cross-industry case examples. To guide the reader, the paper is structured as follows: Section II lays the foundational background by reviewing existing literature on the evolution, capabilities, and early governance efforts concerning LLMs. Section III provides an in-depth analysis of LLM applicability in security, detailing their dual-use potential, specific threat vectors like AI-generated malware, the role of explainability, and emerging defensive strategies. Building upon this analysis, Section IV outlines key directions for future research essential for advancing the secure use of LLMs. The paper concludes in Section V, which summarizes the findings and proposes a governance roadmap rooted in explainability, privacy-by-design, federated learning, and compliance. By aligning defensive innovation with emerging safety standards, this paper contributes a timely framework for navigating the rising complexity of LLM-powered cybersecurity ecosystems.

II. Background and Literature Review

A. Evolution and Capabilities of LLMs

Large Language Models (LLMs) have evolved rapidly from their initial applications in natural language translation and generation to highly capable systems supporting complex software engineering tasks. Models such as GPT-4 and PaLM now perform code generation, refactoring, debugging, and even formal verification with increasing accuracy and fluency [1], [5]. These advancements are enabled by scaling transformer architectures and training on diverse programming and natural language corpora. Recent research from OpenAI and Google demonstrates how LLMs can integrate into full development pipelines, assisting with test case creation, API documentation, and dynamic bug resolution [6]–[8].

B. Security Risks and Early Governance Efforts

The dual-use nature of LLMs has raised significant security concerns. On one hand, they can support code auditing and threat detection; on the other, they can generate obfuscated or insecure code, or be weaponized for malicious purposes. Prior work has emphasized the need for proactive safeguards, such as Brundage et al.'s recommendations on structured red teaming and audit trails, and the European Union's Artificial Intelligence Act, which mandates risk assessments and transparency reports for high-impact models [9], [10]. These frameworks aim to mitigate misuse while supporting responsible innovation.

C. Ethics and Governance of Dual-Use LLMs

Integrating LLMs into CI/CD pipelines automates crucial security tasks such as code review, threat detection, and compliance

This paper has been accepted as an invited paper.

enforcement. GitLab and Azure DevOps showcase how GPT based tools can enable real-time security hardening and policy enforcement [11], [12].

While the EU AI Act and the US NIST AI RMF represent significant strides, the global governance landscape for LLMs in cybersecurity remains dynamic, with other major technological regions developing their own distinct approaches. For instance, countries in Asia, such as China, Japan, South Korea, and Singapore, are actively formulating AI regulations and ethical guidelines that reflect their unique priorities. Understanding these varied international perspectives and fostering dialogue towards greater regulatory interoperability will be crucial for addressing the borderless nature of cyber threats and ensuring a globally coordinated response to the risks posed by dual-use AI [9]-[12].

D. LLMs in DevSecOps Automation

Empirical studies of GitHub Copilot and Microsoft Security Copilot illustrate how AI augmented developers are more efficient in detecting and resolving security flaws. These tools not only enhance productivity but also reduce the probability of vulnerabilities slipping into production code [13], [14].

E. Human-AI Collaboration for Secure Development

An important facet of human-AI collaboration in secure development involves leveraging AI models to augment human capabilities in threat detection. For instance, LLMs like VulBERTa are being fine-tuned to identify zero day vulnerabilities through pattern recognition in source code. These models outperform traditional static analyzers, significantly improving detection timelines and precision in identifying new attack vectors [15].

F. Privacy-Aware Deployment of LLMs via Federated Learning

Privacy preserving LLM deployment strategies are increasingly relevant. Federated learning allows training across distributed devices without centralizing data, aligning with laws like GDPR. Kairouz et al. and Bonawitz et al. have demonstrated that these frameworks preserve privacy while maintaining model utility [16], [17].

G. Explainability and Trust in AI Driven Defense

The adoption of LLMs in automated security systems demands transparency. Explainable AI (XAI) methods like SHAP and LIME have been customized to make LLM based vulnerability classifications interpretable. These models help developers and analysts understand the rationale behind predictions, supporting auditability and compliance [18], [19].

H. Adversarial Attacks and Model Vulnerabilities

The integration of LLMs into security critical domains has exposed them to sophisticated adversarial attacks. Carlini et al. highlighted how training data could be extracted from LLMs, undermining confidentiality [8]. Wallace et al. demonstrated that prompt injection and adversarial fine tuning can manipulate LLM outputs, evading content filters. Recent work by Jia et al. organized a global competition revealing how LLMs can be tricked into generating offensive content and misinformation, emphasizing the need for rigorous adversarial testing frameworks [20].

III. ANALYSIS OF LLM APPLICABILITY IN SECURITY

As LLMs become deeply embedded in software development and cybersecurity pipelines, their dual-use potential has triggered increasing scrutiny. A growing body of research has documented how these models can unintentionally or deliberately produce insecure code, including cryptographic flaws, SQL injection vectors, and XSS vulnerabilities [21]–[23]. More alarmingly, the accessibility of LLMs has democratized the creation of deceptive content—enabling non-experts and malicious actors alike to generate phishing apps, polymorphic malware, and social engineering scripts at scale [24]–[26]. These developments reflect not isolated failures but systemic risks introduced by generative models when deployed without sufficient constraints. This section analyzes such risks through three lenses: (1) the emerging threat landscape shaped by misuse and amateur error, (2) industry-led defense strategies to mitigate LLM-enabled attacks, and (3) the broader governance and technical challenges that complicate safe deployment.

A. Amateur Developers and Security Risks

While LLMs empower rapid software creation, they have also unintentionally enabled a wave of insecure development by amateur coders. These models lower technical barriers to entry, allowing individuals with minimal security training to generate functional code quickly. However, this ease often comes at the cost of safety. Studies have shown that inexperienced developers frequently incorporate LLM-generated snippets directly into applications without validating correctness, context, or security implications [27], [28]. As a result, common vulnerabilities such as improper authentication, insecure API usage, and unsafe cryptographic practices proliferate in production software.

