A Bibliometric Review of Large Language Model Hallucination

Nur Emma Mustaffa[1], Ke En Lai[1], Christopher Nigel Preece[2], Foo Yeu Wong[1]

[1] Department of Quantity Surveying, Faculty of Built Environment, Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia

[2] Department of Construction Management, Faculty of Built Environment, Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia

DOI: https://dx.doi.org/10.47772/IJRISS.2025.909000409

Received: 08 September 2025; Accepted: 13 September 2025; Published: 13 October 2025

ABSTRACT

Hallucination, defined as plausible but factually incorrect outputs generated by Large Language Models, poses risks to knowledge reliability. Despite the growing use of AI in higher education, most research has concentrated on healthcare, leaving academic practices underexplored. This paper reviews research on hallucination in LLMs, with a focus on its implications for academia. While AI-generated content is increasingly used in education and research, limited attention has been given to how hallucinations affect academic credibility and integrity. A bibliometric analysis was conducted using Scopus, retrieving and analysing a total of 2,491 documents with VOSviewer. Keyword co-occurrence mapping, supported by a thesaurus file, was used to identify research trends and thematic clusters. This study provides a structured overview of hallucination-related research, highlights underexplored domains such as education and non-healthcare industriessss, and identifies priorities for future research. Hallucination is a central concern in LLM research, yet discussion is concentrated in healthcare. Academic contexts and technical fields like construction and law remain under-investigated, indicating a significant gap in awareness and application. Educators should integrate AI literacy and hallucination awareness into academic integrity training. Future studies should examine domain-specific causes and impacts of hallucination in academia and beyond healthcare. Raising awareness of LLM hallucinations can safeguard knowledge integrity, reduce misinformation, and promote ethical AI adoption. Further studies are needed in education, law, and industry-specific settings, alongside the development of robust detection and mitigation strategies.

Keywords— large language models, hallucination, academic integrity, AI in education, GPT-4

INTRODUCTION

Large Language Models (LLMs) have transformed artificial intelligence and natural language processing through their ability to generate fluent, human-like text. LLMs are increasingly integrated into academia, supporting tasks such as essay writing, summarization, coding, translation, and literature reviews [1]

Despite their utility in enhancing efficiency and accessibility, LLMs possess a critical limitation: hallucination. This phenomenon refers to the generation of content that, while syntactically sound and contextually plausible, is factually incorrect, unsubstantiated, or entirely fabricated [2]. For example, LLMs might invent academic references, misattribute statements to authors, or confidently present false data. Such inaccuracies are particularly problematic in academic and scientific settings, where precision, trustworthiness, and the integrity of citations are paramount.

Hallucinations stem, in part, from the extensive and imperfect datasets used to train LLMs. Inaccuracies, omissions, or ambiguities within the training data, coupled with imprecise user prompts, can result in outputs that appear convincing but lack factual grounding [3, 4]. This issue is especially critical in sensitive applications such as the summarization of medical records [5, 6], financial reporting, and legal analysis [7]. Notably, a case in New York involved attorneys submitting a legal brief containing fabricated citations generated by ChatGPT, which resulted in judicial sanctions [8].

The academic community has voiced significant concerns. Research indicates that while models like GPT-3 can produce plausible responses, they frequently generate statements that are factually questionable or misleading [9]. Furthermore, scholars have highlighted the potential for misuse in disseminating propaganda or disinformation. These instances underscore the need for critical evaluation of LLM outputs, particularly in scholarly, publishing, and educational contexts.

This paper addresses these challenges by conducting a bibliometric analysis to map the intellectual landscape of LLM hallucination research. The study aims to examine how hallucination is discussed in the literature, identify prevailing themes, and spotlight under-researched areas, including academic practices and non-healthcare industries. Consequently, this research contributes to ongoing discourse on preserving research integrity and promoting the responsible adoption of AI in higher education and professional fields.

Understanding Hallucinations In LLMs

Large language models exhibit "hallucination," a phenomenon where they produce outputs that are coherent and contextually relevant but factually inaccurate, unverified, or fabricated. This means LLMs may generate statements that contradict established knowledge (intrinsic hallucinations) or create entirely fictitious information (extrinsic hallucinations). This tendency stems from LLMs' reliance on statistical prediction of word sequences rather than a grounding in factual data, prioritising fluency over accuracy, a limitation that persists even in advanced models like GPT-4 [10].

Several factors contribute to this issue: LLMs lack real-time access to factual databases and may invent plausible-sounding details when prompts are vague or incomplete [11]. Additionally, training data that is biased, sparse, or lacks diversity can lead to flawed knowledge representations and consequently, inaccurate outputs [12]. Ambiguous prompts can also prompt LLMs to generate fabricated information to fill perceived gaps. Unlike human imagination, AI hallucination is an algorithmic outcome of statistical associations rather than conscious thought. The risks and implications of hallucination are evident across multiple sectors. In organizational settings, overreliance on LLMs without appropriate safeguards can undermine knowledge-sharing practices. Alhusban et al. (2025) [13] highlight that effective integration of generative AI requires both technical measures, such as privacy protection and verification protocols, and human-centered skills, including AI literacy and responsible use. In healthcare, the risks are even more pronounced, as hallucinations can compromise diagnosis, clinical decision-making, and patient safety. Prior studies have demonstrated the dangers of hallucinated outputs in summarising medical records [5, 6], and in generating scientific abstracts [3]. Similarly, in law and finance, hallucinations can have significant reputational and economic consequences. A widely reported case involved attorneys submitting a legal brief filled with fabricated citations generated by ChatGPT, which resulted in judicial sanctions [8].

While much of the early discussion has centered on healthcare and law, recent studies in educational technology illustrate how hallucinations disrupt both teaching and learning. For example, doctoral supervisors remain sceptical of LLM-assisted writing, noting that although GPT-4 chatbots can support early-stage academic tasks, the risk of fabricated or misleading outputs continues to undermine their credibility [14]. Similarly, student-facing studies reveal ambivalent perceptions. Learners view generative AI as a supportive tool for brainstorming, proofreading, and structuring assignments, yet they also report confusion and mistrust when hallucinations or misleading outputs occur during writing tasks [15].

