

**Cyber Security Threat Hunting Through Log Analysis Using LLM –
Research Methodology , Graduation Project 2 – second semester**

Student : OMAR SALEM ALMUTAIRI

Student ID: 251002148

MIDOCEAN UNIVERSITY

FOR MASTER OF AI PROGRAM

GRADUATION PROJECT – second semester

Research Methodology: Systematic Literature Review on Cyber Security Threat Hunting Through Log Analysis Using Large Language Models with Agentic Action

I. Foundational Design and Review Protocol

A. Review Type and Justification

The study utilized a Systematic Literature Review (SLR) methodology to rigorously synthesize findings from over 350 finalized research papers. This comprehensive approach was chosen specifically because the domain of Large Language Models (LLMs) and Agentic Artificial Intelligence (AI) in cybersecurity is nascent and rapidly evolving, requiring a precise and replicable method to establish the state-of-the-art.

A conventional narrative review would lack the structure necessary to manage the extensive volume of identified literature (tens of thousands of initial records) and would be susceptible to selection bias. Conversely, the SLR, by adhering to a predefined protocol, ensures that the investigation into LLM-powered agentic systems remains meticulous and impartial. This methodology ensures the resulting comprehensive understanding of LLM-driven autonomous threat hunting systems is both credible and reproducible for other researchers in the computer security domain. The justification for this rigorous secondary research approach rests on the need to address specific, empirical questions regarding the effectiveness, architectural characteristics, and operational limitations of these complex systems.

B. Protocol Registration and Reporting Guidelines

To maximize transparency and reliability, the review adhered strictly to established systematic review guidelines, primarily leveraging Kitchenham’s framework, widely accepted in software engineering , and integrating the standards set by the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines.

The commitment to transparent reporting mandated that the complete review protocol—detailing the exact search strings, eligibility criteria, and data extraction forms—was pre-specified and documented (e.g., in a dedicated repository). This measure is critical in preventing the *post hoc* manipulation of methods based on observed outcomes, reinforcing the integrity of the findings. The subsequent report explicitly uses the PRISMA flow diagram format to transparently document the flow of records identified, screened, and included in the final synthesis corpus.

C. Research Question Formalization (PICOs Adaptation)

The core research questions were formalized using a modified Population, Intervention, Comparison, Outcome, and Study Type (PICOs) framework. Given the technical nature of Agentic AI and cybersecurity, direct application of the traditional clinical PICO framework is insufficient; therefore, adaptation was necessary to capture relevant technical granularity.

The following table summarizes the adaptation of the PICOs framework for this specific systematic review:

Category	Definition for this SLR	Methodological Implication
Population (P)	Cybersecurity data, specifically log analysis data (e.g., SIEM events, network traffic, system logs, cloud Indicators of Compromise (IoCs)).	Focuses empirical research on data-centric threat hunting, excluding studies limited to policy or broad governance frameworks where specific log data is absent.
Intervention (I)	Large Language Models (LLMs) used in an Agentic framework, demonstrating "Agentic Action" through interaction with external tools, APIs, and persistent memory (VectorDB/Long-Term Memory - LTM).	This narrow definition excludes simple LLM prompt engineering without autonomous interaction, focusing the review exclusively on genuinely autonomous defense systems.
Comparison (C)	Traditional cybersecurity models, including signature-based detection, anomaly-based detection, behavioral models, or human security analysts.	Ensures the selection of studies that benchmark the comparative advantages and disadvantages of LLM-driven agents over existing security models.
Outcome (O)	Detection efficacy (e.g., rule generation correctness, precision/recall, operational viability, adversarial robustness) and architectural characteristics (e.g., memory management, multi-agent coordination complexity).	Guarantees that the data extraction phase captures both quantitative performance metrics and qualitative structural details necessary for taxonomy development.

