From Promise to Peril: Rethinking Cybersecurity Red and Blue Teaming in the Age of LLMs

ALSHARIF ABUADBBA, CSIRO's Data61, Australia
CHRIS HICKS, The Alan Turing Institute, United Kingdom
KRISTEN MOORE, CSIRO's Data61, Australia
VASILIOS MAVROUDIS, The Alan Turing Institute, United Kingdom
BURAK HASIRCIOGLU, The Alan Turing Institute, United Kingdom
DIKSHA GOEL, CSIRO's Data61, Australia
PIERS JENNINGS, Loughborough University, United Kingdom

Large Language Models (LLMs) are poised to transform the cybersecurity landscape by augmenting both offensive (red team) and defensive (blue team) operations. With the capacity to automate and enhance tasks such as threat detection, intelligence synthesis, adversary simulation, and incident response, LLMs promise a new era of efficiency and scalability. Red teams can leverage LLMs to plan attacks, craft phishing content, simulate adversarial behaviors, and generate exploit code, while blue teams can deploy them to aggregate threat intelligence, assist with root cause analysis, and streamline security documentation. However, these capabilities come with significant caveats.

This position paper examines the implications of LLMs across key cybersecurity frameworks, including the MITRE ATT&CK and NIST Cybersecurity Framework (CSF), offering a structured analysis of where and how LLMs intersect with critical red and blue team functions. We explore both the strengths and limitations of current LLM capabilities, emphasizing the urgent need for governance, standardization, and real-world evaluation benchmarks.

Our analysis indicates that while LLMs demonstrate fluency, versatility, and utility across cybersecurity tasks, they remain brittle in high-stakes, context-rich environments. Limitations such as constrained context retention, hallucinations, reasoning deficiencies, and prompt sensitivity undermine their reliability in complex operational settings. Furthermore, integrating LLMs into real-world workflows might introduce significant concerns—including dual-use risks, adversarial misuse, and the erosion of human oversight. Malicious actors can potentially exploit these models to rapidly scale cyber operations, automate reconnaissance, and obfuscate attack paths, thereby lowering the technical barrier to launching sophisticated threats. To guide responsible and safer adoption, we outline strategic recommendations that include maintaining human-in-the-loop oversight, improving explainability, employing privacy-aware and secure integration practices, and building systems resilient to adversarial use. As organizations move toward AI-enabled cyber defence, a nuanced understanding of LLMs' risks and operational impacts is essential to ensure these tools enhance, rather than compromise, cybersecurity posture.

CCS Concepts: • **Security and privacy → Security services**.

Additional Key Words and Phrases: Active Cyber Defence

1 INTRODUCTION

Large Language Models (LLMs) are rapidly becoming pivotal tools in cybersecurity. Their ability to interpret and generate code, reason over complex inputs, and operate across natural language interfaces has opened new frontiers for both attackers and defenders. Once restricted to resource-rich organizations, LLMs are now widely accessible, making advanced cyber capabilities available at unprecedented scale. This position paper examines the implications of this shift for red and blue teaming operations, exploring both the opportunities LLMs create and the risks they introduce.

Among the areas where LLMs are already having visible impact is in *red Vs blue teaming*–a practice central to cybersecurity preparedness. In these adversarial exercises, red teams emulate real-world attackers, probing for vulnerabilities by simulating tactics, techniques, and procedures (TTPs) [3, 19, 20]. The goal is not simply to break in, but to rigorously test defenses and provide actionable feedback to improve detection and response capabilities. Blue teams, meanwhile, are responsible for defending infrastructure: monitoring systems, assessing risks, analyzing threats, and implementing hardening and recovery measures. These roles often align with widely adopted frameworks such as the NIST Cybersecurity Framework (CSF) [26] and MITRE ATT&CK [23].

As interest in LLMs accelerates, both red and blue teams have begun integrating them into their workflows. Red teams are experimenting with LLMs to automate tasks such as phishing campaigns, generating exploit code, simulating adversaries, and supporting reconnaissance. On the defensive side, blue teams are exploring LLMs for threat intelligence synthesis, incident documentation, and root cause analysis. These early use cases demonstrate the models' promise, but also their limitations. Most current applications remain narrow, focused on language-based tasks, and fall short in addressing the broader, data-intensive, and action-oriented demands of many cybersecurity operations—such as lateral movement, dynamic response, and real-time adaptation.

In this position paper, we explore the evolving role of LLMs in cybersecurity through the lens of red and blue teaming. We examine their contributions and constraints across practical scenarios, map their alignment with established frameworks such as MITRE ATT&CK and the NIST CSF, and identify challenges and risks that remain underexplored. Our aim is to guide a responsible and strategic integration of LLMs, one that enhances cybersecurity posture without eroding human oversight, operational resilience, or ethical safeguards.

