

EVALUATING THE OFFENSIVE UTILITY OF LARGE LANGUAGE MODELS IN CYBER THREAT GENERATION

Mukhammadbek Bakhtiyorov

Student of Academic Lyceum under Bukhara State Technical University

Jamolbek Mattiev

Computer Science and AI Technologies department, Urgench State University named after Abu Rayhan Biruni

Abstract:

Large Language Models (LLMs) are increasingly integrated into cybersecurity workflows, creating both defensive opportunities and new forms of offensive enablement. Existing research demonstrates capabilities in threat modeling, attack-path construction, Cyber Threat Intelligence (CTI) processing, adversarial text generation, and automated penetration testing; however, these capabilities are commonly evaluated in isolation. This review synthesizes recent evidence and develops a methodological framework for evaluating offensive utility as a multidimensional construct encompassing technical capability, correctness, contextual grounding, reliability, operational relevance, temporal efficiency, and human-expertise compression. The review distinguishes linguistic plausibility from cyber validity and separates model capability from system-level utility. Evidence from recent studies indicates that LLMs can support attack-technique mapping and multi-stage reasoning, while retrieval-augmented architectures can improve contextualization. At the same time, hallucination, inconsistent reasoning, latency, safety restrictions, and retrieval dependence constrain practical utility. The proposed framework therefore treats offensive utility as an emergent property of the interaction among model, context, retrieval, tools, and human oversight rather than as a single model attribute. The framework provides a basis for reproducible comparative evaluation without requiring unauthorized real-world offensive activity.

Keywords: Large Language Models; cyber threat generation; offensive cybersecurity; Cyber Threat Intelligence; threat modeling; attack-path reasoning; retrieval-augmented generation; reliability; operational utility; cybersecurity automation.

Introduction

The rapid diffusion of generative artificial intelligence has altered the relationship between natural-language processing and cybersecurity. LLMs can summarize heterogeneous technical information, map

natural-language descriptions to cybersecurity concepts, generate structured explanations, and support multi-step reasoning. A substantial body of recent literature now examines these capabilities across threat detection, CTI, vulnerability assessment, incident response, and security evaluation [1].

The same properties that make LLMs useful to defenders can also reduce the cognitive cost of offensive preparation. This does not imply that every LLM is an autonomous attacker. Rather, offensive utility may arise when a model helps transform fragmented information into threat characterizations, technique mappings, attack-path hypotheses, or deceptive intelligence. For instance, recent work on penetration testing shows that LLM-based systems are capable of assisting with tool use, output interpretation, and subsequent-action planning, while still struggling with integrated long-horizon reasoning.

It is noteworthy to mention that threat-modeling research provides a complementary indication. Salek et al. claim that relevant attack techniques with high expert validation are identified by their TraCR-TMF framework which reproduced elements of a real-world attack path [2]. It is also demonstrated by the attack-graph research that LLMs can help convert unstructured CTI into structured representations of attack progression. These findings are very vital since attack-path reasoning is closer to operational planning than isolated cybersecurity question answering.

A second-dimension concerns information manipulation. Shafee et al. illustrate that cybersecurity text, generated on the basis of LLM, can be used to attack components of text-based CTI pipelines. Another research on synthetic CTI shows that generated reports are, sometimes, sufficiently plausible to challenge both human and automated detection mechanisms. Thus, offensive utility is not limited to exploitation or code generation; it can also operate through manipulation of the information environment on which security decisions depend [3].

The central problem addressed by this review is therefore definitional and methodological: how should researchers measure whether an LLM materially contributes to cyber threat generation? Existing evaluations often report task accuracy, benchmark scores, or successful completions, but these measures do not necessarily capture grounding, stability, human intervention, latency, or downstream usefulness. The present review addresses this gap by synthesizing recent evidence and proposing a multidimensional framework for evaluating offensive utility [4].

Review Scope and Research Gaps

The review focuses on peer-reviewed journal and conference literature, together with clearly identified preprints where a recent study is directly relevant and no verified archival publication was located. Priority is given to 2024–2026 work because LLM capabilities, safety behavior, and agentic architectures evolve rapidly [5]. Earlier foundational studies are retained where they provide benchmark, model, or methodological context.