Table 1: PICOs Framework Adaptation for LLM Agentic Threat Hunting SLR

The careful tailoring of the PICOs structure is essential for this domain. Specifically, defining the 'Intervention' to require "Agentic Action" (i.e., tool use and persistent memory) allows the reviewers to differentiate between exploratory LLM applications and fully orchestrated, autonomous systems. This methodological distinction drives the specificity of the search and the subsequent data extraction form (Section V), ensuring that the captured evidence aligns precisely with the goal of analyzing autonomous threat hunting capabilities.

II. Systematic Search Strategy and Information Retrieval

A. Selection of Target Databases and Repositories

To ensure maximum comprehensiveness across computer science, engineering, and specialized security literature, the systematic search targeted six primary academic databases, complemented by grey literature searches :

1. IEEE Xplore
2. ACM Digital Library

3. Scopus
4. Web of Science (WoS)
5. Springer/ScienceDirect (Grouped)
6. arXiv (Recognizing the crucial role of pre-prints for timely research in fast-moving fields like LLM and AI security)

The database selection reflects a strategic effort to capture peer-reviewed literature alongside organizational reports and conference proceedings, which often contain early insights into technology adoption in cybersecurity.

B. Iterative Keyword Elicitation and Refinement

The search terms were developed iteratively based on established protocols for high-volume SLRs. The process began with the identification of a Quality Guiding Set (QGS) of approximately 41 highly relevant studies, which were manually curated and validated against the PICO's criteria. Word frequency analysis performed on the titles and abstracts of the QGS papers provided the foundation for eliciting the final, comprehensive set of search keywords.

Keywords were stratified into three mandatory groups linked by the Boolean operator AND, while terms within each group were linked by OR to maximize recall:

1. **AI/LLM Core:** Terms included, but were not limited to: "Large Language Model," LLM, GPT-*, ChatGPT, T5, Llama, AIGC, and AGI.
2. **Agentic/Autonomous Action:** Key terms reflecting the intervention definition: "Autonomous Agent," "Agentic AI," "Cognitive Autonomy," "Chain-of-Thought," CoT, and "Multi-Agent Coordination".
3. **Cybersecurity Domain:** Terms focusing on the specific application: "Threat Hunting," "Log Analysis," "Cyber Security," "Threat Detection," CTI, and SIEM.

C. Boolean Query Construction and Syntax Translation

The comprehensive search string combined these groups using Boolean logic to ensure precision. Wildcards (e.g., * for truncation) were applied judiciously to capture grammatical variations and model versioning (e.g., model* or GPT-*).

A hypothetical general Boolean query structure illustrates the required complexity:

```
((("Large Language Model*" OR LLM OR GPT-* OR "Language Model") AND ("Threat Hunting" OR "Log Analysis" OR "Cyber Security" OR "Threat Detection" OR CTI OR SIEM) AND ("Agentic AI" OR "Autonomous Agent*" OR "Chain of Thought" OR CoT OR "Multi-Agent*")))
```

The formulation of precise search strings is pivotal to a successful systematic review, but syntax varies significantly across databases. The process involved meticulously translating the optimal Boolean query structure into the specific query language required by each database (e.g., WoS vs. IEEE Xplore). While LLM tools can be leveraged as an aid to generate refined search strings and translate

Boolean queries, this assistance required careful manual cross-referencing and verification. Researchers must confirm the accuracy of LLM-generated translations, particularly regarding database-specific field codes and the potential limitation posed by the LLM’s knowledge cut-off date, which could lead to outdated domain terminology.

D. Search Execution, Range, and Supplementary Techniques

The search was executed with a defined date range, beginning in January 2018 (to capture the emergence of modern transformer models) and running through a hard cut-off date (e.g., August 2024). This initial comprehensive search resulted in an estimated corpus of 38,112 papers across all targeted databases.

To address the risk of missing papers with incomplete abstracts or missing keywords, supplementary techniques were mandatory. Snowballing was performed both backward (reviewing the references of included papers) and forward (identifying subsequent papers that cited the included papers) on the final 350+ studies. This process significantly augmented the initial set, helping to ensure that relevant, highly cited works were not inadvertently overlooked due to search term constraints.