Hallucination concerns extend beyond traditional academic writing to applied fields such as entrepreneurship education. In this context, GPT-4 has been shown to generate plausible but unreliable outputs, creating risks when students rely on fabricated data for business decision-making, thereby reinforcing the urgency of prompt engineering and retrieval-augmented methods to mitigate these issues [4]. Together, these findings suggest that hallucination is not confined to isolated tasks but instead constitutes a cross-sector challenge that directly impacts the credibility of AI-assisted education.

In academic and educational settings, hallucinations pose significant challenges by compromising the accuracy of citations and the reliability of research, thereby threatening the integrity of scholarly endeavours. Studies by Wu & Dang (2023) indicate that a mere 10% of references produced by ChatGPT were accurate, and

Anghelescu et al., (2023) noted the tool's propensity to fabricate bibliographic entries, presenting them as legitimate scholarly sources. Concurrently, evidence suggests that students leveraging generative AI tools may experience enhanced research efficiency and more profound engagement in learning activities [18, 19]. Nevertheless, the uncritical adoption of AI-generated outputs heightens the risk of unintentional plagiarism, the submission of inaccurate work, and the degradation of academic integrity. The scholarly publishing landscape is also encountering new difficulties, with journals already receiving manuscripts containing fabricated references [20]. Editors and reviewers face challenges in identifying AI-generated content when hallucinated references are present, and the absence of uniform disclosure requirements further complicates quality assurance processes. Carabantes et al. (2023) [21] contend that the incorporation of AI into peer review workflows introduces additional ethical and methodological considerations that are yet to be resolved.

To counter these risks, researchers are exploring various mitigation strategies. Improving the quality and diversity of training data is paramount to ensuring that models develop a more accurate and balanced representation of knowledge. Verification mechanisms, such as retrieval-augmented generation, can decrease hallucinations by anchoring outputs to external sources. Furthermore, adept prompt engineering—involving clear, specific, and structured instructions—has demonstrated efficacy in enhancing output reliability. Human oversight remains indispensable, as AI-generated content should invariably be cross-verified against authoritative sources before its application in sensitive or scholarly contexts. Practical approaches include utilising tools solely for their designated purposes, dissecting tasks into manageable segments, providing explicit parameters within prompts, and systematically validating generated outputs.

METHODOLOGY

To systematically map the intellectual landscape and thematic trends in hallucination research within large language models (LLMs), this study employed a bibliometric review approach. Bibliometric analysis is particularly suitable for emerging and rapidly evolving fields such as AI hallucination, as it enables researchers to assess large volumes of literature, identify publication trends, and uncover key themes and gaps through quantitative methods. This method is increasingly recognised as a rigorous and reproducible means of scientific knowledge mapping, particularly where the research base is expanding but lacks conceptual consolidation.

Data Search Strategy

Scopus was selected as the sole data source for this bibliometric review for both technical and methodological reasons. VOSviewer, the primary analysis tool used in this study, is designed to process one bibliographic database at a time, which restricts the possibility of merging datasets from multiple sources. Among the major indexing databases, Scopus provides extensive coverage of peer-reviewed journals, conference proceedings, and book chapters in the fields of computer science, engineering, and education, making it a reliable and comprehensive choice. Moreover, there is a significant overlap between Scopus and Web of Science, as most publications indexed in Web of Science are also available in Scopus. Given this overlap, and the established reliability of both databases, Scopus was chosen to ensure broad coverage and consistency in analysis while aligning with the technical constraints of the bibliometric software.

A search query was executed on 29 April 2025 using a Boolean combination of terms to capture literature related to hallucination phenomena within the context of AI and LLMs. The final query string included variations of the term “hallucination” (e.g., hallucinati*), as well as related terms such as “false information” and “wrong information”, in combination with “artificial intelligence”, AI, “large language model*”, and LLM. The precise search string is as shown in Table 1.

Table 1 Search string

Source	Search String
Scopus	(TITLE-ABS-KEY (hallucinati*)OR ("false information")OR ("wrong information") AND TITLE-ABS-KEY ("artificial intelligence" OR ai OR "large language model*" OR llm*))

This query yielded 2,491 documents, encompassing journal articles, conference proceedings, reviews, and other scholarly contributions.

All types of publications indexed in Scopus, including journal articles, reviews, conference proceedings, book chapters, editorials, and short communications, were included in the analysis. This decision was based on the need to develop a complete and representative overview of the research landscape surrounding AI hallucinations. As Ball (2018) [22] noted, bibliometric studies traditionally interpret "publication" in a broad sense, encompassing a wide array of scholarly outputs such as journal articles, conference papers, and book chapters. Given the novelty of the topic and the rapid pace at which research in AI evolves, restricting the dataset to only certain document types could lead to an incomplete understanding of the field.

Therefore, the inclusive approach adopted in this study ensures both comprehensive coverage and thematic breadth, allowing for a more robust analysis of emerging trends and research directions in the field of hallucinations in AI and LLMs.

Data Analysis

Cluster analysis was employed in this study to systematically identify and interpret the major thematic structures within the literature on hallucinations in large language models (LLMs). This method is particularly valuable in bibliometric research as it enables the grouping of frequently co-occurring keywords, thereby facilitating the recognition of dominant research areas and emerging topics.

The analysis was conducted using VOSviewer, a widely used tool for bibliometric visualisation and network analysis [23]. Developed by van Eck & Waltman (2010) [24], VOSviewer is capable of extracting and mapping bibliographic networks such as co-authorship, co-citation, and keyword co-occurrence, directly from databases like Scopus and Web of Science. Its strength lies in its ability to visually represent large datasets, where node size indicates frequency or importance, node colour represents thematic clusters, and line thickness shows the strength of association between terms[25]. Unlike word clouds, which depict isolated terms, VOSviewer creates networked visualisations that reveal the interrelationships between concepts, offering deeper insight into how themes like hallucination, ethics, misinformation, and trust are interconnected [26].