2 MOTIVATION AND THREAT LANDSCAPE

The cybersecurity threat landscape in 2024 reached unprecedented intensity, marked by the accelerating volume, velocity, and sophistication of attacks targeting global infrastructure. A synthesis of legal analysis and data projections from 2024–2025 estimates cybercrime costs at $9.5 trillion in 2024, with forecasts exceeding $10.5 trillion by 2025 [24]. High-profile incidents such as the Ticketmaster breach—impacting 560 million users—illustrate the scale and societal impact of these threats. Additionally, Amazon's Chief Security Officer, CJ Moses, reported (Wall Street Journal, Nov 2024) that Amazon now faces nearly 1 billion cyber threats per day—a dramatic surge attributed in part to the growing use of artificial intelligence in both offensive operations and defensive needs [29].

This rapidly evolving threat environment demands a fundamental rethinking of cybersecurity paradigms. In particular, the convergence of LLMs with red and blue teaming introduces powerful new capabilities—alongside novel risks. As LLMs are increasingly integrated into both offensive tools and defensive frameworks, understanding their dual-use nature becomes imperative. This requires re-examining conventional red teaming methodologies, anticipating emerging attack surfaces, and proactively addressing the challenges, risks, and governance considerations associated with their adoption. A nuanced understanding of these dynamics is critical to building secure and resilient systems in the LLM age.

3 FOUNDATIONS: LLMS AND CYBERSECURITY OPERATIONS

To understand how LLMs may be safely and effectively integrated into cybersecurity workflows, we first provide a brief overview of how LLMs function and how red and blue team operations are structured.

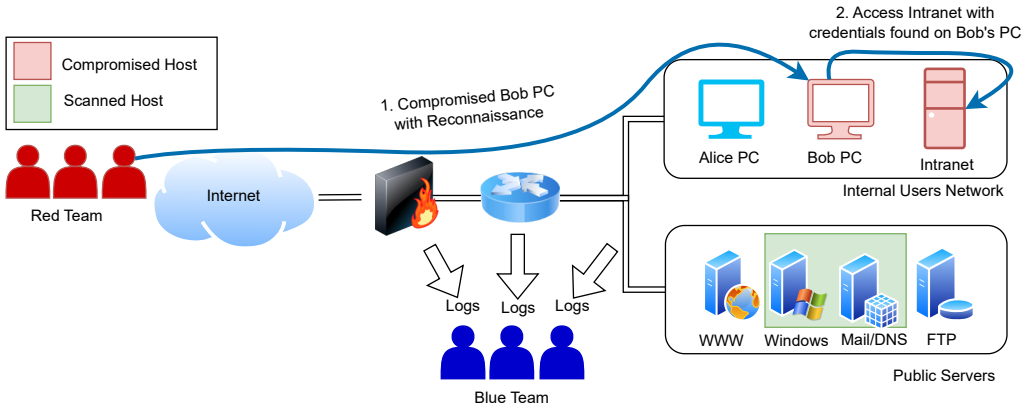

Fig. 1. An illustrative scenario of red team intrusion and corresponding blue team response pathways.

3.1 LLMs from First Principles

LLMs are built on the transformer architecture [33], a foundational breakthrough in deep learning that enables models to process and generate natural language with remarkable fluency. These models are pretrained on vast corpora of text using autoregressive objectives [28], predicting the next word in a sequence, which allows them to capture complex statistical patterns in language, logic, and even code.

To make them useful in practical settings, LLMs are often fine-tuned on domain-specific tasks [9] and aligned with human intent through techniques such as reinforcement learning from human feedback (RLHF) [40]. These post-training strategies improve performance and usability but do not eliminate core limitations. They play a critical role in improving safety, particularly in high-stakes applications like cybersecurity or bioterrorism prevention. They also reflect broader research efforts to manage trade-offs in model capability, safety, and scalability, as highlighted by emerging studies on architectural limits and scaling laws [18].

3.2 Red and Blue Team Key Activities

Figure 1 presents an overview and illustrative example of the roles and activities involved in red and blue teaming. We next outline the common operational activities of red and blue teams.

3.2.1 Common Red Team Activities. We utilise the MITRE ATT&CK framework as a comprehensive reference for adversary tactics, techniques, and procedures derived from real-world observations. This framework has been widely adopted as a foundation for red teaming and threat modeling across various sectors, including government, academia, and industry [2, 23]. As illustrated in Table 1, the framework categorizes adversarial behavior into distinct tactics, organized as rows. In this position paper, we highlight seven representative tactics (out of 14 on MITRE ATT&CK framework) to demonstrate the varying levels of difficulty in automating these tasks using LLMs in Section 4.

3.2.2 Common Blue Team Activities. We adopt the NIST CyberSecurity Framework (CSF) 2.0 [26], a widely adopted approach for cybersecurity risk management, which complements the MITRE ATT&CK framework with a set of security controls used by blue teams to counter threats and mitigate risks. As shown in Table 2, the CSF organises blue team

Table 1. Simplified MITRE ATT&CK Red Teaming Framework with Examples.