III. Study Selection Process and Eligibility Criteria

The transition from an initial corpus of over 38,000 records to the final set of 350+ included papers required a rigorous, multi-stage selection process, strictly monitored using principles adapted from PRISMA and detailed in the following flow diagram. The process adhered to the four stages of screening: Plan, Pilot, Screen, and Log.

A. Managing the Initial Corpus

Criterion Category	Inclusion Criteria (IN)	Exclusion Criteria (EX)
Publication Type	Full research papers from peer-reviewed journals, archival top-tier conferences, or preprints on arXiv.	Short papers, poster papers, tool demonstrations, editorials, books, or theses.
Thematic Relevance	Empirical or architectural study detailing the application of LLM/Agentic systems to threat hunting or log analysis.	General CTI frameworks, policy discussions, or studies focused solely on LLM security risks without log analysis application.

Criterion Category	Inclusion Criteria (IN)	Exclusion Criteria (EX)
Methodological Scope	Must present a clear methodology (experimental setup, simulation, or formal architectural proposal).	Opinion pieces, commercial white papers, existing review articles, or grey literature that cannot be fully verified.
Language	Full text must be available and published in English.	Non-English publications.

All records retrieved from the database searches were immediately imported into specialized systematic review software (e.g., Rayyan) for efficient management. Automated and manual deduplication was performed to ensure that papers indexed in multiple sources were counted only once, eliminating thousands of redundant entries.

B. Development of Hierarchical Inclusion and Exclusion Criteria

Clear, unambiguous, and hierarchically organized eligibility criteria were defined to facilitate rapid exclusion during the screening phase.

C. Multi-Stage Screening Protocol

The selection process was conducted independently by two reviewers to minimize individual bias, with all conflicts resolved through consensus meetings or arbitration by a third senior reviewer.

1. Automated/Semi-Automated Filtering (Stage 1)

Automated exclusion rules were initially applied to metadata, removing records that met clear cut-off criteria, such as papers with less than eight pages or those labeled as workshop proceedings. This process efficiently reduced the corpus by thousands of records.

2. Title and Abstract Screening (Stage 2)

The remaining papers were assessed based on title and abstract using a pre-specified, hierarchical abstract screening tool. The protocol benefited from the use of an LLM-assisted system (e.g., based on GPT-4). This system was utilized to recommend initial inclusion/exclusion decisions and provide tentative rationales for exclusion. The reviewers then focused their manual efforts on cases where the LLM was uncertain or where reviewer decisions conflicted, thereby reducing the manual screening workload. This "human-in-the-loop" design allowed for real-time refinement of the PICO's criteria based on observed disagreements between the LLM and human reviewers, ultimately targeting a high inter-rater reliability, quantified by a Cohen's κ score of ≥ 0.70 .

Before initiating the main screening, multiple pilot tests were conducted. The review team screened the same 20 to 30 abstracts and repeated this process until a

high level of consensus was consistently achieved. This step was necessary because the interpretation of concepts like "Agentic Action" or "Threat Hunting" can vary, and pilot testing ensures that all reviewers apply the criteria consistently, maintaining methodological rigor across a large review scale.

3. Full-Text Review (Stage 3)

Papers passing abstract screening proceeded to full-text review. This final rigorous assessment ensured that the papers met all inclusion criteria, particularly the definition of "Agentic Action" and the quality requirements (Section IV). This phase finalized the corpus of the 350+ included studies.

