ISSN 2055-0286
Advances in Social Sciences Research Journal; Vol. 13 No. 05 (2026) (36-51)
https://doi.org/10.14738/assrj.1305.2402

A Thematic Analysis on Large Language Models: Hallucination, Trust, and Governance Challenges in AI Systems

Bruno Ribeiro Bastos[1] & Murillo de Oliveira Dias[2]

1. Universidade Federal do Rio de Janeiro (UFRJ), Escola Politécnica, Rio de Janeiro, Brasil
2. Universidade do Estado do Rio de Janeiro (UERJ); Escola Superior de Desenho Industrial (ESDI), Programa de Pós-Graduação da Escola Superior de Desenho Industrial (PPDESDI); Rio de Janeiro, Brasil

Abstract: Large language models (LLMs) are becoming increasingly popular and powerful in terms of applications. They can be used for text and image generation, multimodal models, and reasoning. Despite the numerous potential uses, there is a significant problem with hallucinations and the generation of false, misleading, or inaccurate information by the models. This paper conducts a thematic analysis of topics such as hallucination types, mitigations, and the effects on multilingual models. There is also an analysis of attempts to regulate the power of models. The solutions to the problem of hallucinations vary from purely technological solutions to an interdisciplinary approach, identifying both potential benefits and future risks. The sources analyzed are around 320 references, including preprint papers from the fastest-growing field of AI. This paper focuses on equipping the reader with a basic conceptual understanding of the field and directions for future research, to establish knowledge and build trustworthy AI systems.

Keywords: large language models, thematic analysis, AI governance, trustworthiness, hallucination detection.

INTRODUCTION

Large language models (LLMs) have been advancing at unprecedented speed and have become a major focus of AI research. While recent breakthroughs like BERT's (Devlin et al., 2019) bidirectional pre-training architecture, multi-modal learning capability for both text, vision, and audio (Bai et al., 2023; Chen et al., 2023), etc., bring unprecedented capability to generate human language at scale, perform complex reasoning tasks, and capture very specific contexts or scenarios, there are serious concerns regarding their trustworthiness, explainability, and governance. To address the issues mentioned, in the literature and in real-world applications, there is a growing interest in studying the so-called hallucinations; i.e., generated outputs that are factually incorrect, misleading, or even entirely fake. Studying hallucinations in LLMs is not only important from a technical perspective, but also has implications for the ethical, legal, and social consequences of LLMs generating inaccurate output, in domains like healthcare, law, and governance. The thematic analysis enables us to map out the major themes and gaps in the literature, and to provide an integrated overview of current work, highlighting directions for future research.

As LLMs are deployed globally by organizations of all shapes and sizes, it is increasingly important to understand their behavior. As governments worldwide introduce regulations on AI, this is no longer a topic reserved for researchers alone. As these areas begin to emerge, they raise important questions about the equity and justice of current AI

systems and their implications for linguistic justice and the fair deployment of technologies, such as LLMs. Analyzing hallucinations in LLMs is important not only from a technical perspective but also from an ethical, legal, and social perspective, especially as LLMs generate output for an ever-increasing range of applications and generate false output for any input. Hallucinations in LLMs that process medical input, for example, could have severe consequences on patient safety. Dokumacı (2024) noted that there is pending liability or accountability for harmful LLM speech. There has been a lot of writing about hallucinations in LLMs in the last few months. Since the field is so new, I decided to do a thematic analysis on the topic, drawing on ~320 references (including preprint articles). The LLM literature is incredibly vast and diverse, encompassing many different types of analyses (technical, descriptive, normative, etc.). This method of analysis allows major themes and research gaps to be mapped and all relevant literature to be described in a clear and comprehensive form with reference to future areas of research.

The references analyzed in this study comprise a research portfolio of 320 reports from the academic literature. These references include peer-reviewed journal articles, conference papers and proceedings, and preprints stored in open repositories. The inclusion of preprints enables the study to capture the very latest research in this fast-moving field, while peer-reviewed articles provide established theory and evidence on large language models and hallucination. A thematic analysis was conducted and presented as a conceptual map/guideline for future research into language model hallucination.

BACKGROUND

Thematic analysis is becoming increasingly used as a methodological approach to thematic synthesis. In this paper, we discuss how thematic analysis can be used to identify and describe how a set of key conceptual patterns (‘themes’) recur within and across diverse sources, including technical documentation, empirical research, normative papers, and preprints. Thematic analysis offers the researcher a structured approach to making sense of the rapidly changing space of large language models and where future research might fruitfully go (Dias et al., 2026a, 2026b; Lopes & Dias, 2026a, 2026b; Quintão et al., 2026).

Large language models (LLMs) have become the central axis of artificial intelligence (AI) research. From a technical, ethical, and governance-oriented perspective, this article discusses the relevance of LLMs. The rise of LLMs was mainly enabled by architectures that learn to capture long-range dependencies in text, a task that can be addressed using long short-term memory networks and attention mechanisms. In this context, pre-trained and fine-tuned models like BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) demonstrate the great potential of such models. 2020), transformers (Vaswani et al. 2017), BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), ALBERT (Lan et al. 2019), Longformer (Bayer et al. 2020), Reformer (Kootstra et al. 2020).

In addition to the work by Arya et al. (2021), who provided an explanation of LLMs and their interpretation, AI Explainability 360, published by Bai et al. (2021), questioned the so-called ‘interpretability’ of attention mechanisms. The works stressed that attention is not the same as explanation. To explain the behavior of LLMs, thematic approaches that combine an analytical treatment of technical aspects with a critical discussion of the respective underlying concepts are required. Previous work by Chopra et al. (2005) and Hahn (2020) already theorized the limitations of self-attention. As LLMs become more ubiquitous

and enter more aspects of society, people are starting to think about the governance of these models. This includes how external stakeholders, such as policymakers, might influence LLM design decisions. Conversely, LLMs could also raise new governance challenges and opportunities.