Tactic	Techniques	Illustration/Example
Reconnaissance	Active Scanning	Scanning open ports using tools like Nmap
	Phishing for Information	Sending spear-phishing emails with malicious attachments
	Search Open Technical Databases	Crawling Shodan for exposed devices
Privilege Escalation	Abuse Elevation Control Mechanism	Exploiting Windows UAC to gain higher privileges
	Access Token Manipulation	Token impersonation using tools like Mimikatz
	Create or Modify System Process	Modifying a system service to execute malware
Defense Evasion	Abuse Elevation Control Mechanism	Disabling antivirus using administrative privileges
	Debugger Evasion	Detecting and avoiding debugging environments
	Deobfuscate/Decode Files or Info	Decrypting encoded payloads to execute malicious code
Credential Access	Adversary-in-the-Middle	DNS spoofing to intercept credentials
	Brute Force	Repeatedly attempting passwords against a service
	Steal Application Access Token	Extracting OAuth tokens from a compromised device
Lateral Movement	Exploitation of Remote Services	Exploiting RDP or SMB to move laterally
	Lateral Tool Transfer	Moving malicious binaries to a remote host
	Remote Services	Gaining control of remote systems via SSH or RDP
Command and Control	Application Layer Protocol	Using HTTP/S for communication with a C2 server
	Content Injection	Injecting malicious scripts into web applications
	Traffic Signaling	Using DNS tunneling for covert communication
Exfiltration	Automated Exfiltration	Uploading data to a cloud storage service
	Exfiltration Over Alternative Protocol	Using FTP instead of HTTP to exfiltrate data
	Exfiltration Over Web Service	Using Google Drive or Dropbox to exfiltrate files

activities into six core functions: govern, identify, protect, detect, respond, and recover. All of the CSF functions relate to one another, with govern informing the implementation of the other core functions and referred to in Section 4.

4 STRENGTHS, CHALLENGES, AND ETHICAL IMPLICATIONS

4.1 Emergent Strengths of LLMs in Cybersecurity Operations

LLMs are beginning to demonstrate tangible benefits across cybersecurity operations. Here we highlight key areas where LLMs are enhancing the effectiveness and efficiency of red and blue teams, drawing on recent empirical findings and practical use cases.

4.1.1 Enhanced Red Teaming Efficiency. LLMs have enhanced red team operations' in cybersecurity across various TTPs [8]. (1) Planning Support: LLMs can analyze tasks, devise plans, and recommend actions such as executing penetration testing tools or generating executable code to achieve specific objectives [30]. (2) Execution Support: With appropriate guidance, LLMs assist in conducting introductory-level penetration testing, particularly in targeted tasks. For instance, they can perform website hacking techniques such as SQL injection and cross-site scripting [13], execute and evaluate LLM-generated commands for network threat testing [25], and exploit known Common Vulnerabilities and Exposures (CVEs) when provided with instructions [12]. (3) External Tool Integration: LLMs can interact with external tools, such as web search engines, enhancing the breadth and depth of information accessible during testing. (4) Rapid Iteration: LLMs can iteratively refine their actions based on feedback from previous outputs, improving effectiveness over time with human assistance. (5) Automated Reporting: LLMs expedite time-consuming tasks like summarizing findings and documentation, streamlining the reporting process and allowing Red Teams to focus on more complex aspects of penetration testing.

Table 2. NIST CSF Core Functions with Example Risk and Security Controls.

Core Function	Purpose	Example Risk Management Activities and Security Controls
Govern	Establish and monitor risk management strategy, expectations, and policy	Understand legal, regulatory, and contractual requirements Determine expected capabilities, outcomes, and services Standardise methods for calculating and prioritising risks
Identify	Understand current cybersecurity risks	Maintain an accurate model of authorised network communication Identify vulnerabilities and threats to assets, people, and processes Establish, maintain, and communicate incident response plans
Protect	Use security controls to manage risks to assets	Maintain a policy defining access permissions and authorisations Establish and apply configuration management policies Create, protect, maintain, and test data backups
Detect	Find and analyse possible attacks and compromises	Monitor networks and assets for potential adverse events Analyse potential incidents or suspicious behavior Integrate cyber threat intelligence
Respond	React to detected cybersecurity incidents	Execute incident response plans Perform root cause analysis Stakeholder incident notification and information sharing
Recover	Restore assets and operations after an incident	Select, scope, prioritise, and perform recovery actions Verify the integrity of restored assets, systems, and services Communicate progress in restoring capabilities

4.1.2 Enhanced Blue Teaming Efficiency. Blue teams spend a large proportion of time collecting, parsing, and documenting information about security risks [5, 22, 32], but compared to red team activities there is relatively little research documenting the effectiveness of LLMs in blue team operations [35]. Nevertheless, LLMs have the potential to greatly enhance blue team operations across a range of activities [37]. We illustrate this by identifying five key LLM capabilities with the potential to productively augment one or more of the NIST CSF [26] core functions introduced in Section 3. From each of these capabilities, LLMs can: (1) Information Aggregation and Analysis: assist in parsing lengthy legal, regulatory, and contractual frameworks to extract mutual risk management requirements (govern), as well as aggregate cyber threat intelligence from diverse sources during the identification phase (identify). (2) Automated Documentation and Reporting: facilitate the drafting of information-sharing communications and incident notifications for stakeholders in response to security incidents (respond). (3) Policy and Configuration Management: support the development and ongoing maintenance of asset and network configuration policies (protect). (4) Decision support: act as co-pilots alongside blue team members, aiding in the execution of incident response plans (respond). (5) Process Automation: write scripts to automate tasks such as backup creation and validation (protect and recover).

