

**Cyber Security Threat Hunting Through Log Analysis Using LLM –
Literal Review Analysis**

Student : OMAR SALEM ALMUTAIRI

Student ID: 251002148

MIDOCEAN UNIVERSITY

FOR MASTER OF AI PROGRAM

GRADUATION PROJECT 1

I. Executive Summary

This report provides a comprehensive analysis of the application of Large Language Models (LLMs) in cybersecurity threat hunting, specifically through log analysis. It explores the transformative potential of LLMs while critically examining the inherent challenges and proposing strategic directions for future research and practical implementation.

LLMs represent a significant advancement in automating and enhancing log analysis for proactive threat detection, moving beyond the limitations of traditional rule-based systems. Current research demonstrates promising methodologies, including advanced anomaly detection, efficient log parsing and embedding, sophisticated multi-agent collaboration frameworks, and novel system state analysis techniques, showcasing superior performance in specific metrics. These advancements indicate a fundamental change in how log data can be analyzed, shifting from rigid, signature-driven security to a more flexible, behavior-based anomaly detection.

Despite their immense promise, substantial challenges persist. These encompass traditional concerns such as scalability, the prevalence of false positives, and complexities associated with integration into existing cybersecurity frameworks. Furthermore, critical issues like vulnerabilities to adversarial attacks (e.g., prompt injection, data poisoning), inherent data privacy risks stemming from LLM operation, and crucial performance considerations related to real-time latency in dynamic cybersecurity environments demand urgent attention. The opaqueness of LLM decision-making also highlights a critical need for interpretability and explainability.

To ensure responsible and effective LLM integration, this report emphasizes the imperative of developing Human-in-the-Loop (HITL) systems, establishing robust and standardized benchmarking frameworks, prioritizing Explainable AI (XAI) from the design phase, and conducting comprehensive cost-benefit analyses to inform strategic investment and adoption decisions. These measures are essential for fostering trust, ensuring reliability, and enabling the secure deployment of LLM-based solutions in real-world threat hunting operations.

II. Introduction to LLMs in Cyber Threat Hunting

Contextualizing the Evolving Threat Landscape

The contemporary cyber threat landscape is characterized by an escalating sophistication of attack methodologies, an increasing volume of malicious activities, and highly dynamic threat actors. This environment necessitates a fundamental shift from reactive security measures, which respond to incidents after they occur, to proactive strategies such as threat hunting. Threat hunting involves actively searching for potential threats that may have bypassed initial security defenses, aiming to identify malicious activities before they fully emerge and cause significant damage.¹

Traditional log analysis methods, which have long been a cornerstone of cybersecurity, are increasingly proving inadequate in this evolving landscape. These methods typically rely heavily on predefined rules and manual scrutiny, making them time-consuming, prone to errors, and largely incapable of detecting novel, polymorphic, or evolving threats. The sheer volume and complexity of log data generated by modern IT infrastructures further exacerbate these limitations, rendering conventional approaches inefficient for comprehensive threat detection.¹

The Transformative Potential of LLMs in Log Analysis for Cybersecurity

Large Language Models (LLMs) are uniquely positioned to revolutionize log analysis within cybersecurity due to their unparalleled capacity to process and interpret vast quantities of unstructured and semi-structured data. This capability allows LLMs to discern subtle patterns, identify anomalies, and generate actionable insights that often surpass the limitations of conventional methods.¹

The natural language processing (NLP) capabilities inherent in LLMs enable them to understand the semantic meaning embedded within chaotic and complex log entries. This understanding facilitates automated and significantly improved threat detection, vulnerability assessment, and incident response processes. A significant advantage of LLMs is their proficiency in managing the diverse and intricate nature of log data, which can vary considerably in format, content, and structure across system, network, and application sources. This adaptability is crucial for comprehensive threat hunting, where insights must be drawn from disparate data streams to uncover potential threats that might otherwise remain hidden.¹

The application of LLMs in log analysis for threat hunting represents a fundamental change in how security operations can be conducted, rather than merely an incremental enhancement. Traditional methods are typically rigid, rule-based, and rely on explicit signatures, making them less effective against polymorphic or novel attacks. LLMs, with their advanced semantic understanding and pattern recognition across diverse data, move towards a more flexible, behavior-based anomaly detection. This signifies a shift from a deterministic, signature-driven security posture to a more probabilistic and adaptive one. This profound change implies that cybersecurity professionals will need to adapt their skill sets, moving from solely engineering explicit detection rules to understanding, fine-tuning, and effectively interacting with LLMs, as well as interpreting their nuanced outputs. This capability holds significant promise for detecting zero-day exploits or highly evasive threats that do not conform to pre-defined patterns, thereby substantially enhancing proactive defense capabilities.

