Vantage: Journal of Thematic Analysis
(Vantage J. Them. Anal.) ISSN(E): 2582-7391 Vol. 7(1), April, 2026, 36-59

Review Article

Security Vulnerabilities in Large Language Models: Bibliometric Analysis

Sanjay Singh[1], Anamika Gupta[2],

[1,2] Shaheed Sukhdev College of Business Studies, University of Delhi, India.

Corresponding Author: anamikargupta@sscbsdu.ac.in

Received: 19/02/2026 | Accepted: 29/04/2026 | Published: 05/05/2026

ABSTRACT

The sudden rise of large language models (LLMs) has increased concerns about their security, particularly regarding to weaknesses such as adversarial attacks related to prompting, jailbreaking attacks, and data leakage. However, a comprehensive understanding related to the developments in this area of research has been restricted despite numerous publications being generated along these lines. The current research encompasses a bibliometric analysis related to security weaknesses associated with LLMs based on 410 peer-reviewed publications from Scopus and Web of Science between 2021 and 2025. Although the eligibility window begins in the year 2021, no publications were found to be eligible for that year. A few early-access publications and preprints from the year 2026 are included provided they are indexed in that year, as this is a rapidly growing area of literature. The study uses the Bibliometrix package in the R programming language to integrate performance measures with concept and theme mapping methods such as keyword co-occurrence graphs, theme development measures, trend topic analysis, and three-way plots that relate sources to authors and keywords. Findings show an explosion in research work post-2023 driven by the rapid development in prompt attacks and jailbreak methods. Conceptual analysis also reveals that adversarial threats and jailbreak threats are central to LLM security studies, which are largely tied to conceptual probes on foundational models. Data privacy risks such as data leakage and membership inference attacks emerge as technologically rich but relatively niche topics, while system security perspectives are gradually gaining prominence. Also, analysis shows that there is a paradigm shift from foundational AI studies and NLP towards security applications and safety regulation. Thus, this work illustrates the rapid maturity as well as the fragmentation of the LLM security body of work. By integrating the literature that has been conducted to date, the results have aided in the foundation of a resource that will enable the building of more comprehensive models to secure large language models.

Keywords: Large Language Models, Jailbreak Attacks, Prompt Injection, Model Inversion, Security Vulnerability

1. INTRODUCTION

Large Language Models (LLMs) have significantly impacted the AI landscape by enabling state-of-the-art natural language understanding and generation capabilities in various application fields such as software development, health care, education, and security (Brown et al., 2020; Devlin et al., 2019; Vaswani et al., 2017). LLMs have been developed based on transformer architecture and large-scale text datasets to achieve excellent learning capabilities in the few-shot and zero-shot learning regimes to make them applicable in various systems ranging from consumer systems to enterprise systems. Recent

© The Author(s) 2026. This work is licensed under a Creative Commons Attribution 4.0 International License which permits its use, distribution and reproduction in any medium, provided the original work is cited. https://vantagejournal.com

advances in LLMs in terms of the ability of models such as GPT-4 to perform emergent reasoning have significantly enhanced their adaptability to various complex tasks (Bubeck et al., 2023; OpenAI, 2023a). Nevertheless, LLMs still seem to be prone to a set of security risks, as has become evident from current literature. Traditional software systems differ from LLMs in their inference mechanism and representation learning approach; making it difficult for classical security measures to effectively counter these risks. Existing works have clearly shown that safety alignment functions can easily be sidestepped using crafted malicious inputs, also known as "jailbreak attacks," resulting in the LLMs being able to produce prohibited or harmful outputs (Carlini et al., 2023; Zou et al., 2023). Prompt Injection Attacks are among the most dangerous threats posed to the integrity of LLM-based models. In Prompt Injection Attacks, the attacker uses the input prompts or context cues that can bypass system-level defensive tools without needing access to the inner workings of the models (Greshake et al., 2023). The danger posed by this type of attack is even more serious in real-world settings in which the models are layered in autonomous agents, R-A pipeline models, or external script models that support remote prompt injection attacks (Greshake et al., 2023; Naito et al., 2023). Research has suggested that even the most carefully calibrated models might be forced into dangerous actions through the adoption of carefully crafted input prompt attack methodologies (Qi et al., 2024).

In addition to these prompt-based attacks, LLMs are vulnerable to other privacy-related threats such as data leakages, model extraction attacks, and membership inference attacks, as well as model inversion attacks. Previous works have demonstrated that protective data could be partially reconstructed from trained LLMs, especially when the models were over-parameterized or under-regularized (Andriushchenko et al., 2020; Carlini et al., 2021; Song et al., 2017). Membership inference attacks further show that attackers could infer whether a set of data samples were used during training (Shokri et al., 2017). These threats could be of great concern when working with regulated sectors such as healthcare or finance, as data confidentiality becomes a core priority (Menz & Spinaci, 2024; Tsouplaki, Kalloniatis, & Mikros, 2026). To counter these, methods like defensive approaches related to adversarial training, prompt filtering, decoding time defenses, and post-generation moderation using language models have been proposed (Goyal et al., 2024; Xie et al., 2023). However, most of these methods still fall within the category of being reactive, limited, or more susceptible to adaptive attacks. Also, more recent results indicate that even when there is no ill intention on the part of the user, using fine-tuning on well-aligned language models can lead to the compromise of safety properties (Qi et al., 2024; Zhang et al., 2024).

Despite the wide growth of studies focusing on vulnerabilities in LLMs, literature is highly scattered in various academic disciplines, publication types, and methodological types. The current state of surveys provides information only about certain categories of attacks or certain types of defenses but does not provide a comprehensive understanding of overall research trends in the literature (Al Kuwaiti & Ismail, 2026; Das et al., 2025; Huang et al., 2025; Kumar et al., 2024; Mittal & Srinivasan, 2026; Pantserev, 2020). This leads to a gap in understanding the overall development of security vulnerabilities in LLMs, the prevailing threats in the overall research goals, and relationships among vulnerability themes.

The current problem can be addressed by undertaking an in-depth investigation on the security vulnerabilities in large language models from 2021 to 2025 in the form of an in-depth bibliometrics study of the related research published

in peer-reviewed journals indexed in Scopus and Web of Science. The current research studies methods and techniques used in the analysis of performances in the studies published in peer-reviewed journals indexed in Scopus and Web of Science in order to give an evidence-based account on the development of security studies in LLMs.

Accordingly, this study addresses the following research questions:

- RQ1. How has the volume and distribution of scholarly publications on security vulnerabilities in large language models evolved between 2021 and 2025?
- RQ2. Which publication venues and document types play the most prominent roles in disseminating research on LLM security vulnerabilities?
- RQ3. Which studies have exerted the strongest scholarly influence within this field, as reflected through global citation impact?
- RQ4. What are the dominant conceptual themes and keyword structures underpinning research on security vulnerabilities in large language models?
- RQ5. How have key research topics and thematic priorities related to LLM security evolved over time?
- RQ6. What research gaps and future directions emerge from the conceptual and temporal structure of the existing literature?

2. METHODS

2.1 Research Design

This study has adopted a bibliometric methodology for researching literature related to security vulnerabilities in LLMs. Bibliometric analysis is used to examine scientific literature from a quantitative and structural point of view, where scientists can also identify the development or advancement patterns in literature related to a certain field of science. This methodology will therefore enable the estimation of growth in LLM security literature from an interdisciplinary field.

The scope with respect to time was constrained to papers that were indexed during 2021 through 2025. No studies from the year 2021 were found to meet the eligibility criteria which was expected, as it marked an early stage when research on LLMs was still in its infancy. The analysis concentrates on peer-reviewed articles that appeared between 2022 and 2025 since these years relate to a rapid deployment of large language models and development of a new type of threat through prompting. The process of methodology implementation is in line with best practices when doing bibliometrics and science mapping. The existing records from the databases showed a publication year of 2026 but contained content that had already been indexed because of early-access publishing, which existed at the time of data collection. The dataset maintained these publications because they contained the latest research findings, which Scopus and Web of Science provided at the time of data retrieval. Early-access labeling exists as a standard practice within artificial intelligence research because scientists need to access their work before they officially assign their final publication date.

2.2 Data Sources and Search Strategy

For this research, two prestigious bibliographic citation indices, Scopus and Web of Science, shall be employed. These citation indices are two of the most commonly employed citation indices in bibliometric analysis because of their ability to cover a wide variety of fields across their metadata standards (Al Kuwaiti & Ismail, 2026), aside from their stringent inclusion standards (Mittal & Srinivasan, 2026). For the purpose of identifying appropriate publications, search queries were designed uniquely for each database, keeping in mind that some databases have

different fields of indexing and ways of crafting search queries. The search query was aimed at publications that reported on vulnerabilities in large language models, paying attention to prompt attacks and exploitation.