The evidence base covers five intersecting domains: (i) general LLM cybersecurity capability; (ii) offensive security and penetration testing; (iii) threat modeling and attack-path reasoning; (iv) CTI generation, extraction, credibility, and adversarial manipulation; and (v) evaluation, grounding, reliability, and safety–utility trade-offs [6]. This scope is intentionally broader than code generation because cyber threat generation can occur at informational, analytical, planning, and orchestration levels.

Three gaps are particularly evident. First, the literature remains fragmented across threat modeling, penetration testing, CTI, intrusion analysis, and safety evaluation. Second, operational variables such as latency, correction burden, and stability are less consistently reported than task accuracy [7]. Third, there is limited agreement on what constitutes offensive utility at the system level. Both the breadth of the field and the methodological heterogeneity of existing studies are confirmed by the recent systematic reviews.

A further gap concerns longitudinal evaluation. Model capabilities, safety policies, retrieval systems, and agent frameworks change rapidly. Static benchmark results may therefore become outdated. Future evaluations should preserve model versions, prompts, retrieved evidence, tool configurations, and evaluation scripts to support reproducibility [8].

Proposed Methodological Framework

The proposed framework evaluates the offensive utility of LLM-enabled cybersecurity systems through multiple dimensions, including technical capability, correctness, contextual grounding, attack-path coherence, reliability, operational relevance, human-effort reduction, temporal efficiency, and safety. It assesses LLM performance across threat characterization, technique mapping, attack-path reasoning, CTI generation, and controlled workflow assistance using automated metrics and expert evaluation. Thus, offensive utility is considered a system-level property resulting from the interaction of the LLM, contextual evidence, retrieval mechanisms, tools, and human oversight. The Proposed Methodological Framework is illustrated in Figure 1.

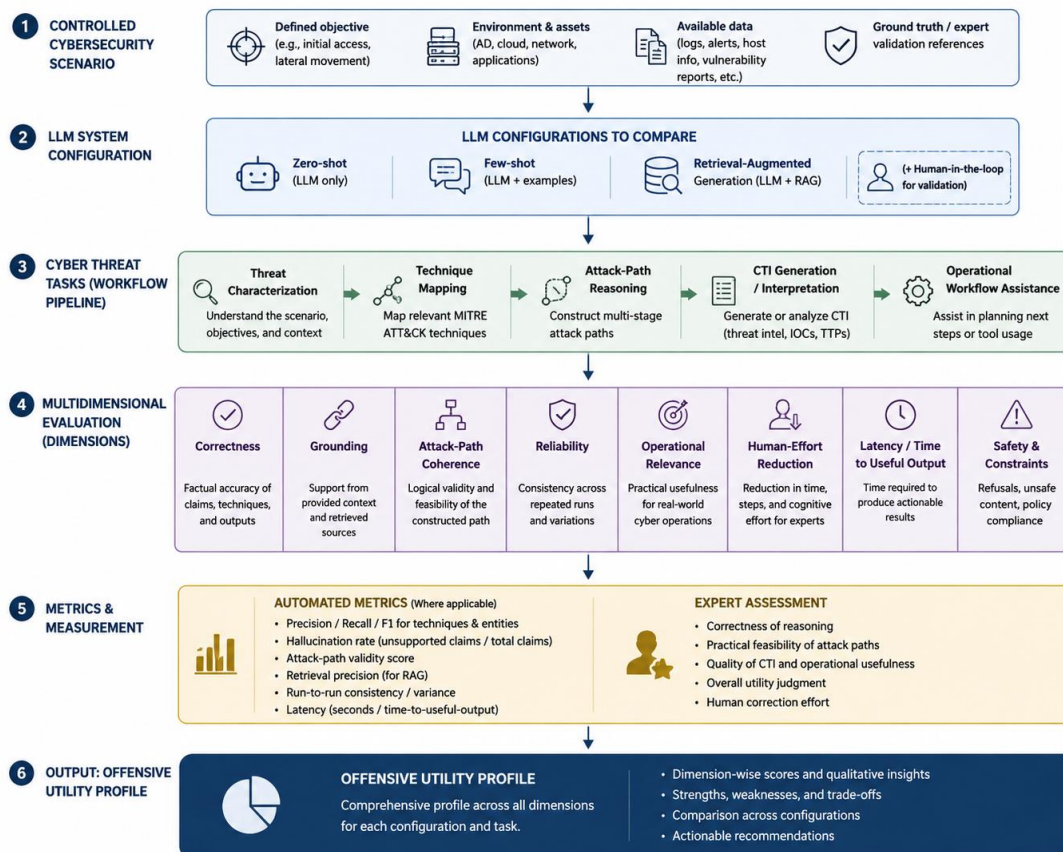


Figure 1. Proposed Methodological Framework

Review Design

This study adopts an evidence-based narrative review with structured comparative synthesis. The purpose is not to claim a new empirical experiment, but to integrate heterogeneous empirical findings into a common analytical framework. The unit of analysis is therefore the reported capability or limitation of an LLM-enabled cybersecurity system, rather than an individual model response.