Regulation of LLMs within the European Union has begun to take shape, with the European Parliament and the European Council (2024) passing the Artificial Intelligence Act, which imposes rules on AI use across all member states. In the short term, this may cause developers to align with broader societal interests, but the European powers are now facing their own 'alignment problem' for the governance of LLMs - i.e., how to ensure proper levels of transparency, licensing, and auditability. In the UK justice system, HM Courts & Tribunals Service and the Ministry of Justice (2025) intend to audit, restrict, and develop new guidelines on the use of AI within justice institutions. In Singapore, the Infocomm Media Development Authority and Personal Data Protection Commission (2025) have launched AI Verify, which certifies organizations that adopt trustworthy LLMs within their operations. Meanwhile, a variety of industry and sector-specific standards and guidelines for AI and LLMs are emerging, including the recently released ISO/IEC 42001:2023 guidelines for organizational excellence. Another theme highlighted in the literature is Healthcare applications. Faust et al. (2018) reviewed the current state of deep learning applied to physiological signals in healthcare.

METHODOLOGY

This study follows an interpretivist, inductive approach, as suggested by Saunders et al. (2009). The study has an interpretivist epistemology and inductive methodology. The study seeks to understand the phenomena of LLMs and how researchers, developers, and institutions perceive them through the meanings they ascribe to them. An inductive approach is particularly relevant here. As the field advances rapidly, themes will emerge from the literature that may be considered outdated by prior theoretical constructs. The Synthesis was conducted according to the principles of thematic analysis. Strauss and Corbin (1998) stated that qualitative research must be conducted through an iterative process of coding and constant comparison to develop conceptual categories that explain the phenomena under study. The materials from the conducted contributions were read and inductively coded. The categories were repeatedly revisited and refined throughout the synthesis. The report contains the following themes: technical architectures for big data offerings, interpretability in big data offerings, governance frameworks for big data offerings, and ethical challenges related to big data offerings.

Analytical Approach

This analysis follows a three-step process: gathering sources, reading and coding them, and analyzing the results to identify thematic clusters among concepts and gaps in the literature. Saunders et al. (2009) state that rigor and transparency are as important in qualitative research as they are in quantitative research. We ensured that the analysis findings were reliable by recording coding decisions throughout the coding process, revisiting them, and re-reading where necessary. Each thematic cluster was supported by

evidence from multiple sources within the data, and the themes were rigorously developed and checked against the data, with recourse to the transcripts where necessary.

Data Sources

We processed documents retrieved from academic databases and data repositories, including Web of Science, Scopus, Google Scholar, and PubMed, as well as the open-source trading data repository OpenAIex. Semantic Scholar was also used to enhance retrieval. We searched for keywords related to LLMs and their applications, including references to seminal work, such as Hochreiter & Schmidhuber (1997). Bibliometric maps were developed using VOSviewer 1.6.20 (Eck & Waltman, 2010) to create co-authorship and keyword co-occurrence maps and to calculate citation profiles of individual documents, thereby identifying clusters of current research activity. Web-based software such as Voyant (voyant-tools.org) was used to calculate frequency distributions and to derive further context patterns from the abstracts and the keywords of all analyzed documents. These results served as input for the thematic coding process. Normalization was performed using VOSviewer 1.6.20 (Eck & Waltman, 2010). Thematic analysis was conducted using Voyant. The methods for quantitative visualization of relations between bibliographic objects and qualitative thematic analysis correspond to the methods of the interpretivist and inductive research logic (Saunders et al. 2009). This combination of methods for analyzing LLM research, to ensure rigor, was based on the emergent nature of the research. These visualizations revealed clusters of research activity and thematic convergence within the LLM literature. In addition, Voyant Tools (Sinclair & Rockwell, 2026) was used to analyze abstracts and keywords, producing frequency distributions and contextual patterns that enriched the thematic coding process.

Methodological Rigor

Using an interpretivist, inductive approach (Saunders et al., 2009), the study employs a variety of methods, including grounded theory (Strauss & Corbin, 1998) and thematic analysis, and draws upon bibliometric mapping techniques. The analysis of current themes corroborates existing knowledge and informs potential future research directions on LLMs, complemented by the structural insights from the bibliometric mapping.

FINDINGS AND ANALYSIS

In our study, we perform inductive thematic coding of research on LLMs and combine it with a bibliometric map. We also perform additional textual analysis using Voyant Tools. The results of our study provide insight into the current state of affairs and dynamics of technical innovations, reliability issues, ethical concerns, and governance frameworks that are dealt with in research on LLMs and are organized in distinct thematic clusters, as illustrated in Figure 1:

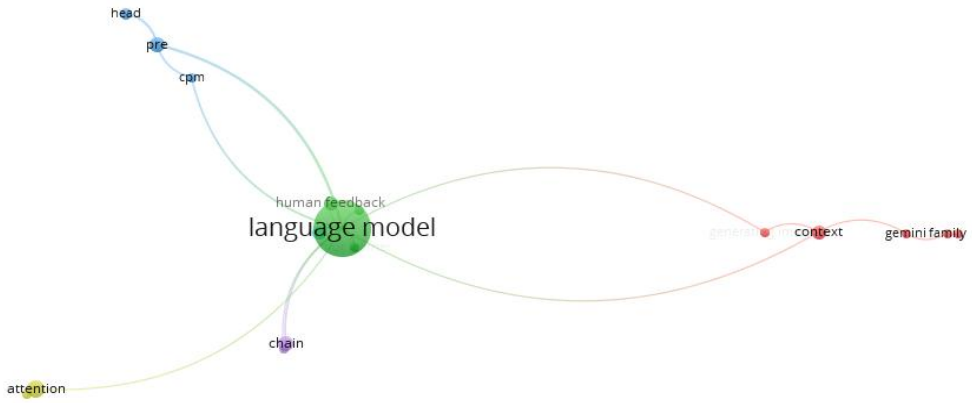

Figure 1: Network map

Source: VOSviewer (version 1.6.20). Adapted from van Eck and Waltman (2010)

Figure 1 shows the network diagram of language models. The center node is the concept of “language model” and is connected to related concepts. Each node and each color represent a different set of categories/tags. For example, “human feedback” is green, “attention” is yellow, “chain” is purple, and “Gemini family” is red. The focus is on the central language model concept, which connects to many different concepts involved in language models. In addition, Figure 2 depicts a network of connections among various researchers in artificial intelligence. Each node represents a researcher, and an edge connects two nodes if those researchers co-authored a paper. The nodes are grouped into clusters and colored according to the group or community they belong to. Radford et al., Raffel et al., Wang et al., and Zhang et al. are highly connected. This means they have significant influence in their field and are likely to be leading figures. The graph also shows relationships among the different research communities. Some individuals work across several areas of AI research, bringing together distinct threads. The graph also conveys the proximity between individual researchers and how ideas flow within and between different research communities.

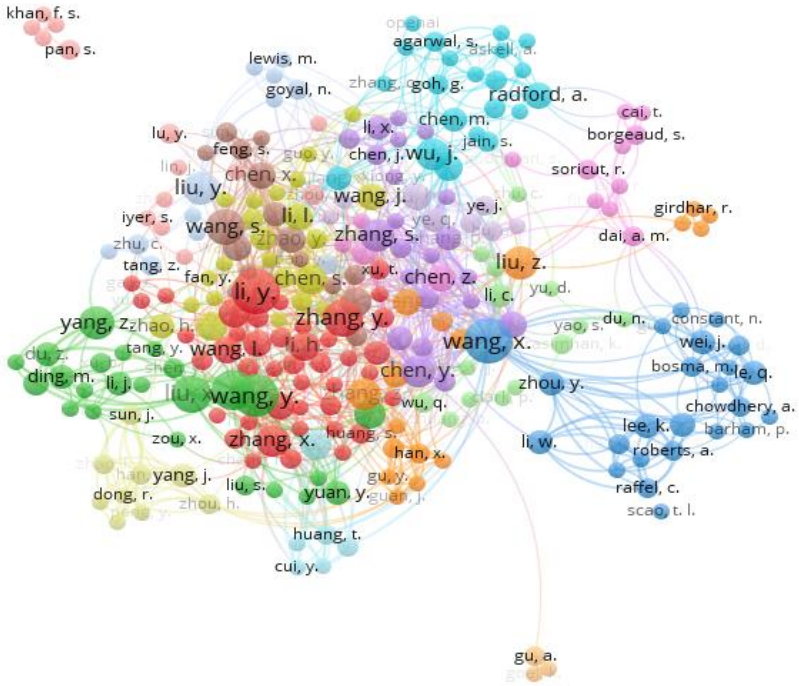

Figure 2: co-authorship network map

Source: VOSviewer (version 1.6.20). Adapted from van Eck and Waltman (2010)

Figure 3 shows the words for large language models. The largest words—language, models, large, and model— are the most important and most frequently referenced words. The next set of smaller words—hallucination, learning, multimodal, training, generation, vision, reasoning, understanding, text, review, image, video, neural, and intelligence— are significant and frequently referenced. The size of words figures illustrates the distribution of concepts that are currently discussed in the field of large language models. While language models are the central topic of this field, many additional concepts are discussed and explored, including multimodal models, hallucination, and factuality.

Figure 3: Conceptual Word Cloud of Large Language Models

Source: Voyant 2.0. (Sinclair & Rockwell, 2016)

The chapter also covers related topics explored in visualizing collaboration networks and conceptual mappings for applications in natural language generation. The first part of this section covers hallucination in large language models, including its causes, detection, and mitigations. The next part discusses multimodal learning with text, images, video, and audio, and how these different modalities can be incorporated into deep learning architectures. The third part discusses training and optimization, specifically how pre-training, fine-tuning, and other techniques can be leveraged to improve efficiency. The final part discusses issues in evaluation and benchmarking, specifically the use of faithfulness, factuality, and reliability metrics for assessing generated output.

Theme One: Hallucination & Reliability

Hallucinations in LLMs pose a challenge for researchers. Xu et al., 2024; Ye et al., 2023; Zhang et al., 2023; Zheng et al., 2025; Zhu et al., 2025; Varshney et al., 2024; Vladika et al., 2025. As shown in Figure 3, the term “hallucination” appeared repeatedly, which indicates that the issue of hallucinations in LLMs has been a challenge in the field. Reliability further requires governance to address the risk that users and applications misuse data and models, leading to incorrect processing and inaccurate or misleading results.

Theme Two: Multimodality

While there is still a prevalent focus on using the primary input architecture in LLMs, there is a growing interest in integrating additional ages,video, and audio within a single architecture. In this direction, Rasheed et al. (2023), Sun et al. (2023), Tang et al. (2023), Tian et al. (2024), Wu et al. (2023), Ye et al. (2023), Zhang et al. (2023), and Zhu et al. (2023, 2024). Figure 3 shows that there is a particularly growing interest in LLM research on “multimodal”, “vision”, and “video”.

