International Journal of Advanced Engineering and Management Research

Vol. 11, No. 03; 2026

ISSN: 2456-3676

Bridging the Blind Spots: A Holistic Risk Model for Secure Deployment of Large Language Models

Yosef Nethaniel Ozeri[1], Ilan Schreiber[2], Amir Schreiber[3]
[1]Ashkelon Academic College (AAC), Department of Computer Science,
12 Ben Tzvi St. PO Box 9071. Ashkelon 78211. Israel
[2]Ashkelon Academic College (AAC), Department of Computer Science,
12 Ben Tzvi St. PO Box 9071. Ashkelon 78211. Israel
[3]Ashkelon Academic College (AAC), Department of Computer Science,
12 Ben Tzvi St. PO Box 9071. Ashkelon 78211. Israel
All authors contributed equally to this work.

doi.org/10.51505/ijaemr.2026.11333 URL: http://dx.doi.org/10.51505/ijaemr.2026.11333

Received: Jun 02, 2026 Accepted: Jun 08, 2026 Online Published: Jun 19, 2026

Abstract

Academic research on securing Large Language Models (LLMs) in cybersecurity currently exists in silos. To address this fragmentation, this study develops the 'Holistic Deployment Risk Model' (HDRM) through a qualitative thematic synthesis of nine 'cornerstone' articles, selected through purposive sampling to ensure a representative cross-section of technical, ethical, and organizational perspectives, consolidating technical vulnerabilities, autonomous agentic risks, and adversarial misuse with organizational governance and human elements to identify critical 'blind spots'. The model clusters associated risks into five holistic, interdependent layers - Governance, Data and Privacy, Model Behavior, Operational Security, and Integrations and Infrastructure - while illustrating how vulnerabilities can propagate across these interdependent layers. To help demonstrate concrete applicability for regulatory readiness and real-world uses, components of the model are qualitatively mapped to current AI risk management frameworks: NIST AI RMF 1.0 and ISO/IEC 23894. While the model is currently theoretical and requires further investigation to confirm its efficacy, it offers organizations an actionable checklist to approach secure LLM deployment. It emphasizes the critical need for established governance before technical implementation and the importance of human-in-the-loop management for high-risk workflows. Ultimately, the work highlights key areas of 'ethical security' necessary to responsibly develop and manage AI systems, including the mitigation of probabilistic decision-making impacts such as bias, misinformation, and privacy violations.

Keywords: Cybersecurity, Risk Model, AI Risk, LLM, Governance Paper type Research paper

Vol. 11, No. 03; 2026

ISSN: 2456-3676

1. Introduction

1.1 Background: Emergence of LLMs in Workflows

To gain perspective on how these risks arise, this section briefly outlines background on LLM deployment and associated challenges. The development of Artificial Intelligence, especially Large Language Models (LLMs), is quickly transforming organization workflows in the information age. Models like ChatGPT, Claude, Gemini are being integrated into operations that were once primarily human cognitive tasks (Schreiber & Schreiber, 2025). While their ability to understand and reason with natural language enables vast improvements in efficiency, they also present novel security concerns not found in traditional software.

1.2 Unique Risks Inherent to LLM Technology

LLMs don’t work like traditional deterministic software programs. Their outputs are probabilistic functions of billions of parameters learned during training, and finely grained changes to their input prompts can cause large differences in output. This means they can occasionally produce confident but incorrect or nonsensical output, accidentally disclose information contained in their training data, produce unexpected behavior when "prompt engineered" by a malicious actor, or fail in subtle ways that humans don’t notice. These qualities invalidate some of our basic assumptions about how we consider software to be safe and require organizations to re-evaluate how they think about risk.

1.3 Risks to Sensitive and High-Stakes Domains

The potential for problems escalates when these models are utilized within sensitive settings like finance, healthcare, and national security. One bad or deceptive answer can have outsized operational, ethical, or legal consequences. Organizations frequently commit a categorization error by treating LLMs as traditional deterministic applications, however their dual nature as both a predictive engine and conversational interface creates additional attack surfaces and systemic vulnerabilities that should be addressed.

1.4 Faster Adoption by Organizations Leads to New Threat Vectors

The increasing issue is further exacerbated by the rapidity that organizations are beginning to adopt LLMs. Fueled by both the race to keep up with competitors and work more efficiently within their organizations, many integrating them into workflows prior to establishing robust security guardrails. For example, employees can accidentally leak private information during routine prompts. Automation can become over-reliant on information from the model or engineers can create integrations that enhance the functionality of the model while unaware of corresponding vulnerabilities.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

1.5 Siloed Studies: Identifying the need for an Integrated Perspective

Existing work tends to address different slices of the problem individually; holistically there has been little work to unify these disparate efforts. This prior work includes segmenting risks into taxonomies posed by LLM risks (Yao et al., 2024); Enumerating abuse operations that LLMs could be used within cybersecurity such as phishing, malware generation, and social engineering attacks (Jaffal et al., 2025); Extracting training data from LLMs posing severe privacy risks should any personal data have been memorized during training (Chen et al., 2025); Parsing related but conceptually different risk categories such as safety vs security vs privacy (Zhang et al., 2025); Mechanisms to score LLMs risks(Zhou and Lin, 2025); Distinguishing unique types of LLMs risks such as hallucinations and innate inaccuracies from the others (Jiao et al., 2025); Identification of risks within Agentic LLMs systems that interact with the world (Brohi et al., 2025); Socio-ethical implications of LLMs such as bias, hallucinations, privacy (Liu et al., 2025); Security of Autonomous and Collaborative LLM Agents (Sun et al., 2026).

Unfortunately, a comprehensive framework to guide companies through the interwoven security challenges of LLM deployment is still absent. Building on previous efforts, this paper aims to provide a consolidated, organization-facing view of risk for LLM systems. We seek to balance granularity with overviews that allow corporations to begin identifying and addressing these risks.

1.6 Study's Contribution: Holistic Deployment Risk Model

Closing this gap was the inspiration for our study. Through a comparative examination of carefully selected final pivotal articles, we pinpointed the core dimensions that were either expressly stated or subtly suggested throughout the existing body of work. Combining these findings, we were able to create a cohesive model that organizes these risk domains and related considerations: The Holistic Deployment Risk Model (HDRM), a 5-layer checklist that organizations can reference when securely adopting LLM systems.

Its 5 layers - Governance, Data and Privacy, Model Behavior, Operational Security, and Integrations and Infrastructure - cover all angles of risk that organizations need to consider. In contrast to prior work, this model's scope extends beyond technical vulnerabilities. It focuses on the policies that dictate language model use, human involvement, integration into workflows, and more. This guides organizations to view LLM deployment as a multi-faceted problem, not just a problem of security.