The synthesis proceeds through four analytical stages: identification of relevant research domains; extraction of comparable constructs and reported metrics; cross-study comparison of capability and limitations; and development of a multidimensional offensive-utility framework. As the existing literature uses heterogeneous datasets, models, task definitions, and evaluation protocols, making direct meta-analysis inappropriate without substantial harmonization, this design is appropriate.

Analytical Coding Dimensions

Each selected study is interpreted according to seven dimensions: technical capability, correctness, contextual grounding, attack-path coherence, reliability, operational relevance, and human-effort reduction. Additional contextual variables include retrieval augmentation, tool access, task complexity,

safety constraints, and temporal cost. This coding structure allows studies with different primary objectives to be compared without treating their numerical metrics as directly interchangeable.

Dimension	Interpretive question	Representative evidence
Capability	What cyber task can the system perform?	Threat identification, TTP mapping, penetration testing
Correctness	Is the output technically valid?	Expert validation, F1, exact match
Grounding	Is the output supported by supplied evidence?	RAG, evidence retrieval, groundedness
Coherence	Can the system connect multiple stages?	Attack paths, attack graphs, long-horizon planning
Reliability	Does performance remain stable?	Repeated runs, hallucination, variance
Operational relevance	Does output materially assist a workflow?	Time, task completion, analyst workload
Safety–utility	How do safeguards affect useful completion?	Refusal rate, unsafe completion, FRR

Proposed Offensive Utility Construct

Offensive utility is defined as the extent to which an LLM-enabled system reduces the cognitive, informational, or coordination cost of generating or refining a cyber threat representation. The construct deliberately excludes unauthorized execution as a necessary condition. A system may possess offensive utility if it improves reconnaissance interpretation, technique selection, attack-path reasoning, CTI manipulation, or planning support within a controlled research environment.

The framework distinguishes three related but non-equivalent properties. LLM capability concerns what the model can generate or infer. Reliability concerns whether those outputs remain accurate and stable under repeated or altered conditions. Offensive utility concerns whether validated outputs produce a meaningful reduction in the effort required to complete a cyber threat-generation task. This distinction prevents fluent but unreliable generation from being interpreted as operational effectiveness .

Evidence-based Evaluation Dimensions

Utilising task-specific precision, recall, F1, exact-match, or benchmark success rates, technical capability can be assessed. SECURE, for example, was developed specifically to evaluate cybersecurity knowledge extraction and understanding in realistic industrial-control-system scenarios. CyberBench can also provide a multi-task cybersecurity benchmark, including several NLP-oriented security tasks.

Attack-path coherence ought to be evaluated separately from individual technique accuracy. On the other hand, AttackKG+ demonstrates the relevance of LLMs to attack-graph construction from CTI, while TraCR-TMF is tend to link vulnerability information to attack techniques and paths. A coherent path requires technically compatible transitions and contextual consistency rather than a high number of techniques.

Grounding is advised to be measured since domain plausibility is not equivalent to factual support. RAG-based CTI architectures express how external retrieval can provide dynamic context, while credibility-verification research highlights the criticality of multi-step evidence retrieval and reasoning.

Reliability consists of hallucination, contradiction, and run-to-run variability. Recent benchmarking work emphasizes that cybersecurity-specific assessment should account for truthfulness and hallucination rather than relying on general language benchmarks.

Operational relevance ought to include time-to-useful-output, human correction burden, and task completion. PentestGPT shows both the potential and the limitation: LLMs can assist with individual

penetration-testing subtasks, yet maintaining an integrated understanding of the complete scenario keeps being difficult.

Safety must be treated as a separate analytical axis. CYBERSECEVAL and CYBERSECEVAL 2 explicitly are able to examine the cybersecurity risks of LLMs and document a safety–utility trade-off in which stronger refusal behavior can also reject benign security requests. CYBERSECEVAL 3 further expands evaluation toward automated social engineering and scaling or autonomy of offensive cyber operations.

