A Systematic Literature Review on LLMs Optimization in Cybersecurity via Prompt Engineering

Nur Ezzad Haqim Ali[1], Nur Hadirah Abdul Razak[2], Muhammad Aiman Zulkifli[3]

[1,2,3]University Teknologi Mara (UiTM), Kuala Lumpur, Malaysia

DOI: https://doi.org/10.47772/IJRISS.2026.101000397

Received: 19 October 2026; Accepted: 24 October 2026; Published: 03 November 2026

ABSTRACT

Cyber threats have been lingering around the Internet for a while now but still ways to overcome it remains a challenge to cybersecurity practitioners. The struggles to adapt through various attacks is not an easy task especially as a beginner in this field. AI presence has lifted some weights to banishing these threats, but it is all depends on how well it is prompted in the Large Language Model that the practitioner used. This systematic literature review investigates the optimization of Large Language Models via prompt engineering to execute cybersecurity task effectively. The optimization will follow the NIST Cybersecurity Framework (CSF 2.0) core functions which are Govern, Identify, Protect, Detect, Respond and Recover so that it reflects the real-world usage. Moreover, the review methodology will be through data collection to find related articles, article screening to analyses the content relevance and results analysis for answering research questions. There are 17 articles that were used for the review to ensure this literature review is concise for assignment fulfilment. Main findings from this review are about how prompt engineering improves LLMs performance in cybersecurity functions. The research gaps will be on evaluation bias, adversarial prompts that comes with LLMs used and legal compliance. Future direction for this review will focus on how prompt engineering can automate tailored recovery planning and develop techniques to reduce stereotype biases in model outputs.

Keywords: Large Language Models (LLMs), Cybersecurity, Prompt techniques, NIST CSF 2.0

INTRODUCTION

Cybersecurity in the modern days is held under constant pressure from evolving complex attacks which includes a surge in malware variants and zero-day exploits. The variable variants are getting out of hand since there are declining number of skilled cybersecurity professionals each year that makes human error is the main cause for roughly 95% of security breaches (Basbas et al., 2025) Prompt engineering in Large Language Models used acts as a saviour to overcome these numbers. However, the effective use of the LLMs needs to be studied because the right prompts will increase its performance in executing tasks in cybersecurity. There are manual techniques for optimizing the prompts effectiveness such as Chain-of-Thought (CoT) and Few-Shot Prompting (FSP) to boost accuracy in complex cybersecurity tasks (Nguyen & Sayadi, 2024) Moreover, there is also an automated optimization frameworks like DSPy and RAG that allow smaller models to bring performance gap closer with massive, closed-source enterprise models (Basbas et al., 2025; Meda et al., 2025).

Current research on LLMs in cybersecurity remains incomplete and often lacks the alignment with established professional standards even though technology advancement is present. The accuracy of LLMs heavily depends on how well the user write their prompts where bad prompts may lead to incomplete information. This paper will be organized with sections numbered in roman characters and the sub section will be sorted with alphabetical order. The section will continue as follows. Section II details the methodology which includes research questions, search strategy, the inclusion and exclusion criteria, article selection process, quality assessment and data extraction as its sub section. Section III will dive on this review findings from the articles collected. Section IV will provide research discussion on the efficacy of prompt engineering in LLMs. Next, Section V reveals the research gaps and future directions based on the findings while Section VI concludes the review as final conclusion.

METHODOLOGY

This section explains how the systematic literature review is conducted.

Research Questions

The research questions that drive this study is as follows:

- **RQ1:** What are the contributions of LLMs to cybersecurity research when mapped across the six core functions of the NIST CSF 2.0?
- **RQ2:** How far can advanced prompt engineering optimization narrow the performance gap between Open-Source LLMs and massive proprietary models?
- **RQ3:** What are the existing biases, knowledge limitations, and adversarial risks (the "Dark Side") associated with LLM optimization in cybersecurity?

Search Strategy

For starters, related articles or proceedings that have the potential for exploring the prompting effectiveness in LLMs were analysed through many academic databases to ensure its relevance with the review. Initial searches returned with 150 records across all databases. Through duplicate removal, abstract screening and full-text eligibility assessment leads to the final study set with has been reduced to 17 studies. The search techniques used to find the papers for this systematic review is depicted in the Table 1 as follows.

Inclusion and Exclusion Criteria

The eligibility criteria are used to ensure the review have quality and relevance to the study. The inclusion criteria help in emphasizing the subject which for example in this case is Large Language Models, prompt engineering and cybersecurity. More than that, the inclusion also includes the year range of the articles was published and types of materials used for instance journal articles and proceedings only. Exclusion criteria, on the other hand, helps filtered out research paper that does not align with review's objectives. So, primary focus to study prompt engineering in LLMs on cybersecurity can be found because the results are more relevant and not too broad. The inclusion and exclusion criteria were described in Table 2.