VOSviewer was selected as the primary analytical tool due to its established effectiveness in bibliometric mapping and its capacity for advanced network visualisation. The software utilises a modularity-based clustering algorithm with a resolution parameter, allowing for the detection of both broad and fine-grained thematic clusters [27]. In the visualisation output, node size reflects the frequency of keyword occurrence, node colour indicates cluster affiliation, and the thickness of connecting lines represents the strength of co-occurrence relationships. These features allow for a nuanced understanding of how concepts related to hallucinations in AI are distributed and interconnected across the scholarly landscape.

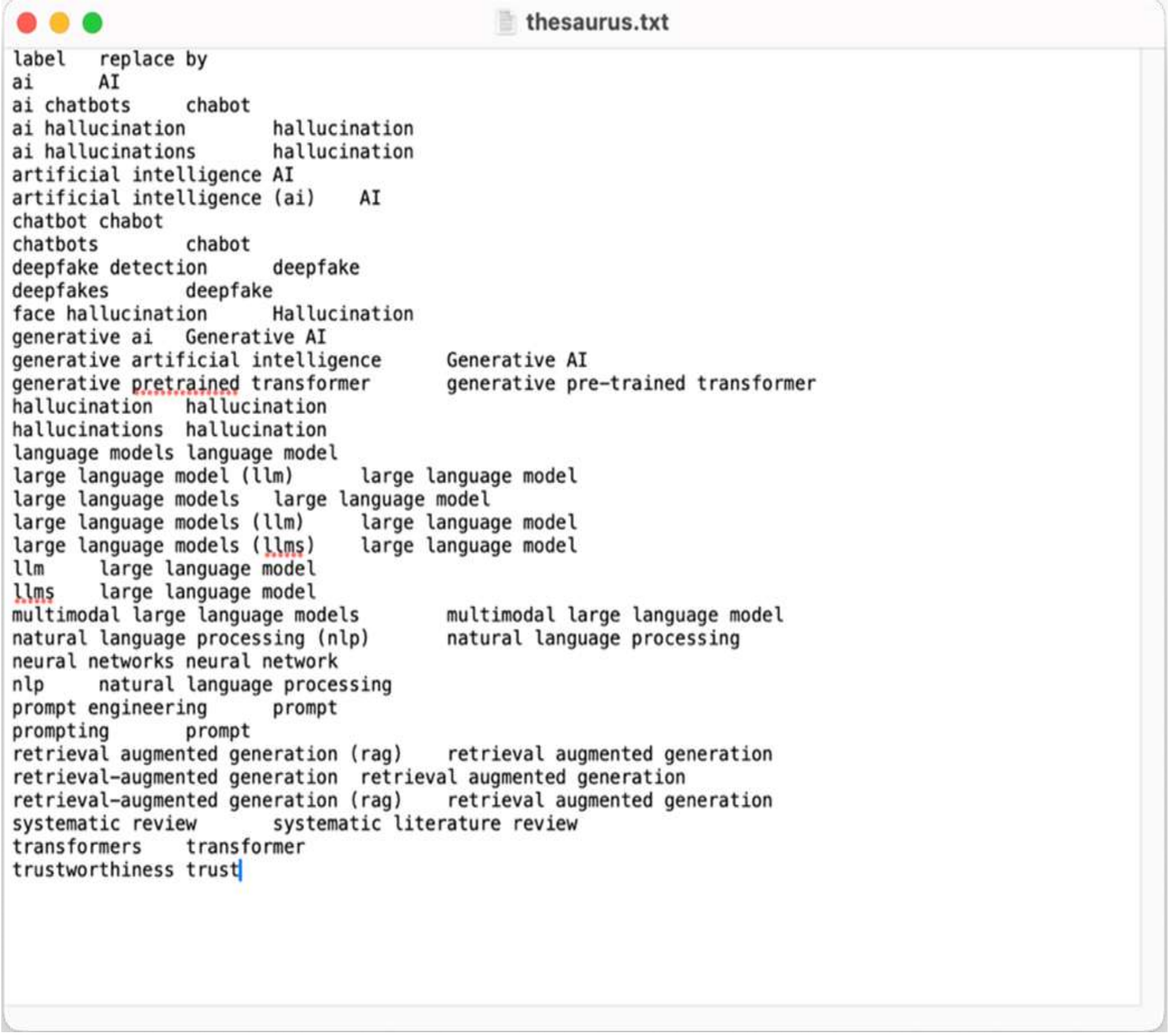
Following the automated clustering, thematic labels were assigned manually by reviewing the most frequent keywords and representative articles within each cluster. This step ensured that the cluster themes accurately reflected the underlying content and context of the publications, as recommended in interpretative bibliometric analysis [28]. The combination of algorithmic clustering and manual validation enhances both the objectivity and interpretive depth of the findings.

Thus, the use of VOSviewer in this study is both methodologically sound and practically appropriate, supporting a clear and structured analysis of thematic trends within a rapidly developing research domain.

The bibliographic data were exported in CSV format and subsequently imported into VOSviewer software to conduct a keyword co-occurrence analysis. A map was generated based on bibliographic data using the "co-occurrence" analysis type, specifically focusing on "author keywords" as the unit of analysis. To enhance the coherence and precision of the bibliometric analysis, a controlled vocabulary was applied through the use of a thesaurus file (see thesaurus.txt). The thesaurus grouped semantically equivalent terms under unified labels. For instance, terms such as “AI chatbots”, “chatbot”, and “chatbots” were standardised to “chatbot”, while “generative ai” and “generative artificial intelligence”

were harmonised as “Generative AI”. Similarly, variants of large language model (e.g. LLM, LLMs, large language models (llm)) were all merged under the umbrella term “large language model” (refer Figure 1). This step was essential to reduce semantic redundancy and increase the reliability of co-occurrence analysis, thereby improving the interpretability of cluster mapping and visualisations.