Furthermore, to enhance its practical relevance, the proposed model is subsequently positioned in relation to established AI risk management frameworks, including NIST (National Institute of Standards and Technology) AI RMF (AI Risk Management Framework) and ISO/IEC 23894. See NIST 2023, ISO 2023.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

1.7 Paper Outline

In this introduction we've laid out our motivation for the remainder of the paper: We'll start by looking back at previous research to form a solid ground of current scholarly knowledge. In the methodological section we describe how those cornerstone pieces were rigorously selected and clarify how we built our HDRM, followed by a close look at the model itself. Finally, we discuss its impact and context within the broader research landscape. We aim to create not just another definition of known and emerging risks but rather provide a guiding foundation for both organizations and future work to understand the unseen complexities at work behind today's implementations of LLMs.

With this in mind, we first examine how the academic community has treated LLM security up until this point. In the following section we'll review key contributions and current gaps.

2. Related work

The work done regarding security, privacy, and safety aspects of LLMs have grown exponentially in the past few years. As organizations integrate LLMs into their workflows, understanding the potential risk has become a focal point of research. Although this is an emerging topic, there are some cornerstone papers that help to organize the landscape of security/privacy/safety work for LLMs. In this section we walk through a representative cross-section of current research, observing different angles of the subject, such as taxonomies of threats, operational misuse, privacy threats and vulnerabilities, differentiations of terms, scoring methods etc. We demonstrate how siloed the current understanding is and the need for a unified framework.

2.1 Offensive/Defensive Dynamics

Yao et al. (2024) provide one of the most comprehensive taxonomies of risks surrounding LLMs. It divides LLM-related phenomena into three categories: (1) The Good - legitimate defensive uses such as code analysis, malware detection, and anomaly identification; (2) The Bad - offensive misuse including phishing, social engineering; and adversarial manipulation; and (3) The Ugly - inherent model vulnerabilities such as memorization of sensitive data, hallucinated facts, and susceptibility to prompt injection. This taxonomy is useful because it illustrates both sides of the cybersecurity landscape - attack and defense - while also housing internal weaknesses that traditional security tools cannot easily detect - LLMs are systems with complex behavioral properties. Although Yao et al. (2024) form an exhaustive taxonomy of things that can go wrong when operating LLMs, there isn't any insight into how an organization would evaluate these risks or manage them on a scale.

Jaffal et al. (2025) examine the operational applications of LLMs in cybersecurity, along with their potential for misuse, demonstrating how they can help analysts through use cases like summarizing threat intelligence, finding malicious code patterns, parsing logs, and aiding in forensic analysis. At the same time, bad actors can use these capabilities to create polymorphic

Vol. 11, No. 03; 2026

ISSN: 2456-3676

malware, improve phishing emails, or automate threat reconnaissance. In contrast to the first piece, they include concrete examples that illustrate the potential for both beneficial and harmful applications of LLMs. This practical lens on risk is helpful: it demonstrates how shortcomings are introduced through actual usage of LLMs (versus hypothetical threats). That being said, it fails to capture risk within a multi-layered framework. There are no links between how operational abuse can interact with privacy misuse, problematic governance, or architecture vulnerabilities. This provides further justification for an integrative multi-layer model.

Brohi et al. (2025) explore the emergence of Agentic AI powered by LLMs. They explain how these systems can independently work towards complicated objectives across disciplines such as education, health, and cybersecurity, relying on 84 academic sources. The authors divide Agentic AI risks into two groups: (1) Continuum of Risks - shows how discovered in LLMs risks extend into autonomous agents, such as biases, disinformation, and privacy; (2) Unique Agent Risks - details risks introduced with agency, including goal misalignment, lack of insight into decision-making, and agent coordination. The researchers include a roadmap for governance and accountability. However, as it focuses on mapping challenges and presenting a roadmap for the future, it suffers from omissions for organizations looking for practical frameworks to the unique risks posed by Agentic AI powered by LLMs.

2.2 Vulnerability and Failure Modes

Jiao et al. (2025) outline the types of ethical concerns LLMs raise that are unique to them (Inherent risks) versus traditional “AI ethics” issues (such as privacy and fairness). They lay out threats posed by bias, hallucinations, mis/disinformation, and copyright issues and how organizations can navigate those risks with interdisciplinary groups and customizable “guardrails". Cost of Control Tools – the researchers point out that one of the methods of controlling these models is through auditing them which can be expensive and requires technical expertise. While the researchers provide a broad and well-structured overview of potential concerns and ethics, it doesn’t leave the reader with any type of actionable checklist that a manager could reference when deciding how to organize these mitigation efforts.

Chen et al. (2025) relate to privacy. As LLMs are trained on increasingly large datasets that include sensitive or personal data, privacy becomes a major concern. The researchers show that given the right prompts, an LLM can remember phone numbers, email addresses, or chunks of text unique to its training data. This creates significant challenges around data privacy regulations like GDPR and HIPAA. One of their best contributions is how leakage occurs: (1) Overfitting to rare patterns; (2) Inadvertent memorization of training data; (3) Inadequate filtering during dataset creation; (4) Vulnerability to extraction attacks via adversarial prompts. However, because the scope was strictly focused on privacy, it was difficult to conceptualize how these risks fit into the bigger picture around an organization’s processes. This research shines a light on one important slice of the multi-layered LLM risk.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

Zhou & Lin (2025) suggest applying the "CVSS" (Common Vulnerability Scoring System) to LLM weaknesses. This is significant because CVSS is the industry standard for evaluating traditional software vulnerabilities. By adapting it to LLM behavior, the authors attempt to quantify risk severity and exploitability. The authors argue that several LLM security threats pose a greater risk than typical software flaws due to unpredictability, adversarial exploitation, and widespread replication. On the other hand, they highlight CVSS's shortcomings in scoring behavioral vulnerabilities that emerge unpredictably at runtime such as hallucinations, chain-of-thought exposure, or role-based context manipulation. The authors provide a "how big" measurement but offer limited insight into where and how these vulnerabilities manifest across an organization.

2.3 Governance and Risk Framing

Zhang et al. (2025) give useful clarification on common terms that are often used interchangeably: safety, security, and privacy. While not expecting most professionals to make these distinctions, they provide specific breakdowns for each: (1) Safety concerns prevent harmful or unethical output; (2) Security focuses on attack vectors; (3) Privacy involves leaking of sensitive information. This distinction is essential because conflation often leads organizations to adopt incorrect mitigation strategies or assume that solving one type of problem addresses another. For example, improving safety alignment does not prevent prompt injection attacks, and implementing encryption does not prevent hallucinated harmful recommendations. Though contributing to defining these concepts, the authors do not provide a framework for risk assessment, highlighting the need for a unified structure that distinguishes and layers them for actionable items.