III. Core Applications and Methodologies of LLMs in Log Analysis

The Fundamental Role of LLMs in Processing Diverse and Unstructured Log Data

LLMs' advanced natural language processing (NLP) capabilities are pivotal in parsing and interpreting the inherently chaotic and complex nature of log data. This enables the analysis of varied log formats, content, and structures originating from a multitude of sources, including system, network, and application logs.¹ This

comprehensive data ingestion and understanding is critical for uncovering subtle indicators of compromise in threat hunting operations, as it allows security analysts to derive meaning from seemingly disparate log entries.

Detailed Examination of Key Frameworks

Various methodologies and frameworks have been proposed to harness LLMs for log analysis within threat hunting operations. These approaches can generally be classified into three main categories: anomaly detection, log parsing and embedding, and multi-agent collaboration.¹

Anomaly Detection

This category focuses on identifying unusual patterns or behaviors within log data that may signal a potential threat.¹

- **LogLLM:** This framework leverages pre-trained LLMs, specifically BERT for extracting semantic vectors from log messages and Llama for classifying log sequences. LogLLM has demonstrated superior performance over traditional deep learning methods in capturing the semantic information embedded in log data, leading to enhanced anomaly detection capabilities. It outperforms state-of-the-art methods in semantic understanding and anomaly detection.¹
- **LogRAG:** A semi-supervised framework built upon Retrieval-Augmented Generation (RAG) principles. LogRAG performs phased detection using both log tokens and log templates to effectively mitigate the impact of log parsing errors. It further employs a single-class classifier to model normal system behavior, which significantly aids in reducing false positives. Experimental results indicate a 15% improvement in F1 Score on the BGL dataset and a remarkable 60% improvement on the Spirit dataset compared to previous leading semi-supervised learning algorithms.¹

Log Parsing and Embedding

These are foundational steps in preparing raw log data for subsequent analysis, where LLMs offer significant enhancements.¹

- **LLMeLog:** This novel anomaly detection approach utilizes LLMs to enrich log events through in-context learning techniques. It fine-tunes a pre-trained BERT model with these enriched log events and subsequently trains a transformer-based anomaly detection model. Evaluation results on three public log datasets show that LLMeLog achieves F1-scores exceeding 99%, even when using only 10% of labeled data for training, highlighting its efficiency in data utilization.¹
- **Paper 17's Approach (AuditLLM-Hybrid Approach in Table):** This methodology employs LLMs for generating regular expressions (REGEX) for accurate log parsing and utilizes a pre-trained LLM for robust event embedding. It also incorporates an autoencoder with an attention mechanism for unsupervised learning of log patterns. This hybrid

framework achieved 96% accuracy on a public dataset for anomaly detection.¹

Multi-Agent Collaboration

An emerging and sophisticated strategy in threat hunting that involves the coordination of multiple intelligent agents to execute complex tasks.¹

- **Audit-LLM:** This multi-agent framework is specifically designed for log-based insider threat detection. It comprises three collaborating agents: the Decomposer, which breaks down complex tasks using Chain-of-Thought (COT) reasoning; the Tool Builder, which creates reusable tools for sub-tasks; and the Executor, which generates the final detection conclusion by invoking constructed tools. The framework also integrates a pair-wise Evidence-based Multi-agent Debate (EMAD) mechanism to refine conclusions through iterative reasoning exchange. Extensive experiments demonstrate its superiority over existing baselines, with significant improvements in the faithfulness of explanations generated by LLMs.¹

System State Analysis and Fingerprinting

Moving beyond traditional log analysis, this approach leverages LLMs to create abstract, information-rich representations of system states for advanced change detection and anomaly identification.

- **From Log to Fingerprint:** This novel approach transforms raw log data and other system metrics into a unique "fingerprint" – a compact, information-rich representation that captures the essence of a complex system at any given moment.⁵ This method moves beyond simple data dumps to intelligent compression, enabling a new level of abstraction for working with system states.⁵ It employs multidimensional projection to map hundreds or thousands of metrics into an optimized multidimensional space, revealing hidden correlations and dependencies.⁵ The generated fingerprints are optimized for efficient storage and retrieval, which is crucial for real-time analysis.⁵ The core innovation lies in comparing these fingerprints to instantly detect even the tiniest differences between two system states, eliminating the need to manually pore over gigabytes of logs.⁵ This breakthrough enables real-time anomaly detection, tracking of system drift, performance optimization, and accelerated debugging and troubleshooting by providing a precise picture of changes that may have caused a failure.⁵ It is described as a paradigm shift in monitoring and managing complex systems, leading to increased stability, predictability, and efficiency.⁵