Fig. 1 Thesaurus file



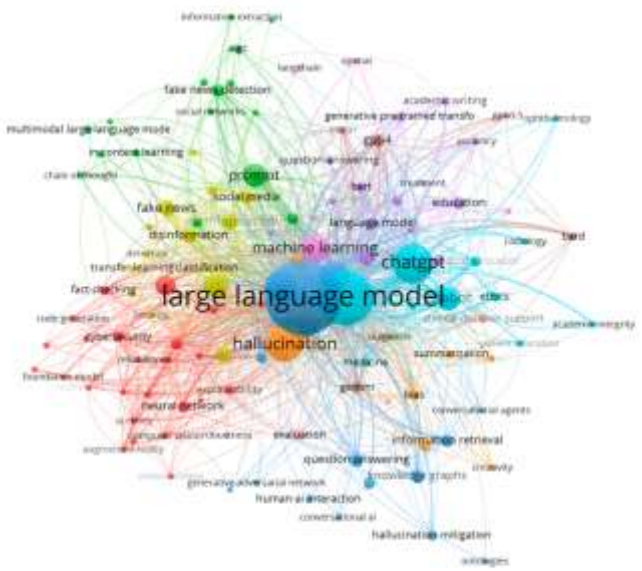
```
label   replace by
ai      AI
ai chatbots     chabot
ai hallucination        hallucination
ai hallucinations       hallucination
artificial intelligence AI
artificial intelligence (ai)    AI
chatbot chabot
chatbots        chabot
deepfake detection      deepfake
deepfakes       deepfake
face hallucination      Hallucination
generative ai   Generative AI
generative artificial intelligence      Generative AI
generative pretrained transformer       generative pre-trained transformer
hallucination   hallucination
hallucinations  hallucination
language models language model
large language model (llm)      large language model
large language models   large language model
large language models (llm)     large language model
large language models (llms)    large language model
llm     large language model
llms    large language model
multimodal large language models        multimodal large language model
natural language processing (nlp)       natural language processing
neural networks neural network
nlp     natural language processing
prompt engineering      prompt
prompting       prompt
retrieval augmented generation (rag)    retrieval augmented generation
retrieval-augmented generation  retrieval augmented generation
retrieval-augmented generation (rag)    retrieval augmented generation
systematic review       systematic literature review
transformers    transformer
trustworthiness trust
```

The cleaned dataset was subjected to co-occurrence analysis using VOSviewer, with keyword co-occurrence serving as the primary analytical unit. The resulting clusters were interpreted qualitatively, guided by cluster composition, central keywords, and thematic coherence. Particular attention was given to the centrality of hallucination within the network, as well as its connections to technical terms, societal themes, and application domains.

RESULT AND FINDINGS

Fig. 2 reveals the keyword “large language model” as the dominant and most central node, reflecting its foundational role in the contemporary discourse on generative AI and its applications.

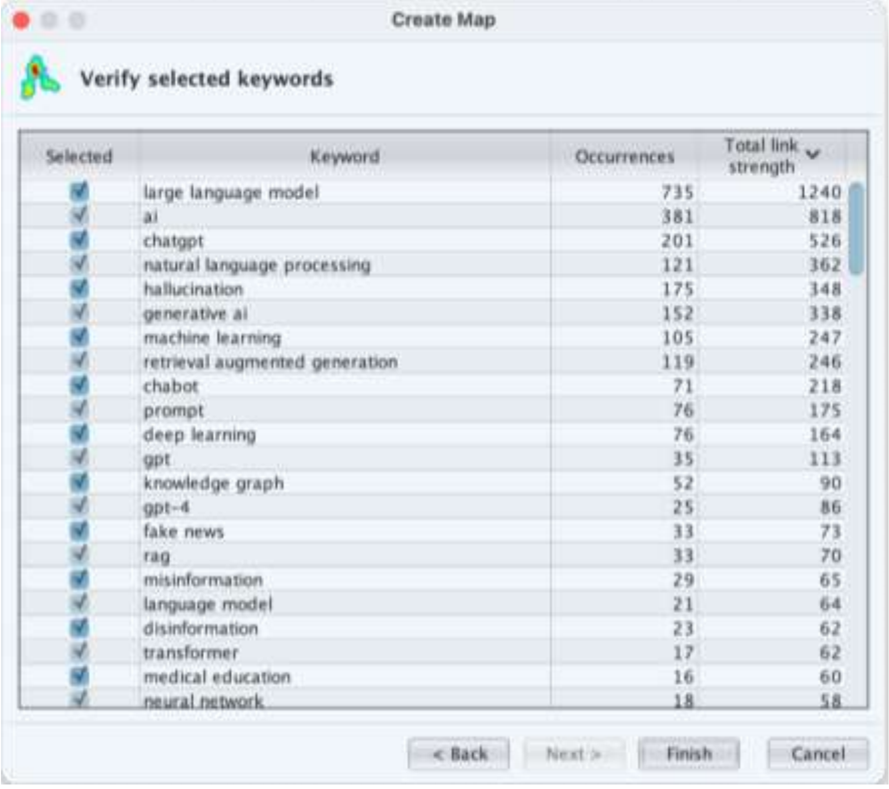
Fig. 2 Keyword co-occurrence network map of literature related to hallucination and artificial intelligence



Notably, “hallucination” emerges as a highly prominent and centralised keyword, signified by a large node and numerous inter-cluster connections. This visual prominence underlines hallucination as a mainstream concern in large language model (LLM) research, with researchers across various domains converging on its critical implications.

Accompanying terms such as “hallucination detection” and “hallucination mitigation” are also present, represented by smaller and less central terms. Their comparatively smaller size suggests that while these areas are emergent, future research may progressively shift towards proactive mitigation strategies rather than mere detection, aligning with a growing emphasis on responsible AI deployment.