Theme Three: LLM Mechanisms & Foundations

Theme three reveals the foundations of LLMs: Models of Reasoning, Coherence, and Scalability. Vaswani et al. (2023), Raffel et al. (2020, 2023), Wei et al. (2022, 2023), Yao et al. (2022). Structural relations between the three components, attention, human feedback, and context, in Figure 1 are crucial to advance LLMs, and are treated as structural pillars to develop LLMs in Figure 1.

Theme Four: LLM Research Ecosystem

The success of LLMs is not only the result of novel technological concepts and innovative approaches but also of a very active global scientific community that collaborates to advance the field. The Co-authorship networks in this contribution give a deeper view into the structures of these communities. Touvron et al. (2023), Wu et al. (2023), Vázquez et al. (2025) et Zhang et al. (2020, 2021). The cross-talk among groups was extensive, and several of them contributed very substantially, including Radford, Rozenberg, and Guryanov, as well as many others, including Brown and Radford et al. Figure 2 illustrates the extent of the links, which were, in many cases, both dense and extensive. The theme of this collection of

papers is LLM research: global, collaborative, with plenty of ideas and people crossing disciplinary and national borders.

Finally, Table 1 summarizes the four emerging themes in our study, as follows:

Table 1: Summary of Thematic Findings: Focus, Representative References, and Key Insights

Theme	Focus	Representative References	Key Insights
Hallucination & Reliability	Detection, mitigation, and governance of hallucinations in LLMs	Xu et al. (2024); Ye et al. (2023); Zhang et al. (2023); Zheng et al. (2025); Zhu et al. (2025); Varshney et al. (2024); Vladika et al. (2025)	Hallucinations are inevitable but can be managed through detection benchmarks, uncertainty modeling, correction methods, and governance frameworks.
Multimodality	Integration of text, image, video, and audio in LLMs	Rasheed et al. (2023); Sun et al. (2023); Tang et al. (2023); Tian et al. (2024); Wu et al. (2023); Ye et al. (2023); Zhang et al. (2023); Zhu et al. (2023, 2024)	Multimodal LLMs expand capabilities by combining different modalities, enabling richer understanding and generation across text, vision, and video.
LLM Mechanisms & Foundations	Core mechanisms that enable reasoning and generation in LLMs	Vaswani et al. (2023); Raffel et al. (2020, 2023); Wei et al. (2022, 2023); Yao et al. (2022)	Foundational elements such as attention, chain-of-thought reasoning, pretraining, and reinforcement learning with human feedback are the building blocks of LLMs.
LLM Research Ecosystem	Collaboration networks and global communities driving LLM innovation	Touvron et al. (2023); Wu et al. (2023); Vázquez et al. (2025); Zhang et al. (2020, 2021)	Co-authorship graphs (Figure 2) reveal how researchers form clusters around LLM development, with hubs like Zhang et al. and Radford et al. bridging communities. These networks show the collaborative and global nature of LLM research.

Source: elaborated by the authors

DISCUSSION

Large language models (LLMs) are rapidly evolving in four important ways. First, while current techniques for detecting "hallucinations" (i.e., model output not contained in input, Xu et al. 2024, Ye et al. 2023, Zhang et al. 2023, Zheng 2025, Zhu 2025) have significantly improved (Varshney 2024, Vladika 2025), there is much to consider regarding whether it is desirable or feasible to entirely avoid generating hallucinations, and if not, how to "mitigate" hallucinations (i.e., allow a model to express uncertainty as to whether it is generating output that is compatible with input, while still providing the best output). For both discourse and trust, the LLMs, especially those deployed in out-of-settings, evoked the theme of “hallucination” vividly, which was thus highlighted as highly prominent in the thematic map (Figure 3).

In addition to the input dimensionality, the language models have undergone transformation along the second axis, namely multimodality. Rasheed et al. (2023), Sun et al. (2023), Tang et al. (2023), Tian et al. (2024), Wu et al. (2023), Ye et al. (2023), Zhang

et al. (2023), and Zhu et al. (2023, 2024) have explored merged text, vision, audio, and video-based generalized LLMs, which support higher dimensionality of input and output modalities and enhance the models with capabilities of reasoning and synthesis that go beyond typical language models. The third dimension of progress in this article is the underlying mechanisms and principles that have enabled these developments. In this dimension, Vaswani et al. (2023), Raffel et al. (2020, 2023), Wei et al. (2022, 2023), and Yao et al. (2022) discuss novel model architectures that LLMs have leveraged, including attention, chain-of-thought reasoning, pretraining methods, and reinforcement learning with human feedback. Figure 1 illustrates these concepts emanating from the center of the node “language model”, highlighting the importance of these novel developments within deep learning for human and animal languages, as highlighted in this article.

Language models can be easily fine-tuned for many downstream NLP applications, such as text classification, question answering, and object detection. These require specialized models that have already been pre-trained on large amounts of text data. The research ecosystem of the advances in this study finally places them within a global context of collaborations and co-authorships. Touvron et al. (2023), Wu et al. (2023), Vázquez et al. (2025), and Zhang et al. (2020, 2021) provide information on co-authorships to characterize the underlying cluster structure and reveal the main protagonists and centers of activity, ranging from the simple, single hub of Zhang to the highly active and highly creative Radford and companions. Figure 2 shows a snapshot of the extremely high density and connectivity of these LLM-related networks, and it is clear that the development of LLMs is an international, highly collaborative field, with contributions from many disciplines and countries, and significant knowledge sharing and exchange.