The descriptions of LogLLM, LLMeLog, and Paper 17's approach clearly illustrate that LLMs are not merely standalone solutions. Instead, they are being integrated as powerful components that augment and enhance existing deep learning and machine learning techniques. LLMs act as sophisticated feature extractors, semantic interpreters, or reasoning engines that feed into or complement more traditional anomaly detection or classification models. This indicates a hybrid approach where LLMs provide advanced understanding of unstructured text, while

other ML models handle specific classification or pattern recognition tasks. This synergistic trend suggests that the most effective future LLM-based cybersecurity solutions will likely be hybrid architectures. These will leverage the unique strengths of LLMs, such as complex pattern recognition and natural language understanding, alongside specialized ML models for optimized classification, real-time performance on structured data, or specific statistical anomaly detection. For organizations, this implies that existing ML infrastructure may not need to be entirely replaced but can be significantly enhanced by integrating LLMs as a powerful new layer for deeper contextual analysis of log data.

Audit-LLM explicitly highlights significant improvements in the faithfulness of explanations generated by LLMs, achieved through its Evidence-based Multi-agent Debate (EMAD) mechanism.¹ This directly addresses one of the major challenges in LLM deployment: the need for interpretability and explainability. The fact that a multi-agent framework is designed with explainability as a core outcome, rather than an afterthought, signifies a growing recognition within the research community that complex AI systems deployed in high-stakes domains like cybersecurity must be transparent. The EMAD mechanism, by promoting iterative reasoning exchange, inherently builds a more traceable decision-making process. This trend indicates a crucial shift in the evaluation criteria for future LLM-based cybersecurity research and development. Beyond traditional performance metrics, there will be an increasing emphasis on the trustworthiness, transparency, and actionability of LLM outputs. For threat hunting, where human security analysts make critical, high-impact decisions, understanding the AI's reasoning is paramount. This enhanced explainability will foster greater trust in LLM-based systems, reduce "alert fatigue" by providing context, and enable analysts to more effectively validate and act upon detected threats, ultimately improving overall incident response efficacy.

Table: Comparison of LLM-Based Log Analysis Frameworks

Framework	Key Features	Performance Metrics	Citation
LogLLM	Leverages BERT for semantic vector extraction and Llama for classification	Outperforms state-of-the-art methods in semantic understanding and anomaly detection	¹
LogRAG	Semi-supervised framework with RAG and single-class classifier; uses log tokens and templates for phased detection	15% improvement in F1 Score on BGL dataset, 60% improvement on Spirit dataset	¹

Framework	Key Features	Performance Metrics	Citation
LLMeLog	Enriches log events with in-context learning and fine-tunes BERT; trains a transformer-based anomaly detection model	Achieves F1-scores exceeding 99% on public datasets, even with 10% labeled data	1
Audit-LLM / Paper 17's Approach (Hybrid)	Multi-agent collaboration with Decomposer, Tool Builder, and Executor agents; combines LLM-based parsing with autoencoder and attention mechanism	Demonstrates superiority over existing baselines with improved explanation faithfulness; Achieves 96% accuracy on public dataset	1
From Log to Fingerprint	Generates unique "fingerprints" from system snapshots; uses multidimensional projection and efficient storage/retrieval for real-time analysis; compares fingerprints for change detection	Enables real-time anomaly detection, tracking system drift, performance optimization, and faster debugging; described as a "breakthrough" and "paradigm shift" for efficiency, deeper insight, scalability, and proactive management	5

IV. Critical Assessment of Challenges and Limitations

Despite the promising advancements in applying LLMs to log analysis for threat hunting, several significant obstacles and limitations persist. These challenges can broadly be categorized into core operational issues and deeper, emerging concerns related to the inherent nature of LLMs.

In-depth Analysis of Core Challenges

- Scalability:** A primary hurdle is the ability of LLMs to handle the immense and ever-increasing volumes of log data generated in modern cybersecurity environments. As log data expands, the computational resources, including processing power and memory, required for effective analysis by LLMs increase proportionally. This poses a significant limiting factor for

organizations with restricted resources.¹ The inherent computational intensity of LLM inference and training, especially for continuous, real-time analysis of vast log streams, presents a substantial practical barrier to widespread adoption.