Fig. 3 Keywords, occurrences, and the total link strength


Create Map

Verify selected keywords

Selected	Keyword	Occurrences	Total link strength
☑	large language model	735	1240
☑	ai	381	818
☑	chatgpt	201	526
☑	natural language processing	121	362
☑	hallucination	175	348
☑	generative ai	152	338
☑	machine learning	105	247
☑	retrieval augmented generation	119	246
☑	chabot	71	218
☑	prompt	76	175
☑	deep learning	76	164
☑	gpt	35	113
☑	knowledge graph	52	90
☑	gpt-4	25	86
☑	fake news	33	73
☑	rag	33	70
☑	misinformation	29	65
☑	language model	21	64
☑	disinformation	23	62
☑	transformer	17	62
☑	medical education	16	60
☑	neural network	18	58

The keyword co-occurrence analysis reveals the central themes within the literature on hallucination and artificial intelligence, with “large language model” (735 occurrences, 1240 link strength) standing out as the most dominant term, reflecting its pivotal role in current discourse. Closely linked terms include “AI” (381 occurrences, 818 link strength), “ChatGPT” (201 occurrences, 526 link strength), and “natural language processing” (121 occurrences, 362 link strength), indicating a strong focus on generative AI technologies. Notably, “hallucination” (175 occurrences, 348 link strength) emerges as a key concern, often explored in relation to model reliability and factual consistency. Terms such as “retrieval augmented generation,” “generative AI,” and “prompt” further highlight growing research into mitigation strategies and model behaviour. The appearance of “misinformation,” “disinformation,” and healthcare-related terms like “medical education” suggests a broadening concern for the societal implications and domain-specific applications of hallucination in AI systems.

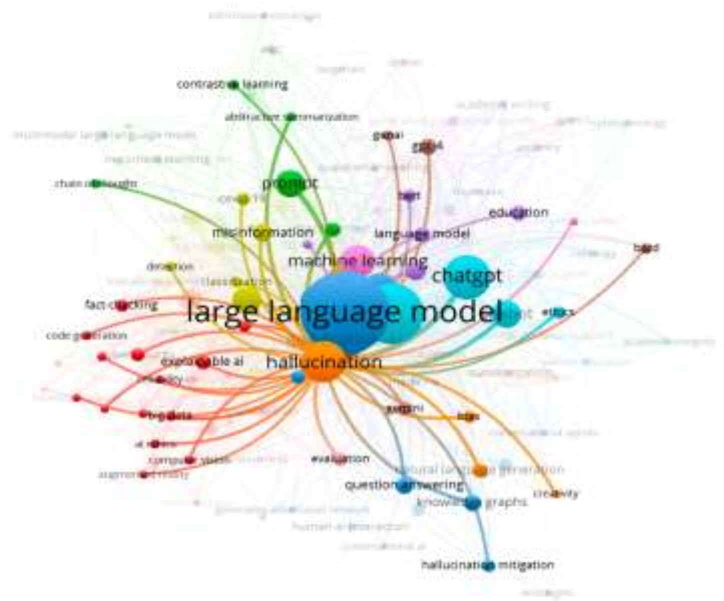
Based on network analysis and the colour-coded visual clustering, ten thematic clusters have been identified and shown in Table 2.

Table 2 Thematic clusters of ai and hallucination research based on keyword co-occurrence analysis

Cluster (colour)	Theme	Description
Cluster 1 (red)	Foundational AI Technologies	Includes core technical terms such as neural networks, explainable AI (XAI), and cybersecurity. This cluster represents the technological substrate enabling LLM development.
Cluster 2 (green)	Prompting and social medial	This cluster intersects linguistic prompting methods, highlighting the socio-technical challenge of guiding LLM outputs from social medias.
Cluster 3 (blue)	Human-AI Interaction and Dialogue Systems	Focused on communication frameworks and interfaces such as conversational agents and knowledge graphs.
Cluster 4 (yellow)	Misinformation and social media	Addresses societal concerns such as disinformation, fake news, and the role of platforms like CNN, signifying the interaction between AI-generated content and public discourse.
Cluster 5 (purple)	AI in Education	Encompasses educational applications and concerns, including academic writing, datasets, and systematic literature review methodologies.
Cluster 6 (light blue)	AI in Healthcare Education	Captures the intersection between generative AI, healthcare, medical training, and ethical concerns.
Cluster 7 (orange)	Hallucination and Information Retrieval	This cluster houses keywords like RAG, summarisation, and hallucination.
Cluster 8 (brown)	Named Models and LLM Brands	Contains model-specific nodes such as GPT-4, Bard, Gemini, and other branded LLMs.
Cluster 9 (pink)	Machine Learning and Associated Infrastructure	Contains foundational computational topics including computer science, machine learning, and OpenAI.

The thematic clustering in Table 2 reveals distinct yet interconnected domains within the literature on AI and hallucination, with several clusters exerting notable influence. Cluster 1 (Red) anchors the analysis by framing foundational AI technologies—such as neural networks, explainable AI, and cybersecurity—as the core enablers of LLM development. Cluster 7 (Orange) is particularly critical, as it directly addresses hallucination and summarisation, encapsulating the epistemic risks associated with AI-generated content. Clusters 2 (Green) and 4 (Yellow) underscore the socio-technical implications, linking prompting strategies and misinformation dynamics, respectively, to the broader discourse ecosystem shaped by AI outputs. Equally impactful is Cluster 5 (Purple), which frames AI's role in educational contexts, signalling growing scrutiny of academic integrity and research practices. Meanwhile, Clusters 6 (Light Blue) and 10 (Peach) indicate the expanding footprint of LLMs in healthcare, although their peripheral links to hallucination suggest an underexplored yet potentially consequential research avenue. Finally, Clusters 8 (Brown) and 9 (Pink) reflect the commercial and infrastructural landscape, situating branded models and computational systems as both influencers and outcomes of current research trajectories.

Fig. 4 Keyword co-occurrence network centred on "hallucination" in artificial intelligence literature



The keyword "hallucination" demonstrates strong interdisciplinary relevance by connecting with nodes from nearly all ten clusters, indicating a multidisciplinary convergence of research. These mixed-colour connections in the co-occurrence map reflect diverse academic interest in hallucination across technical, ethical, societal, and domain-specific domains—particularly healthcare and education. Notably, "hallucination" consistently co-occurs with critical terms such as ethics, misinformation, disinformation, fake news, bias, and trust, evidencing its significant epistemological and ethical implications, especially in contexts demanding high levels of accuracy and reliability.

Despite the abundance of healthcare-related keywords across clusters—e.g., "medicine", "radiology", "diagnosis", "clinical decision support"— direct co-occurrence links between these terms and "hallucination" remain limited. Instead, their association with hallucination is mediated through shared terms such as ChatGPT and large language model, suggesting a thematic but not lexical proximity. This pattern implies that while hallucination is a recognised concern in healthcare AI applications, it is often discussed indirectly rather than being explicitly foregrounded in the discourse.