4.2 Key Technical Challenges of LLMs in Cybersecurity Operations

Technical challenges caveating current LLM capabilities include context length limitations, prompt brittleness, hallucinations, behavioral alignment, and unreliable evaluation practices [17, 21, 34]. Here, we contextualize these challenges from the perspective of red and blue team cyber operations.

4.2.1 Short term memory (i.e., context length). Despite advancements in context length, LLMs continue to struggle with efficiently processing and maintaining focus on relevant parts of the input [16]. This limitation poses several

challenges in cybersecurity operations, where contextual awareness is crucial for effective decision making. The most significant issue arising from context limitations is the loss of context over time. During long interactions (i.e., long token sequences) LLMs may fail to recall important information from earlier stages, leading to misaligned outputs or degraded performance. For red teamers, this is especially problematic in multi-stage attack simulations involving tactics such as *Reconnaissance (Phishing)* Table 1, where tracking the progression of an attack through various phases requires consistent memory of prior steps. Similarly, blue teams relying on LLM-based *detection* tools, Table 2, may miss threats that are distributed across large volumes of benign data due to the model's inability to maintain relevant information across long contexts. One notable consequence of this limitation is task repetition. When LLMs cannot remember prior interactions, they may redundantly reanalyze previously completed tasks or suggest actions that have already been taken. This may have directly deleterious effects (e.g., encrypting data twice could have consequences ranging from wasting resources to permanent denial of access) and could also reduce effectiveness, particularly in red teaming scenarios like simulating *Command and Control (Content Injection)*, where an LLM-based system might fail to recognize that certain injection techniques have already been attempted (i.e., increasing the risk of detection). Blue teams might encounter compounding inefficiencies as both the execution and documentation of incident response procedures become prone to redundant steps.

4.2.2 Hallucinations. When a language model generates false or misleading information with high confidence, particularly when it lacks access to verified or current data, it is said to be suffering from a hallucination. This poses significant challenges for red and blue teaming, where the consequences of acting on false information may be particularly severe. First, inaccurate outputs and false positives can mislead security teams—for example, an LLM might fabricate an attack path involving *Tactic (Persistence) and Technique (Hijack Execution Flow)* when no such exploit is feasible, leading to unnecessary mitigation efforts. Second, undetected errors and flawed conclusions can skew threat models and result in poor security recommendations. In the context of *redential Access (Exploitation for Credential Access)*, Table 1, an LLM may suggest fictitious credentials as attack vectors, diverting attention from actual threats. Hallucinations in blue team incident reporting could hinder or subvert downstream recovery, incident response planning, and stakeholder notification activities. Without robust, real-time validation mechanisms, these hallucinations can undermine the effectiveness of cyber operations and decision making.

4.2.3 Reasoning limitations. Despite their fluency, LLMs often lack robust "reasoning", especially in multi-step logic, contextual adaptation, and strategic decision-making, limiting their effectiveness across red and blue team activities. In particular: (1) Challenges in End-to-End Automation: LLMs struggle to coordinate multi-stage attack sequences, affecting coherence across phases like reconnaissance, exploitation, lateral movement, and exfiltration. For instance, during reconnaissance (*Active Scanning*), poor prioritization may lead to inefficient exploitation in Remote Services, disrupting the attack chain (Table 1). Similarly, their inability to adapt tactics undermines adversary emulation in Phishing, where nuanced, context-aware actions are critical. On the defensive side, limited reasoning during root cause analysis may lead to missed causal links, resulting in incorrect conclusions and wasted analyst effort. (2) Inconsistent Adversary Simulation: LLMs often fail to adapt as scenarios evolve, weakening simulation realism. In Resource Development (*Establish Account*), adversaries pivot credentials for persistence, yet LLM-driven red teams may not escalate privileges appropriately (e.g., Abuse Elevation Control Mechanism to Session Hijacking), limiting fidelity. Without strategic adaptation, LLM simulations diverge from real-world threat behavior. While some studies use Catch the Flag (CTF) challenges to evaluate LLMs in red teaming [1, 8, 39], performance drops sharply beyond the easiest levels, mainly due

to "reasoning and knowledge limitations" [14]. These gaps nonetheless present new opportunities for advancing LLMs in automated cyber operations.

4.2.4 Prompt and Tuning Sensitivity with Coverage Gaps. LLMs evidently demonstrate strong natural-language capabilities but remain highly sensitive to prompt phrasing and fine-tuning parameters, which can result in inconsistent behaviour and limited generalization. Small changes in input or model configuration can lead to significant shifts in output, posing reliability risks. This is further compounded by limitations in information coverage during training (e.g., about specific vulnerabilities), which may collectively undermine the robustness and reliability of LLM-driven cybersecurity solutions.