- **False Positives:** While LLMs demonstrate impressive capabilities in anomaly detection, their reliance on patterns and contextual understanding can occasionally lead to the generation of erroneous alerts. These false positives flag benign activities or normal system behaviors as potential threats, leading to "alert fatigue" and consuming valuable time and resources for security teams investigating non-existent threats. This ultimately diminishes the overall effectiveness and efficiency of threat hunting efforts.¹
- **Integration Complexity:** The incorporation of LLM-based solutions into existing cybersecurity frameworks is often complex. Many organizations have well-established workflows, tools such as Security Information and Event Management (SIEM) and Security Orchestration, Automation, and Response (SOAR) platforms, and existing infrastructure for log analysis and threat detection. Integrating new LLM components demands substantial effort and resources, including potential re-architecture of data pipelines, development of custom connectors or Application Programming Interfaces (APIs), and ensuring seamless communication and data flow between disparate systems.¹
- **The Critical Need for Interpretability and Explainability (XAI):** The decisions and reasoning processes of complex LLM-based models are often opaque, making them akin to "black boxes." This lack of transparency makes it difficult for security analysts to comprehend precisely why a particular alert was generated or how the LLM arrived at its conclusion. This opaqueness can erode trust in the LLM-based system, leading to hesitation in acting on recommendations and hindering the ability of analysts to refine, debug, or improve the models over time.¹

Expanded Challenges from Broader Research

Beyond the challenges identified in the initial review, broader research highlights several critical concerns that are paramount for the secure and effective deployment of LLMs in cybersecurity.

- **Vulnerabilities to Adversarial Attacks:** LLMs are susceptible to various adversarial attacks that can compromise their integrity and effectiveness.
 - **Prompt Injections:** Attackers can manipulate LLM input prompts to bypass security restrictions, generate unauthorized or malicious outputs, or leak sensitive information. This form of attack targets the core logic and instruction following of the model, making it a significant threat to applications relying on LLMs.²
 - **Data Poisoning:** Malicious or biased data can be subtly introduced into the training datasets of LLMs. This can compromise the integrity

and behavior of the model, leading to flawed outputs, reduced accuracy, and the potential leakage of sensitive information embedded within the dataset, particularly during the training or fine-tuning phases.²

- **Model Inversion Attacks:** Through carefully crafted queries, attackers can exploit LLM vulnerabilities to infer sensitive information from the model's outputs, potentially reconstructing private data that was part of the training set.²
- The impact of these adversarial inputs on threat detection is profound: they can be designed to evade LLM-based threat detection systems, allowing malicious activities to go unnoticed. Conversely, they can also lead to degraded decision-making, such as misclassifying benign activities as malicious or vice-versa.³ Research indicates alarmingly high False Positive Rates (FPRs) for ML classifiers (up to 97%) and even for advanced LLMs like ChatGPT-4o (up to 75%) when subjected to evasion attacks, underscoring their susceptibility.⁴ This demonstrates a direct causal relationship where adversarial attacks significantly exacerbate the existing problem of false positives in LLM-based threat hunting. Furthermore, beyond merely increasing false positives, these attacks introduce new, fundamental security risks to the LLM itself, such as data exfiltration, compromised model integrity, and the ability to remotely manipulate or control model behavior. These are not merely operational challenges but deep-seated vulnerabilities of the AI model that can undermine its trustworthiness and effectiveness. The inherent susceptibility of LLMs to adversarial attacks means that relying solely on them for critical threat hunting functions without robust, integrated defense mechanisms is inherently risky. This shifts the traditional cybersecurity attack surface from network and system vulnerabilities to the AI model itself, necessitating the development and deployment of a new generation of defensive strategies, including adversarial training and the integration of additional verification components within the cybersecurity intelligence pipeline.³
- **Data Privacy and Ethical Considerations:** The very nature of LLMs introduces significant data privacy and ethical concerns.
 - **Sensitive Data Leakage:** Due to their vast internal memory and the nature of their training on extensive datasets, LLMs can unintentionally generate content that closely resembles their training data. This poses a significant risk of leaking personal, proprietary, or otherwise sensitive details during both the training and inference phases.⁶
 - **Model Bias:** LLMs can inadvertently perpetuate and amplify biases present in their vast training data, which can be difficult to detect and mitigate, especially in "black box" models.⁷ While some forms of

bias are functional and desirable, such as an LLM being biased towards generating grammatically correct sentences, harmful biases, like racial assumptions in the analysis of legal case studies or favoring certain demographic groups in recruitment scenarios, demand active and continuous mitigation.⁸