Liu et al. (2025) connect technical insecurity topics with sociological/ethical concerns under the theme of "ethical security" as it relates to LLMs, relying on 74 academic sources. It provides a mapping of threats and countermeasures with a broader international view. It brings these points: (1) Holistic view of "ethical security": information security entails more than technical exploitation, including technical (social engineering, malware) as well as socio-ethical issues (bias, hallucinations, privacy); (2) Language Barrier: organizations operating in languages other than English will find gaps in protection methods when dealing with multilingual communities, with no clear methods of implementing security that can be applied to cultures/languages other than English.

Sun et al. (2026) shift the focus to a System Engineering Security viewpoint, treating agents as intricate software with memory, tools, and social abilities. This is accomplished through: (1) Creating a taxonomy of Security Threats to Agents, separated into attacks on interaction with the world (interfaces/perceptual tools), attacks against internal thought processes/memory, and attacks against communication and cooperation with other agents; (2) Holistic System for Securing Agents: defenses that match threats, including sanitization of input, hardening of cognition, and secure consensus; (3) Agents Security Verification and Metrics: comparison of existing verification and metrics for quantifying resistance to both current and future threats.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

However, while the researchers give a high-level taxonomy of agent security concerns, it doesn't translate to "what do I actually do" answers, for example how loose or strict security should be vs performance or resource concerns.

2.4 Synthesis and Gaps in Literature

Taken together, those representative cross-sections of current papers exhibit these trends:

1) Each paper addresses one layer of risks with LLMs (taxonomy layer, operation layer, privacy layer, conceptual layer, scoring layer).
2) None of the papers put these concepts into an integrated architecture for risk evaluation.
3) There isn't an approach to bridge the gap between research and implementation decisions for practitioners.
4) There aren't considerations for interactions between layers (e.g.: how privacy considerations affect operational abuse).
5) Current framework doesn't give advice for enterprises using LLMs as a part of larger systems involving APIs, databases, other software agents, and humans.

For these reasons, the current study proposes a HDRM that bridges these isolated contributions. Rather than replacing prior work, the model positions each core contribution as representing one part of a larger puzzle. By doing so, it transforms fragmented academic insights into a coherent structure that organizations can use to secure LLM deployments.

3. Methodology

While prior studies have shed important light on individual aspects of LLM security, they still lack a unified analytical framework. This motivates the methodological approach presented below, which integrates insights from multiple domains into a single structured model. The goal of this research was to develop a unified risk model for the secure deployment of LLMs based on insights derived from the final selected key academic articles. Because the existing literature is siloed - each study examining only one aspect of the security landscape - the methodology focused on synthesizing these isolated contributions into a cohesive, multi-layered framework. The process consisted of three primary phases: collection and selection of sources, comparative thematic analysis, and model construction, refinement, and Conceptual Positioning.

3.1 Selection of Sources

3.1.1 Purposive Sampling Strategy Overview

In this study, we employed a purposive sampling strategy to select the foundational literature for the synthesis. Following Benoot et al. (2016) and Ames et al. (2019), we adopted the same approach to prioritize qualitative studies rather than attempting exhaustive coverage of all available literature. Unlike a systematic review that seeks exhaustive quantity, purposive sampling aims for theoretical saturation by selecting 'cornerstone' sources that represent maximum variation within the field (Hennink and Kaiser, 2022; Kaur et al., 2024).

Vol. 11, No. 03; 2026

ISSN: 2456-3676

This strategy allowed us to identify a representative cross-section of the current research landscape, focusing on information-rich cases that cover distinct dimensions of LLM risk. The selection process prioritized articles that do not fully overlap, ensuring that each source provides a unique piece of the 'larger puzzle' - from technical vulnerabilities to organizational governance.

3.1.2 Purposive Sampling Implementation

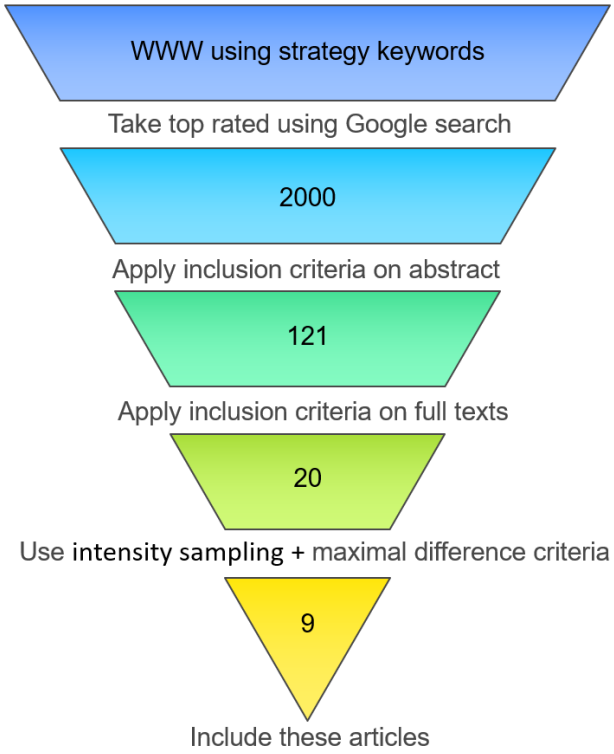


Figure 1. Research Article Selection Process Using Purposeful Sampling

Figure 1 illustrates the overall process which includes four stages to select the final articles:

1) Top Rated Google Scholar Database Results

Initially, we compiled a database of potentially relevant articles based on a scoping review. Scoping is an exploratory and systematic way of mapping the literature available on a topic (Hannes et al., 2013). Scoping exercises are perceived as the ideal way of doing preparatory work for an exhaustive systematic review. In our case, we have used them for building an archive of data for our qualitative evidence synthesis. We searched Google Scholar database. A search string was used as a methodological filter to extract qualitative research articles. For example, the research string we used was "Cybersecurity and LLM* and qualitative". Studies had to be published in English between 2016 and present day 2026, this was for pragmatic reasons. We only took the first 2000 URLs.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

2) Inclusion Criteria Application on Abstracts

The qualitative studies retrieved were qualitative studies matched against the following inclusion criteria:

A. Type of studies - we considered qualitative designs of survey and review to have a large-scale representation of the subject and practical insights. Opinion pieces and editorials were excluded. The study reports should be qualitative in nature.

B. Phenomenon of interest - studies should focus on the risk aspects of developing, integrating, operating LLMs.

C. Type of Results – studies should introduce real-world cases, conceptual clarity, or relevant guidance for organizations.

First, we applied the inclusion and exclusion criteria to the retrieved abstracts, resulting in 121 articles.

3) Inclusion Criteria Application on Full Text

A full text was requested for each of the relevant studies. Second, these studies were further assessed, rechecking them against the same inclusion and exclusion criteria. As can be seen in Fig. 1, a total of 23 articles were included in our pool/archive of data. The pool of 23 articles was then used to initiate purposeful sampling.