RESEARCH IMPLICATIONS

Following the guidance laid out in the report, the four axes of thought on models remain valuable. By inevitably generating hallucinations, large language models force us to move from a total-loss-focused prevention mindset (Ye et al. 2023, Zhang et al. 2023, Zheng 2025, Xu et al. 2024, Zhu et al. 2025, Varshney et al. 2024) to a governance-and-mitigation-based model. Hallucinations or not, reliability is not a binary property that can be designed into a model and expected to achieve 100% reliability — rather, it can only be treated as a spectrum that can be managed through benchmarks, models of uncertainty, and a healthy research ecosystem. Multimodality goes beyond language modeling to include perception and even generation in vision, audio, and video. Medicine, education, and creative industry applications. Language modeling is becoming a general framework for intelligence that operates across multiple modes of human input. This rapid growth in the field is driven by scalability, reasoning, new architectural ideas such as attention, chain-of-thought reasoning, pretraining, and reinforcement learning with human feedback. A key implication of our work for the future is that there is more to be gained by further improving these mechanisms to achieve a good enough balance of efficiency, interpretability, and alignment. Finally, the research ecosystem in Figure 2 (Touvron et al. 2023, Wu et al. 2023, Vázquez et al. 2025, Zhang et al. 2020, 2021) shows that the LLMs' progress is a global collaboration. Thus, the effective governance, ethics, and innovation will depend on an international collaborative research ecosystem. Turning to the future, I see challenges with hallucinations, multimodality, foundations, and collaboration all at the same time. But

viewed at a broad enough level, these challenges suggest that the future of LLMs will be about managing them all at once to build more robust, more powerful, more governing LLMs — in other words, to transform LLMs from technological systems to socio-technical systems.

CONCLUSION

Besides hallucination, reliability, multimodality, and the mechanisms behind them, we are also interested in understanding the ecosystem around large language models. Multimodality has significantly diversified LLMs' functionality, from a linguistically focused model to an umbrella of multimodal intelligence models that hybridize linguistic and perceptual intelligence to address a wide range of novel and increasingly complex problems in human life and industrial development. Such models fuse human intelligence and machine processing power to propose innovative solutions to real-world problems that go beyond simple classification to sophisticated generation, fusion, and manipulation. The LLMs have furthermore shifted from static media to dynamic analytics. Lastly, the scope of LLM applications has increased exponentially. While early LLMs were confined to processing vast amounts of static text, primarily textual data, current models can process data and information of immense scale and complexity, encompassing dynamic, time-based media such as video and real-time communication channels such as audio and video. Vaswani et al. (2023); Raffel et al. (2020, 2023); Wei et al. (2022, 2023) describe key mechanisms underlying scalability and reasoning in current models, along with the accompanying architectural innovations, including attention, chain-of-thought reasoning, pretraining, and reinforcement learning with human feedback. These modules form the core of current work and will likely shape the field’s future developments. The analysis of the co-authorship network reveals a clear global pattern, with dense clusters of individuals and organizations.

FUTURE RESEARCH

Future challenges for LLMs include reducing hallucinations, embracing multimodality, better understanding and improving the basic mechanisms, and creating sustainable socio-technical systems that enable long-term collaborative research ecosystems. As LLMs transition from experimental technology to platforms for information generation, dissemination, and consumption, they will have socio-technical implications that must be considered in design and governance.

Competing Interests: We declare that we have no competing interests.

Funding: This research has received no funding.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process: During the preparation of this work the authors used Grammarly, in order to in order to improve grammar accuracy, ensure clarity of expression, and refine sentence flow and enhance readability. Microsoft Copilot was used to support idea organization, provide suggestions for strengthening academic style, and to compose the cover letter. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

REFERENCES

Abdelrahman, M. (2024). Hallucination in low-resource languages: Amplified risks and mitigation strategies for multilingual LLMs. Journal of Applied Big Data Analytics, Decision-Making, and Predictive Modelling Systems, 8(12), 17–24.

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. Humanities and Social Sciences Communications, 11(1568).

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. Humanities and Social Sciences Communications, 11(1568).

Al-kfairy, M. (2025). Strategic integration of generative AI in organizational settings: Applications, challenges and adoption requirements. IEEE Engineering Management Review, 1–14.

Al-kfairy, M. (2025). Strategic integration of generative AI in organizational settings: Applications, challenges and adoption requirements. IEEE Engineering Management Review, 1–14.

Al-Kfairy, M., Mustafa, D., Kshetri, N., Insiew, M., & Alfandi, O. (2024). Ethical challenges and solutions of generative AI: An interdisciplinary perspective. Informatics, 11, 58. Multidisciplinary Digital Publishing Institute.

Alsadie, D. (2024). A comprehensive review of AI techniques for resource management in fog computing: Trends, challenges, and future directions. IEEE Access, 12, 118007–118059.

Alsadie, D. (2024). A comprehensive review of AI techniques for resource management in fog computing: Trends, challenges, and future directions. IEEE Access, 12, 118007–118059.