Result

The reviewed evidence indicates that LLMs can support cybersecurity activities across threat modeling, attack-path reasoning, penetration testing, CTI generation, and retrieval-supported analysis [9]. However, their practical utility is limited by hallucination, inconsistent reasoning, retrieval errors, latency, safety restrictions, and difficulties with long-horizon tasks. Overall, the evidence suggests that LLMs provide stronger value in cognitive and planning assistance than in fully autonomous offensive operations [10].

Threat Modeling and Attack-Path Reasoning

Threat modeling studies provide some of the strongest evidence that LLMs can perform reasoning relevant to offensive planning without demonstrating autonomous exploitation. TraCR-TMF integrates retrieval, in-context learning, and fine-tuning with cybersecurity knowledge sources to identify attack techniques and potential attack paths. Reported expert validation and real-incident case analysis demonstrate that LLM-supported systems is capable of connecting vulnerabilities, techniques, and multi-stage consequences [11].

Attack-graph research reinforces this observation. AttackKG+ targets the transformation of CTI narratives into structured attack graphs, by addressing the generalization and expertise requirements of earlier approaches. Collectively, these studies suggest that offensive utility can emerge at the planning layer even when direct exploitation is absent [12].

Offensive Security and Penetration Testing

Penetration-testing research provides a more direct view of operational assistance. Deng et al. report that PentestGPT improved task completion over a baseline while identifying difficulties in maintaining global context. Happe and Cito similarly characterize LLM-assisted penetration testing as increasingly capable but constrained by reliability, stability, cost, safety, and long-horizon planning.

These findings support a layered interpretation of offensive utility. LLMs may be highly useful for local reasoning while remaining less reliable for sustained autonomous operation. Consequently, benchmark success on individual tasks should not be interpreted as equivalent to autonomous offensive capability [13].

CTI Generation and Information Manipulation

The CTI literature expands offensive utility into the information domain. Shafee et al. state that LLM-generated text can be utilized to manipulate text-based CTI processing pipelines. It is reported by Jiang et al. that synthetic CTI is difficult for human annotators to distinguish from genuine CTI, illustrating the potential for information-quality attacks [14]. SynthCTI similarly demonstrates that LLM-generated synthetic CTI can be of use for controlled data augmentation and MITRE technique mapping, showing that the same generative mechanism has both defensive and potentially adversarial applications.

The implication is that threat generation should be evaluated at two levels: generation of cyber actions and generation of cyber information. The second level matters because modern security operations depend heavily on automated extraction, classification, scoring, and prioritization of textual intelligence [15].

Retrieval, Grounding, and Reliability

RAG and evidence retrieval represent an important transition from static model knowledge toward adaptive cyber reasoning. Tellache et al. integrate semantic retrieval with external CTI platforms to contextualize incident response, while Tang et al. prefer multi-step evidence retrieval for CTI credibility verification. These architectures indicate that system-level utility can exceed the capability of the base model when relevant external knowledge is dynamically supplied.

However, retrieval introduces a new dependency: incorrect or irrelevant evidence can propagate through the generation process. Therefore, grounding should be evaluated independently from generation quality [16]. A fluent response supported by irrelevant evidence should not receive the same utility score as a technically equivalent response supported by directly relevant evidence.

Benchmarking and Safety–Utility Trade-offs

The growing number of cybersecurity-specific benchmarks provides an important methodological foundation. SECURE evaluates knowledge extraction, understanding, and reasoning; CyberBench evaluates multiple cybersecurity NLP tasks; and CYBERSECEVAL evaluates cybersecurity risks and capabilities, including offensive-security-related behavior.

These benchmarks also reveal why a single offensive-utility score can be misleading. A model can perform strongly on technical knowledge while being unstable, overly cautious, or slow. Conversely, a model can provide useful assistance on benign security tasks while refusing a subset of offensive requests. Evaluation should therefore report a profile rather than a single undifferentiated number [17].

Discussion

The evidence supports a cautious but substantive interpretation of LLM-enabled offensive potential. The strongest current evidence concerns cognitive assistance rather than unrestricted autonomous attack execution. LLMs can compress parts of the expertise required to interpret vulnerabilities, map techniques, construct attack paths, process CTI, and formulate next-step reasoning [17]. The principal risk is therefore not necessarily a single breakthrough capability, but the cumulative reduction in coordination and knowledge-transfer costs across multiple stages.