4) Applying Purposive Sampling Approaches

We applied two common purposive sampling techniques: intensity sampling and maximum variation sampling (Patton, 1990).

1) **Intensity sampling** focuses on identifying cornerstone papers. Cornerstone papers were chosen because they represent examples of the target phenomenon of LLMs risk happening to an extreme degree. This helped to form the foundation for our literature review. We wanted to ensure that our cornerstone papers would offer a large quantity of information to learn from, such as Yao et al. (2024)'s massive review of over 280 papers, but we didn't want to choose outliers that would skew our model.

2) **Maximum variation sampling** was then used to identify other cornerstone papers which each focus on a silo of LLM security research that doesn't overlap significantly with another: We considered threats and their classification, how LLMs are misused, privacy issues, and the differences between conceptual definitions and scoring systems. Within each of these silos, there are unique vulnerabilities, but we also discovered that overlap existed in major trends across all spaces.

By synthesizing insights from these intensive heterogeneous sources, we were able to build a solid ground for the suggested HDRM which provides a comprehensive framework that bridges fragmented academic perspectives and identifies critical 'blind spots' in organizational LLM deployment.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

Final Selected Articles

Each article represents a distinct research direction within the broader topic of LLM security:
1) Yao et al. (2024) - Threat Taxonomy
2) Jaffal et al. (2025) - Operational Misuse and Cybersecurity Applications
3) Chen et al. (2025) - Privacy and Memorization Risks
4) Zhang et al. (2025) - Safety-Security-Privacy Distinctions
5) Zhou and Lin (2025) - Quantitative Vulnerability Scoring
6) Jiao et al. (2025) - LLM Ethics and Governance
7) Brohi et al. (2025) - Agentic AI Risks
8) Liu et al. (2025) - Ethical Security Perspective
9) Sun et al. (2026) - Agent Security Frameworks

These nine articles were intentionally selected not because they fully overlap, but because they do not. Their differences create the intellectual space needed for a holistic framework that organizes diverse insights into a functional deployment model.

3.2 Comparative Thematic Analysis

Based on this academic solid ground, patterns across articles were identified using comparative thematic analysis. Themes were identified using a process consistent with reflexive thematic analysis (Braun and Clarke, 2019; Braun and Clarke, 2021; Braun and Clarke, 2023), whereby key concepts and ideas from articles were coded into potential thematic categories, which were then compared within and across articles using a constant comparative method (Glaser and Strauss, 1967) to develop a set of analytical dimensions across which data converged and diverged. This resulted in our set of analytic dimensions that summarize patterns across articles, consistent with approaches to thematic synthesis in qualitative literature reviews (Thomas and Harden, 2008). The following section describes a four-step systematic thematic analysis used in this research:

Step 1: Extraction of Key Concepts

For each article, central ideas, definitions, risk types, and proposed mitigation strategies were identified. Examples include:
1) Offensive misuse (e.g. phishing automation).
2) Defensive applications (e.g. log analysis).
3) Internal vulnerabilities (e.g. memorization of private data).
4) Conceptual distinctions (safety vs. security).
5) Quantitative severity scoring.

The goal was to capture not only explicit findings but also implicit assumptions about how LLMs behave in organizational systems.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

Step 2: Coding and Categorization

Each extracted concept was grouped according to its functional domain. During this process, early patterns emerged:

1) Some concerns were technical (e.g. prompt injection, model extraction).
2) Others were behavioral (e.g. hallucinated outputs, unaligned reasoning, misalignment).
3) Others were organizational (e.g. misuse by employees, lack of policies).
4) A few were architectural (e.g. insecure APIs, weak access control).

The categorization stage was crucial in revealing that LLM security cannot be understood through technical analysis alone - it requires examining human, structural, and procedural components as well.

Step 3: Cross-Article Comparison

The categories defined above were then used to map each study's contributions. Looking at the research in light of the identified gaps immediately exposes the remaining shortcomings. For example:

1) Privacy was well-developed in the Chen et al. (2025) article but almost entirely absent in the Jaffal et al. (2025) operational study.
2) The Yao et al. (2024) threat taxonomy article highlighted internal vulnerabilities but did not discuss governance or infrastructure.
3) Zhou and Lin (2025) quantified risks but did not contextualize them within real organizational workflows.
4) The Zhang et al. (2025) defined safety-security-privacy conceptual categories but did not explain how they interact.

Analyzing the limitations of each article helped us uncover shortcomings in the existing research, ultimately paving the way for a model designed to illustrate these connections.

Step 4: Identification of Core Dimensions

From the comparison, five recurring dimensions emerged as essential to understanding LLM deployment risk:

1) Governance - policies, oversight, user behavior, organizational procedures.
2) Data and Privacy - training data, memorization, leakage, regulatory compliance.
3) Model Behavior - hallucinations, internal vulnerabilities, adversarial susceptibility, misalignment.
4) Operational Security - misuse by attackers or employees, workflow integration risks.
5) Integrations and Infrastructure - APIs, multi-agent systems, system-level access, architecture.

Although not all final selected articles explicitly referenced all dimensions, each one provided a strong insight into at least one of them. Their combination creates a truly comprehensive picture.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

3.3 Model Construction

After identifying the five core dimensions, the next step was to construct the HDRM. The construction followed three principles:

Principle 1: Multi-layered Representation

Risk is rarely siloed to singular events across layers. For example, a privacy leak (Data and Privacy layer) may result from insecure APIs (Infrastructure layer) or from misaligned behavior (Model Behavior layer). The model treats layers as interdependent.

Principle 2: Organizational Applicability

Academic models tend to end up useless because they are purely theoretical. To ensure real-world usability, each layer focuses on questions that organizations can directly evaluate, such as:
1) Who can prompt the model?
2) What data does the model store or recall?
3) How predictable is the model's reasoning?
4) Are API calls audited and restricted?
5) What governance exists for misuse scenarios?

Principle 3: Integration of Prior Work

Each of the academic articles contributes to specific layers:
1) Yao et al. (2024): Model Behavior + Operational Security
2) Jaffal et al. (2025): Operational Security
3) Chen et al. (2025): Data and Privacy
4) Zhang et al (2025): Governance + Conceptual Foundations
5) Zhou and Lin (2025): Severity Scoring Across All Layers
6) Jiao et al (2025) - LLM Ethics and Governance: Governance + Conceptual Foundations.
7) Brohi et al. (2025) - Agentic AI Risks: Model Behavior + Conceptual Foundations.
8) Liu et al. (2025) - Ethical Security Perspective: Operational Security + Conceptual Foundations.
9) Sun et al. (2026) - Agent Security Frameworks (Preprints Org): Governance + Model Behavior.