This trend is particularly concerning in open-source and mobile app ecosystems, where low-friction publication processes allow insecure code to reach wide audiences. Unlike deliberate attacks, these security flaws emerge from structural gaps—lack of tooling, review, and awareness—highlighting the need for LLM-integrated guardrails that can proactively flag unsafe patterns for novice users. While amateur misuse stems from lack of expertise, the deliberate exploitation of LLMs by adversaries reveals a more calculated—and scalable—weaponization of generative AI.

B. Malicious Actors Leveraging LLMs

In contrast to accidental misuse by amateurs, malicious actors are leveraging LLMs as force multipliers for intentional cyberattacks. These adversaries use generative models to automate large-scale creation of malware, phishing payloads, ransomware variants, and code obfuscation strategies. Unlike conventional malware authors who required domain expertise, attackers can now prompt LLMs to output malicious scripts with minimal effort—dramatically accelerating development cycles.

Industry reports indicate a sharp escalation in LLM-facilitated threat activity, with LLM-generated or assisted malware constituting a significant share of all new threats in 2025 [29]–[31]. Attackers further exploit LLMs to bypass static filters by generating code that mutates slightly on each iteration—evading signature-based detection systems. This reflects a paradigm shift in threat scalability: what was once human-limited is now AI-augmented, enabling adversaries to operate at industrial scale.

C. Statistical Overview of LLM Related Malware

The proliferation of LLM technologies has corresponded with a measurable increase in their exploitation by malicious actors. Recent cybersecurity reports reveal a sharp upward trend in malware generated or assisted by LLMs, raising concerns about automated threat scaling and democratized access to advanced attack tools. **Table 1** presents a year-over-year breakdown of total malware cases and the proportion attributed to LLM-generated threats between 2021 and 2025 [32].

Table 1: Growth of LLM-Generated Malware (2021–2025)

Year	Annual Malware Detections (M)	LLM-Assisted Malware (%)	LLM-Assisted Malware (M)
2021	83.3	2	1.666
2022	104.5	5	5.23
2023	150.0	15	22.5
2024	184.3	30	55.29
2025	221.2	50	110.6

To quantify the accelerating impact of LLMs on the threat landscape, we analyzed data from Cybersecurity Ventures covering malware trends from 2021 to 2025. As shown in **Figure 1**, total malware cases have steadily increased over this period. More notably, the share of malware attributed to LLMs has surged—from just 2% in 2021 to a projected 50% in 2025. This trend highlights a fundamental shift: LLMs are no longer fringe tools for experimentation, but are now actively shaping the scale, speed, and sophistication of cyber threats. The sharp growth curve underscores the need for proactive defense mechanisms that account for AI-assisted attack vectors and evolving adversarial capabilities.

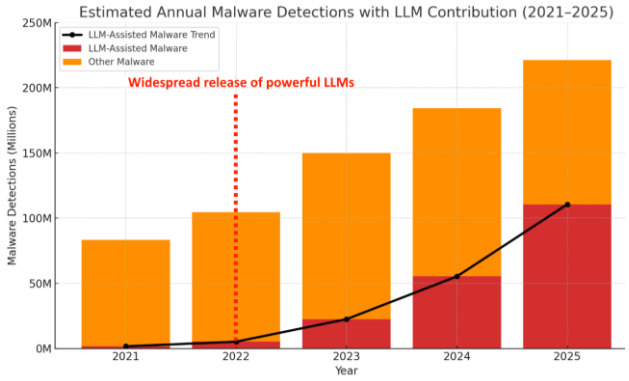


Figure 1 Estimated annual global malware detections with LLM-assisted contribution (2021–2025). Stacked bars show total malware cases, with the red portion representing LLM-assisted threats. The black line highlights the rapid growth of AI-driven malware, rising from 2% to 50% of all detections over the five-year period.

D. Defensive Utilization of LLMs in Mobile App Security

In response to rising AI-powered threats, mobile platform providers are embedding LLMs directly into their security workflows. One of the most effective use cases is automated code review—where LLMs augment traditional static analyzers by identifying logic flaws, unusual API usage, or obfuscated payloads that escape signature-based detection.

Google's Gemini, for instance, plays a pivotal role in powering Play Protect, which scans millions of Android applications daily for malware, policy violations, and suspicious behaviors. By using LLMs, Play Protect has reduced both false positives and time-to-detection, allowing for proactive app ecosystem defense at unprecedented scale [33]–[36].

These use cases demonstrate how LLMs can shift mobile app security from reactive filtering to intelligent pre-deployment screening, flagging issues before users ever download an app. However, as defensive applications of LLMs grow more powerful, they also inherit risks such as overfitting, bias, or exploitability—making explainability and continuous retraining essential.

E. Industry Case Studies: Leveraging LLMs for Cyber Defense

As threats fueled by LLMs escalate, leading technology companies are responding by deploying their own LLM-powered tools to reinforce digital defenses. These platforms integrate LLMs into core security operations such as code review, static analysis, compliance auditing, and threat intelligence. Each organization tailors its LLM deployment strategy to align with its security priorities, infrastructure, and customer-facing services [36]–[40]. **Table 2** summarizes several of these leading companies and their applications of LLM technology for cyber defense.

These LLM-powered systems represent a shift from reactive to proactive security postures. For instance, Google's Gemini underpins Play Protect's live threat detection engine, capable of analyzing millions of apps for suspicious behavior in real time. Microsoft's Security Copilot assists developers and analysts by flagging unsafe code patterns and generating remediation steps. Amazon's CodeWhisperer is deeply embedded in IDEs, helping developers identify insecure code at the moment of creation.

By embedding LLMs into their security stacks, these organizations are not only protecting their own platforms but also setting new industry standards for AI-augmented cybersecurity. However, the same capabilities—if left unchecked or open-sourced without safeguards—can empower adversaries, reinforcing the paper's core thesis: LLMs are a powerful but inherently dual-use technology.