Table 2 illustrates the study selection process:

Table 2: Study Selection and Screening Flow (PRISMA Adaptation)

Stage	Process Step	Number of Records	Exclusion Rationale
Identification	Records retrieved through database search (including supplementary methods)	N = ~38,112 (Estimate)	N/A
Screening 1	Automated Removal (Duplicates, Non-research/Short papers)	N = 7,277 (Example)	Duplication, less than 8 pages, tool demos, editorials, conference abstracts
Screening 2	Title/Abstract Screening (Dual Review + LLM-Assisted)	N = 6,507 (Example)	Scope misalignment (e.g., non-cyber security focus, no LLM component)
Eligibility	Full-Text Review	N = 450 (Example)	Failure to meet required quality criteria, lack of empirical data, or missing definition of Agentic Action.
Included	Final Studies for Synthesis	N = 350+	N/A

IV. Quality Assessment and Risk of Bias (RoB) Evaluation

A critical phase of the SLR involved evaluating the methodological rigor of the included 350+ studies. This step ensures that the conclusions drawn during synthesis are weighted appropriately based on the quality and robustness of the underlying evidence.

A. Instrument Selection and Adaptation

A customized Quality Assessment Instrument (QAI) was developed, adapted from established systematic review checklists used in Software Engineering. Crucially,

this instrument incorporated additional items specifically designed to address the unique reproducibility and security risks inherent in research involving emerging technologies like LLMs and Agentic AI. Two reviewers independently assessed the quality of each paper, followed by consensus meetings to resolve discrepancies.

B. Assessment Criteria for Reproducibility and Validation

Since reproducibility is often a significant challenge in LLM research, the QAI heavily weighted criteria related to technical transparency. Key assessment points included:

1. **Code and Prompt Availability:** Confirmation that the source code, LLM-specific parameters, and the exact methodology for prompt engineering were transparently described or made publicly available. This is essential for verifying LLM output.
2. **Dataset Integrity:** Verification that the cybersecurity dataset utilized (especially log data) was either publicly available or described in sufficient detail (e.g., source, volume, labeling methodology) to allow for replication. This mitigates the risk of bias introduced by proprietary, non-auditable training or evaluation data.
3. **Operational Viability and Adversarial Robustness:** Assessment extended beyond simple functional correctness. Studies were graded on whether they evaluated their detection rules holistically, incorporating real-world corpus testing to surface operational noise thresholds, maintainability, and performance guardrails. Furthermore, the study must have addressed the brittleness commonly associated with security detection rules by measuring vulnerability to trivial adversarial bypasses.

Studies using LLM-generated output were also required to use standardized evaluation metrics. The QAI assessed whether the paper clearly differentiated between a Perfect Match (case- and punctuation-insensitive equivalence) and a Partial Match, defining the latter via a high lexical and semantic similarity threshold (e.g., a $\geq 70\%$ token_set_ratio score).

C. Risk of Bias (RoB) Scoring and Weighting

The QAI included a specific dimension dedicated to assessing the Risk of Bias (RoB). This included evaluating potential biases across the AI lifecycle—specifically design, development, deployment, and maintenance—as identified in government assessments of AI security.

A crucial component was the assessment of Reporting Completeness in the technical domain. Prior systematic reviews focusing on AI models found that reporting completeness in the technical assessment domain can be low (historically measured around 39.9%). Studies that failed to fully report on LLM configuration, data handling, or evaluation metrics were flagged as having a high RoB.

The Quality Assessment score was explicitly integrated into the evidence synthesis process. This methodological link ensures that studies assessed as robust, highly

reproducible, and having low RoB—particularly those demonstrating rigorous operational testing and adversarial evaluation—received a greater weighting when

Assessment Domain	Criteria (Yes/No/Partial)	Weighting	Source/Rationale
Reproducibility	Full source code and prompts publicly available?	High	Essential for technical validation of LLM outputs
Data Integrity	Dataset (log data) publicly available or detailed for replication?	High	Addresses RoB from inaccessible or biased training data
Operational Viability	Evaluated against real-world corpus/noise thresholds?	Medium	Crucial for assessing practical utility in threat hunting
LLM Security/RoB	Does the study analyze internal security risks (e.g., adversarial robustness, over-agency)?	Medium	Addresses inherent security risks of the technology being studied
Reporting Completeness	Full reporting on technical assessment domain (e.g., metrics, architecture)?	High	Reduces bias associated with poor technical reporting

establishing trends, drawing conclusions, and constructing the final taxonomy. Conversely, studies with high RoB, such as those relying on proprietary data or lacking transparency regarding LLM prompting, contributed less significantly to the final evidence base.