The search query used specific vulnerability-related keywords to identify documented attack methods that LLM security researchers have described in their emerging research. The terms prompt injection," "jailbreak," "model inversion," "membership inference," and "model extraction" represent established vulnerability classes that researchers investigate in adversarial machine learning and AI security studies. The query development process included additional terms that researchers used to create their own security-related content. The preliminary testing of the query showed that security-related content about LLM systems retrieved excessive research articles that did not relate to LLM security vulnerabilities. The research team developed their final search method by choosing attack-related terms that directly linked to LLM adversarial exploitation as their primary focus.

Search Query:

TS = ("large language model" OR "large language models" OR "LLM" OR "LLMs")

AND

("prompt injection" OR "jailbreak" OR "jailbreaking"

OR "data leakage" OR "model leakage"

OR "model extraction" OR "membership inference"

OR "model inversion")

AND

(security OR vulnerability OR vulnerabilities)

AND PY = (2021-2025)

The query was phrased with an attempt to combine precision and breadth:

- Model identifiers ("LLM," "large language model") captured general LLM-related work.
- The terminologies used in this paper, such as attack and vulnerability terms, ensured that the datasets focused on security-relevant research.
- Year restrictions allowed the analysis to reflect the modern LLM era, beginning when transformer-based generative systems first saw widespread adoption.

The query was framed broadly in order to capture various terminologies applied across AI, cybersecurity, and privacy research studies. English language publications only. The document types were journal articles and conference papers, as these tend to be the dominant dissemination formats for computer science and artificial intelligence research studies.

2.3 Data Collection and Preprocessing

After gathering the raw datasets, each collection was cleaned and standardized into a single Scopus and WoS collection in the R package Bibliometrix using the Biblioshiny interface (Aria & Cuccurullo, 2017). The process involved several well-framed steps to make the metadata field consistent and remove duplicate entries that had appeared in both databases.

Step 1: Raw Exports Import

Both the Scopus and Web of Science files were downloaded separately in CSV format. In Biblioshiny, each file was uploaded individually as an external file with the built-in import tool from the application that automatically converted each one into an RData file in a compatible format with Bibliometrix. The results were standardized with regard to field tagging, including title (TI), authors (AU), publication year (PY), encoding, and special characters. Once imported, each dataset was exported from Biblioshiny as a clean one. RData file. This ensured that both the

collections followed the same field structure before merging.

Step 2: Merging Collections and Removing Duplicates

Next, the Scopus and WoS .RData were combined using Biblioshiny through "Merge Collection." According to this tool, duplication findings depend on the title, DOI, and author matching. The selection and deduplication workflow is summarized in the PRISMA flow diagram as shown in Figure 1.

- Initial combined size: 475 records
- Duplicates removed: 65
- Final merged dataset: 410 distinct records

When examining datasets, a few entries were found which were published in the year 2026. These articles were selected since they were indexed as early access/preprint publications in the 2025 indexing cycle. In LLM security research, where topics mature rapidly, indexing, not final print publication, is used in bibliometrics as the inclusion criterion.

Step 3: Keyword Cleaning and Normalization

Next, a keyword preprocessing operation was carried out to eliminate the disparities among the keywords contributed by the authors and to normalize thematic correctness. The keyword normalization within the combined keyword field proceeded according to rule-based methods and regular expression processing to structure the keywords systematically. The procedure required transforming all keywords into lowercase letters while eliminating duplicate terms that existed within the records and fixing all spelling errors and expanding all necessary abbreviations and uniting all different yet similar meanings of words. The dataset required unification through the combination of different forms, which included both singular and plural versions and all possible parenthetical forms and all acronym-based versions that represented the same concept. This normalization process was conducted before performing keyword-based analysis in order to keep matching terms together as one unit in the bibliometric networks. A particular focus was placed on the normalization of variants of basic concepts such as "large language model," "jailbreak," and "prompt injection" to specifically assign each concept with a keyword. This significantly reduced fragmentation in the use of keywords, thus improving the validity and reliability of keyword frequency, co-occurrence, and thematic mapping.

2.4 Application of Bibliometric

The purified dataset was processed for the study of performance as well as science mapping through the Bibliometrix instrument. By the application of performance indicators, publication increase, source productivity, and citation impact were analyzed to provide a descriptive view of the scenario. The intellectual foundation of the field of study was explored by analyzing the conceptual structure by keyword co-occurrence networking and thematic mapping. These methods identify groups of co-occurring keywords, which demonstrate the underlying themes in the research area and their connections. The trends that were studied related to the themes in which developments were taking place in relation to those that are under development in LLM security research.

In addition, a three-field graphic presentation has been developed to represent the connections between authors, sources, and keywords. In this way, it becomes clear how the scientific output of single scientists or teams of scientists is allocated among scientific publications and domains because such a presentation enables the integrated consideration of the scientific landscape (Al Kuwaiti & Ismail, 2026).

2.4.1 Performance Analysis

Performance analysis was used to determine the characteristics and growth trend of literature. The following indicators were assessed:

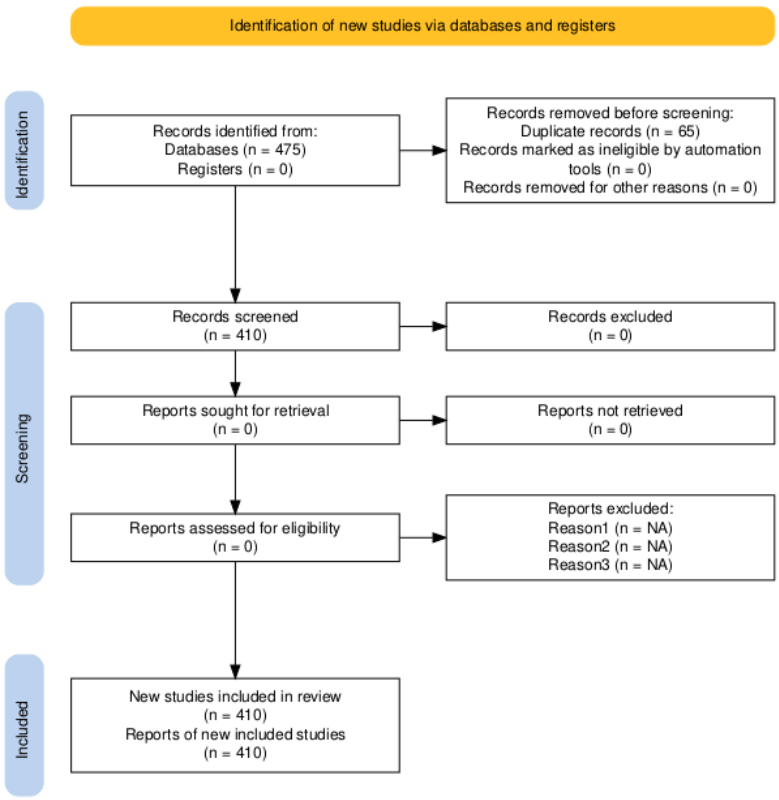

Figure 1: PRISMA flow diagram.

- Annual Scientific Production: It gives a view of the temporal development of the publications from 2021 to 2025.
- Most relevant sources: those journals and conference proceedings that have the highest number of publications in the dataset are identified.
- Most cited documents by showing the studies receiving great scholarly attention in the analyzed corpus.
- Keyword frequency analysis based on the keywords and index terms provided by the authors for pinpointing the concepts of vulnerabilities that keep cropping up.

These indicators give a descriptive overview of how research activity on the security of LLMs has developed in time and where it is mostly diffused.

2.4.2 Conceptual Structure Analysis

Various techniques of conceptual structure were applied in order to analyze the intellectual content and thematic organization of the field. First, the keyword co-occurrence analysis was done based on the availability of author keywords and index terms. Co-occurrence networks were visualized and analyzed for clusters representing a close concept, which then showed some dominant research topics related to jailbreak attacks, prompt injection, adversarial prompting, privacy leakage, model inversion, etc. Furthermore, a thematic map was built that classified the themes based on their centrality and density. This allowed identification of central and well-elaborated motor themes, emerging or declining themes, niche topics, and foundational themes. Temporal dynamics were captured by the performance of trend topic analysis concerning how key security-related terms evolved in a temporal sense, underlining the shift in research focus due to the emergence of new LLM capabilities, attack techniques, and defensive strategies.