Table 1: Search Strategy

Database	Search String	Year Range	Number of Results
Google Scholar	"prompt engineering" AND "cybersecurity" AND "Large Language Model"	2021-2026	18 000
Scopus	"Large language model" OR "LLM" AND "cybersecurity" OR "security incident"	2021-2026	111
ACM Digital Library	"AI" OR "Large Language Model" AND "Prompts" AND "cybersecurity" OR "security analysis"	2021-2026	4702
IEEE Xplore	"prompt engineering" OR "prompting" AND "LLMs" AND "cybersecurity"	2021-2026	417

Table 2: Inclusion and Exclusion Criteria

Criteria Type	Description
Inclusion	Articles published within 2021-2026
Inclusion	Articles related to prompting techniques in LLMs to optimize its effectiveness for cybersecurity tasks
Inclusion	Journal articles or Proceedings

Exclusion	Non-English articles
Exclusion	Articles without full text or not related to the topic
Exclusion	Broad topic on prompt engineering without cybersecurity focus

Study Selection Process

This process was conducted in multiple stages to ensure that only high quality and relevant studies were included in the review. In identification stage, the literature was retrieved from selected databases which is shown on Table 1 using the predefined search strings. All gathered literatures were exported into Mendeley Reference Manager for better management so that duplicate literatures can be removed. These literatures were then going through automated screening with Google NotebookLM where studies unrelated to cybersecurity, prompt engineering, or LLM optimization were excluded in this stage. After that, the full text that potentially relevant with the study were further manually examined to confirm if they met the inclusion criteria.

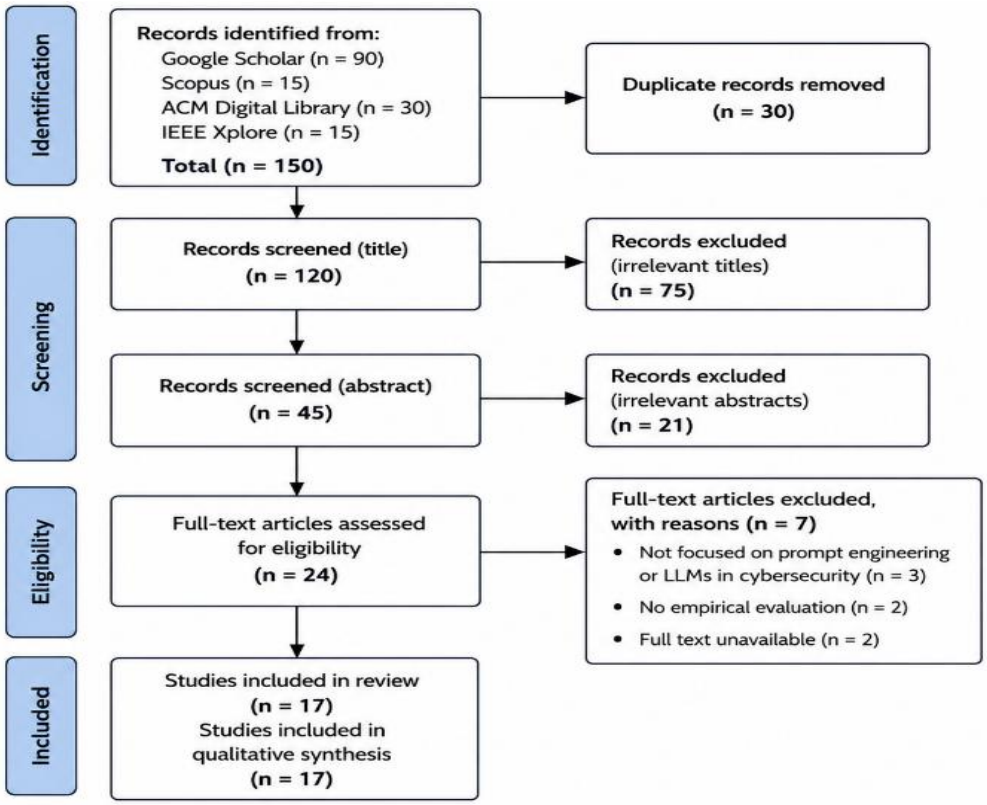

Figure 1: Study Selection Process

Figure 1 shows the PRISMA study selection process. Initial search of 150 records has been through screening and eligibility assessment which finally reduced the dataset to 17 studies. This process helps to choose the appropriate articles and proceedings for the systematic literature review.

Quality Assessment

In order to verify validity of the selected studies, this assessment was made using specific evaluation criteria based on evaluated articles methodologies. Each study was assessed through these questions:

1. Does the study clearly state its research objectives?
2. Is the prompt engineering methodology described in detail?
3. Does the study provide empirical evaluation or experimentation?
4. Are cybersecurity applications clearly identified and explained?
5. Are evaluation metrics and performance results reported?
6. Does the study discuss limitations or threats to validity?