Table 3 details the Custom Quality Assessment Instrument (QAI) for LLM4Security:

Table 3: Custom Quality Assessment Instrument (QAI) for LLM4Security

V. Data Extraction Framework and Categorization

Following the inclusion of the final 350+ papers, the data extraction phase commenced. This process systematically collected relevant information from the studies and organized it into a structured format to enable robust analysis and synthesis.

A. Design of the Structured Data Extraction Form

The data extraction form was carefully designed, pilot-tested, and standardized. It was structured around four key thematic sections: General Characteristics, LLM Configuration, Agentic Architecture, and Evaluation Metrics, directly reflecting the adapted PICO criteria (Section I.C).

Data extraction was conducted independently by two reviewers to minimize data

Category	Specific Data Points Extracted	Analytical Significance
LLM Core	Base Model (e.g., GPT-4, Llama), Fine-tuning Strategy (e.g., RAG, Instruction Tuning), Input Modality.	Necessary to link specific model capabilities and augmentation techniques to observed security outcomes.
Agentic Architecture	Agent Roles (e.g., Investigator, Responder), Orchestration Mechanism (Dynamic vs. Hard-coded pipeline), Tool Use (External APIs, Security Tools like SIEM), Use of Chain-of-Thought (CoT) reasoning.	To classify the system’s operational autonomy and complexity, particularly specialized coordination.
Context/Memory	Memory Type (Persistent VectorDB/LTM vs. Episodic/Stateless), Context Window Management, Knowledge Base source.	Essential for assessing the agent’s ability to conduct long-term, sequential threat hunting and retaining context.
Cyber Domain Focus	Log Source Used (Network, Endpoint, Cloud IoCs), Target Threat Class (e.g., LotL, APT, malware analysis).	To map technological applications to domain-specific security challenges and generalizability.
Automation/Reproducibility	Prompt detail, reproducibility of LLM output, source code and dataset availability, strategies for applying LLMs in the study.	Mandatory fields for technical validation and quality assessment integration.

entry errors and subjective interpretation, a best practice in systematic reviews. Discrepancies between reviewers were investigated and resolved through mandated consensus meetings, guaranteeing data fidelity before final synthesis.

B. Extraction Fields for LLM/Agentic Specifics (The Core Data Set)

The data collection was designed to capture granular architectural and performance details necessary for developing a robust taxonomy of LLM agentic systems.

The extraction process was uniquely challenging due to the need to capture qualitative, architectural descriptions—such as how "dynamic meta-agents" function or the mechanisms for "role-specialized coordination". Unlike quantitative metrics, these data points require interpretive reading of complex system designs. To mitigate the risk of subjective interpretation, the reviewer training focused heavily on precisely defining specialized terminology related to

agent orchestration and memory management, ensuring consistency across the dual-reviewer team. The successful extraction of these specific items acts as the bridge between reported study characteristics and the final derived taxonomy.

VI. Data Synthesis, Analysis, and Taxonomy Generation

A. Qualitative Synthesis Approach

The synthesis phase transformed the collected data into a cohesive body of knowledge. Given the heterogeneity of the studies—combining empirical performance data with complex architectural proposals—a purely statistical meta-analysis was inappropriate. Therefore, the review employed an Interpretive Thematic Analysis. This approach allowed researchers to aggregate and interpret qualitative findings, leading to the development of a new theoretical understanding and the construction of a comprehensive taxonomy for LLM-based agentic threat hunting systems. The synthesis focused on identifying recurring concepts, shared architectural patterns, and emerging challenges across the 350+ studies.