2.5 Limitations of the Method

The study offers a comprehensive bibliometric overview of research on security vulnerabilities in large language models, yet multiple research constraints must be recognized.The research only considers documents that have been indexed into Scopus and Web of Science databases. The two databases provide extensive coverage of high-quality materials, but the research failed to include studies that only appear in arXiv or ACM Digital Library or industry white papers. The research method encountered challenges because emerging security research in LLMs does not have established peer-reviewed sources yet.The dataset construction process used a query-driven inclusion methodology instead of conducting manual screening procedures. The bibliometric research method establishes objectivity and reproducibility through its standard procedure, yet it permits the inclusion of papers that mention LLM security concepts but do not use them as their main research topic. The research study probably did not find studies which discuss the same security vulnerabilities because of their use of different terms.

The secondary metadata fields contained numerous fields which had incomplete information. The merged dataset lacked complete DOI information and keyword fields while the cited reference data remained completely absent. The research process required intellectual structure analyses through reference co-citation networks and historiographic mapping which did not occur. The analysis process focuses on performance metrics and conceptual frameworks that researchers established through their examination of titles and abstracts and keywords.

The database shows early-access publications from 2026, which reflect indexing methods instead of showing finished publication periods. The records were kept to show current research patterns, but researchers must be careful when interpreting citation-based indicators for the most recent years. The research methodologies of bibliometric analysis examine scholarly communication patterns and relationships between different research works instead of evaluating particular security weaknesses or their protective solutions. The research results explain how research themes develop and connect with each other, but they do not evaluate the actual protection abilities or danger levels of particular LLM security threats. The dataset provides enough information despite its limitations to enable strong conceptual studies and trend-based research about emerging developments in large language model security research.

3. RESULTS

A total of 410 peer-reviewed publications concerning the vulnerabilities of large language models were culled for the final dataset, which covered the years 2021 to 2025 and was indexed in 103 different sources of publications, although no records existed for the year 2021. There are a number of papers indexed as early access/preprints with an estimated publication year of 2026; these records remain indexed because they fall within the range of publication years 2025. The field shows a steep rate of growth with an average annual rate of 114.07%, which shows how much the field has grown with the mass adoption of generative AI technology. The average annual growth rate of publications was calculated using the compound annual growth rate (CAGR) formula commonly applied in bibliometric analyses (Equation 1):

$$\text{Growth Rate} = (N0/Nt)1/t - 1 \qquad (1)$$

where $N0$ represents the number of publications in the initial year, Nt represents the number of publications in the final year, and t denotes the number of years in the observation period. This metric provides a standardized estimate of the rate at which research output has expanded

over time and follows common bibliometric growth analysis methods implemented in the Bibliometrix framework (Aria & Cuccurullo, 2017). These publications were written by 1,493 individual authors, which indicates a wide range of people involved in research. Joint writing is common with a mean of 4.84 co-authors of each document; as a matter of fact, only 12 were solo-authored publications. Cross-national collaboration has remained lower with 4.878% of all publications authored by people from different countries.

The conceptual and thematic environment of the study field is quite vibrant, as evident from the 771 unique author-provided keywords, which demonstrate both the technical variety and the shifting domains of LLM security study. Despite the relative freshness of most research work, it indicates that the mean number of citations per document is 4.093, pointing towards a beginning level of intellectual impact. Keeping in view the emergent character of the research domain, the mean document age is 0.337 years, implying that most contributions to the research domain have been produced during the past year. Taken together as descriptive criteria, these confirm that the body of work related to the study of vulnerabilities in large language models is young and growing.

3.1 Annual Scientific Production and Growth Trend

Figure 2 below shows the annual scientific output on security vulnerabilities in large language models from 2022 through 2026. No documents produced in 2021 met the specified search criteria. From the results, there has been a visible positive trend with respect to the number of publications. Although there are not very many papers from 2022, this indicates that research on LLM security vulnerabilities was still in its initial phases at that time. Generally, most papers during this phase related to security concerns in adversarial machine learning and AI safety and not specifically LLMs (Goodfellow et al., 2015; Madry et al., 2018). There is a marked increase in the number of publications in 2023, which corresponds to the widespread adoption of LLA-style large-scale generative models. It also represents a turning point in which researchers started targeting attacks via prompts, jailbreaking, and aligning failures in LLMs (Greshake et al., 2023). It clearly continues to rise sharply in 2024 and 2025, symbolizing steady and expanding interest in the topic. This is because there has been rampant emergence of new vulnerability types such as prompt injection, privacy leakage, and system-level exploitation of LLM-integrated applications (Carlini et al., 2021; Greshake et al., 2023; Qi et al., 2024). Publications emerging in the year 2026 also indicate that LLM security is no longer an emerging trend but a research area with constant activities being conducted. Finally, based on the annual production trend, it has been observed that security flaws identified in large language models have continued to emerge as an area of research being driven by technological improvement as well as adoption.

Note: No publications were indexed in 2021. A small number of early-access studies labeled as 2026 appear in the dataset but were indexed during 2025.

3.2 Leading Sources of Publication

Table 1 shows the most pertinent sources that have helped generate at least one result on security vulnerabilities for large language models. The findings indicate that there is a significant publishing trend among conference-tailored sources that include Lecture Notes in Computer Science (LNCS), Lecture Notes in Networks and Systems (LNNS), and Communications in Computer and Information Science (CCIS). Taken together, these sources make a major contribution to the overall number of publications, which

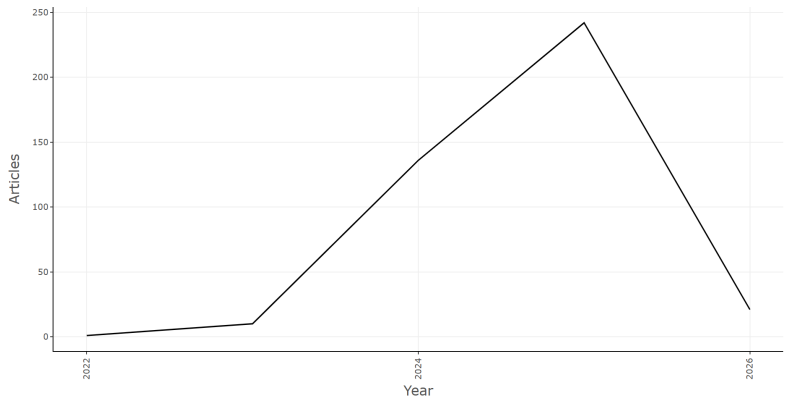

Figure 2: Annual scientific production (2021–2025)

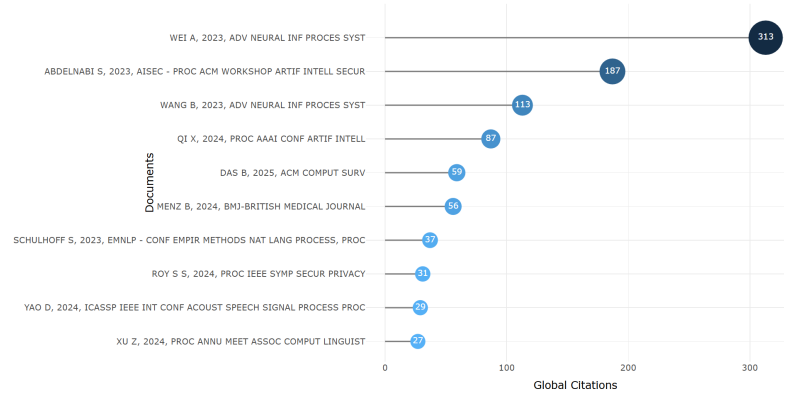

Figure 3: Most Globally Cited Documents

underscores the strong presence of conferences in sharing research on the security of LLMs. This is in line with how AI and cybersecurity research are published in academic circles, where fast-rising issues are mostly addressed in conferences with shorter review periods (Al Kuwaiti & Ismail, 2026; Mittal & Srinivasan, 2026). Conferences on artificial intelligence and natural language processing have a significant presence. The Proceedings of the Annual Meeting of the Association for Computational Linguistics have 24 entries in this blend of papers that highlight the importance of the NLP research community in the field of research in manipulating prompts, jailbreak attacks, and robustness in the NLP community. Advances in Neural Information Processing Systems and Proceedings of the AAAI Conference on Artificial Intelligence follow suit with 10 publications each.

Besides a series of conferences, there are other peer-reviewed journals such as Neural Networks, Pattern Recognition, and Expert Systems with Applications, which also qualify as highly relevant sources. Contributions found within these journals tend to be relatively mature studies involving formalized frameworks for attacks, experiments, and strategies for defense against LLM vulnerabilities (Qi et al., 2024; Xie et al., 2023). In sum, based on the source distribution, research into LLM security vulnerability is essentially prompted by the rapidly developing setting of academic conferences, in which novel security threats and countermeasures are introduced and discussed at a breakneck pace, with journal articles playing a supplementary role in building upon these discoveries.