Each criterion was scored as 1 = Yes, 0.5 = Partially and 0 = No. There are 17 sources in total for the assessment. The quality assessment helped reduce the risk of bias and to show prove that conclusions were made with careful considerations. The evaluation results were presented in Table 3 below.

Data Extraction

Data extraction was held systematically to collect related information from each selected study. This process is necessary to identify the objectives, methods, datasets, key findings and limitations that correlate with the optimization of Large Language Models (LLMs) in cybersecurity through prompt engineering. The extracted data was managed into a structured table to show comparison among studies while also ease to synthesize the findings. These attributes were chosen because it helps to answering the research questions regarding LLMs role in cybersecurity, prompt engineering techniques effectiveness and the challenges and risks that comes with it. The data extraction table is shown in Table 4.

Table 3: Quality Assessment

Authors and Date	QA1	QA2	QA3	QA4	QA5	QA6	Total Score
Bai et al. (2025)	1	1	1	0.5	1	1	5.5
Villegas-Ch et al. (2025)	1	0	1	1	1	1	5
Ali & Ghanem (2025)	1	0.5	0	1	0.5	1	4
Nguyen & Sayadi (2024)	1	1	1	1	1	1	6
Wahréus et al. (2025)	1	1	1	1	1	1	6
Balaji et al. (2024)	1	1	1	1	1	1	6
Meda et al. (2025)	1	1	1	1	0.5	1	5.5
Foundjem et al. (2026)	1	1	1	1	1	1	6
Priescu et al. (2025)	1	1	1	1	1	1	6
Ali & Mohamed (2026)	1	0.5	0	1	0.5	1	4
Hannon et al. (2024)	1	1	1	1	1	1	6
Jajodia et al. (2026)	1	1	1	1	1	1	6
Gao et al. (2025)	1	0	1	0.5	0	0	2.5
Chen et al. (2025)	1	1	1	1	0.5	1	5.5
Alawida et al. (2024)	1	0.5	1	1	0.5	1	5
Hilal et al. (2024)	1	0	0	1	0	1	3
Basbas et al. (2025)	1	1	1	1	1	1	6

Table 4: Data Extraction Table

Author /Year	Author /Year	Method	Dataset /Tool	Key Findings	Limitations
Priescu et al. (2025)	Explore the role of structured prompting techniques, such as Chain-of-Thought (CoT), in optimizing LLM interactions for cybersecurity workflows.	Experimental Study	GPT-4o, Multimodal CoT Prompting.	Structured prompts significantly improve LLM performance in network log analysis, phishing detection, and automated security intelligence reporting.	Reliability depends heavily on prompt quality; models remain susceptible to biases and adversarial manipulation.

Basbas et al. (2025)	Assess cybersecurity knowledge of open-source LLMs and evaluate the impact of automated prompt engineering using NIST CSF 2.0 alignment.	Benchmark Evaluation	21 OSLLMs (e.g., Qwen2, Llama-3, Phi-3), DSPy framework, CyberMetric, SecEval.	Automated prompt engineering via DSPy-CoT boosted performance significantly, with smaller models like Phi-3-mini achieving up to 18.37% gains.	Significant positional and forced-choice biases identified in standard MCQA datasets; LLMs exhibited low self-awareness of knowledge gaps.
Nguyen & Sayadi (2024)	Perform a comparative evaluation of ChatGPT and Gemini Pro in cybersecurity education, specifically measuring the impact of prompt engineering.	Comparative Analysis	ChatGPT-4, Gemini Pro 1.5, Phishing Emails Dataset, Web Page Phishing Detection Dataset.	ChatGPT outperformed Gemini in code generation (96% vs 76%); prompt engineering improved Gemini's modular arithmetic accuracy from 8% to 85%.	Prompt engineering cannot fundamentally enhance core capabilities for high-complexity tasks like AES-256 de-obfuscation.
Chen et al. (2025)	Provide a comprehensive review of prompt engineering methodologies and efficacy in improving LLM performance, multimodal integration, and AI security.	Survey	GPT-4, Gemini, Claude, Llama, DALL-E, CLIP, GSM8K, SVAMP, RefCOCO.	Advanced techniques like CoT and self-consistency enhance logical reasoning; prompt engineering is effective in mitigating adversarial vulnerabilities like jailbreaking.	Automated prompt generation methods often lack generalizability; there is a critical need for standardized security defense benchmarks.
Balaji et al. (2024)	Evaluate the effectiveness of LLM answers to cybersecurity questions and improve them through specialized prompt engineering.	Experimental Study	ChatGPT-4, Llama 2 (70B), Proofpoint's Cybersecurity Awareness Training Modules, together.ai API.	The 'Motivating' prompt template improved ChatGPT-4 scores by 15.28% on average, significantly increasing the actionability of security advice.	Small initial question set; relied on technical researchers acting as personas for annotation rather than actual end-users.
Foundjem et al. (2026)	Systematically characterize machine learning security threats and evaluate the scalability of current defenses across RAG pipelines.	Framework Development	GPT-4o, Claude-3.5, Gemini-1.5, LLaMA-3.2, MITRE ATLAS, GitHub/PyPI.	A multi-agent RAG system lifted classification macro-F1 from 0.71 to 0.87; larger models were found more vulnerable to introspection-driven exploits.	The HGNN model produces probabilistic severity estimates rather than formal proofs and requires labelled anchors for training.