By explicitly linking each study to the model, the framework does not replace prior work but unifies it into something actionable.

3.4 Model Refinement

After building the initial model from five layers, we iteratively examined the literature and adjusting the model until we were satisfied that:

Vol. 11, No. 03; 2026

ISSN: 2456-3676

1) The layers didn't overlap conceptually (each was discrete).
2) Collectively they span the risks raised in all final selected articles.
3) There was enough breadth in each layer to apply to different LLM use cases (internal tools, public-facing chatbots, agentic systems, etc.)

The end result was a HDRM grounded in theory while also speaking to practice. The unified model brought to light some interesting takeaways, including areas where we're doing well and significant gaps in current LLM deployment practices. You can find those insights below.

3.5 Conceptual Positioning

As an additional conceptual positioning step, the resulting model was qualitatively aligned with established AI risk management frameworks, namely NIST AI RMF 1.0 and ISO/IEC 23894:2023. This alignment was not intended as formal empirical validation, but rather as a structured comparative exercise aimed at assessing external coherence and practical applicability within recognized governance paradigms.

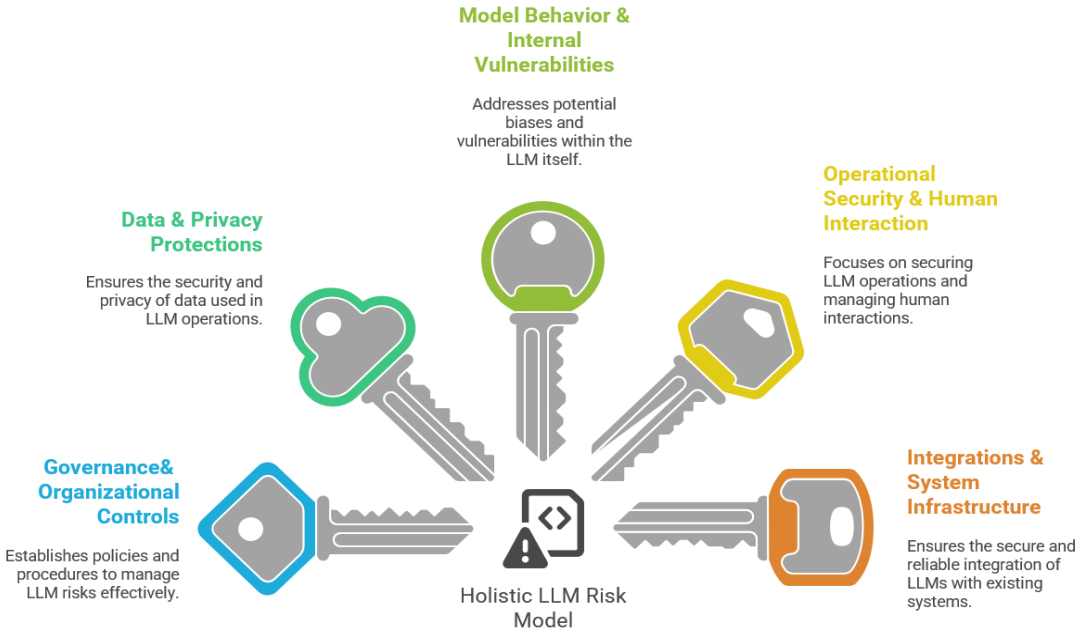
4. Findings

The comparative analysis of the final selected academic articles revealed a notable fragmentation in the way LLM security risks are understood. While each study provides valuable insights, none offers a comprehensive framework that reflects the complexity of real-world LLM deployment. In response, this research proposes the HDRM, a five-layer structure that organizes risks into interconnected domains: Governance, Data and Privacy, Model Behavior, Operational Security, and Integrations and Infrastructure. Together, these layers offer a complete perspective on the challenges organizations face when deploying LLM systems. Figure 2 illustrates the model, followed by elaboration on each layer. Key insights from each of the final selected papers will be highlighted where they become relevant to the layer's discussion.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

Figure 2. Holistic Deployment Risk Model (HDRM)



4.1 Layer 1 – Governance and Organizational Controls

Governance is the outermost layer, as it essentially dictates the policies, roles, and intended purpose that govern an LLM's interactions within the organization. Things we found during our literature review: almost every vulnerability can be traced back to some kind of governance failing even if the model itself is not technically vulnerable.

Key Risk Areas

1) No usage policies - someone could inadvertently share private information through a prompt.
2) Inadequate access controls - leaving LLM access open to all means someone will abuse/misuse it or become overly reliant on it.
3) Lack of monitoring or auditing - impossible to flag unsafe/inappropriate/responsible uses.
4) Risk assessment does not align with actual vulnerabilities - Most organizations and stakeholders still confuse safety, security, privacy. (Refer to Zhang et al. (2025) article)
5) Risk assessment does not align with emerged Agentic and Multi Agent threats – if implemented, organizations should govern unique threats of Agentic AI and Agent-to Agent interactions.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

Connection to Prior Work

1) The Zhou and Lin (2025) conceptual paper discusses confusion between safety, security, and privacy, reinforcing the need for governance structures that precisely define these terms.
2) The Yao et al. (2024) taxonomy article emphasizes that misuse often arises from human behavior, not just model vulnerabilities.
3) The Jiao et al. (2025) LLM ethics and governance emphasizes the uniqueness of those risks.
4) The Sun et al. (2026) agent security frameworks discuss threats and defenses of Agentic AI and Agent-to Agent interactions.

Implications

If this layer goes corrupt/governance fails, we risk all other layers being exposed. E.g.: we can have all these data protections, but a careless employee accidentally overrides these through his use. We emphasize that LLM security needs a strong governance base, not just some administrative tasks.

4.2 Layer 2 - Data and Privacy Protections

The Data and Privacy layer addresses how LLMs interact with sensitive information - both in their training data and during real-time use. This was one of the strongest themes in the literature, particularly in the Chen et al. (2025) article.

Key Risk Areas

1) Training Data Memorization - Possibility of recalling rare or sensitive phrases present in training dataset.
2) Data Leakage During Inference - Potential leakage of private or proprietary data through prompt injection.
3) Lack of Proper Data Filtering - Weak data preprocessing increases chance of memorizing private data.
4) Regulation Risks - Possibility of non-adherence to GDPR, HIPAA, etc. privacy regulations.

Connection to Prior Work

1) The Jaffal et al (2025) operational Security article provides empirical evidence showing that LLMs can leak personal data.
2) The Yao et al. (2024) taxonomy article's "Ugly" category aligns with inherent memorization vulnerabilities.
3) The Zhou and Lin (2025) scoring study identifies privacy leaks as high-severity vulnerabilities within a CVSS- style framework.
4) The Liu et al. (2025) ethical security relates ethical security to privacy as well.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

Implications

Even organizations with strong governance may fall victim to privacy-related risks if they misunderstand how LLMs handle and recall information. Effective data protection must include dataset curation, prompt-level controls, and privacy-enhancing mechanisms such as differential privacy- none of which are sufficiently addressed in current frameworks.