Table 1: Most Relevant Sources as per the current review

Sources	Articles
Lecture notes in computer science	34
Proceedings of the annual meeting of the association for computational linguistics	24
Advances in neural information processing systems	10
IEEE access	10
Proceedings of the AAAI conference on artificial intelligence	10
Proceedings - international conference on computational linguistics, coling	9
Electronics	8
Applied sciences-basel	7

3.3 Most Globally Cited Documents

Figure 3 illustrates the most globally cited documents within the data set analyzed above, highlighting publications that have significantly influenced research on security vulnerabilities in large language models. The work that has been cited the most is that of Wei et al. (2023), which appeared in "Advances in Neural Information Processing Systems," with a total of 313 citations globally. This study can also be regarded as one of the pioneering works in the study of the vulnerabilities associated with large language models, mostly owing to its safety aspects.

The second most frequently cited reference is the work of Greshake et al. (2023), presented at the ACM Workshop on Artificial Intelligence Security that has garnered 187 citations. This paper highlights the importance of adversarial risk and security assessments focusing on LLM-based models in the domain of application security research.

Another highly cited contribution includes Carlini et al. (2023), published in Advances in Neural Information Processing Systems, with 113 citations, and Qi et al. (2024), published in the Proceedings of the AAAI Conference on Artificial Intelligence, with 87 citations. These papers further emphasize the importance of top-level AI conferences in influencing early discussions in the literature concerning vulnerabilities of LLMs, with specific regard to robustness and degradation of safety. Additionally, survey-oriented and interdisciplinary literature is recognized as among the most frequently cited documents. Notably, as illustrated in ACM Computing Surveys (Das et al., 2025) and British Medical Journal (Menz & Spinaci, 2024) in texts by Das et al. (2025) and Menz and Spinaci (2024), discussions about LLMs' concerns of security, privacy, and misuse have reached beyond traditional AI conference settings into broader computing and application fields.

Analytically, the citation distribution indicates that a small number of initial, rather influential studies have contributed greatly towards informing current work. The fact that conferences are the most dominant group of the most cited documents indicates that the area of LLM security is rapidly evolving, where quick knowledge diffusion/engagement is essential.

3.4 Keyword Analysis and Thematic Structure

In order to analyze the conceptual structures under which research on security vulnerabilities in large language models is carried out, keyword co-occurrence analysis and thematic mapping were undertaken based on author keywords and index terms.

3.4.1 Keyword Co-occurrence Network

Figure 4 displays the keyword co-occurrence network, which researchers created using author keywords and index terms. The network uses node size to show how often keywords appear while the edge thickness shows how strongly keywords occur together. The resulting structure shows multiple distinct thematic groups which together create the research framework used to analyze security vulnerabilities in large language models.

The network uses the keyword ”large language model” as its central node which links various thematic groups throughout the system. This node exhibits strong linkages with terms such as ”jailbreak,” ”prompt injection,” ”adversarial attack,” and ”AI safety,” indicating that prompt-based exploitation techniques form one of the most dominant research directions in the literature (Carlini et al., 2023; Greshake et al., 2023; Qi et al., 2024; Zou et al., 2023). The high connectivity of these keywords suggests that adversarial prompting and alignment bypass mechanisms are deeply integrated with broader discussions on LLM safety and security.

The second major research group investigates privacy protection through data safety research which includes the following keywords: ”data leakage” and ”membership inference attack” and ”differential privacy” and ”training data.” The internal relationships between these terms show strong connections, while their links to the prompt-based attack cluster remain less strong. The field of privacy research has reached its current state through dedicated technical development, which has now become its permanent research path within LLM security research. The keywords of this research area connect internally within their cluster while displaying weaker connections to the prompt-based attack cluster. The network shows limited inter-cluster connections that create a structural divide between privacy research and prompt-based vulnerability research because researchers study these fields through different methods that exist within LLM security publications (Carlini et al., 2021; Fredrikson et al., 2015; Shokri et al., 2017; Song et al., 2017).

The system-level and cybersecurity domain uses another group of keywords, which includes ”network security,” ”cybersecurity,” ”risk assessment,” and ”artificial intelligence.” The nodes link the adversarial attack cluster to AI-related topics because system-level security concepts enable researchers to study model-related vulnerabilities and operational deployment use cases (Naito et al., 2023; Tsouplaki et al., 2026; Wahréus, Hussain, & Papadimitratos, 2026).

A distinct but smaller cluster exists that connects natural language processing with its linguistic base through elements that include the keywords ”natural language processing” and ”language processing” and related terms used in NLP systems. The cluster shows LLM research methods as its foundation but links to security research through its conceptual framework. The network shows that NLP methods form the fundamental basis for research, although current studies focus on adversarial and security aspects of LLM systems.

The co-occurrence network shows a multi-cluster structure which contains three major thematic groups that investigate adversarial prompting and privacy leakage and system-level cybersecurity issues. The privacy cluster shows weak connections to the prompt-based attack cluster which demonstrates that the literature has been divided into different research areas that study distinct vulnerability types without using integrated threat assessment frameworks. The existing structural pattern shows researchers an important chance to use knowledge from multiple fields to create complete security solutions for large language models.

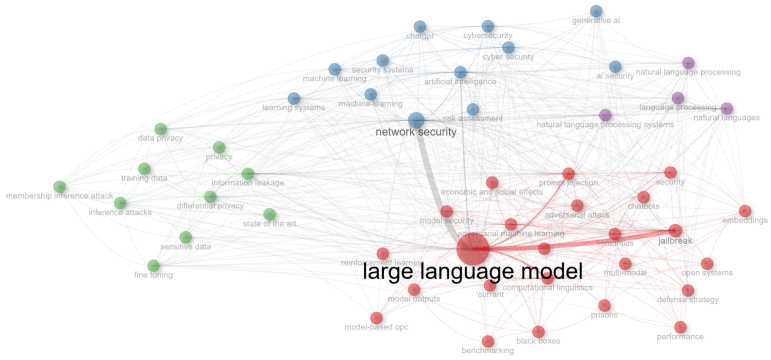

Figure 4: Keyword Co-occurrence Network

3.4.2 Thematic Map Analysis

The thematic map shown in Figure 5 contains two components that display research themes' importance through keyword centrality measurements and their development stage through density measurements. Centrality shows how much a research theme connects to other themes in the field, while density shows how much a theme develops its internal structure and existing connections. The motor themes quadrant (upper-right) contains themes that display both high centrality and high density, which demonstrates that these research topics exhibit both strong development and major impact on the research field. The study shows that network security, artificial intelligence, and cybersecurity themes, which dominate this quadrant, demonstrate that security-centered research about AI systems drives most research on large language model security vulnerabilities. The high connectivity between these themes demonstrates that they establish strong connections with other research areas throughout the field (Menz & Spinaci, 2024; Tsouplaki et al., 2026).

The basic themes quadrant (lower-right) includes topics that exhibit high centrality but relatively lower density. These themes function as essential concepts which link different research areas, yet they maintain their broad scope instead of developing into specific expertise. The analysis uses large language models and computational linguistics and black-box models as basic themes, which serve as principal conceptual elements that support LLM security and adversarial behavior research studies (Brown et al., 2020; Bubeck et al., 2023; Devlin et al., 2019; OpenAI, 2023a; Vaswani et al., 2017). The upper-left quadrant of niche themes contains themes which demonstrate complete internal development yet lack strong ties to the research network. The region includes research topics which study information leakage and differential privacy and data privacy. The research field shows advanced methodological progress through its high research density, but privacy research remains restricted to specific areas within the broader LLM security analysis framework (Fredrikson et al., 2015; Shokri et al., 2017; Song et al., 2017). The lower-left quadrant of the emerging or declining themes contains themes that exist at both low centrality and low density because the themes represent either fresh topics or subjects that are diminishing in importance. The current investigation shows that prompt injection and adversarial machine learning and natural language processing systems occupy this particular area. The research network has not yet reached complete integration of these topics, which continue to develop based on their increasing visibility in recent studies about LLM vulnerabilities (Carlini et al., 2023; Greshake et al., 2023; Qi et al., 2024). The research map shows an academic field which contains three main research areas about LLM technology and new attack methods and privacy

protection research. The theme distribution across quadrants indicates that the research field is moving from basic model studies toward security research which focuses on practical applications.