Jajodia et al. (2026)	Establish a rigorous agentic evaluation framework and dataset for benchmarking LLM capabilities in automated security incident analysis (SIA).	Benchmark Evaluation	SIABENCH, GPT-5, Claude-4.5-Sonnet, Llama 3.1, Volatility, Tshark, Snort.	GPT-5 and Claude-4.5-Sonnet outperformed predecessors but only solved 8 of 25 scenarios; performance declined as technical depth increased.	Models exhibited 'shallow investigation' tendencies and infinite loops; agent scaffolding requirements varied significantly by model maturity.
Hannon et al. (2024)	Evaluate the resilience and cybersecurity robustness of leading LLMs against innovative adversarial prompts designed to bypass filters.	Experimental Study	ChatGPT-4, Claude, Bard, Microsoft Copilot, LLaMa 2, FreedomGPT.	Advanced LLMs remain vulnerable to sophisticated jailbreaking techniques like 'Condition Red', though ChatGPT-4 managed response severity better than 3.5.	Study limited to a select group of models; ethical constraints prevented testing with highly offensive content.
Wahréus et al. (2025)	Present CySecBench, a comprehensive dataset for benchmarking LLM resilience against jailbreaking in the cybersecurity domain.	Benchmark Evaluation	ChatGPT, Gemini, Claude, CySecBench (12,662 prompts), AdvBench.	Prompt obfuscation jailbreaking achieved success rates of 65% with ChatGPT and 88% with Gemini, while Claude was most resilient at 17%.	Approximately 5.8% of prompts in the dataset remain non-malicious or logically nonsensical despite filtering efforts.
Ali & Ghanem (2025)	Investigate the impact of LLMs on key cybersecurity technical areas including threat detection, autonomous defense, and adversarial simulation.	Survey	GPT family, PaLM, LLaMA, Falcon, Mistral, CodeBERT, BERT, T5.	LLMs enable zero-shot and few-shot threat hunting and automated vulnerability patching, shifting security from reactive to proactive.	Issues include model hallucination, adversarial misuse (prompt injection), and a lack of grounding in specific domain knowledge.
Meda et al. (2025)	Develop a specialized LLM training system using RAG and embedding techniques to educate users on cybersecurity concepts.	Framework Development	Mistral Large 2, Google Gemini 1.5, RAG, Python, Gradio.	Integration of RAG and embedding models for clustering improved contextual accuracy in training scenarios like phishing and malware awareness.	Potential for hallucinations and model bias; performance delays occurred with large-scale API queries.

Gao et al. (2025)	Explore the feasibility of applying prompt engineering techniques to optimize LLM performance for data analysis in the public service sector.	Case Study	DeepSeek, Doubao, table summarization datasets.	Precise prompt design and iterative optimization significantly improved data processing efficiency and reduced error rates in complex table summarization.	Reliance on large-scale deep learning models results in lower interpretability compared to traditional rule-driven NLP methods.
Ali & Mohamed (2026)	Examine prompt engineering for LLMs in security analysis by synthesizing research across digital twin security and adaptive defense.	Comparative Analysis	Generative models, Deep Learning architectures, ARMOR framework.	Multi-layered adaptive defense frameworks (ARMOR) learn from attack attempts to strengthen defenses beyond conventional systems.	The dual-use nature of AI presents challenges as advanced signal processing can be exploited by adversaries for evasion.
Villegas-Ch et al. (2025)	Enhance digital forensic analysis accuracy and efficiency by integrating advanced AI techniques like CNNs, machine learning, and NLP.	Framework Development	GPT, BERT, MVTec AD, DAGM 2007, MNIST, CIFAR-10.	NLP techniques achieved 88% precision in extracting relevant text information, while CNNs achieved 92% precision in identifying image patterns.	Reliance on institutional data may introduce bias; inherent complexity of AI models can obscure internal workings critical for legal contexts.
Bai et al. (2025)	Introduce prompt-based measures to reveal implicit stereotypical biases in value-aligned, proprietary LLMs.	Experimental Study	GPT-4, GPT-3.5-turbo, Claude-3, Llama-2/3, LLM Word Association Test.	Pervasive implicit biases were found across 8 models regarding race, gender, and religion, despite models passing explicit safety tests.	Measures lack a mechanistic interpretation due to the black-box nature of proprietary models; relative decisions are more biased than absolute ones.
Alawida et al. (2024)	Investigate adversary tactics in leveraging ChatGPT for cyberattacks and evaluate user awareness of these threats.	Survey	ChatGPT, OpenAI API, Nmap.	Researchers demonstrated ChatGPT's ability to create full malware infection strains; over 80% of participants agree it is used for malicious purposes.	Difficult to determine if specific malware in the wild was constructed using ChatGPT; non-expert awareness remains low.
Hilal et al. (2024)	To propose a strategic conceptual framework for integrating AI-driven security	Framework Development	Secondary sources from academic databases (IEEE Xplore, Scopus) and	AI-driven security facilitates a shift from reactive to proactive defense by automating real-time threat	Critical limitations include the challenge of managing citizen privacy within large datasets and a significant research