- **Hallucinations:** A notable characteristic of LLMs is their propensity to "hallucinate" – generating plausible-sounding but inaccurate, nonsensical, or entirely fabricated information. In a domain like cybersecurity threat hunting that demands absolute accuracy, inventing facts, misrepresenting concepts, or fabricating details can have severe operational consequences.⁷
- **Accountability and Attribution:** Ethical concerns also arise from the lack of clear accountability for content generated by LLMs and their potential failure to properly attribute original sources or even fabricate fake sources. This can undermine the credibility of threat intelligence reports and analyses.⁷ These are not merely "good practices" but fundamental challenges stemming directly from the nature and operation of LLMs. Data leakage is a direct consequence of the model's memory and training data characteristics. Bias is inherent to any machine learning model, and hallucinations are a known behavioral trait of generative LLMs. These issues directly impact the trustworthiness, reliability, and legal/ethical compliance of LLM-based systems, especially when processing highly sensitive cybersecurity log data. Addressing these profound issues necessitates a "privacy-by-design" and "ethics-by-design" approach throughout the entire LLM development and deployment lifecycle. This includes implementing data anonymization, secure model serving, and privacy penetration tests.⁸ It also implies the need for clear disclaimers about the probabilistic nature and potential inaccuracies of LLM outputs, alongside robust data governance policies, particularly for organizations operating in highly regulated industries.
- **Real-time Performance and Latency Challenges:** While LLMs offer sophisticated analytical capabilities, they face significant challenges in maintaining quick and responsive performance, which is a critical requirement for real-time threat hunting and incident response scenarios.⁹ Delayed responses can lead to missed opportunities, suboptimal outcomes, or even unsafe or incorrect behavior in dynamic cybersecurity environments where rapid decision-making is paramount.¹⁰
 - The overall latency in LLM applications is a multifaceted concept influenced by several key factors. These include the length of the input prompt, network conditions and geographic location, the model's size and complexity, the length of the generated response, the complexity of the task and prompt, system load, and any post-processing requirements, such as security checks or guardrails.⁹ Different tokenization strategies employed by various models can

also impact the effective processing time for the same input.⁹ The initial review primarily flags "scalability" as a challenge, which pertains to handling the volume of data. However, the speed of response, or "latency," is a distinct challenge. A system can be scalable, able to handle massive data volumes, but still suffer from high latency, meaning slow response times. For threat hunting, which inherently requires proactive and often rapid decision-making to preempt or mitigate threats, low latency is paramount. If an LLM takes too long to analyze a critical log stream or generate an actionable insight, the threat might have already materialized, escalated, or caused significant damage. This represents a direct causal relationship where high latency directly leads to degraded operational performance and missed opportunities in a time-sensitive domain. Optimizing LLMs for real-time performance in cybersecurity demands specialized architectural considerations beyond merely increasing raw computational power. This includes adopting strategies like concise prompt engineering, breaking down complex tasks into smaller, manageable chunks, smart context management, and potentially leveraging distributed processing or optimized inference techniques.⁹ It also suggests that for certain high-frequency, ultra-low-latency threat hunting tasks, more specialized, smaller-scale models might be more appropriate than large, general-purpose LLMs.

Table: Key Challenges and Mitigation Strategies for LLM-based Threat Hunting

Challenge	Description	Impact on Threat Hunting Operations	Proposed Mitigation Strategies / Future Research Directions
Scalability	Difficulty handling vast, ever-increasing volumes of log data due to high computational resource demands.	Limits widespread adoption, especially for organizations with restricted resources; hinders comprehensive, real-time analysis.	Explore distributed processing, parallel computing, optimized pre-training strategies; develop more efficient LLM architectures.

Challenge	Description	Impact on Threat Hunting Operations	Proposed Mitigation Strategies / Future Research Directions
False Positives	LLMs generate erroneous alerts based on patterns/context, flagging benign activities as threats.	Leads to "alert fatigue," wastes security team time and resources, diminishes overall effectiveness of threat hunting.	Refine anomaly detection algorithms (e.g., LogRAG's single-class classifier); integrate human feedback loops; improve contextual understanding .
Integration Complexity	Substantial effort and resources required to incorporate LLMs into existing cybersecurity workflows, tools (SIEM, SOAR), and infrastructure.	Slows adoption, requires significant re-architecture, custom development, and seamless data flow adjustments.	Develop standardized APIs and connectors; offer modular LLM components; emphasize interoperability with existing security ecosystems.
Interpretability & Explainability (XAI)	Opaque decision-making processes ("black box" nature) make it difficult for analysts to	Erodes trust in LLM-based systems; hinders validation and refinement of models; makes acting on	Develop more transparent LLM architectures; apply XAI techniques (e.g., LIME); design

Challenge	Description	Impact on Threat Hunting Operations	Proposed Mitigation Strategies / Future Research Directions
	understand LLM conclusions.	recommendations difficult.	systems for "faithfulness of explanations" (e.g., Audit-LLM's EMAD).
Adversarial Attacks	LLMs vulnerable to prompt injection, data poisoning, model inversion, leading to manipulated outputs or data leakage.	Threat detection evasion, compromised model integrity, degraded decision-making, increased false positives/negatives.	Implement adversarial training; integrate additional verification components in CTI pipelines; robust input validation and sanitization.
Data Privacy & Ethical Concerns	Risk of sensitive data leakage from training/inference; perpetuation of harmful biases; generation of fabricated info (hallucinations); lack of accountability.	Compromises data confidentiality; leads to unfair or inaccurate threat assessments; undermines credibility of intelligence.	Adopt "privacy-by-design" and "ethics-by-design"; data anonymization; secure model serving; continuous bias monitoring; clear disclaimers on LLM outputs.