3.5 Trend Topics Over Time

Figure 6, given below, illustrates how major research areas associated with security vulnerabilities in large language models have evolved over time. The size of the marks in this figure (Figure 6) is based on the relative use of a term in a particular year, and the position of a mark corresponds to its extent in the dataset. In the initial phase of the dataset, keywords like computational linguistics, adversarial machine learning, and cybersecurity gained prominence with a relatively high number of appearances. These keywords can be attributed to initial research contexts encompassing LLM security concerns and not a specific research field (Goodfellow et al., 2015; Madry et al., 2018). Within this phase of research and development in LLMs and cyber threats, vulnerabilities related to LLMs received coverage as an extension of traditional adversarial learning as well as NLP concerns. Since 2023, the popularity of large language models has risen significantly, which marks a transition towards specific security analysis models. Notably, this transition aligns with the rising usage of large generative models and the advent of new attack primitives specific to follow-instructions and chat models (Bubeck et al., 2023; OpenAI, 2023a). In recent years, ”jailbreak” seems to be one of the most highlighted trend topics, representing an ever-growing amount of attention given to methods involving rapid exploitation and jailbreak alignment bypassing (Carlini et al., 2023; Greshake et al., 2023). Also, its annual entry into the lists with highlighted trend topics brings out the fact that this topic of jailbreak attacks is now being highlighted at the level of academic studies rather than merely being treated as a trend. Other topics include AI safety and network security that emerged as matters of concern during the later stages of development, representing an escalating interest in the field of system-level safety, security, and governance. This reflects a gradual widening of the research agenda from model-level vulnerabilities to security aspects associated with system-level LLM integration. (Naito et al., 2023; Tsouplaki et al., 2026; Wahréus et al., 2026). Thus, the topic trend analysis shows a clear movement from basic AI and adversarial learning themes to more applied and LLM-related issues of cybersecurity. The increase in the focused span of temporally proximate topics to jailbreaks, safety, and system integration issues indicates the increasing maturity of the topic area in relation to practice.

3.6 Three-Field Plot

Figure 7 is a three-field plot, where fields are related by linking publication sources (SO) with authors (AU) and merged keywords (KW), enabling an overall representation of how research topics about the security of large language models are related to their contributing authors and publication channels. Concerning authors, although a small group of researchers are active repeatedly across publications, most of them have published once or maybe two times at most, taking into account the fact that rapidly developing areas of research have initial participants defining methodological frameworks for future research followed by a broader group of people who examine attack paths/points of attack (Al Kuwaiti & Ismail, 2026). The fact that authors are connected to each other proves new competence groups emerge for jailbreak attacks and adversarial prompts, as well as privacy threats. The source dimension reveals the importance of conference-related publishing forums and proceedings collections such as Lecture Notes in Computer Science,

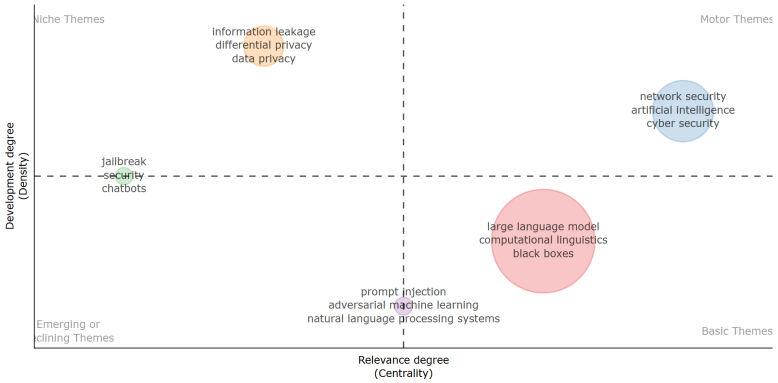

Figure 5: Thematic Map

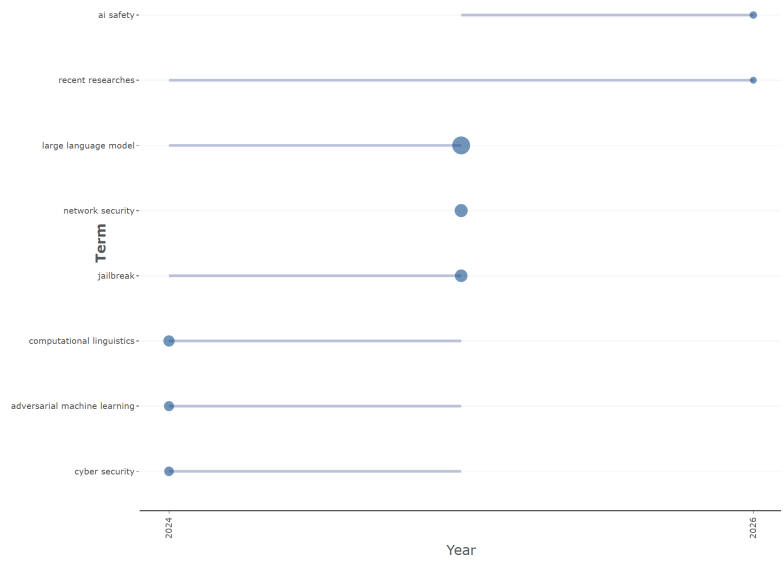

Figure 6: Trend Topics

Lecture Notes in Networks and Systems, and Communications in Computer and Information Science. The sources are used as a primary means of dissemination between producers of research works and main research topics. High connectivity indicates the priority of quick publication cycles, with early participation by peers in research on LLMs' security issues, particularly in newly discovered vulnerabilities and mitigation methods (Al Kuwaiti & Ismail, 2026; Carlini et al., 2023). Analyzing these keywords, words like "large language model," "jailbreak," "prompt injection," and "AI security" are very strong themes that help in connecting disparate writers and articles. This indicates that these are key terminology issues that are addressed at different forums. Keywords associated with issues related to privacy concerns like "membership inference," "data leakage," and others are indicative of selective patterns that are not very scattered at different forums (Carlini et al., 2021; Fredrikson et al., 2015; Shokri et al., 2017; Song et al., 2017). Altogether, the three-field layout presents an integrated relationship between the thematic investigations, the primary publication sources, and an expanding yet constantly shifting and scattered group of authors. Such an environment represents a developing field, with themes of vulnerability condensing despite diversification with respect to inquiries and application domains.

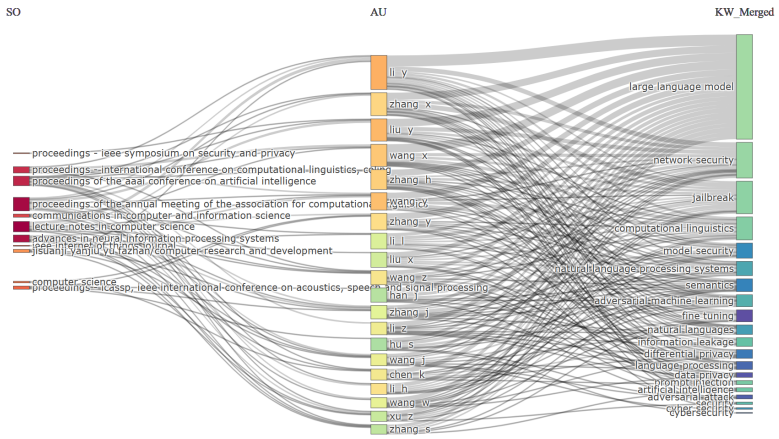

Figure 7: Three-Field Plot

3.7 Word Cloud Analysis of Author Keywords

Figure 8 depicts word clouds based on the keywords given by the authors for the 410 publications in the dataset. The relative font size of each term represents its relative frequency, thus providing an intuitive overview at first sight of which concepts prevail in the research activity on security vulnerabilities of large language models. Not surprisingly, given this central role, ”large language model” is the clear and dominating term underlying its function in tying together the wide range of literature reviewed here. Coupled in with that core concept is the suite of security-relevant terms ”jailbreak,” ”prompt injection,” ”network security,” and ”model security,” all focusing the mind on adversarial manipulation and system-level exploitation as points of critical interest. The prominence of ”artificial intelligence” reflects the degree to which this LLM-focused work is embedded within the wider AI security discourse.

Another cluster of high-frequency terms indicates the thematic differentiation of the text. Words and phrases such as ”computational linguistics,” ”natural language processing systems,” and ”semantics” reflect the continuing impact of research traditions in language technology (Brown et al., 2020; Devlin et al., 2019; Vaswani et al., 2017). In contrast, terms such as ”information leakage,” ”data privacy,” and ”differential privacy” indicate another strand of publications concerned with privacy-related hazards and confidential data disclosure (Carlini et al., 2021; Fredrikson et al., 2015; Shokri et al., 2017; Song et al., 2017).

Overall, the word cloud provides further visual support to both the co-occurrence and thematic analyses on the leading LLM-centric security concerns, including jailbreak and prompt-based attacks but also identifies entwined research strands of privacy-focused and NLP-oriented work (Carlini et al., 2021; Carlini et al., 2023; Greshake et al., 2023; Qi et al., 2024; Zou et al., 2023).