	within digital governance to redefine and enhance Security Information Systems (SIS)		reports from international organizations including NIST, ITU, and ENISA	detection, reducing human error, and processing vast datasets for pattern recognition.	gap concerning AI implementation in government-led systems in developing nations like Malaysia

FINDINGS

This section synthesizes the findings across the studies which answers the research questions of this review.

LLMs Contributions Across the NIST CSF 2.0 Functions

Based on the reviewed studies, Large Language Models (LLMs) has contributed to the six core functions of the NIST Cybersecurity Framework (CSF) 2.0. Cybersecurity tasks such as governance and risk assessment to incident response and recovery planning have benefited from the LLMs no matter they are open-source or not. According to Basbas et al. (2025), their studies have shown that open-source LLMs can effectively support cybersecurity operations when it is aligned with NIST CSF 2.0 functions. In addition to that, Ali & Ghanem (2025) highlighted that LLMs plays it role through automating threat detection, vulnerability analysis and security monitoring activities.

Within the **Govern** function, LLMs helps in refining policy generation, compliance documentation, and suggests cybersecurity awareness initiatives for an organization. Through **Identify** function, LLMs assist in discovering system vulnerability, threat intelligence analysis and even analyses the risk assessment activities so that no threats are overlooked. The **Protect** function benefits from automated security awareness training, phishing detection and generate manuals for users to navigate through LLM-based systems. These studies also show LLMs strong capabilities in **Detect** function where it can assist in anomaly detection, network log analysis and threat hunting in which is much efficient than human response.

Jajodia et al. (2026) have pointed out the **Respond** function through the SIABench development to evaluate LLM performance in analysing security incident. Their findings indicate that advanced models can better assists security practitioners in incident investigation and response workflow. Lastly, the **Recover** function remains to be less explored because most studies only provide general recommendations for recovery planning and post-incident. Therefore, the literature proves that LLMs have practical applications with some NIST CSF 2.0 functions even though their effectiveness relies on task complexity and prompt quality.

Prompt Engineering Effectiveness for LLM Optimization

LLM performance in cybersecurity applications is heavily influenced by prompt engineering. Several studies reported rising performance when structured prompting strategies were employed.

Chain-of-Thought (CoT) prompting is seen as the most frequently adopted approaches. Priescu et al. (2025) found that tailored CoT prompts significantly improved network logs analysis, phishing detection and automated security reporting from the LLMs. Apart from that, Basbas et al. (2025) demonstrated that integrating DSPy-based automated prompt optimization with CoT prompting increased multiple open-sources LLMs with smaller models performance achieving up to 18.37% performance improvements. Few-shot prompting also depicted positive results in cybersecurity education and detection tasks. Nguyen & Sayadi (2024) uncover that prompt engineering improved Gemini Pro's performance in complex reasoning tasks, increasing modular arithmetic accuracy from 8% to 85%. In another research, Balaji et al. (2024) reported that a specialized "Motivating" prompt template improved the quality and actionability of cybersecurity advice generated by ChatGPT-4.

Recent studies further explained Retrieval-Augmented Generation (RAG) and embedding-based approaches. Meda et al. (2025) showed that combining RAG with embedding techniques improved contextual understanding and user training in cybersecurity awareness scenarios. These findings suggest that prompt

engineering is a cost-effective alternative to model fine-tuning while enabling smaller open-source models to achieve performance levels that on par with larger models.

Biases, Limitations, and Adversarial Risks of Optimized LLMs

Along with prompt engineering benefits comes some drawbacks. The reviewed studies mentioned several limitations and risks related with LLM deployment in cybersecurity environments.

Vulnerability to adversarial prompting and jailbreaking attacks is one of the significant concerns when optimizing the LLMS. Hannon et al. (2024) demonstrated that advanced LLMs remain susceptible to tailored adversarial prompts made to bypass safety controls. Coherent to that, Wahréus et al. (2025) have developed CySecBench framework that showed prompt obfuscation techniques achieved jailbreak success rates as high as 88% for some models.

Next, the major challenges are hallucination and inaccurate information generation. Multiple studies reported that LLMs occasionally generate plausible but incorrect response to cybersecurity recommendations. This can be seen when dealing with unfamiliar attack scenarios or insufficient context. These hallucinations negatively affect security operations specifically in decision-making processes.