4.3 Layer 3 - Model Behavior and Internal Vulnerabilities

This layer covers risks that originate inside the model itself- vulnerabilities that persist regardless of user intentions or system design. These risks are emphasized heavily in the taxonomy article, which presents "The Ugly" as the most dangerous category.

Key Risk Areas

1) Hallucinations / False but Plausible Outputs - The model can output incorrect information that sounds believable.
2) LLMs are susceptible to prompts designed to elicit harmful responses.
3) Fringe situations often trigger unpredictable and inconsistent outputs.
4) Flaws in Foundation - Due to their probabilistic nature, LLMs will never be able to be 100% deterministic or verified.
5) Multi-agent systems and Agentic agents may go out of control if not carefully designed and monitored.

Connection to Prior Work

1) The Yao et al. (2024) taxonomy paper stresses the importance of understanding inherent model risks.
2) The Jaffal et al (2025) operational study demonstrates real-world examples of manipulated outputs.
3) The Zhang et al (2025) conceptual article distinguishes between safety failures (harmful model outputs) and security failures (attacker-induced manipulation), both of which originate here.
4) The Jiao et al (2025) LLM Ethics and Governance study – discuss specifically hallucinations and bias.
5) The Brohi et al. (2025) agent security frameworks – relates autonomous agents

Implications

Model behavior is a layer of risk that organizations sometimes fail to consider. There's a risk that technical teams feel safer just by upgrading to a more advanced model, even though larger models can make internal vulnerabilities more difficult to pinpoint.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

4.4 Layer 4 - Operational Security and Human Interaction

Operational Security examines how LLMs are used in real workflows- by employees, customers, and attackers. The Alshahrani (2025) article demonstrates clearly that the operational environment shapes how safe or dangerous an LLM becomes.

Key Risk Areas

1) Employee abuse - for instance depending on LLMs to make decisions outside their control.
2) Automated attacks - Malicious users of LLMs to create phishing attacks, malware, reconnaissance.
3) Over-Reliance on Output - Employees may believe the model is correct when it is not.
4) Inconsistent operating environments - The same prompt can return different results, making it difficult to secure.

Connection to Prior Work

1) The Jaffal et al (2025) article is the primary foundation of this layer.
2) The Yao et al. (2024) taxonomy article supports this by showing that LLMs can strengthen or weaken cybersecurity depending on usage.
3) The Zhou and Lin (2025) scoring model indicates that many operational vulnerabilities score higher than traditional software vulnerabilities.

Implications

Operational failures often happen not because the model is malicious, but because human interaction introduces unpredictable risk. This layer helps organizations understand the real-world environments in which LLMs operate, something missing from earlier frameworks.

4.5 Layer 5 - Integrations and System Infrastructure

The final layer reflects the technical environment surrounding the LLM- the APIs, databases, pipelines, agents, and systems with which the model interacts. This layer is surprisingly absent in most prior research yet is critical for real-world deployment.

Key Risk Areas

1) Insecure API Integrations - Models connected to backend systems can unintentionally trigger actions.
2) Multi-Agent Systems - Chains of LLM-driven agents can escalate errors.
3) System-Level Access - LLMs embedded in workflows may interact with sensitive environments.
4) Lack of Infrastructure Isolation - Shared resources increase risk of cross-system leakage.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

Connection to Prior Work

Although none of the final selected articles focus solely on systems architecture, each implies it indirectly:

1) The Yao et al. (2024) examples show LLMs acting as security assistants inside tools.
2) The Chen et al. (2025) privacy paper underscores risks in data pipelines.
3) The Zhou and Lin (2025) scoring model highlights infrastructure weaknesses as severity multipliers.

Implications

Without securing infrastructure, even the most advanced LLM is exposed to systemic vulnerabilities. This layer provides the final piece of the holistic model, illustrating that LLM risk is never confined to the model alone- deployment context is equally important.

4.6 Brief Discussion of Results

As illustrated by the five-layer model, securing LLMs is a cross-domain problem - one which requires systematized thinking around security. Each layer represents a unique class of risks, and shortcomings in one layer can propagate to another. Moreover, each of the five layers is pulled directly from literature: there is a theme present in each academic article that can be mapped to a specific layer of the model. The utility of this model is that it allows organizations to think about securing LLMs from all angles rather than focusing through siloed lenses.

When considering all of these results, there are larger implications about how organizations should approach securing LLMs. The discussion will analyze these findings and provide insight into real-world deployment.

5. Discussion and Conclusion

Based on the insights discussed above, several overarching conclusions can be drawn about the future of secure LLM deployment. The purpose of this study was to synthesize fragmented academic insights into a unified model that reflects the real-world complexity of deploying LLMs. While previous research has offered valuable contributions - ranging from taxonomies of risks to empirical demonstrations of privacy leakage and attempts to quantify vulnerabilities - these studies remain limited in scope. Each focuses on one part of the problem, leaving organizations without a comprehensive framework for assessing the full spectrum of LLM-related risks. The HDRM developed in this research addresses this limitation by organizing risks into five interconnected layers: Governance, Data and Privacy, Model Behavior, Operational Security, and Integrations and Infrastructure.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

5.1 The Significance of a Multi-Layered Perspective

Possibly the key takeaway from this mapping exercise is that LLM risks do not exist in silos. For example unclear policies about acceptable prompts (Governance failure) may lead employees to expose sensitive data (Data and Privacy), which attackers could then exploit through prompt manipulation (Model Behavior), ultimately affecting downstream systems (Integrations and Infrastructure). This chain reaction demonstrates why a single-layer view cannot capture the true nature of LLM risk.

The findings from the final papers are synthesized to demonstrate their overlaps, distinctions, and individual strengths (see Table 1 with include a representative sample of the core articles for comparison). As illustrated, each study contributes a unique perspective to the broader landscape of LLM security; however, the fragmentation among these works reinforces the need for a unified framework - a gap this research addresses through the proposed HDRM.