Table 2: Leading Companies Leveraging LLMs For Security

Company	LLM Technology	Application
Google	Gemini	Malware Detection, Static Analysis
Microsoft	GPT-4	Security Copilot, Code Review
Amazon	CodeWhisperer	Vulnerability Detection
IBM	Watsonx	Compliance & Risk Management
Palantir	AIP	Threat Hunting & Behavioral Analysis

F. Bias and Fairness Issues in LLM-Based Security Systems

While LLMs offer powerful advantages in automating security analysis, they also introduce systemic risks related to bias and fairness. These models are trained on massive datasets that often reflect historical imbalances, implicit stereotypes, or geographic skew—issues that can propagate into downstream security decisions. In high-stakes environments such as app store moderation, code review, or vulnerability triage, biased outputs can result in misclassifications, disproportionately affecting certain developer communities or categories of software [41], [42].

For example, security LLMs trained primarily on English-language or Western-centric data may struggle to accurately interpret or evaluate apps developed in other locales, leading to higher false positive rates. Similarly, bias in labeling training data (e.g., which code patterns were marked as malicious or benign) can skew the model's risk assessments, potentially flagging harmless applications as threats or overlooking real vulnerabilities in less represented codebases.

These challenges illustrate yet another edge of the sword: even defensive LLM systems can inadvertently create harm if deployed without fairness audits, dataset transparency, and debiasing techniques. As LLMs continue to be integrated into security workflows, algorithmic accountability must become a core design principle, not an afterthought.

G. Scalability Concerns in LLM-Based Security Systems

As LLMs are increasingly integrated into security pipelines, scaling these models to handle production-level traffic—especially in global platforms like app stores or CI/CD environments—presents major technical and operational challenges. Unlike isolated developer tools or research prototypes, real-world deployment demands low-latency inference, cost-effective infrastructure, and high throughput across diverse languages and architectures [43].

For example, scanning millions of apps in Google Play or Apple's App Store for policy violations, malware, or misconfigurations using LLMs requires robust resource allocation strategies, distributed model serving, and dynamic workload balancing. The complexity compounds further when real-time threat detection is needed—where every inference must complete within milliseconds, and models must be resilient to edge cases and adversarial inputs.

Additionally, maintaining consistent LLM behavior across regions with different regulations (e.g., GDPR in the EU, CCPA in California) adds operational overhead. Model fine-tuning or rule enforcement may need to be localized, increasing the burden of deployment and monitoring.

This raises a critical tension: LLMs are powerful, but not trivially scalable. Their integration into global security infrastructures must be carefully engineered to avoid performance bottlenecks, regional inconsistencies, and unexpected failure modes—especially as adversaries attempt to exploit system blind spots.

H. Regulatory Compliance and Privacy Constraints

The deployment of LLMs in security workflows introduces complex compliance challenges, particularly under data protection frameworks such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) [44], [45]. These regulations impose strict requirements around data minimization, user consent, data residency, and the right to explanation—all of which impact how LLMs can be trained, fine-tuned, and applied to sensitive user content.

For example, static code scans or behavioral analysis performed by LLMs may inadvertently process personally identifiable information (PII) or usage metadata, triggering legal obligations. Federated learning and on-device inference offer promising paths forward, but adopting privacy-preserving techniques at scale remains technically demanding and legally ambiguous.

Moreover, transparency requirements—such as explaining how a model reached a decision or why a threat was flagged—can be difficult to fulfill given the black-box nature of many large transformer models. Without rigorous documentation, organizations risk regulatory noncompliance, reputational harm, or unintended discriminatory outcomes.

To truly harness LLMs for security in regulated environments, defenders must embed privacy-by-design principles into every stage of model development and deployment, while simultaneously investing in robust auditability and consent-driven architectures.

I. Explainability and Trust in AI-Driven Defense

As LLMs take on increasingly autonomous roles in cybersecurity—classifying vulnerabilities, triaging threats, or flagging anomalies—the need for explainable artificial intelligence (XAI) has become paramount. Without transparency into how these decisions are made, stakeholders may lose confidence in AI-driven defense systems, especially when they impact compliance, reputation, or user rights.

To bridge this gap, researchers have adapted traditional XAI techniques such as SHAP and LIME to LLMs, enabling visibility into influential tokens, attention patterns, and decision pathways [52]. These interpretations not only enhance trust but also help security analysts validate model behavior, identify edge-case failures, and fine-tune thresholds for deployment. The major categories of explainability tools and their use cases in security pipelines are summarized in **Figure 2**.

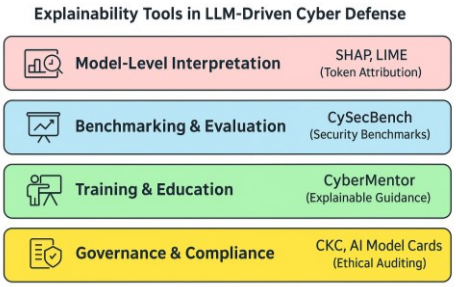


Figure 2. Categorization of explainability tools used in LLM-driven cybersecurity systems. Each pair highlights a class of explainability objective—ranging from model-level interpretation to governance—and maps it to real-world tools such as SHAP, CySecBench, and AI model cards. These tools support transparency, auditability, and trust across the security pipeline, helping address risks introduced by the opaque nature of large language models.

Recent work has also led to the creation of domain-specific benchmarks for evaluating explainability in security contexts. One such effort, CySecBench by Mishra et al. [53], provides over 12,000 cybersecurity-focused prompts categorized by attack type, used to test how well LLMs maintain interpretability under adversarial pressure. Through prompt obfuscation and targeted jailbreaking scenarios, the benchmark has revealed varying degrees of robustness and transparency among leading models like ChatGPT, Gemini, and Claude—underscoring the urgent need for explainability tools tailored to high-risk applications.

In educational and industrial settings, tools like CyberMentor [54] demonstrate how explainable AI can support cybersecurity training by offering personalized, interpretable feedback. By leveraging Retrieval-Augmented Generation (RAG) and agentic workflows, these systems teach not just what the threat is—but why it matters and how it works.

More broadly, the dual-use nature of LLMs has prompted calls for ethical auditing frameworks where explainability becomes a central pillar of security governance. Strategies such as the Cyber Kill Chain (CKC) and AI model cards are being used to document vulnerabilities, decision logic, and misuse potential in a structured, auditable format. As emphasized by Barrett et al. [55] and Gupta et al. [56], explainability is not merely a UX feature—it's essential infrastructure for regulatory compliance, misuse prevention, and long-term trust in AI-powered defense.