Challenge	Description	Impact on Threat Hunting Operations	Proposed Mitigation Strategies / Future Research Directions
Real-time Performance & Latency	LLMs can have slow response times, critical for proactive threat hunting and incident response.	Missed opportunities, suboptimal outcomes, inability to respond rapidly to dynamic threats.	Optimize prompt engineering (concise prompts, task breakdown); smart context management; efficient token management; explore low-latency inference techniques.

V. Future Directions and Strategic Recommendations

The application of LLMs in threat hunting through log analysis is a rapidly evolving field, with several promising avenues for future research and strategic recommendations for practical implementation.

Advancing Interpretable and Explainable LLM Models (XAI)

The development of more interpretable and explainable LLM models is crucial for fostering trust among security analysts and enabling them to effectively refine and improve these models over time.¹ LLMs themselves can be leveraged to enhance XAI by transforming complex machine learning outputs into easy-to-understand narratives, thereby bridging the gap between sophisticated model behavior and human interpretability.¹¹ Future efforts must focus on key XAI principles: transparency, which involves making the operations and decision-making processes of AI systems clear; interpretability, referring to the extent to which a model's behavior can be comprehended by humans; and explainability, which is about providing clear reasons for decisions. These principles are essential to address the "black box" problem and bolster model security.¹² Techniques like LIME (Local Interpretable Model-Agnostic Explanations) are relevant tools for achieving local interpretability by explaining individual predictions.¹²

Developing Domain-Specific and Robust LLM Architectures

While general-purpose LLMs have shown impressive capabilities in log analysis, tailoring them for cybersecurity-specific tasks by fine-tuning on relevant datasets and adapting them to the unique requirements of log analysis can significantly

boost performance and ensure greater relevance of results.¹ This includes the exploration and development of specialized LLM architectures and models explicitly optimized for various cybersecurity functions, moving beyond generic language understanding to deep domain-specific intelligence.¹³

Enhancing Scalability and Real-time Processing Capabilities

It is essential to design and develop more robust and scalable frameworks that can efficiently process and analyze the ever-growing volume and complexity of log data while maintaining high performance.¹ This requires exploring advanced techniques such as distributed processing, parallel computing, and optimized pre-training strategies to enhance the scalability and efficiency of LLM-based systems.¹

Addressing the latency challenges for real-time threat hunting involves optimizing factors like input prompt length, model size, network conditions, and post-processing requirements. Strategies include keeping prompts concise, breaking down complex tasks into smaller, manageable chunks, smart context management, and efficient token management.⁹

Integrating Human-in-the-Loop (HITL) for Enhanced Decision-Making

Human-in-the-Loop systems are critical for integrating human expertise into automated processes, thereby improving threat detection, response accuracy, and overall decision-making in cybersecurity.¹⁴ In HITL models, human analysts are involved at key decision points: validating alerts generated by LLMs, approving automated responses, labeling training data to improve model accuracy, and overriding automation when necessary. This ensures that automated systems are informed, contextual, and accountable.¹⁴ This approach is particularly vital in Security Operations Center (SOC) environments and Security Orchestration, Automation, and Response (SOAR) platforms, serving as a crucial bridge between automated tooling and indispensable human judgment.¹⁴

The initial literature review lists interpretability and explainability as a "future direction." However, when considering the expanded challenges identified in this report—specifically, the high rates of false positives, the severe threats posed by adversarial attacks, and the inherent ethical concerns like bias and hallucinations—XAI and HITL cease to be merely desirable enhancements and become operational imperatives. If LLM-based systems remain opaque "black boxes" or are susceptible to manipulation, human security analysts cannot effectively trust, validate, or act upon their outputs. HITL directly addresses this by building essential human oversight and intervention points into the system, while XAI aims to make the black box transparent and understandable. The direct causal link is clear: without robust XAI and HITL, the challenges of untrustworthy outputs, alert fatigue, and vulnerability to adversarial attacks become insurmountable barriers to real-world, effective deployment. This profound shift means that future research and development in LLM-based threat hunting must prioritize human-AI collaboration and transparency from the earliest design stages, rather than treating them as optional add-ons. It advocates for a paradigm where AI augments human intelligence, enabling security teams to leverage LLMs' power while maintaining critical human judgment and accountability, moving away from purely automated "lights-out" security.