This confirms that the field has evolved around a small set of shared conceptual anchors, even as researchers explore diverse technical and applied perspectives on LLM security (Fredrikson et al., 2015; Naito et al., 2023; Shokri et al., 2017; Song et al., 2017; Tsouplaki et al., 2026; Wahréus et al., 2026).

4. DISCUSSION

This bibliometric analysis helps to systematically examine the development of research on security vulnerabilities in large language models over the years from 2021 to 2025. The findings indicate that the research area is expanding rapidly with increasing diversification of research at a time when LLMs are being widely used with new attack mechanisms being developed to defy

Figure 8: A Word Cloud

existing AI safety and security paradigms.

4.1 Rapid Growth and Expanding Event-Driven Research

The lack of publication in 2021 emphasizes further the relative recency of LLM research in security, which emerged as a mature area of study in 2022 and beyond. The steep jump in the publication volume from 2023 onwards reveals that this field of research has been event-driven in terms of the publication of research work in this field, right around the time of deployment of the best models in this field. This refutes the idea of a research field undergoing incremental increases in publication volumes over a mature research field in terms of a long-term research agenda. LLM security research exhibits a surge pattern in publication volumes, which is due to the attack vectors triggered by research work after being triggered by new discoveries that happen immediately after their occurrence. It can be noted that the preponderance of papers published in conferences also supports this observation. Conferences are rapid-response publications, which allow the rapid dissemination of results related to the discovery of jailbreak methods, attacks, and defensive tools with a latency of a few months at most (Carlini et al., 2023; Greshake et al., 2023).

The dataset contains a limited number of records which show a 2026 publication date because Scopus and Web of Science use early-access indexing. The analysis included these papers because they were already present in the databases at the time of data collection. The most recent year requires cautious interpretation because early-access articles have not yet built their citation counts and their publication cycle is still in development. The inclusion of early-access records will have a minor effect on the accuracy of trend visualization and growth rate estimation.

In addition, the rapid dissemination of papers also leads to an unbalanced level of consolidation. Essentially, the inclusion of early-access 2026 papers represents the ever-shorter time spans for publications.

4.2 Centrality of Prompt-Based Attacks and Jailbreak Vulnerabilities

To sum up, jailbreak attacks and prompt injection attacks are the most central and important themes of current research work on all levels. Their status as motor themes not only reflects the amount of active work but also the strong conceptual relationship with the broader discourse on the security of LLMs. This also reflects a tension between the ability to follow instructions and security alignment in LLM designs.

Contrary to typical adversarial attacks on machine learning models, prompt attacks tend to be extremely successful, even without the need for model access and, therefore, are very practical. The continued focus on jailbreak attacks does indicate that the current mechanisms for alignment are still inadequate, especially when faced with the possibility of intent-hiding attacks as expressed in a variety of heavily cited sources (Greshake et al., 2023; Zou et al., 2023).

4.3 Fragmentation Between Privacy and Prompt-Based Research Streams

According to this structure, there is a degree of separation between research on prompt-based attacks and the research associated with privacy. Though these two types of studies pertain to security, overall, these two streams of research tend to progress in disconnected ways with not much conceptual overlap. Research pertaining to privacy, such as membership inference and model inversion attacks, is very technical and of good quality (Carlini et al., 2021; Fredrikson et al., 2015; Shokri et al., 2017; Song et al., 2017) but is somewhat alien to the recent studies on jailbreak attacks. This could come about through a mismatch in methodologies related to performing evaluations and threat modeling. Privacy attacks tend to depend on formal adversarial models and statistical guarantees. Prompt-based attacks primarily rely on manipulation of behavior and misuse of systems. A confluence of these two representative research streams is a good research opportunity in the future, particularly in aspects of hybrid attack structures mixing privacy and manipulation attacks (Guo et al., 2026; Liu et al., 2024; Papernot et al., 2016; Tramèr et al., 2017).

4.4 Shift to System-Level and Deployment-Aware Security

Nevertheless, it is significant that themes related to network security and application-level cybersecurity began to appear in later years. This is because there is evident progression from analysis related to models to an analysis that is aware of deployment. This is reminiscent of LLMs being used in autonomous systems and cloud services, increasing vulnerabilities not limited to LLM outputs (Na et al., 2026; Naito et al., 2023; Tsouplaki et al., 2026; Wahréus et al., 2026).This transition shows that LLM security is no longer a problem isolated in AI alignment. Rather, this is a system security perspective in which software architecture, access control, and operation context must also be considered.

The presence of ”interdisciplinary venues” among the influential sources also reinforces this transition into consideration of security in a more general sense, which is application driven (Menz & Spinaci, 2024; Tsouplaki et al., 2026).

4.5 Maturity Signals & Challenges - Consolidation Research

The co-existence of highly influential early works and an ever-expanding number of new publications defines the state of partial maturity in the research field. In fact, pioneering works remain the foundation upon which research streams are based, and the rate at which new works are being published could potentially lead to redundancy and overlapping ideas. This effect is particularly evident in jailbreak research, where a method to subvert the system is often rediscovered or is an expanded form of earlier methods (Deshpande et al., 2023).

The level of integration in the three-field plot also suggests that the level of specialization is still largely among a few regular authors and journals. This quickens the initial pace, while later integration will be enabled by standardized benchmarks, protocols for evaluating research, and standardized taxonomies for attack/defense mechanics (Al Kuwaiti & Ismail, 2026; Mittal & Srinivasan, 2026). The field development stage is shown through two indicators which include the cross-national collaboration rate of 4.878% and the average citation count per document of 4.093. The research field on LLM security vulnerabilities needs international collaboration because research groups work on different attack methods and defense strategies which have not yet been explored. Research fields that have achieved maturity tend to experience increased international research collaboration when standardized methods and common research goals become established. The study results show that

recent research works do not meet the minimum requirement for citation counts. The majority of publications, which appeared between 2022 and 2025, have not yet received sufficient time to build up their citation counts. The citation-based indicators should be used with caution because they only show initial academic influence instead of showing permanent academic value.

4.6 Implications for Future Research

Taken together, these results support the knowledge that the state of LLM security research is increasingly shifting from a mode of exploration to one of problem structuring, focusing on risk prioritization. Nevertheless, the ongoing literature work appears disunified with respect to types of attack and method of approach as well as application scope. Otherwise, the defensive effort is expected to continue being overwhelmed by the innovation velocity of the attackers.

In terms of application and implications, the results provide a need for research related to the security analysis of model and context behavior together. For researchers, the bibliometric patterns signal possibilities of synthesis across isolated research streams, while for practitioners, they underscore the importance of following a principle of layered defense-strategies that take care of risks at the level of prompts, models, and systems. The bibliometric findings enable the development of a security vulnerability taxonomy that defines security vulnerabilities for large language models through the primary thematic clusters that emerged from keyword co-occurrence analysis and thematic mapping analysis (Figures 4 and 5). Three broad categories of vulnerabilities can be distinguished.

The most important and rapidly developing area of research focuses on prompt-based behavioral attacks which function as the primary attack method. The instruction-following behavior of LLMs gets exploited through jailbreak attacks and prompt injection and adversarial prompting and alignment bypass techniques (Carlini et al., 2023; Greshake et al., 2023; Qi et al., 2024; Zou et al., 2023). The research on these vulnerabilities maintains strong ties with AI safety research and model alignment research.The research area on privacy and data leakage vulnerabilities investigates methods that enable attackers to extract sensitive information from trained models. The category encompasses membership inference attacks and model inversion attacks and training data extraction techniques (Carlini et al., 2021; Fredrikson et al., 2015; Shokri et al., 2017; Song et al., 2017). The internal methodological development of these topics shows strong progress, but their connection with prompt-based attack research remains weak.

The research area examines multiple deployment scenarios which examine system-level vulnerabilities that emerge when LLMs connect with different software systems and networked environments. The research investigates cybersecurity issues and network security problems and operational deployment contexts which involve LLMs connecting with external tools and data pipelines (Menz & Spinaci, 2024; Naito et al., 2023; Tsouplaki et al., 2026; Wahréus et al., 2026). Thus, this taxonomy highlights that LLM security vulnerabilities span multiple layers of the AI ecosystem, including model behavior, training data exposure, and system-level integration risks. Understanding how these categories interact will be essential for developing comprehensive security frameworks for future LLM deployments.

5. FUTURE RESEARCH DIRECTIONS

The findings from the bibliometric analysis and discussion have shown a number of gaps both structurally and thematically with regards to security vulnerabilities of large language models. Although research into this field is increasing exponentially, this section identifies some of the most prominent gaps and sheds

light on directions for future research accordingly. The future research directions identified in this study are derived directly from the bibliometric findings presented in the Results section. The keyword co-occurrence network (Figure 4) shows structural fragmentation between major vulnerability research streams while the thematic map (Figure 5) displays different centrality and development levels of key research themes. The trend topic analysis (Figure 6) shows which new topics have become more important during the past few years. The research directions which follow this statement analyze underrepresented clusters together with low-centrality themes and emerging research trends to find research areas that will boost progress in large language model security.