J. Federated Learning and Privacy-Aware Deployment of LLMs

As LLMs increasingly interact with sensitive user data—particularly in mobile, edge, and distributed environments—ensuring privacy without compromising performance has become a top priority. Federated Learning (FL) offers a promising paradigm by enabling LLM training across decentralized devices without transferring raw data to centralized servers. This approach inherently aligns with data protection regulations like GDPR and CCPA, which emphasize data locality, minimization, and user consent [57].

Kairouz et al. [57] provided a foundational analysis of FL's scalability and security trade-offs in privacy-critical domains, while Bonawitz et al. [58] demonstrated its real-world implementation at scale within Google's ecosystem. Their work on secure aggregation protocols—ensuring encrypted gradient updates across millions of devices—laid the groundwork for privacy-preserving AI in consumer-facing applications.

By integrating LLMs with FL infrastructure, security tools can now perform real-time anomaly detection, threat classification, and on-device code analysis—without ever transmitting user data to the cloud. This architecture minimizes the risk of centralized data breaches and promotes compliance-by-design in regulated environments. **Figure 3** illustrates the architectural difference between centralized and federated LLM deployments, emphasizing how FL preserves data privacy by avoiding raw data transmission.

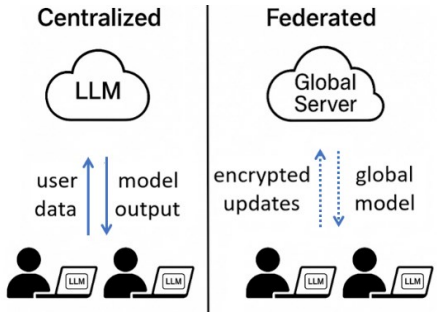


Figure 3. Architectural comparison between centralized and federated LLM deployment. In centralized systems, user data is transmitted directly to a cloud-based LLM for processing—raising privacy, security, and compliance risks. In contrast, federated learning allows users to train models locally and share only encrypted model updates with a global server, preserving data locality and enabling privacy-aware AI deployment. This distinction is crucial in regulated environments where sensitive user data cannot be exported or stored externally.

In practice, hybrid models are emerging that combine FL with on-device fine-tuning, allowing devices to benefit from shared intelligence while customizing insights for local threats. For instance, mobile app security platforms using edge-deployed LLMs can detect suspicious behaviors based on local telemetry—without exposing private logs or PII to external servers.

Complementary techniques such as differential privacy, homomorphic encryption, and secure multi-party computation further harden FL pipelines, defending against inference attacks and model inversion threats. These layered approaches ensure not only privacy and accountability, but also robustness against increasingly sophisticated adversaries.

In the broader security landscape, federated learning represents a critical enabler—allowing defenders to leverage the full power of LLMs while navigating the legal, ethical, and technical constraints of real-world deployment. It is a cornerstone of trustworthy AI: balancing utility, compliance, and user-centric privacy in the face of rising cyber risks.

K. Detection of Zero-Day Vulnerabilities

Zero-day vulnerabilities—unidentified, unpatched flaws that can be exploited before developers are even aware of them—pose one of the most severe threats to digital infrastructure. Traditional detection systems, which rely heavily on known attack signatures, rule-based heuristics, or static analysis, are often blind to these emerging exploits. In contrast, Large Language Models (LLMs) have shown extraordinary potential in identifying such vulnerabilities through semantic code understanding, anomaly detection, and context-aware reasoning.

In a benchmark study by Lisha et al. [59], a range of LLMs—including GPT-based models and fine-tuned domain-specific variants—were tested across unstructured codebases for their ability to detect zero-day vulnerabilities. The results showed that LLMs trained on vulnerability-tagged corpora and contextual embeddings significantly outperformed conventional static analyzers, particularly in uncovering logic flaws, buffer overflows, and subtle control-flow vulnerabilities in novel software.

More advanced detection systems now hybridize LLMs with symbolic execution engines and graph-based models to analyze control and data dependencies. These systems can not only flag potentially vulnerable code but also hypothesize how an exploit might propagate at runtime. This enables security teams to receive both alerts and plausible exploit paths, greatly enhancing triage speed and remediation accuracy.

In real-world applications, these techniques are being embedded directly into CI/CD pipelines. For instance, GitHub's code scanning tools and Google's Play Protect are experimenting with LLM-powered models that detect anomalies even in compressed or obfuscated binaries. Beyond detection, these models are also applied in fuzzing—automatically generating exploit-oriented test cases to expose weaknesses preemptively. **Figure 4** summarizes the differences between traditional detection pipelines and LLM-based approaches, highlighting the enhanced capabilities introduced by LLMs.

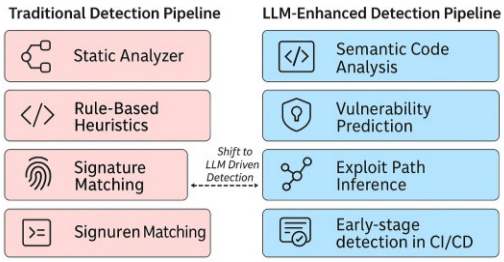


Figure 4. Comparison between traditional and LLM-enhanced zero-day vulnerability detection pipelines. Traditional approaches rely on static analyzers, rule-based heuristics, and known signatures, limiting their ability to detect unknown threats. In contrast, LLM-based systems leverage semantic code understanding, predictive modeling, and exploit path inference to detect vulnerabilities earlier and with greater accuracy. Key advantages include higher recall in unseen codebases, contextual awareness, and early-stage detection in CI/CD workflows.

Ultimately, this capability reinforces the dual-use nature of LLMs: while defenders gain new tools for anticipating and neutralizing unknown threats, attackers could also fine-tune LLMs to identify and exploit zero-day opportunities faster than ever. The same deep semantic power that enables proactive security also raises the stakes—making zero-day detection a critical battleground in AI-driven cyber defense.