Amidi, A., & Amidi, S. (2017). Deep learning cheatsheet. Stanford University. Retrieved from https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-deep-learning-tips-and-tricks

Amidi, A., & Amidi, S. (2017). Deep learning cheatsheet. Stanford University. Retrieved from https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-deep-learning-tips-and-tricks

Anonymous. (2025). Agreeing to disagree: Human-AI collaboration in ethical decision-making. SSRN Electronic Journal. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5262517

Anonymous. (2025). Where is morality on wheels? Decoding large language model (LLM) ethical decision-making in autonomous vehicles. Transportation Research Interdisciplinary Perspectives. https://www.sciencedirect.com/science/article/abs/pii/S2214367X25000572

Anthropic Research Team. (2025, June). Agentic misalignment: How LLMs could be insider threats. Anthropic Research. https://www.anthropic.com/research/agentic-misalignment

Anthropic Research Team. (2025, June). Agentic misalignment: How LLMs could be insider threats. Anthropic Research. https://www.anthropic.com/research/agentic-misalignment

Anthropic. (2025, May). Introducing Claude 4. Anthropic News. https://www.anthropic.com/news/claude-4

Anthropic. (2025, May). Introducing Claude 4. Anthropic News. https://www.anthropic.com/news/claude-4

Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., et al. (2021). AI explainability 360 toolkit. In Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD) (pp. 376–379).

Bai, B., Liang, J., Zhang, G., Li, H., Bai, K., & Wang, F. (2021). Why attentions may not be interpretable? In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining.

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. Improving Image Generation with Better Captions. Available online: https://api.semanticscholar.org/CorpusID:264403242 (accessed on 19 April 2024).

Brock, D. C., & Grad, B. (2022). Expert systems: Commercializing artificial intelligence. IEEE Annals of the History of Computing, 44, 5–7. https://doi.org/10.1109/mahc.2022.3149612

Brock, D. C., & Grad, B. (2022). Expert systems: Commercializing artificial intelligence. IEEE Annals of the History of Computing, 44, 5–7. https://doi.org/10.1109/mahc.2022.3149612

Cass, S. (2016). What would Marvin Minsky read? Key works from the AI Titan's favorite authors. IEEE Spectrum, 53, 22. https://doi.org/10.1109/MSPEC.2016.7439586

Chaudhari, S., Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., Deshpande, A., & Castro da Silva, B. (2024). RLHF deciphered: A critical analysis of reinforcement learning from human feedback for LLMs. ACM Computing Surveys, 58(2), 1–37.

Chen, J., Wang, L., Zhang, Y., & Liu, X. (2024). Towards a better understanding of evaluating trustworthiness in AI systems. ACM Computing Surveys, 57(3), 1–42. https://dl.acm.org/doi/abs/10.1145/3721976

Chen, J., Wang, L., Zhang, Y., & Liu, X. (2024). Towards a better understanding of evaluating trustworthiness in AI systems. ACM Computing Surveys, 57(3), 1–42. https://dl.acm.org/doi/abs/10.1145/3721976

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) (Vol. 1, pp. 539–546). IEEE.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30.

Constitutional AI Project. (2025). Constitutional AI: Tracking Anthropic's revolutionary framework. Constitutional AI Website. Retrieved from https://constitutional.ai/

Databricks Inc. (2025). Databricks AI Governance Framework (DAGF): A comprehensive approach to responsible AI. Technical Report. Databricks, San Francisco, CA. Retrieved from https://www.databricks.com/ai-governance-framework

Databricks Inc. (2025). Databricks AI Governance Framework (DAGF): A comprehensive approach to responsible AI. Technical Report. Databricks, San Francisco, CA. Retrieved from https://www.databricks.com/ai-governance-framework

Dias, M., & Silva Junior, D. S., Oliveira, A.R. (2026a). Innovation at the Core of Competitive Advantage: A Thematic Exploration of Emerging Organizational Pathways. Veredas do Direito,23(6), e235771. https://doi.org/10.18623/rvd.v23.5771

Dias, M., & Silva Junior, D. S., Oliveira, A.R. (2026b). Paradigms and Frontiers in Entrepreneurship: A Systematic Literature Review. The International Journal of Business Management and Technology, 10(4), 1-17; https://doi.org/10.5281/zenodo.19438545

Dokumacı, M. (2024). Legal frameworks for AI regulations. Human Computer Interaction. https://doi.org/10.62802/ytst2927

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Retrieved from https://eur-lex.europa.eu/eli/reg/2024/1689/oj

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Retrieved from https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Fan, Y., Li, R., Zhang, G., Shi, C., & Wang, X. (2025). A weighted cross-entropy loss for mitigating LLM hallucinations in cross-lingual continual pretraining. In ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 1–5). IEEE.

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017), 625–630.

Faust, O., et al. (2018). Deep learning for healthcare applications based on physiological signals: A review. Nature Scientific Reports, 8(1), 1–13.

Faust, O., et al. (2018). Deep learning for healthcare applications based on physiological signals: A review. Nature Scientific Reports, 8(1), 1–13.

Fish, W. (2009). Perception, hallucination, and illusion. Oxford University Press.

Galitsky, B. A., & Rybalov, A. (2025). An information-theoretic model of abduction for detecting hallucinations in explanations. Manuscrito não publicado.

Ganhor, C., Penz, D., Rekabsaz, N., Lesota, O., & Schedl, M. (2022). Unlearning protected user attributes in recommendations with adversarial training. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. https://doi.org/10.1145/3477495.3531820

Gantla, S. R. (2025). Exploring mechanistic interpretability in large language models: Challenges, approaches, and insights. In 2025 International Conference on Data Science and Engineering. Retrieved from https://ieeexplore.ieee.org/abstract/document/11011640/

Guha, N., Lawrence, C. M., Gailmard, L. A., Rodolfa, K. T., Saremi, F., Beaumavant, R., Deborah, I., Raj, M., Covellan, F., Hemingberg, C., et al. (2024). AI regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. George Washington Law Review, 92(5), 1473.