4.2.5 Evaluation Practices & Integration Challenges. Integrating LLMs into operational cybersecurity frameworks such as Security Information and Event Management (SIEM) systems, MITRE ATT&CK, or the NIST CSF presents several challenges. A key issue is the lack of real-world benchmarks and reliability measures. Most security AI evaluations are conducted in controlled environments that fail to capture the complexity of real-world live networks. This makes it difficult to assess how LLMs will perform in production environments and whether they can meaningfully support existing security workflows. Additionally, integration requires compatibility with existing tools, data formats, and protocols, an area still lacking standardization. As a result, organizations risk deploying LLMs without a clear understanding of their limitations or interoperability requirements.

4.3 Risks of LLM-Enhanced Cyber Operations Technology

Here, we detail the contextual risks emerging from the application of LLMs to cyber operations. For a more general overview of generative AI risks we refer the reader to [27].

4.3.1 Dual-Use and the Blurring of Threat Actor Boundaries. Open-source tools lower the barriers to implementing strong cybersecurity measures but might also empower adversaries. Tools originally built for red or blue team operations are increasingly repurposed by threat actors to automate reconnaissance, craft phishing content, and generate evasive code [15]. This reduces both the skill and resource requirements to execute sophisticated attacks, especially when combined with powerful, openly accessible LLMs lacking sufficient safeguards or oversight. Red team frameworks like Cobalt Strike, Metasploit, and PowerSploit, once intended for ethical use, have long been co-opted by state-sponsored and criminal groups. Their open-source nature and resemblance to legitimate traffic complicates detection and attribution. Critically, the integration of LLMs into these toolchains may reshape the threat landscape, enabling low-skilled actors (e.g., script-kiddies) to operate with sophistication once reserved for advanced persistent threats. This convergence blurs traditional distinctions between amateur and state-funded adversaries, challenging existing threat models. CrowdStrike's 2024 Global Threat Report [6] highlights this trend, noting the group *Scattered Spider* likely used an LLM to generate PowerShell scripts in a 2023 attack. Without clear boundaries and responsible practices, ethical deployments risk accelerating the proliferation of offensive cyber capabilities.

4.3.2 Over-reliance on Automation. Dependence on LLMs in security operations risks diminishing human awareness, oversight, and judgment. As LLMs become integrated into automated pipelines, there is a growing tendency to rely on their outputs without sufficient human validation. In high-stakes scenarios this could lead to missed threats, misinterpretations, or overconfidence in flawed assessments—particularly where human security experts are sidelined from decision loops. Running weakly or even entirely non-audited code increases the risks of both unintended consequences

and introducing new vulnerabilities. Analysts' ability to find and fix code vulnerabilities may be hampered by greater unfamiliarity, and there is a risk that the overall quality of code supporting business activities drifts lower over time.

More subtly, prolonged reliance on LLMs may erode the expertise and situational awareness of security analysts. This may ultimately result in a detrimental incapacity among security teams to respond effectively in the event of LLM unavailability, whether due to deliberate service restrictions (e.g., imposed by governmental entities or AI companies) or unrelated operational failures.

4.3.3 Under-reliance on Automation. Conversely, failing to adopt sufficient automation can leave security teams unable to keep pace with the speed, scale, and sophistication of modern threats. As adversaries increasingly leverage AI-powered tools to craft targeted phishing campaigns, generate convincing deepfakes, or execute large-scale probing of infrastructure, manually operated defenses may simply fall short [11]. Security analysts overwhelmed by alert fatigue, repetitive triage, and log analysis may miss high-impact events or respond too slowly. Underutilization of LLMs for automating routine tasks, synthesizing intelligence, or detecting anomalies not only wastes a critical force multiplier but also increases the burden on already stretched human teams. Without strategic integration of AI-driven tools, organizations risk a growing asymmetry in cyber defense, unable to match the speed and variation of LLM-enhanced attackers.

4.3.4 Emerging Technology Risk. A potential risk emerging from an increased pace of technological progress, driven at large by LLM augmentation, is a widening gap between the literature and practice-of cyber security. Both attackers and defenders now benefiting from LLM-enhancement quickly move into uncharted and undocumented territory as they find new ways to apply and integrate emerging AI technologies. This leaves a significant risk of "learning the hard way" without established, or up-to-date, best-practice guidelines available.

4.3.5 Privacy risks. To access the most effective LLMs users typically need to rely on API access to proprietary models. Reliance on proprietary models raises significant privacy concerns as it necessitates sharing the sensitive information for operating LLM-based agents in red and blue team scenarios with the model owners. Even if AI companies provide privacy-preserving solutions for their customers, security vulnerabilities in the inference pipeline could still expose critical information to third parties. Moreover, many organizations have policies that restrict sensitive data from leaving their network. Thus, even if proprietary models deliver substantial performance improvements for red and blue team operations, privacy risks may prevent customers from using them for critical cybersecurity tasks. This situation could leave organizations at a disadvantage by not fully leveraging the potential of LLMs in their cyber operations. Conversely, the growing capabilities of LLMs might tempt individual employees to use proprietary models despite potential non-compliance with their organizations data protection policies.