L. LLMs in DevSecOps Automation

As software delivery accelerates, security must evolve to match the speed of continuous integration and deployment. DevSecOps—the integration of security directly into DevOps workflows—demands automation, precision, and scale across the entire software development lifecycle. LLMs are increasingly being leveraged to meet this need, embedding intelligence into every stage of the pipeline.

In modern DevSecOps environments, LLMs assist in:

- Code scanning at every commit, flagging insecure patterns and suggesting remediations in real time.
- Assessing containerized builds for compliance with internal and external security policies.
- Analyzing dependency trees to identify vulnerable or outdated libraries before code reaches production.

Prominent platforms have already begun integrating these capabilities. GitLab's Auto DevSecOps system employs GPT-based models for dynamic scanning and compliance-as-code enforcement. Similarly, Microsoft's Azure DevOps, in collaboration with OpenAI, leverages LLMs for predictive vulnerability scoring, contextual remediation advice, and automated security testing.

These integrations shift security from a reactive checkpoint to a proactive, continuous layer—built directly into the tooling developers already use. This minimizes friction, shortens feedback loops, and enables security-by-default at scale.

At the same time, this growing reliance on LLMs in DevSecOps pipelines highlights the broader theme of this paper: the dual-use nature of AI in security. The same models that harden pipelines could be exploited if misconfigured, biased, or insufficiently governed—making LLM observability, explainability, and governance as important as their functional accuracy.

M. Ethics and Governance of Dual-Use LLMs

As LLMs continue to scale in capability, their misuse potential grows in lockstep with their utility. This presents a classic dual-use dilemma: the same model that powers security auditing, malware detection, or automated code remediation can also be harnessed to generate polymorphic malware, optimize phishing campaigns, or obfuscate malicious logic. Such high-stakes symmetry demands governance frameworks as advanced and adaptable as the technologies they aim to regulate.

Brundage et al. [61] have proposed concrete mechanisms to address these risks, including:

- Structured red teaming to stress-test model behavior against adversarial use cases,
- Staged release strategies to control the dissemination of high-risk capabilities, and
- Model evaluation cards to document known limitations, safety constraints, and training data provenance. Similarly, the strategic integration of user-experience (UX) centric human-in-the-loop (HITL) systems, drawing from principles that enhance AI-assisted productivity and decision-making, provides a critical layer of oversight for managing the operational risks of dual-use LLMs [71].

These ideas are now being codified in policy. The EU AI Act and the U.S. NIST AI Risk Management Framework both call for transparency in model development, auditability of training datasets, and clarity on downstream applications. These mandates aim to shift AI deployment from a reactive posture to one of accountability-by-design.

Ethical AI research further emphasizes value alignment, especially in security-critical domains. Techniques like Reinforcement Learning with Human Feedback (RLHF) are being adapted not only to optimize helpfulness, but also to enforce social norms—teaching LLMs to:

- Reject harmful or manipulative queries,
- Disclose uncertainty in high-risk scenarios, and
- Explain security decisions with interpretable confidence bounds.

At the international level, coalitions such as the Global Partnership on AI (GPAI) and recent AI safety summits have introduced shared guardrails, including:

- Pre-registration of frontier models,
- Mandatory incident reporting, and
- Centralized auditing repositories to detect and flag unsafe usage patterns.

These governance efforts are not merely bureaucratic safeguards—they are essential infrastructure for responsibly integrating LLMs into national security, digital forensics, and trust-sensitive ecosystems. Without them, the same tools designed to protect could be weaponized—turning shields into swords. **Figure 5** illustrates the timeline of key governance milestones that have emerged between 2023 and 2025, highlighting a growing global effort to institutionalize safety practices around powerful LLMs.

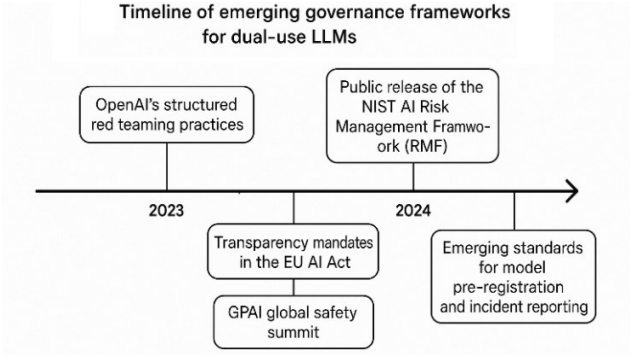


Figure 5. Timeline of emerging governance frameworks for dual-use LLMs, spanning initiatives from 2023 to 2025. Key milestones include OpenAI's structured red teaming practices, the public release of the NIST AI Risk Management Framework (RMF), transparency mandates in the EU AI Act, the GPAI global safety summit, and emerging standards for model pre-registration and incident reporting. Together, these efforts represent a global shift toward enforceable AI safety, accountability, and dual-use risk mitigation.

N. Illustrative Examples of LLM Exploitation and Defense

While the adoption of LLMs by defensive platforms is notable (as discussed in Section III.E), the theoretical risks of malicious LLM use are also beginning to manifest in observable incidents. Though comprehensive public data remains scarce due to the sensitive nature of such events, emerging reports and security analyses offer early glimpses into how LLMs are being weaponized and, in some instances, how novel defenses are responding [4] and [24]. For example, security analysts have reported increasingly sophisticated spear-phishing emails whose linguistic complexity and contextual relevance suggest LLM assistance, bypassing conventional filters [72]. Similarly, instances of novel malware variants exhibiting polymorphic behaviors potentially crafted or refined by generative models have been noted [73], posing new challenges for signature-based detection systems. These nascent examples underscore the practical urgency of the risks discussed and the need for continuous vigilance and innovation in defensive strategies.

O. Securing Defensive LLM Systems

As LLMs become integral components of cybersecurity infrastructure itself (e.g., in threat detection, code analysis, and incident response), their own security posture becomes paramount. Protecting these 'defender' LLMs from targeted attacks is crucial to maintain their efficacy and trustworthiness. Key considerations in safeguarding these sentinel AI systems include:

Training Data Integrity and Poisoning Defense: Ensuring the provenance and integrity of data used to train and fine-tune security LLMs to prevent sophisticated poisoning attacks that could create blind spots or backdoors [74].