Apart from hallucination, bias also a persistent issue. Bai et al. (2025) explained that even safety-aligned LLMs shows implicit biases related to race, gender, and religion even though passing conventional safety evaluations. Basbas et al. (2025), on the other hand, clarified that model performance assessments are influenced by positional and forced-choice biases in cybersecurity benchmark datasets.

Another major challenge is LLM technology can work for both defensive and offensive cybersecurity purposes. Alawida et al. (2024) emphasized that LLMs can facilitate malicious actors in generating malware-related content which encourage cyberattacks when appropriate safeguards are absent. Therefore, prompt engineering improves model effectiveness in cybersecurity tasks, but further research is needed to mitigate the bias and adversarial misuse. These findings are discussed in detail on the following section.

RESEARCH DISCUSSION

The sections synthesize the findings from the papers.

Prompt Engineering as an Alternative to Model Scaling

The reviewed studies share that prompt engineering is a practical alternative to increasing model size for cybersecurity applications. Old LLMs optimization have relied on larger perimeter counts and expensive model training which is not cost effective. These findings have shown that techniques such as Chain-of-Thought prompting, few-shot prompting, Retrieval-Augmented Generation, and DSPy-based optimization can increase cybersecurity task performance drastically without modifying model parameters. This method bound to be useful for organizations with limited computational power because it helps to enhance smaller open-source LLMs to be as good as the larger model. Prompt engineering offers cost-effective option for organization or cybersecurity personnel to use AI systems while reducing the need for better models.

LLM Capabilities Alignment with NIST CSF 2.0

The results have shown that there is a significant correlation between the capabilities of the current LLM and the 2021 NIST Cybersecurity Framework core functions. It is said because the majority of the studies are on the use of prompt engineering to assist in threat intelligence analysis, phishing detection, security awareness training, vulnerability assessment or incident response support. The **Recover** function was not as much discussed as other functions, however, because research efforts have continued to focus on prevention and detection efforts instead of recovery planning after an incident. The optimization of LLM in the future should be directed towards applications of cybersecurity with a focus on recovery. The more comprehensive the coverage of the NIST CSF 2.0 framework, the better.

Security and Ethical Implication of LLM Optimization

Prompt engineering successfully improves LLM effectiveness, but the reviewed studies mentioned that the optimization also leads to security and ethical concerns. The LLMs will respond to practically any question you might ask, including some adversarial questions and jailbreak attack, which is a challenge to reliable deployment. Not to mention the hallucination and biases that can occur in the answers. Therefore, the use of the LLMs in cybersecurity is primarily dependent on the user's intended purpose of using the LLM for. However, by itself, prompting is never enough to ensure that the trustworthy can be deployed in cybersecurity systems. Other factors such as human oversight, security monitoring and continuous evaluation are needed to reduce operational risks.

RESEARCH GAPS AND FUTURE DIRECTIONS

This section addresses several limitations and unresolved issues that remains even though prompt engineering has made LLMs better in cybersecurity applications. Hence, these gaps allow for future research and development. Table 5 summarizes the major research gaps identified from the literature and proposes future directions.

Table 5: Research Gap

Gap Identified	Evidence from Literature	Possible Future Work
Limited support for the Recover function in NIST CSF 2.0	Most studies focus on threat detection, vulnerability assessment, phishing detection, and incident response, while recovery planning receives minimal attention.	Develop prompt-engineered LLM frameworks for disaster recovery planning, business continuity management, and post-incident recovery support.
Lack of standardized cybersecurity evaluation frameworks	Studies use different datasets, benchmarks, and metrics, making direct comparison difficult. Although CySecBench and SIABench developed, no common evaluation standard is adopted.	Establish standardized benchmarking frameworks that evaluate accuracy, robustness, explainability, bias, and security performance across cybersecurity tasks.
Vulnerability to adversarial prompting and jailbreak attacks	CySecBench and adversarial testing studies show that optimized LLMs remain susceptible to prompt obfuscation, prompt injection, and jailbreak attempts.	Design defensive prompt engineering techniques, automated prompt validation systems, and adversarially robust LLM architectures.
Hallucination and unreliable cybersecurity recommendations	Several studies report that LLMs occasionally generate plausible but incorrect security advice, particularly in unfamiliar attack scenarios.	Integrate Retrieval-Augmented Generation (RAG), external knowledge bases, and human-in-the-loop verification to improve reliability.
Persistent bias in optimized LLMs	Research demonstrates that even safety-aligned models exhibit implicit biases related to race, gender, and religion, while benchmark datasets may introduce positional bias.	Develop bias-aware prompting strategies and cybersecurity-specific fairness evaluation methods.
Dependence on prompt quality and user expertise	Performance improvements often rely on carefully crafted prompts, making results inconsistent across users.	Investigate automated prompt optimization frameworks and adaptive prompting systems that can reduce user dependency.