B. Codification and Theme Development (LLM-Augmented Analysis)

To handle the extensive data volume, the thematic analysis utilized an LLM-augmented workflow, maintaining human expertise as the final arbiter of theme validity.

1. **Initial Coding:** Extracted qualitative descriptions (e.g., agent collaboration mechanisms, observed failure modes, RAG implementations) were subjected to initial coding by human reviewers.
2. **LLM-Augmented Grouping:** Text samples corresponding to similar architectural and functional classifications were provided to a Large Language Model, which was prompted for thematic summarization and grouping.
3. **Human Validation and Refinement:** Human coders rigorously reviewed the LLM-generated themes against four qualitative dimensions: alignment, convergence, coherence, and complementarity. This parallel human-LLM workflow demonstrated the feasibility of leveraging AI to efficiently support the organization of thematic data while simultaneously mitigating risks inherent in LLMs, such as overgeneralization or misinterpretation of nuanced technical concepts.

C. Development of a Cyber Threat Hunting Agentic Taxonomy

The ultimate goal of the synthesis was the creation of a multi-dimensional taxonomy, organizing the diverse field of LLM-powered agentic threat hunting systems based on their structural and functional attributes. The taxonomy was anchored by the four primary dimensions derived directly from the specialized data extraction fields:

1. **Level of Autonomy:** Classifying systems based on their decision-making capacity (e.g., from human decision support to fully autonomous detection and response).

2. **LLM Reasoning Process:** Categorizing the core mechanism of intelligence (e.g., direct prompt engineering, Chain-of-Thought (CoT) reasoning, RAG-augmented CTI, or fine-tuned model usage).
3. **Agentic Architecture:** Differentiating systems based on their coordination structure (e.g., single agent versus role-specialized collaborative multi-agent systems, and the implementation of persistent memory via VectorDB or Long-Term Memory (LTM)).
4. **Log Data Domain:** Mapping the application focus (e.g., Cloud IoC hunting, network intrusion detection, or endpoint forensics).

The synthesis revealed a complex operational landscape. For instance, the analysis shows a distinct divergence in reliability: Agentic AI systems are successfully applied to complex, high-level tasks like autonomous patch synthesis, CTI triage, and root cause analysis. Yet, studies focusing on operational security confirm that LLMs often do not consistently produce accurate or reliable queries for fundamental threat hunting tasks, such as detecting Living off the Land (LotL) techniques. This discrepancy suggests a critical gap between theoretical architectural capability and operational reliability in real-world security contexts. To address this finding, the resulting taxonomy included a fifth dimension: **Operational Reliability Score**, linking specific architectural features (like robust RAG implementation or adversarial testing) to consistent performance metrics, thereby providing a necessary nuanced understanding of current LLM limitations in security workflows.

VII. Conclusions

The rigorous systematic literature review methodology employed, encompassing protocol pre-specification, multi-source information retrieval, customized Quality Assessment, and LLM-augmented thematic synthesis, established a robust and transparent foundation for analyzing the emerging field of LLM-powered Agentic Threat Hunting.

The scale of the initial search (over 38,000 records) justified the systematic, hierarchical approach and the use of "human-in-the-loop" AI assistance during screening, ensuring efficiency without compromising the critical quality assessment and consensus requirements. The purposeful adaptation of the PICO framework and the development of a tailored QAI instrument—focused on reproducibility, adversarial robustness, and reporting completeness specific to AI/ML research—enhanced the analytical validity of the review for this specific technical domain.

Ultimately, the derived multi-dimensional taxonomy, generated through interpretive thematic analysis, serves as a crucial framework for navigating the diverse landscape of LLM agentic systems. It synthesizes architectural trends—such as the observable shift toward collaborative, memory-persistent agents—with empirical performance data. The review concluded that while Agentic AI systems demonstrate significant promise in high-level defense and orchestration tasks, attention must be paid to the observed inconsistency in reliability for generating fundamental security detection rules, underscoring the necessity for

continued research focusing on robustness and operational viability in deployment.