In a similar vein, education-related terms, particularly within Cluster 5 (AI in education) and Cluster 6 (AI in healthcare education), show limited direct association with hallucination. Despite the presence of keywords such as academic writing, student learning, and academic integrity, there is minimal direct discussion on hallucination's influence in academic contexts. This indicates a potential lack of focused investigation into how hallucinations affect teaching, learning, and scholarly practices. However, the appearance of hallucination alongside ChatGPT in medical education suggests that awareness is emerging, particularly around the pedagogical reliability and ethical use of AI tools in instructional settings.

The strong proximity of prompt to hallucination in Cluster 2 underscores the critical role of prompt engineering in influencing hallucination outcomes. Given that LLM outputs are heavily dependent on the phrasing and context of user input, this finding highlights a growing interest in technical mitigation strategies rooted in user interaction design.

Overall, the cluster analysis revealed that healthcare-related themes are highly over-represented, as evidenced by multiple clusters centred on clinical decision support, digital health, and healthcare education. This suggests that a significant portion of hallucination-related research is concentrated within the medical and biomedical fields, reflecting growing concerns about the implications of AI-generated misinformation in high-stakes environments.

Conversely, clusters related to education, academic integrity, and research applications—particularly Cluster 5 (AI in education) and Cluster 6 (AI in healthcare education)—appear under-represented in terms of direct connections to the term “hallucination.” Although these clusters include keywords like "academic writing", "education", and "student learning", they are not strongly linked to hallucination-related terms, suggesting a lack of focused investigation on how AI hallucinations affect teaching, learning, or scholarly work.

Overall, the findings highlight an imbalance in the thematic focus of current literature—prioritising healthcare while overlooking the broader academic and professional applications of LLMs where hallucination may also pose significant risks. This gap points to opportunities for future research to explore hallucination phenomena in non-medical domains, especially in academia and industry-specific practices like law, education, and construction.

Academics’ Awareness On Llms And Hallucinations

The thematic analysis of Clusters 5 (AI in Education) and 6 (AI in Healthcare Education) reveals an active interest in the application of large language models (LLMs) within academic and medical education, particularly in areas such as academic writing, dataset management, and training support. However, a notable omission is the absence of hallucination-related keywords within these clusters, despite hallucination being a recognised limitation of LLMs. This suggests a disconnection between the growing adoption of AI tools in academic contexts and the awareness of their potential drawbacks, particularly the generation of false or misleading content, commonly referred to as hallucination. Prior research predominantly situates AI hallucination concerns within high-stakes domains such as medicine [29, 30], law [8], and journalism [31], where factual inaccuracies have clear and immediate implications. This bibliometric finding aligns with recent literature that warns of the over-reliance on AI tools in education without a corresponding increase in user understanding of their limitations, particularly their tendency to produce plausible but false information.

As LLMs become increasingly integrated into scholarly practices—ranging from literature reviews to content generation and even decision-support tools, the risk of academic reliance on fabricated or inaccurate outputs becomes more pronounced. While ethical concerns regarding AI usage in academic work have garnered attention [14, 32–34], particularly around issues of plagiarism and authorship, the phenomenon of hallucination remains underexplored. Evidence suggests that hallucinations diminish students’ ability to critically evaluate AI-generated content. For instance, students were able to identify AI-generated text with only 70% accuracy, dropping to 60% in domain-specific contexts [35]. Similarly, Kim et al. (2025) report that students view generative AI as useful for efficiency but express uncertainty and mistrust when encountering hallucinated outputs. This lack of critical awareness threatens research integrity, academic rigor, and educational outcomes.

The risks are not limited to higher education. Zhu et al. (2025) show that middle school students using LLM-based chatbots produced more assignments with higher word counts and surface-level performance gains, yet scored significantly lower on knowledge tests than their peers. This suggests that hallucination, compounded by overreliance on AI, fosters superficial rather than deep learning. Such outcomes resonate with concerns in higher education, where both students and faculty often struggle to distinguish between reliable and unreliable

AI content [14, 35].

Collectively, these findings confirm that hallucinations undermine academic integrity across multiple educational levels—from doctoral research to secondary school assignments. They underscore the need to embed AI literacy, critical evaluation skills, and transparent verification mechanisms into academic practice, ensuring that AI supports rather than weakens knowledge construction. Despite extensive research on hallucination in clinical and healthcare domains, its implications for education remain underexamined. Addressing this gap is essential for fostering responsible AI use in teaching and research, and for guiding future efforts in mitigation strategies, policy development, and pedagogy.

CONCLUSION

Hallucination in artificial intelligence presents both a compelling phenomenon and a critical challenge in the development and deployment of large language models (LLMs). The findings of this bibliometric review reveal a growing scholarly interest in hallucinations, especially in high-stakes domains such as healthcare, where misinformation can significantly impact clinical decision-making. However, this focus has resulted in an overrepresentation of healthcare-related studies and a marked underrepresentation of research exploring the implications of hallucinations in academic and other professional contexts.

Although LLMs are increasingly adopted in education for academic writing, research support, and learning facilitation, there is limited attention to how hallucinations affect academic integrity, knowledge construction, and student outcomes. The weak co-occurrence of “hallucination” with keywords such as academic writing, academic integrity, and student learning suggests a gap in awareness and critical engagement with the issue. Future research should examine the influence of hallucinated content on scholarly behaviours, misinformed citation practices, and the erosion of academic standards. This also underscores the need for institutional policies and digital literacy strategies to mitigate the risks of unverified AI-generated outputs in academic environments.

The results of this review highlight several actionable implications for practice. For educators, AI literacy must be embedded into curricula to train students in critically evaluating and cross-verifying AI-generated outputs. Institutions should establish clear guidelines on responsible AI use, extend academic integrity frameworks to address fabricated citations, and provide training for faculty in detecting hallucination-related errors. At the policy level, stakeholders should promote standards and audit mechanisms to ensure transparency and accountability in educational AI adoption. Collectively, these measures would safeguard academic integrity while enabling constructive use of LLMs in teaching, learning, and research.