4.3.6 Agentic LLMs. Agentic LLMs are characterized by tool use and multi-agent interactions, enabling standard LLMs to achieve more complex goals with minimal human intervention [10, 31, 36]. In this process, many of the risks posed by standard LLMs are both exacerbated and made more challenging to estimate. The risks of over and under-reliance on LLMs both rise as agentic workflows enable tackling more complex tasks (e.g., end-to-end cyber kill chains) with less human intervention, in turn elevating the pace of threat escalation and magnifying the risks of solutions unable to move at machine speed. Tool-equipped LLMs with API access increase the worst-case risks of LLMs and heighten the privacy risks as control over which information is shared is handed off to the agent. Finally, agentic LLM systems have greater attack surfaces and pose correspondingly enlarged risks. Agentic platforms have already been found to contain exploitable vulnerabilities from which credentials could be leaked by an attacker [7].

5 RECOMMENDATIONS, AND FUTURE DIRECTIONS FOR LLM ADOPTION AND READINESS IN CYBERSECURITY

As LLMs become embedded in cybersecurity workflows, both red and blue teams face a pivotal challenge: *how to harness these tools' capabilities without introducing unacceptable risks.* The integration of LLMs must be approached as a series of carefully judged trade-offs between e.g., speed and accuracy, autonomy and oversight, and innovation and safety. This section outlines strategic principles for responsible adoption, drawing on the technical challenges and operational risks presented earlier.

5.1 Mitigate Adversarial Dual-Use of LLM Technology

Highlighting the need for strong public-private coordination to detect and counter emerging threats, CrowdStrike's 2024 report [6] shows adversaries already adopting LLMs, including commercial variants, to accelerate their attacks. Open-source tools and permissive APIs further lower the technical barriers (e.g., accelerate reconnaissance and exploit development) for advanced and low-skilled actors alike. To mitigate these risks, we recommend: (1) *Develop usage constraints and API-level access control*, especially for security-critical LLM applications. (2) *Establish dual-use governance frameworks* to track and assess AI-enhanced offensive capabilities in cybersecurity, especially within open-source ecosystems. (3) *Foster cross-sector collaboration* among governments, industry, and academia to establish norms and standards, and to enable timely detection and response to emerging threats, including the misuse of red team frameworks and LLM-assisted tooling. Insights from OpenAI's research on LLM safety, Google DeepMind's robust AI systems, and initiatives by the various AI Safety and Security Institutes [4] further indicate the value of a collaborative effort. (4) *Support early warning systems* and shared intelligence platforms informed by LLM misuse trends e.g., the reported use of LLMs by threat actor group Scattered Spider in CrowdStrike's global report.

5.2 Balance Automation with Human Oversight

Both over- and under-reliance on automation can have adverse medium-to-long-term consequences and must be strategically balanced. While integrating LLMs to help in more tasks across incident response, threat modelling and reporting, it is critical to retain some level of human involvement and understanding to ensure safe deployment. To ensure transparency and accountability: (1) *Implement human and LLMs collaboration mechanisms* especially for high-impact, ambiguous or irreversible actions. (2) *Design adaptive automation thresholds*, escalating to human intervention when model confidence is low or contextual anomalies emerge. (3) *Log decisions and generate explanations* for actions taken to support forensic analysis and compliance. (4) *Design interactive, real-time interpretability tools* tailored to the needs of blue team operators. These principles align with NIST's AI Risk Management Framework and the broader effort towards safe and trustworthy AI.

5.3 Ensure Privacy-Conscious and Secure Deployment

Integrating LLMs into security operations demands a privacy-conscious and security-first approach. Many deployments rely on proprietary APIs, raising risks around data exposure, compliance, and vendor lock-in. To mitigate these risks: (1) *Sandbox and isolate deployments:* Run LLM agents in contained environments to prevent unauthorised access or unintended interactions. (2) *Apply least privilege principle:* Restrict model permissions to only what is necessary for the task. (3) *Automate monitoring and logging:* Continuously observe model outputs and tool interactions to detect anomalies, misuse, or drift over time. (4) *Avoid public APIs where sensitive threat data or logs are involved* and advocate

for on-premise or private inference solutions when feasible. These measures align with the Zero Trust Architecture principles advocated by NIST and help reduce both accidental leakage and adversarial abuse.

5.4 Strengthen Guardrails for Agentic LLMs

Agentic LLMs with tool-use capabilities present significant benefits but also entail compound risks due to autonomy, unpredictable misbehaviour, and extended access to systems. To mitigate these risks: (1) *Design tool access brokers* that gate and log each external API or system call made by an agentic LLM for auditibilty when needed. (2) *Apply limiting policies* for task chaining, especially in scenarios where agents attempt multi-step operations without human review. (3) *Simulate fail-closed scenarios* where agentic systems lose access to certain tools or contexts to evaluate response robustness and investigate their unpredicted behaviour. (4) *Require red-teaming and adversarial testing* of agentic LLMs prior to deployment, focusing on unintended tool use, privilege escalation, and emergent behaviors under real-world constraints.