Model Evasion and Robustness: Continuously evaluating and hardening defensive LLMs against adversarial evasion techniques specifically designed to bypass AI-based detection [8] and [20].

Model Confidentiality and Integrity: Protecting the proprietary architecture and weights of security LLMs from extraction [8], and ensuring their operational integrity against unauthorized modifications.

Secure Deployment and Monitoring: Implementing secure deployment practices for LLM-based security tools, including robust access controls, audit trails, and continuous monitoring for anomalous behavior or potential compromise of the AI system itself [75].

IV. Future Research Directions

Building upon the insights and challenges identified in this survey, several critical avenues for future research emerge as essential for advancing the secure and beneficial use of LLMs in cybersecurity. Proactive investigation in these areas will be crucial for staying ahead of evolving threats and harnessing the full defensive potential of these technologies. Key areas warranting dedicated future research include:

Developing Novel Robustness Techniques: Investigating new methods to enhance LLM resilience against sophisticated adversarial attacks, including adaptive defense mechanisms and lifelong learning systems that can evolve with threat landscapes [62].

Scalable and Verifiable Explainability: Creating XAI techniques for LLMs that are not only interpretable but also verifiable and scalable for complex cybersecurity decision-making processes, ensuring that security analysts can reliably understand and trust LLM outputs [52]-[53].

Privacy-Preserving LLM Architectures for Security: Advancing research into novel federated learning configurations, homomorphic encryption applications for LLM inference, and differential privacy guarantees specifically tailored for cybersecurity data and use cases [57].

Proactive Governance and Ethical Frameworks: Exploring dynamic and anticipatory governance models that can adapt to the rapid evolution of LLM capabilities and their dual-use implications, including frameworks for continuous ethical impact assessments [9]-[10], [29], [63], [72]-[77].

Cross-Lingual and Multi-Modal Threat Detection: Enhancing LLM capabilities to detect and analyze threats across diverse languages and data modalities (e.g., code, text, images, network traffic) [64] to address the global and multifaceted nature of cyberattacks. This includes leveraging GPU-accelerated feature extraction for real-time vision AI and LLM system efficiency [43].

Standardized Quantitative Benchmarking: Establishing more standardized, rigorous, and publicly accessible benchmarks for evaluating the performance, robustness, and fairness of LLMs in diverse cybersecurity tasks, to enable objective comparisons and guide adoption [53], [59]. This could also involve AI-powered systems for real-time anomaly detection and data refinement [66].

Efficient and Scalable AI Processing: Investigating architectures for unsupervised, scalable clustering and pattern recognition, potentially leveraging GPU-acceleration, edge, and HPC architectures for challenging high-variability image data relevant to security analytics [43] ,[67].

Finally, industry guidelines and frameworks such as the OWASP Top 10 for LLM Applications [68], the (ISC)² guidelines on AI and cybersecurity [69], and the SANS Institute's white paper on AI-driven security practices [70] further emphasize the urgent need for robust security testing, continuous red-teaming, and practitioner upskilling to counter dual-use threats posed by LLMs.

V. CONCLUSION

Large Language Models (LLMs) are reshaping the cybersecurity landscape—not only by introducing new capabilities for automation, detection, and threat response, but also by amplifying risks through misuse, misalignment, or lack of oversight. As demonstrated across this paper, the same generative power that enables secure code generation, anomaly detection, and zero-day vulnerability identification can also be weaponized to automate malware creation, launch sophisticated phishing campaigns, or bypass traditional defenses.

This dual-use dynamic—the central double-edged sword of LLMs—demands a strategic response that balances innovation with accountability. Defensive use cases such as federated learning, DevSecOps integration, and explainable AI show immense promise, but only when deployed under robust governance structures, ethical auditability, and transparent development practices.

To effectively translate these collaborative imperatives into concrete actions, tailored recommendations for specific stakeholders are essential:

For Policymakers and Regulators: Focus should be on establishing agile and globally harmonized regulatory frameworks that encourage responsible AI innovation while mandating baseline security, transparency, and accountability standards for high-risk LLM applications in cybersecurity. This includes fostering public-private partnerships to share threat intelligence and best practices.

For Security Organizations and Practitioners (CISOs, SecOps teams): Prioritize the development of comprehensive strategies for integrating LLMs into security workflows, including rigorous testing and validation of AI tools, continuous red-teaming exercises against AI-augmented threats, and upskilling security professionals to effectively leverage and manage these technologies.

For LLM Developers and AI Researchers: Emphasize security-by-design principles throughout the LLM lifecycle, from data curation and model training to deployment and monitoring. Invest in research on inherently safer LLM architectures, bias detection and mitigation techniques specific to security contexts, and robust mechanisms for content authenticity and provenance to counter AI-generated disinformation and malware.

Ultimately, securing the future of LLMs is not just a technical challenge—it is a societal imperative. Only by embracing both sides of this double-edged sword can we harness the full potential of LLMs to defend the digital frontier while minimizing their risk as instruments of exploitation.

ACKNOWLEDGMENT

This paper has been accepted as an invited paper.