5.1 Limited Integration Across Vulnerability Classes

Among the largest gaps that have been found from the keyword co-occurrence network (Figure 4) is the lack of cohesion between research streams in major vulnerabilities. Attacks based on prompts, especially jailbreak attacks and prompt injection attacks, in the literature overwhelmingly define research in the field (Carlini et al., 2023; Greshake et al., 2023). Meanwhile, work focused on privacy research in membership inference attacks, inversion attacks, and data leaks, which would seem a parallel research track (Carlini et al., 2021; Fredrikson et al., 2015; Shokri et al., 2017; Song et al., 2017).

Looking forward, it is important to focus research on more inclusive threat models that embrace the potential for a combination of different vulnerability types to occur within the same environment. For instance, manipulation techniques requiring a prompt may be merged with privacy inference attack tactics within a functional application.

5.2 Insufficient Focus on Deployment and System-Level Security

Though recent trend analyses show rising levels of focus on network security and related themes of cybersecurity, system-level issues are presently underrepresented compared with model-related work (Naito et al., 2023; Tsouplaki et al., 2026; Wahréus et al., 2026). The vast majority of current work focuses on evaluation of the effects of LLMs in isolation without consideration of the system-level effects of realistic settings. Future research should employ evaluation methods that are deployment-aware and analyze the spread of vulnerabilities across system boundaries. This encompasses surveying attack surfaces resulting from APIs, plugins, external tools, and data pipelines; assessing access control; monitoring and responding related to LLM-based systems, and other security-related aspects.

5.3 Lack of Standardized Benchmarks and Evaluation Protocols

The fact that the number of influential studies and attack rediscoveries is concentrated suggests the absence of standardized benchmarks and evaluation procedures for LLM security studies. Although a number of benchmarks have been proposed, they are still patchy and disorganized (Jiang et al., 2024; Nambiar & Pöpper, 2026; OpenAI, 2023b). Future work would be made easier with the establishment of a common set of benchmarks to enable comparable analysis of attack and defense techniques across different models and scenarios. Metrics to define success ratio, safety degradation, and defense robustness would also help in avoiding redundant work in defensive techniques against model jailing techniques (Goyal et al., 2024).

5.4 Underexplored Longitudinal and Comparative

The bibliometric data suggests that there is a great deal of literature related to short-term

assessments of freshly launched models or methods of attack. Long-term studies regarding the development of vulnerabilities in model versions are relatively rare.The future work should include longitudinal and comparative studies that analyze the persistence of vulnerabilities, mitigation efficiency, and regression effects. These studies would be very important in determining whether future protection methods are offering permanent security enhancements or are simply moving the point of attack (Zou et al., 2024).

5.5 Limited Attention to Governance and Risk Management Perspectives

Although there are publications that cross over a variety of disciplines and are oriented towards policy, these remain on the periphery of the primary conceptual framework concerning the subject (Bengio et al., 2024; Floridi et al., 2018; Menz & Spinaci, 2024). There is a focus on mitigation and a failure to consider the potential organizational or ethical issues surrounding the issues of LLM security. It will be important for future work to incorporate risk management principles and governance structures that align technology risk with risk of compliance and ethical issues. Indeed, this will be even more true for critical domains such as healthcare and finance, where LLMs can cause deep harm to society (Guo et al., 2024; Shiomi et al., 2026; Shokri et al., 2017).

5.6 Towards Integrated and Preventive Approaches to Security

On the whole, the identified gaps indicate that the area of security studies in LLMs is very reactive in nature since many security measures are developed in response to attacks on the models that have recently been identified. The area will thus benefit from embracing design-oriented security measures in their future studies (Qi et al., 2024; Xie et al., 2023).

Overcoming these challenges will necessitate increased collaboration between AI researchers, computer security experts, and system engineers. By incorporating multiple views related to vulnerability classes and settings, future work can help build more robust trustworthy large language models.

6. CONCLUSION

The study performed an extensive bibliometric study of research about security vulnerabilities in large language models through 410 peer-reviewed articles which Scopus and Web of Science listed between 2021 and 2025. The research study used performance indicators together with conceptual and thematic analysis methods to create a research map which showed how the field developed and its existing structure and upcoming research areas. LLM security research has evolved from being an unimportant topic to becoming a focus area because of two factors: first, researchers found new ways to attack machine learning systems through prompt-based attack methods, which include jailbreaks and prompt injection attacks. Academic research currently uses conference venues as its primary method for sharing research findings because security flaws and protection methods are discovered at an accelerated rate. The research domain maintains its present intellectual base through a limited number of initial studies which established its fundamental research principles. The study conducted a comprehensive bibliometric analysis which examined research on security weaknesses in large language models through a collection of 410 peer-reviewed articles that Scopus and Web of Science listed between 2021 and 2025. The research study used performance indicators together with conceptual and thematic analysis methods to create a research map that showed how the field developed and its existing structure and upcoming research areas.

The field of LLM security research developed

into a major research area because researchers developed new methods for attacking machine learning systems through prompt-based attack techniques, which include jailbreaks and prompt injection attacks. The academic research community uses conference venues as their main method to present research results because researchers identify new security vulnerabilities together with protective solutions at a fast pace. The research domain maintains its present intellectual base through a limited number of initial studies that established its fundamental research principles. The study offers a quantitative and structural assessment of LLM security research because it investigates the research domain through its collection of all available existing literature. The study uses bibliometric methods to identify dominant themes and conceptual clusters and emerging research trends, which create a systematic understanding of field development and future research directions. The results create a structured reference point that helps researchers and practitioners and policymakers understand how large language model risks develop and how to build secure and trustworthy AI systems.

Conflict of Interest

The authors have no conflict of interest to declare.

Funding

Not Applicable

REFERENCES

Al Kuwaiti, M., & Ismail, H. (2026). Adversarial attacks on large language models: A survey. *In V. Bhateja, M. El Barachi, A. T. Azar, & D. K. Sharma (Eds.), Information system design: Big data analytics and data science (ISDIA 2025). Lecture Notes in Networks and Systems (Vol. 1539). Springer.* https://doi.org/10.1007/978-981-96-9248-4_40

Andriushchenko, M., Croce, F., Flammarion, N., & Hein, M. (2020). Square attack: A query-efficient black-box adversarial attack via random search. *In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 484–501). Springer.* https://doi.org/10.1007/978-3-030-58565-5_29

Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-q., Xue, L., Shalev-shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., Mcilraith, S., Gao, Q., Acharya, A., Krueger, D., . . . Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science, 384*(6698), 842–845. https://doi.org/10.1126/science.adn0117

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33,* 1877–1901.https://doi.org/10.48550/arXiv.2005.14165.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712.* https://doi.org/10.48550/arXiv.2303.12712.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. *In Proceedings of the 30th USENIX Security Symposium (USENIX Security '21) (pp. 2633–2650). USENIX Association.* https://www.usenix.org/conference/usenixsecurity21/presentation/carlini.

Carlini, N., Nasr, M., Choquette-choo, C. A., Jagielski, M., Gao, I., Oprea, A., & Tramèr, F. (2023). Are aligned neural networks adversarially aligned? *In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in Neural Information Processing Systems 36 (NeurIPS 2023). Curran Associates.* https://doi.org/10.48550/arXiv.2306.15447.

Das, A., Liu, X., & Zhang, J. (2025). A survey on privacy and security in LLMs. *ACM Computing Surveys, 57*(6), Article 1, 1–39. https://doi.org/10.1145/3712001.

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023). Toxicity in ChatGPT: Analyzing persona-assigned language models. *In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 1236–1270). Association for Computational Linguistics.* https://doi.org/10.18653/v1/2023.findings-emnlp.88.

Devlin, J., Chang, M.-w., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the*

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), 4171–4186. https://aclanthology.org/N19-1423/.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5.

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information. *In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15) (pp. 1322–1333). ACM.* https://doi.org/10.1145/2810103.2813677.

Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *In Proceedings of the International Conference on Learning Representations (ICLR 2015). Computational and Biological Learning Society.* https://doi.org/10.48550/arXiv.1412.6572.