The identified gaps shows that current research on prompt-engineered LLMs in cybersecurity mainly focus on performance improvement rather than long-term reliability, security and governance. Future studies should

revolve around the recovery-oriented cybersecurity applications, standardized evaluation methodologies, address adversarial prompts issue and bias mitigation. Therefore, automated prompt optimization and adaptive prompting mechanisms might improve the LLM scalability and consistency for deployment across diverse cybersecurity environments.

PROPOSED CONCEPTUAL FRAMEWORK

Based on the gaps from the reviewed papers, a conceptual framework can be made to mitigate most of the problems in prompt engineering to increase the LLMs capabilities when answering cybersecurity related answers. The proposed conceptual framework will be **Secure Adaptive Prompt Engineering (SAPE)**. Rather than relying on a single static prompt, SAPE dynamically adapts prompts according to the cybersecurity task while incorporating verification, safety checks, and domain knowledge to improve its capabilities. This framework is intended as a future research direction and has not yet been empirically validated. It means to develop an improved version of DSPy before the prompt is sent to the LLMs for replies.

The evaluation metrics is adapted from the reviewed literature, and future research may be measured through:

- Task Accuracy (%)
- Hallucination Rate (%)
- Attack Success Rate (ASR)
- Bias Score
- Average Response Time (ms or s)
- Prompt Optimization Rate
- Precision, Recall, F1-score
- User Satisfaction Score

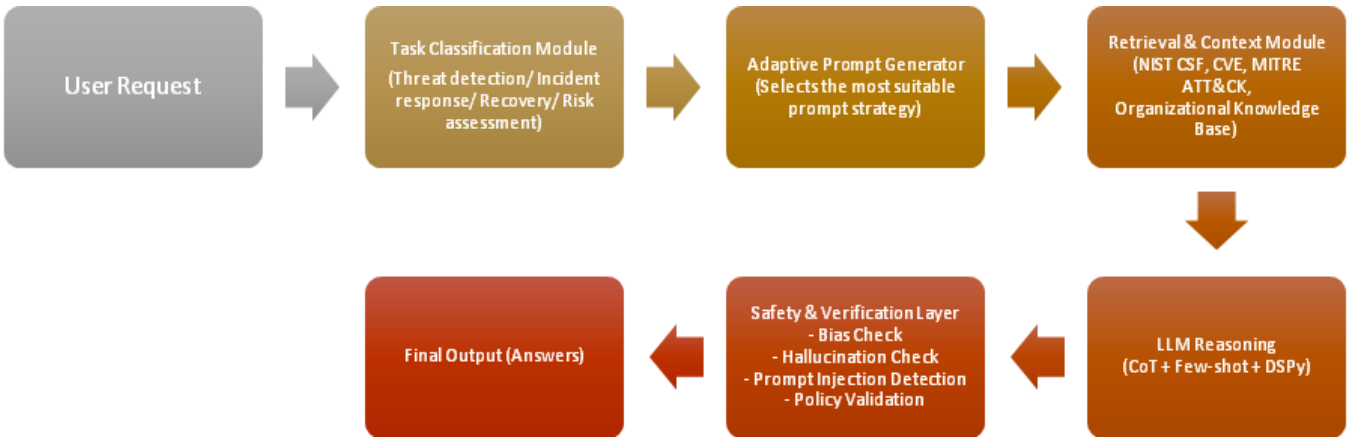

Figure 2: SAPE Conceptual Framework

CONCLUSION

This systematic literature review examined the optimization of Large Language Models (LLMs) in cybersecurity through prompt engineering techniques. The review analyzed 17 selected studies following the NIST Cybersecurity Framework (CSF) 2.0. These studies were published between 2021 to 2026 so it can be considered as still recent. The review was conducted to explore the LLMs contributions in cybersecurity, evaluate prompt engineering approaches effectiveness and identify the limitations along with their deployment.

The findings demonstrate that LLMs have become the valuable tools for some cybersecurity functions which particularly in threat intelligence analysis, vulnerability assessment, phishing detection, security awareness training, and incident response support. The reviewed literature exposes that prompt engineering can improve smaller model LLM performance to the level that matches larger model LLM. Prompting techniques such as Chain-of-Thought (CoT) prompting, few-shot prompting, Retrieval-Augmented Generation (RAG),

embedding-based prompting, and DSPy-based optimization were identified to improve LLMs reasoning capabilities, contextual understanding, and task-specific accuracy.

The review highlighted various limitations in the use of reliable LLMs for cybersecurity applications. These studies revealed that there are several challenges to overcome to ensure reliable LLMs are deployed in cybersecurity environments. Some of the key challenges that continue to plague LLM technologies are adversarial prompting, jailbreak attacks, hallucinations, implicit biases, and the dual-use problem. On critical view, there are some gaps in their research since it focused mainly on recovery-oriented cybersecurity tasks, lacked standardized evaluation frameworks and were subject to vulnerability from adversarial manipulation and continuous issues of bias and inaccurate outputs of the results.