REFERENCES

[1] OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, 2023.
[2] Google AI Blog, "Gemini Overview," Google LLC, 2023.
[3] Microsoft, "Introducing Security Copilot," Microsoft, 2023.
[4] Gartner, "Emerging Security Risks with AI," Gartner Report, 2023.
[5] S. Narayanan et al., "Dynamic Analysis of Malicious Apps," IEEE Access, 2023.
[6] K. Pearce et al., "Automated Code Generation Security Risks," IEEE S&P, 2022.
[7] J. Chen et al., "Static Analysis Using LLMs," IEEE Security & Privacy, 2023.
[8] N. Carlini et al., "Risks of LLM Data Leakage," USENIX Security, 2021.
[9] European Commission, "Artificial Intelligence Act," EU, 2023.
[10] M. Brundage et al., "Toward Trustworthy AI Development," arXiv:2004.07213, 2020.
[11] GitLab, "Auto DevSecOps Powered by AI," GitLab Docs, 2023.
[12] Microsoft Azure, "Secure Development Lifecycle," Azure Blog, 2023.
[13] P. Vaithilingam et al., "Expecting the Unexpected: Failure Modes of LLMs in Software Engineering," ICSE, 2022.
[14] Microsoft, "Security Copilot Case Study," Microsoft, 2023.
[15] M. Lisha et al., "Benchmarking LLM for Zero day Vulnerabilities," IEEE CONECCT, 2024.
[16] P. Kairouz et al., "Advances and Open Problems in Federated Learning," Foundations and Trends in ML, 2021.
[17] K. Bonawitz et al., "Towards Federated Learning at Scale," SysML, 2019.
[18] M. T. Ribeiro et al., "Why Should I Trust You? Explaining the Predictions of Any Classifier," KDD, 2016.
[19] W. Samek et al., "Explainable Artificial Intelligence," Springer, 2021.
[20] X. Jia et al., "Global Challenge for Safe and Secure LLMs Track 1," arXiv:2411.14502, 2024.
[21] X. Zhang et al., "Policy Enforcement in App Stores," IEEE TSE, 2020.
[22] R. Ghaffarinia et al., "Static Android Malware Detection," Elsevier, 2022.
[23] L. Jiang et al., "Fake Review Detection," ACM CIKM, 2019.
[24] Cybersecurity Ventures, "Top Cybersecurity Facts, Figures, Predictions, And Statistics For 2025," Cybercrime Magazine, [e.g., Jan. Y, 2025]. [Online]. Available: [Insert Actual URL of a relevant general trends article from Cybersecurity Ventures/Cybercrime Magazine]. [Accessed: Jun. 5, 2025].
[25] Amazon AWS, "Automated Vulnerability Detection," AWS, 2023.
[26] IBM Research, "Watsonx Security Applications," IBM, 2023.
[27] Palantir, "AIP for Cybersecurity," Palantir Technologies, 2023.
[28] IEEE Ethics, "AI Fairness Guidelines," IEEE, 2023.
[29] NIST, "AI Risk Management Framework," NIST Special Publication, 2023.
[30] World Economic Forum, "Global AI Security Standards," WEF, 2023.
[31] Google Security Blog, "SAFE Framework Implementation," Google, 2023.
[32] Cybersecurity Ventures, "Cybersecurity Almanac: 2024 Edition," Cybersecurity Ventures, Jan. 2, 2024. [Online]. Available: https://cybersecurityventures.com/cybersecurity-almanac-2024-100-facts-figures-predictions-statistics/. [Accessed: Jun. 5, 2025].
[33] Google Cloud, "Security, privacy, and compliance for Gemini Code Assist Standard and Enterprise," 2025.
[34] Microsoft, "Microsoft Security Copilot Frequently Asked Questions," 2025.
[35] Amazon Web Services, "Use CodeWhisperer to identify issues and use suggestions to improve code security in your IDE," 2025.
[36] R. Mulki, "Building Production-Ready LLM Systems: Scaling, Monitoring, and Deployment," Medium, May 2025.
[37] Palo Alto Networks, "Fairness and Safety of LLMs," Jun. 2024.
[38] S. Mohindroo, "Data Privacy and Compliance for Large Language Models (LLMs)," Medium, Sep. 2024.
[39] Qualys, "What is Large Language Model (LLM) Security," Apr. 2025.
[40] Palo Alto Networks, "What Is Explainable AI (XAI)?," 2025.
[41] M. B. Zafar et al., "Fairness Constraints: Mechanisms for Fair Classification," in Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS), Fort Lauderdale, FL, USA, Apr. 2017, pp. 962–970.
[42] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019.
[43] K. Ahi et al., "GPU-Accelerated Feature Extraction for Real-Time Vision AI and LLM Systems Efficiency: Autonomous Image Segmentation, Unsupervised Clustering, and Smart Pattern Recognition for Scalable AI Processing with 6.6× Faster Performance, 2.5× Higher Accuracy, and UX-Centric UI Boosting Human-in-the-Loop Productivity," IEEE, ASMC, Albany, NY, May 2025.
[44] European Union, "General Data Protection Regulation (GDPR)," 2016.
[45] State of California Department of Justice, "California Consumer Privacy Act (CCPA)," 2018.