5.5 Build Real-World Benchmarks to Advance Safe LLM Use and Risk Awareness

The real-world impact of LLM-driven cybersecurity research is often constrained by the lack of realistic benchmarks. Existing methods such as CTF challenges and synthetic datasets [38, 39] fail to capture the complexity and operational nuance of real-world attack and defense scenarios. To responsibly integrate LLMs into red and blue team workflows we should also develop evaluation practices that both validate practical effectiveness, and expose limitations and risk thresholds. This can ensure these tools are not only useful but also deployed with awareness of their boundaries, enabling timely mitigation when needed. To address this gap, we recommend: (1) *Domain-Relevant Benchmarking:* Develop evaluations grounded in real or high-fidelity environments that reflect realistic threat models and operational demands. (2) *Community-Supported Testing Grounds:* Establish open benchmarking frameworks similar to MITRE ATT&CK or MISP, jointly developed by cybersecurity and AI communities. (3) *Transparent Reporting and Reproducibility:* Encourage the release of evaluation results with code, prompts, and configurations to enable replication, validation, and comparison across settings. (4) *Practical, impact-driven metrics:* Draw on lessons from initiatives like Microsoft's AI Safety Research and the AI Incident Database (Partnership on AI), to develop metrics focused on operational performance, not just academic proof-of-concept.

5.6 Design for Real-Time Adaptability and Agility

The threat landscape evolves rapidly and so must the tools used to defend against it. Effective use of LLMs in cyber operations requires agility, which involves: (1) *Live model updating:* Regularly feed the latest threat intelligence into LLMs to maintain relevance. (2) *Incorporate continuous learning pipelines* that allow for safe, rapid model updates.(3) *Enable contextual adaptation* tailoring responses to current attack vectors and threat landscapes. (4) *Adopt MLOps-inspired monitoring frameworks* for operational oversight and rollback when needed.

5.7 Promote Responsible Access While Avoiding Asymmetric Advantage

A healthy cyber ecosystem requires enabling defenders while preventing malicious actors from exploiting LLMs. To that end, we recommend. (1) *Encourage responsible access tiers:* broad access for safe use cases (e.g., threat detection) vs. stricter review for dual-use or offensive capabilities. (2) *Incentivize LLM-for-good initiatives:* (e.g., Blue Team Olympiads, AI4Cyber Defence Grand Challenges) that build capacity in under-resourced teams. (3) *Disincentivise irresponsible model hosting and deployment:* i.e., those failing to implement misuse prevention, especially in high-risk domains.

(4) *Promote open, secure access pathways for public-good applications:* for example providing vetted LLM access for nonprofit, educational, and public-sector cybersecurity efforts, ensuring these groups are not disadvantaged by limited resources while still upholding strong safeguards against misuse.

6 REMARKS

LLMs are reshaping the cybersecurity battlefield, offering new opportunities while blurring traditional lines between defense and offense. This paper outlines the evolving role of LLMs in red and blue teaming, alongside the technical challenges and risks they introduce. As adoption accelerates, our ability to govern, benchmark, and secure these tools will determine whether they become assets for resilience or vectors for risk. We advocate for responsible and secure deployment rooted in human oversight, open collaboration, and continuous adaptation—a vision that must guide cybersecurity innovation in the age of LLMs.

REFERENCES

[1] Talor Abramovich, Meet Udeshi, Minghao Shao, Kilian Lieret, Haoran Xi, Kimberly Milner, Sofija Jancheska, John Yang, Carlos E Jimenez, Farshad Khorrami, et al. 2024. EnIGMA: Enhanced Interactive Generative Model Agent for CTF Challenges. *arXiv preprint arXiv:2409.16165* (2024).

[2] Bader Al-Sada, Alireza Sadighian, and Gabriele Oligeri. 2024. MITRE ATT&CK: State of the art and way forward. *Comput. Surveys* 57, 1 (2024), 1–37.

[3] Evan Anderson. 2023. Red teaming 101: What is red teaming? *IBM* (2023). https://www.ibm.com/think/topics/red-teaming Accessed: 2024-09-26.

[4] Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, et al. 2025. International AI Safety Report. *arXiv preprint arXiv:2501.17805* (2025).

[5] Sarah Brown, Joep Gommers, and Oscar Serrano. 2015. From Cyber Security Information Sharing to Threat Management. In *Proceedings of the 2nd ACM Workshop on Information Sharing and Collaborative Security (WISCS '15)*. Association for Computing Machinery, 43–49. https://doi.org/10.1145/2808128.2808133

[6] CrowdStrike. 2024. 2024 Global Threat Report - Hiding in Plain Sight. (2024).

[7] CVE-2025-31491 2025. CVE-2025-314910. Available from MITRE.. https://www.cve.org/CVERecord?id=CVE-2025-31491

[8] Gelei Deng, Yi Liu, Víctor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, and Stefan Rass. 2024. {PentestGPT}: Evaluating and Harnessing Large Language Models for Automated Penetration Testing. In *33rd USENIX Security Symposium (USENIX Security 24)*. 847–864.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.

[10] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. 2024. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568* (2024).

[11] enisa. 2022. Enisa Threat Landscape 2022. *Europien Union Agency For Cybersecurity* (2022). https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022 Accessed: 2025-05-28.

[12] Richard Fang, Rohan Bindu, Akul Gupta, and Daniel Kang. 2024. Llm agents can autonomously exploit one-day vulnerabilities. *arXiv preprint arXiv:2404.08144* (2024).