Table 1. A representative Sample of the Core Articles for Comparison

	Zhang et al. (2025)	Jaffal et al (2025)	Liu et al. (2025)	Zhou and Lin (2025)	Yao et al. (2024)
Main Focus	Classification of LLM risks into positive, negative, and internal vulnerabilities	Practical use cases and misuses of LLMs in real operations	Privacy leakage, memorization, data exposure	Scoring vulnerabilities using CVSS-like metrics	Conceptual differentiation between safety, security, privacy
Strengths	Broad survey (280+ papers), holistic view, strong typology	Real-world relevance; operational insights	Deep look at data leakage mechanisms	Clear numeric scoring, structured evaluation	Clear taxonomy; clarifies terminology
Weaknesses	Less technical depth per vulnerability	Limited theoretical model	Narrow focus on privacy only	Limited scope; only vulnerability scoring	Lacks empirical evidence
Methodology	Literature review + conceptual framing	Case-based analysis	Technical examination + experiments	Quantitative metric adaptation	Theoretical argumentation
Data Used	Prior academic work only	Real-world operational examples	Synthetic and real prompts; memorization tests	Scoring of known vulnerabilities	Conceptual distinctions only

Vol. 11, No. 03; 2026

ISSN: 2456-3676

Findings	Ugly (internal vulnerabilities) are most dangerous	LLMs help defenders but also attackers	LLMs memorize more than expected	CVSS can rank LLM vulnerabilities	Terms often confused in research
Contribution	A unified security perspective	Operational understanding	Better privacy protections	A numeric scoring baseline	A shared conceptual vocabulary
Limitations	Few real-system experiments	No unified framework	Limited scope	Too narrow for full assessment	Missing integration with other models

5.2 Contributions to Academic Research

From an academic perspective, this study contributes three key advancements:

1) A unified conceptual structure

The model provides the first academically grounded framework that links the entire spectrum of LLM vulnerabilities into a coherent structure. It transforms broad theoretical insights from taxonomy and conceptual articles into an applied, actionable format.

2) A bridge between technical and organizational domains

Previous works focus on technical issues such as adversarial prompts/memorization/etc. Our work expands the scope to integrate governance and human elements, recognizing that security flaws often arise within organizational contexts like culture and decision-making processes.

3) A foundation for future quantitative research

Because the model clearly separates risk layers, future studies can attempt to measure and compare the severity of vulnerabilities within and across layers. The Zhou and Lin (2025) article's adaptation of the CVSS system shows the potential for quantitative scoring, and the multi-layer model can serve as an improved structural basis for such efforts.

5.3 Practical Implications for Organizations

The HDRM is not purely academic. Rather, it is designed to inform decision-making in production settings. Some real-world takeaways include:

1) Governance must precede deployment
   Organizations often deploy LLMs prematurely, relying on general usage guidelines rather than formal policies. The model emphasizes that governance failures frequently trigger technical vulnerabilities.
2) Privacy risks require continuous monitoring
   Since LLMs can sometimes memorize and leak information, organizations should carefully examine training data, incorporate privacy measures, and limit potentially harmful prompts.
3) Model behavior cannot be fully controlled

Vol. 11, No. 03; 2026

ISSN: 2456-3676

LLMs will never be deterministic. They are probabilistic by nature. Even with extensive adjustments, the chance of hallucinations remains constant. Respect the LLM - add human checks to critical workflows.

4) Operational misuse will become more common
   Attackers are already deploying LLMs to power phishing, malware creation, reconnaissance, etc. so defenders need to plan for that use case as well. This is consistent with findings from the Jaffal et al (2025) operational misuse paper.
5) Secure integration is essential
   Many organizational incidents arise not from the LLM itself but from the systems connected to it- APIs, pipelines, agents, and automated workflows. As integration complexity grows, infrastructure security becomes just as important as model security.

In addition to its conceptual contributions, the model was further examined in relation to established AI risk management frameworks in order to assess its external coherence and practical applicability.

5.4 Cross-Framework Alignment with Established AI Risk Management Standards

In order to enhance the validity and applicability of our proposed Holistic Deployment LLM Risk Model, we provide a qualitative mapping to two widely adopted AI governance frameworks - NIST AI Risk Management Framework and ISO/IEC 23894 (see Table 2). This mapping is not intended to replace existing frameworks, but to position our model as a lens for highlighting LLM-specific risks that are not explicitly addressed in broad AI governance approaches.

Specifically, our model focuses on how risk propagates within LLMs across five interdependent layers, highlighting causal escalation paths. While complementary to these frameworks, they do not explicitly decompose LLM-specific risks across the operational lifecycle, particularly in the context of real-world deployment.

Table 2- Mapping to Popular Existing AI Governance Frameworks

Proposed Risk Layer / Framework	**LLM-Specific Risk Characteristics (This Study)**	**Alignment with NIST AI RMF Functions**	**Alignment with ISO/IEC 23894 Principles**
Governance and Organizational Controls	Root-cause layer: absence of policies, unclear accountability, misalignment between perceived and actual risks, lack of governance over Agentic AI and multi-agent systems	Govern – risk governance, accountability, policies, risk culture	leadership, governance structures, integration of AI risk management into organizational processes

Vol. 11, No. 03; 2026

ISSN: 2456-3676

Data and Privacy Protections	Dual exposure: training-time memorization and inference-time leakage, regulatory non-compliance, weak data filtering and prompt-level controls	Map, Measure, Manage – data context mapping, privacy risk identification, mitigation controls	risk identification, analysis, and treatment of data-related risks
Model Behavior and Internal Vulnerabilities	Inherent model risks: hallucinations, probabilistic uncertainty, prompt injection susceptibility, instability in edge cases, agentic unpredictability	Measure, Manage – evaluation of model trustworthiness, robustness, and behavioral risks	AI-specific risk assessment, evaluation, monitoring of model behavior
Operational Security and Human Interaction	Human-centered risks: misuse, over-reliance, adversarial use (phishing, malware), variability in outputs across contexts	Govern, Manage – operational oversight, monitoring, misuse mitigation, incident response	control implementation, human-in-the-loop oversight, monitoring and review
Integrations and System Infrastructure	System-level exposure: insecure APIs, multi-agent chains, backend connectivity risks, lack of isolation, infrastructure-driven escalation of impact	Map, Manage – system context, interface risks, dependency and integration management	lifecycle integration, technical controls, system-level risk management

There is a high degree of complementarity between the two alignments, however there are a couple of notable gaps to call out. NIST AI RMF & ISO/IEC 23894 don't explicitly decompose risk based on the operational lifecycle considerations specific to LLM deployment. Even though both frameworks mention them, the specific risks linked to human interaction, model agency, and system integration aren't really broken down in detail.

Hence, this five-layer model helps augment existing frameworks by providing a decomposition that maps to how LLMs are actually deployed in the real world. It helps highlight that risk can bleed through layers (poor governance can cause unsafe data usage which can be leveraged through model behavior that can be amplified through unsafe system integrations).

As such, this model should be thought of as an additional operational layer on top of existing AI governance frameworks that help organizations map high-level concepts to LLM-specific practical risk identification.

Vol. 11, No. 03; 2026

ISSN: 2456-3676

5.5 Limitations of This Study

The study's methodology relies on qualitative thematic synthesis rather than quantitative experimentation. As a result:

1) The model is conceptual, not empirically tested.
2) Severity and likelihood assessments are not numerically validated.
3) Real-world deployment environments differ significantly across organizations.