Guha, N., Lawrence, C. M., Gailmard, L. A., Rodolfa, K. T., Saremi, F., Beaumavant, R., Deborah, I., Raj, M., Covellan, F., Hemingberg, C., et al. (2024). AI regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. George Washington Law Review, 92(5), 1473.

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). Retrieval augmented language model pre-training. In International Conference on Machine Learning (pp. 3929–3938). PMLR.

Hahn, M. (2020). Theoretical limitations of self-attention in neural sequence models. Transactions of the Association for Computational Linguistics, 8, 156–171. https://direct.mit.edu/tacl/article-abstract/doi/10.1162/tacl_a_00306/43545

Hahn, M. (2020). Theoretical limitations of self-attention in neural sequence models. Transactions of the Association for Computational Linguistics, 8, 156–171. https://direct.mit.edu/tacl/article-abstract/doi/10.1162/tacl_a_00306/43545

Henderson, P., Hashimoto, T., & Lemley, M. A. (2023). Where’s the liability in harmful AI speech? Journal of Free Speech Law, 3, 535–594.

Henderson, P., Hashimoto, T., & Lemley, M. A. (2023). Where’s the liability in harmful AI speech? Journal of Free Speech Law, 3, 535–594.

HM Courts & Tribunals Service & Ministry of Justice. (2025). AI Action Plan for Justice: Transforming the justice system through artificial intelligence. Technical Report. Ministry of Justice, United Kingdom. Retrieved from https://www.gov.uk/government/publications/ai-action-plan-justice

HM Courts & Tribunals Service & Ministry of Justice. (2025). AI Action Plan for Justice: Transforming the justice system through artificial intelligence. Technical Report. Ministry of Justice, United Kingdom. Retrieved from https://www.gov.uk/government/publications/ai-action-plan-justice

Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. Neural Comput. 1997, 9, 1735–1780.

Infocomm Media Development Authority & Personal Data Protection Commission. (2025). AI Verify: Singapore's framework for AI governance and testing. Technical Report. Government of Singapore. Retrieved from https://www.aiverify.sg/

International Organization for Standardization & International Electrotechnical Commission. (2023). ISO/IEC 42001:2023 Information technology—Artificial intelligence—Management system. Retrieved from https://www.iso.org/standard/81230.html

International Organization for Standardization & International Electrotechnical Commission. (2023). ISO/IEC 42001:2023 Information technology—Artificial intelligence—Management system. Retrieved from https://www.iso.org/standard/81230.html

Jacob, C., Brasier, N., Laurenzi, E., Heuss, S., Mougiakakou, S.-G., Cöltekin, A., & Peter, M. K. (2025). AI for IMPACTS framework for evaluating the long-term real-world impacts of AI-powered clinician tools: Systematic review and narrative synthesis. Journal of Medical Internet Research, 27, e67485.

Jacob, C., Brasier, N., Laurenzi, E., Heuss, S., Mougiakakou, S.-G., Cöltekin, A., & Peter, M. K. (2025). AI for IMPACTS framework for evaluating the long-term real-world impacts of AI-powered clinician tools: Systematic review and narrative synthesis. Journal of Medical Internet Research, 27, e67485.

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Javed, H., El-Sappagh, S., & Abuhmed, T. (2024). Robustness in deep learning models for medical diagnostics: Security and adversarial challenges towards robust AI applications. Artificial Intelligence Review, 57(11), 1–55.

Javed, H., El-Sappagh, S., & Abuhmed, T. (2024). Robustness in deep learning models for medical diagnostics: Security and adversarial challenges towards robust AI applications. Artificial Intelligence Review, 57(11), 1–55.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730

Kamoi, R., Zhang, Y., Zhang, N., Han, J., & Zhang, R. (2024). When can LLMs actually correct their own mistakes? A critical survey of self-correction of LLMs. Transactions of the Association for Computational Linguistics, 12, 1417–1440.

Kim, I.; Han, G.; Ham, J.; Baek, W. KoGPT: KakaoBrain Korean(hangul) Generative Pre-Trained Transformer. 2021. Available online: https://github.com/kakaobrain/kogpt (accessed on 19 April 2024).

Kim, M., Jung, H., & Koo, M.-W. (2024). Self-expertise: Knowledge-based instruction dataset augmentation for a legal expert language model. In Findings of the Association for Computational Linguistics: NAACL 2024 (pp. 1098–1112). ACL.

Li, J., Li, G., Li, Y., & Jin, Z. (2025). Structured chain-of-thought prompting for code generation. ACM Transactions on Software Engineering and Methodology, 34(2), 1–23.

Liu, Z. (2024). Cultural bias in large language models: A comprehensive analysis and mitigation strategies. Journal of Transcultural Communication, 3(2), 224–244.

Lopes, R. & Dias, M. (2026a) Boards of Directors and Corporate Governance Outcomes: A Literature Review. Advances in Social Sciences Research Journal, 13(04), 46–66. https://doi.org/10.14738/assrj.1304.20194

Lopes, R. & Dias, M. (2026b) Mapping the Themes of Gender Diversity in Boards: A Thematic Analysis. Veredas Do Direito, 23(5), e235738. https://doi.org/10.18623/rvd.v23.5738

Lund, B., Orhan, Z., Mannuru, N. R., Bevara, R. V. K., Porter, B., Vinaih, M. K., & Bhaskara, P. (2025). Standards, frameworks, and legislation for artificial intelligence (AI) transparency. AI and Ethics, 5, 3639–3655. https://link.springer.com/article/10.1007/s43681-025-00661-4

Maharaj, K., Saxena, A., Kumar, R., Mishra, A., & Bhattacharyya, P. (2023). Eyes show the way: Modelling gaze behaviour for hallucination detection. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 11424–11438). ACL. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.764/

Malin, B., Kalganova, T., & Boulgouris, N. (2025). A review of faithfulness metrics for hallucination assessment in large language models. IEEE Journal of Selected Topics in Signal Processing. https://doi.org/10.1109/JSTSP.2025.XXXXXXX

Manheim, D., Martin, S., Bailey, M., Samin, M., & Greutzmacher, R. (2025). The necessity of AI audit standards boards. AI & Society. https://link.springer.com/article/10.1007/s00146-025-02320-y

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. AI Magazine, 27, 12–14. https://doi.org/10.1609/aimag.v27i4.1904

McDonald, D., Papadopoulos, R., & Benningfield, L. (2025). Reducing LLM hallucination using knowledge distillation: A case study with Mistral Large and MMLU benchmark. Authorea Preprints.