Goyal, S., Hira, M., Mishra, S., Goyal, S., Goel, A., Dadu, N., Db, K., Mehta, S., & Madaan, N. (2024). LLMGuard: Guarding against unsafe LLM behavior. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(21), 23790–23792. https://doi.org/10.1609/aaai.v38i21.30566.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *In Proceedings of the ACM Workshop on Artificial Intelligence and Security (AISec '23) (pp. 79–90). Association for Computing Machinery.*https://doi.org/10.1145/3605764.3623985

Guo, S., Zhai, Y., Zhang, S., Zhao, L., & Wang, Z. (2026). IntentBreaker: Intent-adaptive jailbreak attack on large language models. *In Machine learning and knowledge discovery in databases: Research track (ECML PKDD 2025). Lecture Notes in Computer Science (Vol. 16016, Chapter 14). Springer.* https://doi.org/10.1007/978-3-032-06078-5_14.

Guo, X., Yu, F., Zhang, H., Qin, L., & Hu, B. (2024). COLD-Attack: Jailbreaking LLMs with stealthiness and controllability. *In Proceedings of the 41st International Conference on Machine Learning (ICML 2024). PMLR.*https://arxiv.org/abs/2402.08679.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models. *ACM Transactions on Information Systems, 43*(2), Article 42, 1–55. https://doi.org/10.1145/3703155.

Jiang, F., Xu, Z., Niu, L., Xiang, Z., Ramasubramanian, B., Li, B., & Poovendran, R. (2024). ArtPrompt: ASCII art-based jailbreak attacks against aligned LLMs. *In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024) (pp. 15157–15173). Association for Computational Linguistics.* https://doi.org/10.18653/v1/2024.acl-long.809.

Kumar, S. S., Cummings, M. L., & Stimpson, A. J. (2024). Strengthening LLM trust boundaries: A survey of prompt injection attacks. *In Proceedings of the 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS) (pp. 1–6). IEEE.* https://doi.org/10.1109/ICHMS59971.2024.10555871.

Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., Wang, K., & Liu, Y. (2024). Jailbreaking ChatGPT via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860.*https://arxiv.org/abs/2305.13860.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *In Proceedings of the International Conference on Learning Representations (ICLR 2018). OpenReview.net.* https://doi.org/10.48550/arXiv.1706.06083.

Menz, S., & Spinaci, G. (2024). Risks of generative AI in healthcare. *BMJ, 385,* q1164.https://doi.org/10.1136/bmj.q1164.

Mittal, V., & Srinivasan, M. K. (2026). State-of-the-art artificial intelligence security taxonomies. *In ICT analysis and applications. Lecture Notes in Networks and Systems (pp. 512–526). Springer.*https://doi.org/10.1007/978-3-032-06694-7_48.

Na, H., Kim, H., Yoon, D., & Choi, D. (2026). Countering jailbreak attacks with two-axis pre-detection and conditional warning wrappers. *In V. Nicomette, A. Benzekri, N. Boulahia-Cuppens, & J. Vaidya (Eds.), Computer security – ESORICS 2025. Lecture Notes in Computer Science (Vol. 16053, Chapter 13). Springer.* https://doi.org/10.1007/978-3-032-07884-1_13.

Naito, T., Watanabe, R., & Mitsunaga, T. (2023). LLM-based attack scenarios generator with IT asset management and vulnerability information. *In 2023 6th International Conference on Signal Processing and Information Security (ICSPIS 2023) (pp. 99–103). IEEE.* https://ieeexplore.ieee.org/.

Nambiar, S., & Pöpper, C. (2026). JailFact-Bench: A comprehensive analysis of jailbreak attacks vs. hallucinations in LLMs. *In M. Manulis (Ed.), Applied cryptography and network security workshops (ACNS 2025). Lecture Notes in Computer Science (Vol. 15655, pp. 23–42). Springer.* https://doi.org/10.1007/978-3-032-01823-6_2.

Openai. (2023a). GPT-4 technical report. *arXiv preprint arXiv:2303.08774.*https://doi.org/10.48550/arXiv.2303.08774.

Openai. (2023b). ChatGPT data usage FAQ. Retrieved April 27, 2026, from https://openai.com/.

Pantserev, K. A. (2020). The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. *In H. Jahankhani, S. Kendzierskyj, N. Chelvachandran, & J. Ibarra (Eds.), Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity.* Advanced Sciences and Technologies for Security Applications (pp. 37–55). Springer. https://doi.org/10.1007/978-3-030-38128-9.

Papernot, N., Mcdaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. *In Proceedings of the 2016 IEEE Symposium on Security and Privacy (S&P 2016) (pp. 582–597). IEEE.* https://doi.org/10.1109/SP.2016.41.

Shiomi, K., Lian, Z., Nakanishi, T., & Kitasuka, T. (2026). POSTER: Tricking LLM-based NPCs into spilling secrets. *In G. Yang, S. Liu, C. Su, A. Otsuka, & Z. Lian (Eds.), Provable and practical security (ProvSec 2025). Lecture Notes in Computer Science (Vol. 16172, pp. 483–487). Springer.* https://doi.org/10.1007/978-981-95-2961-2_26.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. *In Proceedings of the 2017 IEEE Symposium on Security and Privacy (S&P 2017) (pp. 3–18). IEEE.*https://doi.org/10.1109/SP.2017.41.

Song, C., Ristenpart, T., & Shmatikov, V. (2017). Machine learning models that remember too much. *In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17) (pp. 587–601). ACM.*https://doi.org/10.1145/3133956.3134077.

Tramer, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & Mcdaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. *In Proceedings of the International Conference on Learning Representations (ICLR 2018). OpenReview.net.* https://arxiv.org/abs/1705.07204.

Tsouplaki, A., Kalloniatis, C., & Mikros, G. (2026). Adopting LLMs in internet of cloud ecosystems: Identifying the key privacy challenges. *IFIP Advances in Information and Communication Technology. Springer.* https://link.springer.com/series/6318.

Qi, X., Zeng, Y., Xie, T., Chen, P.-y., Jia, R., Mittal, P., & Henderson, P. (2024). Fine-tuning aligned language models compromises safety. *In Proceedings of the 38th AAAI Conference on Artificial Intelligence (Vol. 38, pp. 21534–21541). AAAI Press.* https://doi.org/10.48550/arXiv.2310.03693.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30,* 5998–6008.https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wahreus, J., Hussain, A., & Papadimitratos, P. (2026). Jailbreaking LLMs through content concretization. *In J. S. Baras, S. Papavassiliou, E. E. Tsiropoulou, & M. O. Sayin (Eds.), Game Theory and AI for Security: GameSec 2025.* Lecture Notes in Computer Science (Vol. 16223). Springer. https://doi.org/10.1007/978-3-032-08064-6_20.

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? *In Advances in Neural Information Processing Systems 36 (NeurIPS 2023).* Curran Associates. https://doi.org/10.48550/arXiv.2307.02483.

Xie, Y., Yi, J., Shao, J., Curl, J., Lyu, L., Chen, Q., Xie, X., & Wu, F. (2023). Defending ChatGPT against jailbreak attack via self-reminders. *Nature Machine Intelligence, 5*(12), 1486–1496.https://doi.org/10.1038/s42256-023-00765-8.

Zhang, T., Zheng, S., & Li, X. (2024). Jailbreaking LLMs via intent concealment. *In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024) (pp. 8989–9000). Association for Computational Linguistics.* https://doi.org/10.18653/v1/2024.acl-long.481.

Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal adversarial attacks on aligned LLMs. *arXiv preprint arXiv:2307.15043.* https://doi.org/10.48550/arXiv.2307.15043.

Zou, A., Phan, L., Wang, J., Duenas, D., Lin, M., Andriushchenko, M., Wang, R., Kolter, J. Z., Fredrikson, M., & Hendrycks, D. (2024). Improving alignment and robustness with circuit breakers. *In Advances in Neural Information Processing Systems 37 (NeurIPS 2024). Curran Associates.* https://doi.org/10.48550/arXiv.2406.04313.

COPYRIGHT AND LICENSE

By submitting and publishing the article in **Vantage: Journal of Thematic Analysis**, author(s) remains the copyright owner. This is a peer-reviewed multidisciplinary Platinum OA publication of the Centre for Research, Maitreyi College, University of Delhi (https://vantagejournal.com). The author(s) grant the publisher a non-exclusive, worldwide, perpetual license to publish, reproduce, distribute, archive, index, and communicate the work in all media and formats.

This is an Open Access article distributed under the terms of the **Creative Commons Attribution 4.0 International (CC BY 4.0) License**.

This license permits unrestricted use, sharing, adaptation, distribution, and reproduction in any medium or format, provided appropriate credit is given to the original author(s) and the source.

HOW TO CITE?

Sanjay Singh & Anamika Gupta. (2026). Security Vulnerabilities in Large Language Models: Bibliometric Analysis. *Vantage: Journal of Thematic Analysis*, 07(01), 36-59. https://vantagejournal.com/