The Imperative of Standardized Benchmarking and Evaluation

Standardized benchmarking is essential for objectively measuring progress, ensuring reproducibility of results, driving fair comparisons across different LLMs and frameworks, identifying strengths, exposing failure modes, and ultimately guiding real-world deployment decisions.¹⁵ Future efforts must focus on utilizing and further developing specialized cybersecurity LLM benchmarks, such as CTI-Bench and SecEval. These benchmarks evaluate LLMs across diverse tasks, including factual recall, causal mapping, vulnerability severity prediction, and threat actor attribution, drawing from authoritative sources like MITRE ATT&CK and NVD.¹⁵ It is crucial to address existing gaps in log analysis benchmarking, such as the limited availability of diverse datasets, for example, covering both web application and system logs comprehensively, and the narrow range of LLM architectures currently explored in evaluation studies.¹⁶ The initial review vaguely mentions the need for "more robust and scalable frameworks" as a future direction. However, the specifics of benchmarking efforts highlight that benchmarking is not merely an academic exercise for comparing models; it is explicitly about identifying strengths, exposing failure modes, and guiding real-world deployment. The identified limitations of current benchmarks underscore that existing evaluation methodologies are often insufficient to fully capture the complexity and dynamism of real-world cybersecurity threats. Therefore, the development and continuous refinement of robust, standardized benchmarking frameworks become a prerequisite for establishing the reliability, trustworthiness, and practical utility of LLM-based threat hunting solutions, directly bridging the gap between theoretical research and operational deployment. This emphasizes the critical need for collaborative efforts between academia and industry to create comprehensive, diverse, and continuously updated benchmarks. These benchmarks should include realistic, anonymized log datasets representing various attack vectors, system types, and potentially multilingual data. Such rigorous evaluation will be crucial for validating LLM performance in dynamic threat landscapes, fostering confidence in their capabilities, and accelerating their responsible adoption in operational security environments.

Addressing Practical Adoption: Cost-Benefit Analysis for LLM Solutions

Beyond technical feasibility, the economic viability of LLM-based cybersecurity solutions is a crucial factor for organizational adoption. Organizations must conduct comprehensive cost-benefit analyses, quantifying all associated costs, such as hardware and software expenses, employee salaries, training, operational expenses, and potential costs related to incidents, against tangible and intangible benefits, such as risk reduction, compliance benefits, reputation protection, competitive advantage, and incident response cost savings.¹⁷ Performing sensitivity analysis and risk assessment is vital to account for uncertainties in cost and benefit estimates, providing a more robust foundation for strategic investment decisions.¹⁷

VI. Evaluation of the Original Literature Review's Scope and References

The original literature review, "Cyber Security Threat Hunting Through Log Analysis Using LLM" ¹, provides a foundational overview of the topic, demonstrating several strengths while also exhibiting notable gaps.

Strengths in Covering Core LLM Applications and Methodologies

The original literature review successfully identifies and describes the foundational role of LLMs in log analysis for threat hunting. It provides a clear overview of key LLM applications, including specific frameworks like LogLLM and LogRAG for anomaly detection, LLMeLog and Paper 17's approach for log parsing and embedding, and Audit-LLM for multi-agent collaboration.¹ It effectively conveys the potential of LLMs to automate and enhance threat detection by processing unstructured log data and discerning patterns, laying a solid conceptual foundation for the topic.

Identified Gaps in Addressing Emerging Challenges and Practical Considerations

While the review acknowledges important challenges such as scalability, false positives, integration complexity, and the need for interpretability ¹, it notably understates or entirely overlooks several critical challenges that are highly relevant to the real-world deployment and long-term viability of LLM-based threat hunting solutions. The review does not address the significant and growing threats posed by adversarial attacks, including prompt injection, data poisoning, and model inversion, nor does it discuss the inherent data privacy risks associated with LLM training and operation, or the crucial real-time performance and latency requirements that are essential for effective, proactive threat hunting in dynamic environments.² Furthermore, the review lacks discussion on vital practical aspects such as the necessity of Human-in-the-Loop (HITL) integration for robust decision-making, the importance of comprehensive and standardized benchmarking for validating LLM performance, and the imperative of conducting thorough cost-benefit analyses to justify organizational investment in LLM solutions.¹⁴ These considerations are paramount for transitioning LLMs from theoretical potential to deployable, reliable solutions.