While bibliometric analysis offers a valuable macro-level view of research trends and thematic structures, it has inherent limitations. This study used Scopus as the sole data source, which may have excluded relevant literature indexed elsewhere. The predefined keyword search strategy, though comprehensive, could miss studies using varied terminologies. Moreover, bibliometric methods do not assess the quality or rigour of individual studies, and manual cluster labelling, while necessary, introduces a degree of subjectivity. Nonetheless, this approach provides a strong foundation for mapping the current knowledge landscape and identifying research gaps.

Nevertheless, future research would benefit from adopting a multi-database search strategy that includes Web of Science, Google Scholar, or discipline-specific repositories. This would minimise potential database bias and ensure even broader coverage of hallucination-related publications.

Future studies also should extend beyond general bibliometric mapping to provide domain-specific insights into how hallucinations affect research and practice. In education, this includes examining their influence on student learning outcomes, academic integrity, and assessment reliability across higher education, and professional training. In engineering and construction, future work could investigate how hallucinations compromise design accuracy, project management, and compliance within digital tools especially in Building Information Modelling (BIM) systems. Healthcare education offers another critical avenue, where hallucinations may distort training simulations and diagnostic reasoning exercises. Beyond these, targeted analyses in law, policy, and the social sciences are needed to understand how hallucinated outputs may shape legal training, policy drafting, or historical and media narratives. Comparative and interdisciplinary studies across these domains would allow researchers to identify both common risks and unique vulnerabilities, ultimately leading to tailored mitigation frameworks that ensure responsible and trustworthy AI adoption.

As the role of AI in knowledge production expands, addressing hallucination must remain a research and policy priority to ensure its responsible and informed application in both academic and professional domains.

ACKNOWLEDGEMENTS

The authors would like to extend their heartfelt gratitude to Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, for granting permission and providing the necessary support to conduct this study. The university's encouragement and facilitation of academic exploration have been invaluable in advancing out research on LLM and hallucinations and their implications across various domains.

REFERENCES

1. Alhusban, M.I., Khatatbeh, I.N., Alshurafat, H.: Exploring the influence, implications and challenges of integrating generative artificial intelligence into organizational learning and development. Competitiveness Review: An International Business Journal. ahead-of-p, (2025). https://doi.org/10.1108/CR-06-2024-0121
2. Ali Bukar, U., Sayeed, M.S., Amodu, O.A., Razak, S.F.A., Yogarayan, S., Othman, M.: Leveraging VOSviewer approach for mapping, visualisation, and interpretation of crisis data for disaster management and decision-making. International Journal of Information Management Data Insights. 5, 100314 (2025). https://doi.org/10.1016/j.jjimei.2024.100314
3. Alkaissi, H., McFarlane, S.I.: Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus. 15, (2023). https://doi.org/10.7759/cureus.35179
4. Anghelescu, A., Ciobanu, I., Munteanu, C., Anghelescu, L.A.M., Onose, G.: ChatGPT: "To be or not to be" ... in academic research. The human mind's analytical rigor and capacity to discriminate between AI bots' truths and hallucinations. Balneo and PRM Research Journal. 14, (2023). https://doi.org/10.12680/balneo.2023.614
5. Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G.L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A., Hovy, E., Ji, H., Menczer, F., Miguez, R., Nakov, P., Scheufele, D., Sharma, S., Zagni, G.: Factuality challenges in the era of large language models and opportunities for fact-checking. Nat Mach Intell. 6, 852–863 (2024). https://doi.org/10.1038/s42256-024-00881-z
6. Aure, P.A., Cuenca, O.: Fostering social-emotional learning through human-centered use of generative AI in business research education: an insider case study. Journal of Research in Innovative Teaching & Learning. 17, 168–181 (2024). https://doi.org/10.1108/JRIT-03-2024-0076
7. Ball, R.: An Introduction to Bibliometrics. Jonathan Simpson (2018)
8. Borger, J.G., Ng, A.P., Anderton, H., Ashdown, G.W., Auld, M., Blewitt, M.E., Brown, D. V., Call, M.J., Collins, P., Freytag, S., Harrison, L.C., Hesping, E., Hoysted, J., Johnston, A., McInneny, A., Tang, P., Whitehead, L., Jex, A., Naik, S.H.: Artificial intelligence takes center stage: exploring the capabilities and implications of ChatGPT and other AI-assisted technologies in scientific research and education. Immunol Cell Biol. 101, 923–935 (2023). https://doi.org/10.1111/imcb.12689
9. Carabantes, D., González-Geraldo, J.L., Jover, G.: ChatGPT could be the reviewer of your next scientific paper. Evidence on the limits of AI-assisted academic reviews. Profesional de la Informacion. 32, 1–12 (2023). https://doi.org/10.3145/epi.2023.sep.16
10. Cui, Y., Zhang, H.: Can student accurately identify artificial intelligence generated content? an exploration of AIGC credibility from user perspective in education. Educ Inf Technol (Dordr). 30, 16321–16346 (2025). https://doi.org/10.1007/s10639-025-13448-1
11. Dwivedi, Y.K., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., Baabdullah, A.M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M.A., Al-Busaidi, A.S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L., Chowdhury, S., Crick, T.,

Cunningham, S.W., Davies, G.H., Davison, R.M., Dé, R., Dennehy, D., Duan, Y., Dubey, R., Dwivedi, R., Edwards, J.S., Flavián, C., Gauld, R., Grover, V., Hu, M.C., Janssen, M., Jones, P., Junglas, I., Khorana, S., Kraus, S., Larsen, K.R., Latreille, P., Laumer, S., Malik, F.T., Mardani, A., Mariani, M., Mithas, S., Mogaji, E., Nord, J.H., O'Connor, S., Okumus, F., Pagani, M., Pandey, N., Papagiannidis, S., Pappas, I.O., Pathak, N., Pries-Heje, J., Raman, R., Rana, N.P., Rehm, S.V., Ribeiro-Navarrete, S., Richter, A., Rowe, F., Sarker, S., Stahl, B.C., Tiwari, M.K., van der Aalst, W., Venkatesh, V., Viglia, G., Wade, M., Walton, P., Wirtz, J., Wright, R.: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manage. 71, (2023). https://doi.org/10.1016/j.ijinfomgt.2023.102642