In summary, the literatures indicate that prompt engineering is a feasible solution to enhance LLMs' performance in cybersecurity tasks with low cost. The key is to recognize that there are always risks to security and ethics to consider and that prompting deployment is an ongoing process that must be evaluated, monitored and supervised by humans. Future research recommendations should include developing secure, explainable, and adaptive prompt engineering techniques that would enable a wider scope of cybersecurity functions to be supported and remain reliable, fair, and resistant to new cyber threats.

REFERENCES

1. Alawida, M., Abu Shawar, B., Abiodun, O. I., Mehmood, A., Omolara, A. E., & Al Hwaitat, A. K. (2024). Unveiling the Dark Side of ChatGPT: Exploring Cyberattacks and Enhancing User Awareness. *Information (Switzerland)*, *15*(1). https://doi.org/10.3390/info15010027
2. Ali, A., & Ghanem, M. C. (2025). Beyond Detection: Large Language Models and Next-Generation Cybersecurity. *SHIFRA*, *2025*, 81–97. https://doi.org/10.70470/SHIFRA/2025/005
3. Ali, A., & Mohamed, S. (2026). *Prompt Engineering for Large Language Models in Security Analysis*.
4. Bai, X., Wang, A. I., Sucholutsky, I. I., & Griffiths, T. L. (2025). *Explicitly unbiased large language models still form biased associations*. https://doi.org/10.1073/pnas
5. Balaji, A., Duesterwald, L., Yang, I., Priyanshu, A., Alfieri, C., & Sadeh, N. (2024). *Generating Effective Answers to People's Everyday Cybersecurity Questions: An Initial Study*.
6. Basbas, M., Djamaa, B., & Senouci, M. R. (2025). *Unleashing the Power of Open Source Large Language Models for Cybersecurity: Key Insights and Challenges*. https://doi.org/10.22541/au.175825816.67299409/v1
7. Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns*, *6*(6), 101260. https://doi.org/10.1016/j.patter.2025.101260
8. Foundjem, A., Nganyewou Tidjon, L., Da Silva, L., & Khomh, F. (2026). Multi-Agent AI Framework for Threat Mitigation and Resilience in Machine Learning Systems. *ACM Transactions on Software Engineering and Methodology*. https://doi.org/10.1145/3780095
9. Gao, Y., Xiao, L., Shen, Y., & Chen, R. (2025). The Feasibility of Applying Prompt Engineering Based on Large Language Models in the Public Domain. *2025 11th International Symposium on System Security, Safety, and Reliability (ISSSR)*, (2025), 467–468. https://doi.org/10.1109/ISSSR65654.2025.00068
10. Hannon, B., Kumar, Y., Gayle, D., Li, J. J., & Morreale, P. (2024). Robust Testing of AI Language Model Resiliency with Novel Adversarial Prompts. *Electronics*, *13*(5), 842. https://doi.org/10.3390/electronics13050842
11. Hilal, M., Muhammad, B., Bin Abas, Z., Suzastri, A., Ahmad, B., Sufyan, M., Sulaiman, B., Antarabangsa, I., & Mu'adzam Shah, A. H. (2024). *AI-Driven Security: Redefining Security Informations Systems within Digital Governance*. https://doi.org/10.47772/IJRISS
12. Jajodia, S., Sultana, M., Majumdar, S., Taylor, A., & Vandenberghe, G. (2026). *Before You Hand Over the Wheel: Evaluating LLMs for Security Incident Analysis*. http://arxiv.org/abs/2603.06422
13. Meda, K. N., Nara, P. S. C., Bozenka, S., Zormati, T., Turner, S., Worley, W., & Mitra, R. (2025). Integrating Prompt Structures Using LLM Embeddings for Cybersecurity Threats. *Proceedings of the 2025 ACM Southeast Conference*, 180–187. https://doi.org/10.1145/3696673.3723069

14. Nguyen, T., & Sayadi, H. (2024). ChatGPT vs. Gemini: Comparative Evaluation in Cybersecurity Education with Prompt Engineering Impact. *Proceedings - Frontiers in Education Conference, FIE*. https://doi.org/10.1109/FIE61694.2024.10893499
15. Priescu, I., Banu, G. S., Dosescu, T. C., & Banu, M. I. (2025). Prompt Engineering in Cybersecurity – Achieving Technological Edge. *Land Forces Academy Review*, *30*(2), 291–302. https://doi.org/10.2478/raft-2025-0028
16. Villegas-Ch, W., Gutierrez, R., & Maldonado Navarro, A. (2025). Artificial intelligence techniques for enhancing accuracy and efficiency in digital forensic analysis. *Discover Artificial Intelligence*, *6*(1), 19. https://doi.org/10.1007/s44163-025-00729-4
17. Wahréus, J., Hussain, A. M., & Papadimitratos, P. (2025). *CySecBench: Generative AI-based CyberSecurity-focused Prompt Dataset for Benchmarking Large Language Models*. http://arxiv.org/abs/2501.01335