[13] Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, and Daniel Kang. 2024. Llm agents can autonomously hack websites. *arXiv preprint arXiv:2402.06664* (2024).

[14] Andreas Happe and Jürgen Cito. 2025. Benchmarking Practices in LLM-driven Offensive Security: Testbeds, Metrics, and Experiment Design. *arXiv preprint arXiv:2504.10112* (2025).

[15] Stephen Hilt and Aliakbar Zahravi. [n. d.]. Red Team Tools in the Hands of Cybercriminals and Nation States. *Trend Research* ([n. d.]). https://documents.trendmicro.com/images/TEx/articles/Research_Paper-Red-Team-Tools.pdf Accessed: 2025-06-01.

[16] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models?. In *First Conference on Language Modeling*.

[17] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and Applications of Large Language Models. arXiv:2307.10169 [cs.CL] https://arxiv.org/abs/2307.10169

[18] Jared Kaplan, Sam McCandlish, Tom Henighan, et al. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[19] Michael Kouremetis, Marissa Dotter, Alex Byrne, Dan Martin, Ethan Michalak, Gianpaolo Russo, Michael Threet, and Guido Zarrella. 2025. OCCULT: Evaluating Large Language Models for Offensive Cyber Operation Capabilities. *arXiv preprint arXiv:2502.15797* (2025).

[20] Ivan Kovačević and Stjepan Groš. 2020. Red Teams-Pentesters, APTs, or Neither. In *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*. IEEE, 1242–1249.

[21] Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.764

[22] Valérie Lavigne and Denis Gouin. 2014. Visual Analytics for cyber security and intelligence. *The Journal of Defense Modeling and Simulation* 11, 2 (2014), 175–199. https://doi.org/10.1177/1548512912464532

[23] MITRE. 2024. ATT&CK Matrix for Enterprise. *MITRE* (2024). https://attack.mitre.org/ Accessed: 2024-12-18.

[24] Dr Kamshad Mohsin. 2025. Cybercrime and Privacy in the Digital Age: Legal Frameworks, Emerging Challenges, and Future Trends. *Emerging Challenges, and Future Trends (March 07, 2025)* (2025).

[25] Stephen Moskal, Sam Laney, Erik Hemberg, and Una-May O'Reilly. 2023. LLMs killed the script kiddie: How agents supported by large language models change the landscape of network threat testing. arXiv 2023. *arXiv preprint arXiv:2310.06936* (2023).

[26] NIST. 2024. The NIST Cybersecurity Framework (CSF) 2.0. *NIST* (2024). https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf NIST CSWP 29, Accessed: 2025-03-10.

[27] NIST. 2024. NIST Trustworthy and Responsible Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. *NIST* (2024).

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. OpenAI Blog, available at https://openai.com/blog/better-language-models/.

[29] James Rundle. November 21, 2024. The AI Effect: Amazon Sees Nearly 1 Billion Cyber Threats a Day. Wall STreet Journal, available at https://www.wsj.com/articles/the-ai-effect-amazon-sees-nearly-1-billion-cyber-threats-a-day-15434edd.

[30] Kumar Shashwat, Francis Hahn, Xinming Ou, Dmitry Goldgof, Lawrence Hall, Jay Ligatti, S Raj Rajgopalan, and Armin Ziaie Tabari. 2024. A Preliminary Study on Using Large Language Models in Software Pentesting. *arXiv preprint arXiv:2401.17459* (2024).

[31] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2023), 8634–8652.

[32] Faheem Ullah and Muhammad Ali Babar. 2019. Architectural Tactics for Big Data Cybersecurity Analytics Systems: A Review. *Journal of Systems and Software* 151 (2019), 81–118. https://doi.org/10.1016/j.jss.2019.01.051

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[34] Nan Wang, Kane Walter, Yansong Gao, and Alsharif Abuadbba. 2025. Large Language Model Adversarial Landscape Through the Lens of Attack Objectives. *arXiv preprint arXiv:2502.02960* (2025).

[35] Tongze Wang, Xiaohui Xie, Lei Zhang, Chuyi Wang, Liang Zhang, and Yong Cui. 2024. ShieldGPT: An LLM-based framework for DDoS mitigation. In *Proceedings of the 8th Asia-Pacific Workshop on Networking*. 108–114.

[36] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang (Eric) Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Ahmed Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In *COLM 2024*.

[37] Tingmin Wu, Shuiqiao Yang, Shigang Liu, David Nguyen, Seung Jang, and Alsharif Abuadbba. 2024. Threatmodeling-llm: Automating threat modeling using large language models for banking system. *arXiv preprint arXiv:2411.17058* (2024).

[38] John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. 2023. Intercode: Standardizing and benchmarking interactive coding with execution feedback. *Advances in Neural Information Processing Systems* 36 (2023), 23826–23854.

[39] Andy K Zhang, Neil Perry, Riya Dulepet, Joey Ji, Celeste Menders, Justin W Lin, Eliot Jones, Gashon Hussein, Samantha Liu, Donovan Jasper, et al. 2024. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. *arXiv preprint arXiv:2408.08926* (2024).

[40] Daniel M. Ziegler, Nisan Stiennon, et al. 2019. Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593* (2019).