At the same time, reliability remains a limiting factor. Hallucination, contextual inconsistency, retrieval errors, latency, and long-horizon planning difficulties create friction between model capability and operational utility. This supports the distinction proposed in this review: capability is necessary but insufficient for offensive utility.

The socio-technical nature of the problem is also important. Offensive utility is influenced by model choice, prompting, retrieval, external tools, knowledge sources, human oversight, and safety configuration. Consequently, evaluating the base model alone may substantially underestimate or misattribute the capabilities of a complete LLM-enabled cyber system [18].

Based on the synthesis, future studies should evaluate offensive utility through a controlled scenario framework rather than isolated prompts. A scenario should specify a target context, initial condition, threat objective, available evidence, relevant cybersecurity knowledge, and validation criteria. Models should then be evaluated under comparable prompting conditions, such as zero-shot, few-shot, and retrieval-augmented configurations [19].

The evaluation should proceed from lower-complexity tasks to higher-complexity reasoning: threat characterization, technique mapping, multi-stage path reasoning, CTI generation, and controlled workflow assistance [20]. Outputs should be evaluated through a combination of automated ground-truth comparison and expert assessment. The framework should explicitly record refusals, hallucinations, unsupported claims, retrieval errors, latency, and human correction effort.

A composite score should not replace the underlying dimensions. If a composite index is used, weights must be defined before analysis and sensitivity analysis should be reported. This is important because

weighting reliability and technical correctness more heavily than stylistic quality changes the ranking of systems and makes the evaluation more aligned with operational cybersecurity requirements [21].

Limitations

This review has several limitations. First, the underlying literature is heterogeneous in datasets, models, task definitions, and evaluation metrics, preventing direct quantitative aggregation across all studies. Second, rapidly evolving models and agent architectures create a moving target for comparative assessment. Third, some recent evidence remains available primarily as preprints; such sources are explicitly distinguished from peer-reviewed publications [22]. Finally, offensive utility is inherently context-dependent, and a framework developed from published research cannot fully represent every operational environment.

Conclusion

LLMs are increasingly capable of contributing to cybersecurity tasks that have direct relevance to offensive threat generation, but their practical value cannot be inferred from language-generation ability alone. The evidence reviewed here indicates meaningful capabilities in threat modeling, attack-path reasoning, penetration-testing assistance, CTI generation, retrieval-supported analysis, and adversarial information manipulation. These capabilities are constrained by reliability, grounding, latency, safety behavior, and human-validation requirements.

The principal contribution of this review is a multidimensional interpretation of offensive utility that separates capability, reliability, and operational usefulness. The proposed framework provides a structured basis for future empirical benchmarking while avoiding the unsupported assumption that fluent generation is equivalent to autonomous cyber capability. Future work should evaluate LLM-enabled systems across controlled scenarios and report technical validity, grounding, coherence, reliability, human-effort reduction, temporal efficiency, and safety–utility trade-offs together.

References

- [1] M. S. Salek *et al.*, ‘A large language model-supported threat modeling framework for transportation cyber-physical systems’, *IEEE Access*, vol. 13, pp. 163046–163070, 2025, doi: 10.1109/ACCESS.2025.3603580.
- [2] Y. R. Chen, M. J. Cui, D. Wang, and others, ‘A survey of large language models for cyber threat detection’, *Computers & Security*, vol. 145, p. 104016, 2024, doi: 10.1016/j.cose.2024.104016.
- [3] A. Tellache, A. A. Korba, A. Mokhtari, H. Moldovan, and Y. Ghamri-Doudane, ‘Advancing autonomous incident response: Leveraging LLMs and cyber threat intelligence’. 2025. [Online]. Available: <https://arxiv.org/abs/2508.10677>
- [4] Y. Zhang *et al.*, ‘AttacKG+: Boosting attack graph construction with Large Language Models’, *Computers & Security*, vol. 150, p. 104220, 2025, doi: 10.1016/j.cose.2024.104220.
- [5] J. Jiang, S. K. Jha, and others, ‘Can LLM-generated misinformation be detected: A study on Cyber Threat Intelligence’, *Future Generation Computer Systems*, vol. 173, p. 107877, 2025, doi: 10.1016/j.future.2025.107877.
- [6] Z. Liu, J. Shi, and J. F. Buford, ‘CyberBench: A multi-task benchmark for evaluating large language models in cybersecurity’, in *AAAI-24 Workshop on Artificial Intelligence for Cyber Security (AICS)*, 2024. [Online]. Available: <https://arxiv.org/abs/2403.18493>
- [7] M. Bhatt *et al.*, ‘CyberSecEval 2: A wide-ranging cybersecurity evaluation suite for large language models’. 2024. [Online]. Available: <https://arxiv.org/abs/2404.13161>
- [8] S. Wan *et al.*, ‘CyberSecEval 3: Advancing the evaluation of cybersecurity risks and capabilities in large language models’. 2024. [Online]. Available: <https://arxiv.org/abs/2408.01605>