Furthermore, the field of LLM security is still evolving. New attack vectors, alignment techniques, and regulatory frameworks emerge rapidly. The model should therefore be considered a foundation rather than a final blueprint.

5.6 Directions for Future Research

Possible areas of future work based on the observations above include:

1) Empirical validation of the five-layer model.
   Validation could be performed by having an organization use the model to identify risks and prevent them.
2) Quantitative scoring across all layers
   Inspired by the Zhou and Lin (2025) CVSS adaptation, future work may attempt to build a scoring system aligned with this model.
3) Expansion to multi-agent systems
   As LLM-based agent ecosystems grow, new research is needed to map interactions between agents and identify emergent vulnerabilities.
4) Policy and regulatory frameworks
   Governments and industry bodies may use this model as a basis for creating certification standards or compliance guidelines for LLM deployment.

5.7 Conclusion

LLMs are spreading faster than ever before, bringing both unparalleled opportunity and risk. This study showed that by comparing and synthesizing nine cornerstone articles, there are large gaps in academia that only touch on pieces of a much bigger security picture. Our model, the HDRM, is designed to reveal the complexities of real-world deployment. Our model can act as a guideline for practitioners to recognize their own blind spots, and plan for future vulnerabilities in the pursuit of building safer, more responsible AI.

As LLMs start to increasingly impact our decisions, workloads, and infrastructure. It's important we have solid, holistic frameworks to guide us. This study was only meant to be one step in that direction.

References

Ahmed, S. K., Mohammed, R. A., Nashwan, A. J., Ibrahim, R. H., Abdalla, A. Q., Ameen, B. M. M., & Khdhir, R. M. (2025). Using thematic analysis in qualitative research. *Journal of Medicine, Surgery, and Public Health*, *6*, 100198. https://doi.org/10.1016/j.glmedi.2025.100198

Ames, H., Glenton, C., and Lewin, S. (2019). Purposive sampling in a qualitative evidence synthesis: A worked example from a synthesis on parental perceptions of vaccination communication. *BMC Medical Research Methodology, 19*, 26.
https://pubmed.ncbi.nlm.nih.gov/30704402/

Anh-Hoang D, Tran V and Nguyen L-M (2025) Survey and analysis of hallucinations in LLMs: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence* 8:1622292.
https://doi.org/10.3389/frai.2025.1622292

Benoot, C., Hannes, K., and Bilsen, J. (2016). The use of purposeful sampling in a qualitative evidence synthesis: A worked example on sexual adjustment to a cancer trajectory. BMC medical research methodology, 16(1), 21.
https://pmc.ncbi.nlm.nih.gov/articles/PMC4757966/

Braun, V., and Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597.
https://doi.org/10.1080/2159676X.2019.1628806

Braun, V., and Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, *18*(3), 328–352.
https://doi.org/10.1080/14780887.2020.1769238

Braun, V., and Clarke, V. (2023). *Thematic analysis: A practical guide*. SAGE.
https://doi.org/10.1177/14733250231170275

Brohi, S., Mastoi, Q. U. A., Jhanjhi, N. Z., and Pillai, T. R. (2025). A research landscape of agentic ai and LLMs: Applications, challenges and future directions. *Algorithms*, *18*(8), 499.
https://doi.org/10.3390/a18080499

Chen, K., Zhou, X., Lin, Y. *et al.* (2025). Privacy risks and protection mechanisms in LLMs. *Journal of AI Privacy and Security*, 4(1), Article 177.
https://doi.org/10.1007/s44443-025-00177-1

Glaser, B. G., and Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine Publishing Company**.**

Hannes, K., Booth, A., Harris, J., and Noyes, J. (2013). Celebrating methodological challenges and changes: reflecting on the emergence and importance of the role of qualitative evidence in Cochrane reviews. *Systematic Reviews*, *2*(1), 84.
https://link.springer.com/article/10.1186/2046-4053-2-84

Hennink, M., and Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social science and medicine*, *292*, 114523.
https://doi.org/10.1016/j.socscimed.2021.114523

International Organization for Standardization. (2023). *ISO/IEC 23894:2023 - Information technology - Artificial intelligence - Risk management*. ISO.
ISO/IEC 23894:2023 - AI — Guidance on risk management

Jaffal, N. O., Alkhanafseh, M., and Mohaisen, D. (2025). LLMs in cybersecurity: A survey of applications, vulnerabilities, and defense techniques. *AI*, *6*(9), 216.
https://doi.org/10.3390/ai6090216

Jiao, J., Afroogh, S., Xu, Y., and Phillips, C. (2025). Navigating LLM ethics: Advancements, challenges, and future directions. *AI and Ethics*, 1-25.
https://link.springer.com/article/10.1007/s43681-025-00814-5

Liu, F., Jiang, J., Lu, Y., Huang, Z., and Jiang, J. (2025). The ethical security of LLMs: A systematic review. *Frontiers of Engineering Management*, *12*(1), 128-140.
https://link.springer.com/article/10.1007/s42524-025-4082-6

National Institute of Standards and Technology (NIST). (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce.
https://doi.org/10.6028/NIST.AI.100-1

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. SAGE Publications, inc.

Rahimi, S. (2024). Saturation in qualitative research: An evolutionary concept analysis. *International journal of nursing studies advances*, *6*, 100174.
https://doi.org/10.1016/j.ijnsa.2024.100174

Schreiber, A., & Schreiber, I. (2025). AI for cyber-security risk: harnessing AI for automatic generation of company-specific cybersecurity risk profiles. *Information & Computer Security*, *33*(4), 520-546.
https://doi.org/10.1108/ICS-08-2024-0177

Sun, Y., Yu, H., Jiang, W., Yu, X., Zhan, D., Wang, L., ... and Zhu, T. (2026). A Survey on the Unique Security of Autonomous and Collaborative LLM Agents: Threats, Defenses, and Futures.
https://doi.org/10.20944/preprints202602.1655.v2

Thomas, J., and Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*, 45.
https://doi.org/10.1186/1471-2288-8-45

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., and Zhang, Y. (2024). A survey on LLM security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, *4*(2), 100211.
https://doi.org/10.1016/j.hcc.2024.100211

Zhang, R., Li, H. W., Qian, X. Y., Jiang, W. B., and Chen, H. X. (2025). On LLMs safety, security, and privacy: A survey. *Journal of Electronic Science and Technology*, *23*(1), 100301.
https://doi.org/10.1016/j.jnlest.2025.100301

Zhou, B., and Lin, J. (2025). Security vulnerability analyses of LLMs: An extended scoring approach. *Journal of Information Security Research*, 13(2), 1–18.
https://www.scirp.org/journal/paperinformation?paperid=133503