12. Feng, J., Zhao, L., Qin, H., Xu, Y., Wang, Z.: CADLRA: A multi-charge prediction method based on the Criminal Act-Driven Law Retrieval Augmentation. Eng Appl Artif Intell. 134, 108619 (2024). https://doi.org/10.1016/j.engappai.2024.108619
13. Ferhataj, A., Memaj, F., Sahatcija, R., Ora, A., Koka, E.: Ethical concerns in AI development: analyzing students' perspectives on robotics and society. Journal of Information, Communication and Ethics in Society. 23, 165–187 (2025). https://doi.org/10.1108/JICES-08-2024-0111
14. Guo, Y., Guo, M., Su, J., Yang, Z., Zhu, M., Li, H., Qiu, M., Liu, S.S.: Bias in Large Language Models: Origin, Evaluation, and Mitigation. (2024)
15. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Trans Inf Syst. 1, 1–58 (2023). https://doi.org/10.1145/3703155
16. Hwang, T., Aggarwal, N., Khan, P.Z., Roberts, T., Mahmood, A., Griffiths, M.M., Parsons, N., Khan, S.: Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. PLoS One. 19, 2–9 (2024). https://doi.org/10.1371/journal.pone.0297701
17. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Comput Surv. 55, 11–38 (2023)
18. Kim, J., Yu, S., Detrick, R., Li, N.: Exploring students' perspectives on Generative AI-assisted academic writing. (2025)
19. Krumsvik, R.J.: Chatbots and academic writing for doctoral students. Springer US (2025)
20. Lund, B.D., Wang, T., Mannuru, N.R., Nie, B., Shimray, S., Wang, Z.: ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. J Assoc Inf Sci Technol. 74, 570–581 (2023). https://doi.org/https://doi.org/10.1002/asi.24750
21. Mahir, A.: AI Hallucinations and Scientific Integrity: Towards Reliable Language Models for Research and Societal Impact, https://www.linkedin.com/pulse/ai-hallucinations-scientific-integrity-towards-reliable-atheer-mahir-enrie
22. Maynez, J., Narayan, S., Bohnet, B., McDonald, R.: On faithfulness and factuality in abstractive summarization. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 1906–1919 (2020)
23. Merken, S.: New York lawyers sanctioned for using fake ChatGPT cases in legal brief, https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/
24. Moral-muñoz, J.A., Herrera-viedma, E., Santisteban-espejo, A., Cobo, M.J., Herrera-viedma, E., Santisteban-espejo, A., Cobo, M.J.: Software tools for conducting bibliometric analysis in science: An up- to-date review. El profesional de la informa- ción. 29, 1–20 (2020)
25. Salvagno, M., De Cassai, A., Zorzi, S., Zaccarelli, M., Pasetto, M., Sterchele, E.D., Chumachenko, D., Gerli, A.G., Azamfirei, R., Taccone, F.S.: The state of artificial intelligence in medical research: A survey of corresponding authors from top medical journals. PLoS One. 19, 1–15 (2024). https://doi.org/10.1371/journal.pone.0309208
26. Song, C., Chen, K., Jin, Y., Chen, L., Huang, Z.: Visual analysis of research hotspots and trends in traditional Chinese medicine for depression in the 21st century: A bibliometric study based on citespace and VOSviewer. Heliyon. 11, e39785 (2025). https://doi.org/10.1016/j.heliyon.2024.e39785
27. Stornaiuolo, A., Higgs, J., Jawale, O., Martin, R.M.: Digital writing with AI platforms: the role of fun with/in generative AI. English Teaching: Practice & Critique. 23, 83–103 (2024).

https://doi.org/10.1108/ETPC-08-2023-0103

28. Thanasi-Boçe, M., Hoxha, J.: From ideas to ventures: building entrepreneurship knowledge with LLM, prompt engineering, and conversational agents. Springer US (2024)
29. van Eck, N.J., Waltman, L.: Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. 84, 523–538 (2010). https://doi.org/10.1007/s11192-009-0146-3
30. Vicsek, L., Pinter, R., Bauer, Z.: Shifting job expectations in the era of generative AI hype – perspectives of journalists and copywriters. International Journal of Sociology and Social Policy. 45, 1–16 (2025). https://doi.org/10.1108/IJSSP-05-2024-0231
31. Waltman, L., van Eck, N.J., Noyons, E.C.M.: A unified approach to mapping and clustering of bibliometric networks. J Informetr. 4, 629–635 (2010). https://doi.org/10.1016/j.joi.2010.07.002
32. Wu, R.T., Dang, R.R.: ChatGPT in head and neck scientific writing: A precautionary anecdote. American Journal of Otolaryngology - Head and Neck Medicine and Surgery. 44, 6–8 (2023). https://doi.org/10.1016/j.amjoto.2023.103980
33. Zainal Abidin, N.A., Mohd Shariff, K.K., Mohd Yassin, I., Zabidi, A., Saadon, A., Md Tahir, N., Ridzuan, A.R., Amin Megat Ali, M.S.: How Factually Accurate is GPT-3? A Focused Case Study on Helping Malaysia's B40s Through e-Commerce. Qeios. 3, 1–26 (2024). https://doi.org/10.32388/g2gh34
34. Zhu, Y., Zhu, C., Wu, T., Wang, S., Zhou, Y., Chen, J., Wu, F., Li, Y.: Impact of assignment completion assisted by Large Language Model-based chatbot on middle school students' learning. Educ Inf Technol (Dordr). 30, 2429–2461 (2025). https://doi.org/10.1007/s10639-024-12898-3
35. Zhui, L., Fenghe, L., Xuehu, W., Qining, F., Wei, R.: Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: Viewpoint. J Med Internet Res. 26, (2024). https://doi.org/10.2196/60083
36. Zupic, I., Čater, T.: Bibliometric Methods in Management and Organization. Organ Res Methods. 18, 429–472 (2015). https://doi.org/10.1177/1094428114562629