Assessment of the Relevance and Direct Applicability of the Cited Works

The review effectively cites several papers that are highly relevant and directly applicable to LLM-based log analysis for threat hunting. These include papers detailing the specific frameworks and methodologies discussed, such as LogLLM, LogRAG, LLMeLog, and Audit-LLM, which form the core technical discussion of the review.¹ Additionally, the newly added "From Log to Fingerprint" ⁵ is directly applicable as it describes a methodology for leveraging LLMs to read logs and report based on system state analysis. However, a significant portion of the cited references, as explicitly detailed in the analysis ¹, state that they do not specifically address threat hunting using LLMs for log analysis. These papers often cover broader applications of LLMs in cybersecurity, for example, vulnerability detection, malware analysis, digital forensics, or general cyber threat intelligence, or focus on log analysis techniques without the explicit "threat hunting" context. While these broader papers provide valuable background and context for the overall field of LLMs in cybersecurity, their direct applicability to the precise scope of "threat

hunting through log analysis using LLM" is limited. This observation reveals a clear pattern: a substantial number of the cited works explicitly state that their content does not directly align with "threat hunting using LLMs for log analysis." This indicates a potential contradiction between the stated, narrow scope of the original literature review and the broader nature of its supporting evidence. This suggests that while the author aimed to cover the topic, the available specific literature might be limited, leading them to draw from more general LLM-in-cybersecurity or log analysis papers. This implies two main possibilities: either the specific sub-field of "LLM for threat hunting specifically through log analysis" is still nascent, requiring researchers to synthesize insights from broader, related domains; or the original review could have benefited from a more stringent selection of references to maintain a sharper, more focused scope, or a clearer acknowledgment of how the broader context provided by less direct citations informs the specific topic. For an expert-level report, this signals a potential area for improvement in the original review's methodological rigor regarding literature selection and its direct relevance to the stated thesis.

VII. Conclusion

The integration of Large Language Models into cyber threat hunting, particularly through log analysis, undeniably represents a powerful and transformative frontier in cybersecurity. LLMs offer a compelling shift from traditional, often reactive, rule-based systems to automated, proactive, and semantically-driven threat detection capabilities. Current research demonstrates promising advancements across core applications, including sophisticated anomaly detection frameworks, highly efficient log parsing and embedding techniques, innovative multi-agent collaborative systems, and novel system state analysis approaches like "From Log to Fingerprint." These methodologies have shown superior performance metrics in specific tasks, highlighting the significant potential of LLMs to enhance threat intelligence and response.

However, the path to widespread, reliable, and secure deployment of LLM-based threat hunting solutions is fraught with significant, multifaceted challenges. These extend beyond the commonly acknowledged issues of scalability and false positives to encompass critical security vulnerabilities, such as susceptibility to adversarial attacks, inherent ethical dilemmas, including model bias and data privacy risks, and crucial operational constraints, like real-time latency requirements.

To navigate this complex landscape, the following recommendations are crucial for both future research and practical implementation:

- **Holistic Approach:** Future research and development must adopt a holistic approach, simultaneously addressing technical performance and efficiency, mitigating inherent security vulnerabilities, navigating complex ethical implications, and considering practical deployment constraints. A siloed approach will hinder effective and responsible integration.
- **Human-Centric Design:** It is paramount to prioritize the development of interpretable and explainable LLM models (XAI) and to integrate robust

Human-in-the-Loop (HITL) systems. This will foster essential trust in AI-driven insights, enable effective human-AI collaboration, and provide critical human oversight to mitigate the risks associated with autonomous decision-making in high-stakes cybersecurity scenarios.

- **Robust Evaluation:** Significant investment is required in developing and utilizing standardized, comprehensive benchmarking frameworks. These benchmarks must accurately reflect the complexity and dynamism of real-world cybersecurity scenarios, including diverse log types, various attack vectors, and multi-modal data, to rigorously validate LLM performance and identify their true limitations.
- **Strategic Investment:** Organizations considering the adoption of LLM-based threat hunting solutions must conduct thorough and realistic cost-benefit analyses. This analysis should factor in not only direct implementation and operational costs but also the potential for substantial reductions in incident response costs, improved regulatory compliance, and enhanced reputational resilience, thereby providing a clear business case for investment.
- **Adaptive Security Posture:** Finally, it is crucial to recognize that LLM-based threat hunting systems, like any advanced technology, will require continuous monitoring, adaptation, and the implementation of robust defensive mechanisms to counter the ever-evolving landscape of adversarial techniques and emerging cyber threats. This necessitates an agile and adaptive security posture around these AI systems.