Mohammed, M. Y., Ali, S. A., Ali, S. K., Majeed, A. A., & Mohamed, E. H. (2025). AFTINA: Enhancing stability and preventing hallucination in AI-based Islamic fatwa generation using LLMs and RAG. Neural Computing and Applications, 1–26.

Mortaji, S. T. H., & Sadeghi, M. E. (2024). Assessing the reliability of artificial intelligence systems: Challenges, metrics, and future directions. International Journal of Innovation in Management, Economics and Social Sciences, 4(4), 1–15.

Naik, N., Hameed, B., Shetty, D., Swain, D., Shah, M., Paul, R., Aggarwal, K., Ibrahim, S., Patil, V., Smriti, K., Shetty, S., Rai, B. P., Chlosta, P., & Somani, B. (2022). Legal and ethical consideration in artificial intelligence in healthcare: Who takes responsibility? Frontiers in Surgery, 9. https://doi.org/10.3389/fsurg.2022.862322

Nannini, L., Balayn, A., & Smith, A. L. (2023). Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the EU, US, and UK. ACM Conference on Fairness, Accountability, and Transparency, 1198–1212.

National Institute of Standards and Technology. (2024). AI Risk Management Framework (AI RMF 1.0): Generative AI Profile. Technical Report NIST AI 600-1. U.S. Department of Commerce. Retrieved from https://www.nist.gov/itl/ai-risk-management-framework

NVIDIA Corporation. (2017). Deep learning: An introductory guide. Retrieved from https://research.nvidia.com/publication/2017-06_Deep-Learning-Introductory-Guide

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

Pan, Y., Kong, L., Wu, J., Yang, Y., Zuo, H., Xiu, Z., & Wang, X. (2025). Towards reliable large language models: A survey on hallucination detection. In International Conference on Intelligent Computing (pp. 438–451). Springer.

Quintão, H., Dias, M., da Mata, R. & da Silva Jr., D. (2026). Artificial Intelligence In Alternative Dispute Resolution: A Literature Review on Guidelines, Best Practices, and The Role Of Engineering Experts In Complex Construction Contracts. International Journal of Advance Research, 14(3), 821-830, https://doi.org/10.5281/zenodo.19161386

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.

Radford, A.; Narasimhan, K. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://api.semanticscholar.org/CorpusID:49313245 (accessed on 29 March 2024).

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. OpenAI Blog 2019, 1, 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

Raiaan, M.A.K.; Mukta, M.S.H.; Fatema, K.; Fahad, N.M.; Sakib, S.; Mim, M.M.J.; Ahmad, J.; Ali, M.E.; Azam, S. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. IEEE Access 2024, 12, 26839–26874.

Saunders, M., Lewis, P., & Thornhill, A. (2009). Research methods for business students (5th ed.). Harlow: Pearson Education.

Shah, A., Banner, N., Heginbotham, C., & Fulford, B. (2014). Substance use and older people. Substance Use and Older People, 21(5), 9.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2024). Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems, 36.

Sinclair, S., & Rockwell, G. (2026). Voyant Tools (Version 2.6.19) [Text analysis software]. Voyant Consortium. https://voyant-tools.org

Strauss, A., & Corbin, J. (1998). Basics of qualitative research: Techniques and procedures for developing grounded theory (2nd ed.). Thousand Oaks, CA: Sage.

Sultan, M. A., Ganhotra, J., & Astudillo, R. F. (2024). Structured chain-of-thought prompting for few-shot generation of content-grounded QA conversations. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 16172–16187). ACL.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; Hashimoto, T.B. Stanford Alpaca: An Instruction-Following LLaMA Model. 2023. Available online: https://github.com/tatsu-lab/stanford_alpaca (accessed on 29 March 2024).

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523–538

Xu, T., Wu, S., Diao, S., Liu, X., Wang, X., Chen, Y., & Gao, J. (2024). Retrieval and reasoning on KGs: Integrate knowledge graphs into large language models for complex question answering. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 7598–7610). ACL.

Yang, D., Yuan, R., Fan, Y., Yang, Y., Wang, Z., Wang, S., & Zhao, H. (2023). RefGPT: Dialogue generation of GPT, by GPT, and for GPT. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 2511–2535). ACL.

Yuan, S.; Zhao, H.; Du, Z.; Ding, M.; Liu, X.; Cen, Y.; Zou, X.; Yang, Z.; Tang, J. WuDaoCorpora: A super large-scale Chinese corpora for pre-training language models. AI Open 2021, 2, 65–68.

Zhang, H., Liu, X., & Zhang, J. (2023). Summit: Iterative text summarization via ChatGPT. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 10644–10657). ACL.

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., ... & Wu, F. (2026). Instruction tuning for large language models: A survey. ACM Computing Surveys, 58(7), 1-36.