- [9] S. Shafee, A. Bessani, and P. M. Ferreira, 'Evaluation of LLM-based chatbots for OSINT-based cyber threat awareness', *Expert Systems with Applications*, vol. 261, p. 125509, 2025, doi: 10.1016/j.eswa.2024.125509.
- [10] P. R. B. Houssel, S. Layeghy, P. Singh, and M. Portmann, 'eX-NIDS: A framework for explainable network intrusion detection leveraging large language models', *Computers and Electrical Engineering*, vol. 129, p. 110826, 2026, doi: 10.1016/j.compeleceng.2025.110826.
- [11] S. Shafee, A. Bessani, and P. M. Ferreira, 'False alarms, real damage: Adversarial attacks using LLM-based models on text-based Cyber Threat Intelligence systems', *Future Generation Computer Systems*, vol. 184, p. 108603, 2026, doi: 10.1016/j.future.2026.108603.
- [12] P. Balasubramanian *et al.*, 'Generative AI for cyber threat intelligence: Applications, challenges, and analysis of real-world case studies', *Artificial Intelligence Review*, vol. 58, p. 336, 2025, doi: 10.1007/s10462-025-11338-z.
- [13] M. A. Ferrag *et al.*, 'Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities', *Internet of Things and Cyber-Physical Systems*, vol. 5, pp. 1–46, 2025, doi: 10.1016/j.iotcps.2025.01.001.
- [14] M. Uddin *et al.*, 'Generative AI revolution in cybersecurity: A comprehensive review of threat intelligence and operations', *Artificial Intelligence Review*, vol. 58, p. 236, 2025, doi: 10.1007/s10462-025-11219-5.
- [15] H. Alqahtani and G. Kumar, 'Large language models for cybersecurity intelligence: A systematic review of emerging threats, defensive capabilities, and security evaluation frameworks', *Computers, Materials & Continua*, vol. 87, no. 3, 2026, doi: 10.32604/cmc.2026.077367.
- [16] F. Tang, H. Li, M. Zhao, Z. Wu, S. Peng, and T. Yin, 'LRCTI: A large language model-based framework for multi-step evidence retrieval and reasoning in cyber threat intelligence credibility verification'. 2025. [Online]. Available: <https://arxiv.org/abs/2507.11310>
- [17] A. Happe and J. Cito, 'On the surprising efficacy of LLMs for penetration-testing'. 2025. [Online]. Available: <https://arxiv.org/abs/2507.00829>
- [18] G. Deng *et al.*, 'PentestGPT: An LLM-empowered automatic penetration testing tool'. 2023. [Online]. Available: <https://arxiv.org/abs/2308.06782>
- [19] M. Bhatt *et al.*, 'Purple Llama CyberSecEval: A benchmark for evaluating the cybersecurity risks of large language models'. 2023. [Online]. Available: <https://arxiv.org/abs/2312.04724>
- [20] D. Bhusal *et al.*, 'SECURE: Benchmarking large language models for cybersecurity', in *Proceedings of the 2024 Annual Computer Security Applications Conference*, 2024. doi: 10.1109/ACSAC63791.2024.00019.
- [21] Á. Ruiz-Ródenas, J. Pujante Sáez, D. García-Algora, M. Rodríguez Béjar, J. Blasco, and J. Hernández-Ramos, 'SynthCTI: LLM-driven synthetic CTI generation to enhance MITRE technique mapping'. 2025. [Online]. Available: <https://arxiv.org/abs/2507.16852>
- [22] J. Zhang *et al.*, 'When LLMs meet cybersecurity: A systematic literature review', *Cybersecurity*, vol. 8, p. 55, 2025, doi: 10.1186/s42400-025-00361-w.