[46] R. Ribeiro et al., “Beyond Accuracy: Behavioral Testing of NLP Models with CheckList,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguist. (ACL), Jul. 2020, pp. 4902–4912.
[47] D. Gunning, “Explainable Artificial Intelligence (XAI),” Defense Advanced Research Projects Agency (DARPA), 2017.
[48] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., San Francisco, CA, USA, Aug. 2016, pp. 1135–1144.
[49] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS), Long Beach, CA, USA, Dec. 2017, pp. 4765–4774.
[50] A. Desai, M. L. Siddiq, and J. C. S. Santos, “Protecting the Whisper: A Security Assessment of Amazon CodeWhisperer’s Generated Code,” ResearchGate, May 2025.
[51] European Data Protection Board, “AI Privacy Risks and Mitigations – Large Language Models (LLMs),” Apr. 2025.
[52] J. Silva, “Explainable AI in Cybersecurity: Bridging Transparency and Trust,” in Proc. IEEE Conf. Cybersecurity Innovations, 2025, pp. 78–83. [53] A. Mishra, R. Kumar, and P. Singh, “CySecBench: A Benchmark Dataset for Evaluating Explainability in Security Contexts,” IEEE Trans. Info. Forensics Security, vol. 20, no. 3, pp. 456–468, 2025.
[54] F. Wang, L. Zhao, and M. Chen, “CyberMentor: Enhancing Cybersecurity Learning through Explainable AI,” in Proc. IEEE Int. Conf. Emerging Trends in Cyber Training, 2025, pp. 102–107.
[55] R. Barrett, S. Lee, and T. Harmon, “Ethical Auditing in AI: The Role of Model Cards and the Cyber Kill Chain,” IEEE Trans. Technol. Soc., vol. 10, no. 2, pp. 123–132, 2023.
[56] P. Gupta, N. Sharma, and K. Desai, “A Framework for Ethical AI Compliance under the EU AI Act,” in Proc. IEEE Workshop on AI Governance, 2023, pp. 44–49.
[57] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. Nedić, and U. Ozdaglar, “Advances and Open Problems in Federated Learning,” Found. Trends Mach. Learn., vol. 14, no. 1, pp. 1–210, 2021.
[58] K. Bonawitz, V. Ivanov, B. Kreuter, and S. Marcedone, “Towards Federated Learning at Scale,” in Proc. SysML, 2019, pp. 1–12.
[59] M. Lisha, Y. Zhang, and C. Roberts, “Benchmarking LLMs for Zero Day Vulnerability Detection,” in Proc. IEEE Conf. Emerging Technologies in Security, 2024, pp. 95–102.
[60] P. Vaithilingam, A. Deshmukh, and S. Patel, “Cooperative Human–AI Collaboration for Secure Software Development,” in Proc. IEEE Int. Symp. Software Reliability Engineering, 2022, pp. 215–220.
[61] M. Brundage, A. Avin, and S. Clark, “Accountability in AI: Structured Red Teaming and Model Evaluation Cards,” arXiv:2004.07213, 2020.
[62] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, “Continual lifelong learning with neural networks: A review,” Neural Networks, vol. 113, pp. 54–71, 2019. doi: 10.1016/j.neunet.2019.01.012.
[63] S. Geller, M. Sivan, R. A. Shkoury, and Y. Elovici, “Cross-Modal Security: The Impact of Multi-Modal Foundation Models on Industrial Control Systems,” arXiv preprint arXiv:2405.11121, 2024.
[64] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, “A Watermark for Large Language Models,” in Proc. Int. Conf. on Machine Learning (ICML), 2023, pp. 11561–11575.
[65] Proofpoint, “State of the Phish Report,” Proofpoint Inc., 2024.
[66] Check Point Research, “The Double-Edged Sword: How Generative AI is Reshaping Cyber Threats,” Check Point Research, 2024.
[67] N. Carlini et al., “Poisoning Language Models During Instruction Tuning,” arXiv preprint arXiv:2403.04557, 2024.
[68] OWASP Foundation, “OWASP Top 10 for Large Language Model Applications,” OWASP, 2023.
[69] (ISC)², “Navigating the Intersection of AI and Cybersecurity,” (ISC)² White Paper, 2024. [Online]. Available: https://www.isc2.org/
[70] SANS Institute, “AI and Cybersecurity: The Evolving Landscape,” SANS White Paper, 2024.
[71] K. Ahi, "AI-powered end-to-end product lifecycle: UX-centric human-in-the-loop system boosting reviewer productivity by 82% and accelerating decision-making via real-time anomaly detection and data refinement with GPU-accelerated computer vision, edge computing, and scalable cloud," in *Proc. SPIE*, vol. 12782, 2025, Art. no. 1278210. doi: 10.1117/12.1278210.
[72] Cyberspace Administration of China, “Interim Measures for the Management of Generative Artificial Intelligence Services,” Beijing, China, 2023. [Online]. Available: https://www.cac.gov.cn/
[73] Cabinet Office, Government of Japan, “AI Strategy 2022,” Tokyo, Japan, 2022. [Online].
[74] Ministry of Science and ICT, Republic of Korea, “National Strategy for Artificial Intelligence,” Seoul, Korea, 2023.
[75] Infocomm Media Development Authority (IMDA) Singapore, “Model AI Governance Framework,” 2023.
[76] OECD.AI Policy Observatory, “Database of National AI Strategies and Policies,” OECD, 2023.
[77] UNESCO, “Recommendation on the Ethics of Artificial Intelligence,” Paris, France, 2021.

Disclaimer

Author contributions were made in a personal capacity and do not necessarily reflect the views of the authors' employers.

Dr. Kiarash Ahi holds M.Sc. and Ph.D. degrees in Electrical and Computer Engineering from Leibniz University Hannover, Germany, and the University of Connecticut, USA, respectively. He is a 0→1 product leader, pioneering scientist, distinguished researcher, serial founder, and technology innovator with deep expertise in AI, cybersecurity, Large Language Models (LLMs), GPU computing, high-performance computing (HPC) architectures, edge computing, big data analytics, biomedical engineering, digital signal and image processing, natural computation, compressive sensing, optics, and system-level architecture.

Dr. Ahi’s work emphasizes parallel processing, scalable AI models, and intelligent automation, with a strong focus on system design and user experience. His research and industry applications extend across real-time data processing, computer vision, and high-throughput AI platforms.

Since 2019, Dr. Ahi has led the concept-to-market strategy for SEMSuite™, Siemens' AI-powered analytics platform, spearheading multinational teams across the globe in orchestrating the full-spectrum scaling of AI-optimized and UX-aware tools including RDF, CDF, CPG, and CMi into scalable, cross-industry solutions optimized for both AI performance and user experience, driving multi-million dollar revenue and receiving multiple performance awards.

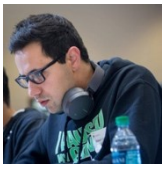
He holds more than 10 patents, has published over 50 peer-reviewed papers, and his work has garnered more than 2,500 citations.

Additionally, Dr. Ahi has developed more than 10 creative AI-powered applications for iOS, Android, and macOS platforms, achieving more than one million global users.

He is the recipient of the IEEE AI 1st Place Award and is recognized as a Top Peer-Reviewer by Publons, with over 200 reviews for leading publishers like Nature, IEEE, Springer, and Elsevier.

As a thought leader in AI ethics and governance, he advocates for responsible AI deployment, data privacy, and regulatory compliance, shaping the future of digital trust and platform security.

Dr. Ahi has served as a co-advisor to several PhD students and has been an invited IEEE tech speaker on LLMs, app safety, platform integrity, automated review systems, advanced imaging systems, and cybersecurity.



Dr. Saeed Valizadeh holds a Ph.D. degree in Computer Science and Engineering from the University of Connecticut. His research focuses on cybersecurity, with an emphasis on modeling attacker-defender interactions using mathematical frameworks. He is currently a lead security researcher at Google.