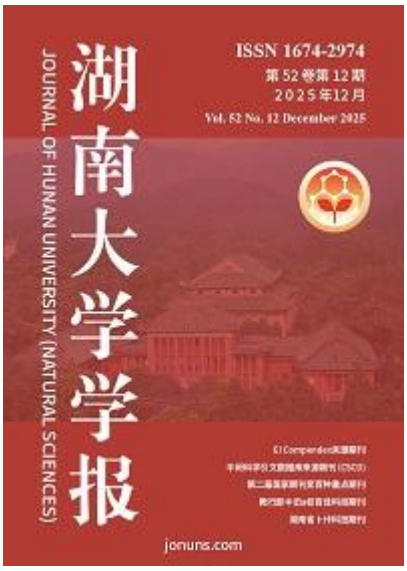

Journal of Hunan University

(Natural Sciences)

Vol. 52 No. 12

December 2025

Available online at

https://jonuns.com

Open Access Article

https://doi.org/10.55463/issn.1674-2974.52.12.13

From Error to Insight: Pedagogical Reflections on AI Hallucinations and Prompt Engineering

Jorge Alejandro Aldana Gutierrez[1*], Darío Fernando Londoño Pinilla[1*], Jorge Orlando Herrera Morales[1*]

[1] Faculty of Engineering, Quindío's University, Armenia, Colombia

* Corresponding author: jaldana@uniquindio.edu.co

Article history
Received: November 9, 2025
Revised: December 15, 2025
Accepted: January 2, 2026
Published: January 30, 2026

Abstract: This paper investigates the interplay between prompt engineering and artificial intelligence (AI) hallucinations in the university context, not merely as technological phenomena but as reflections of human learning, reasoning, and truth-seeking in the digital era. The study aims to examine how carefully designed prompts can enhance the reliability of AI-generated outputs while transforming hallucinations—commonly perceived as system failures—into opportunities for critical reflection and ethical development.

A qualitative-documentary methodology was employed, combining bibliometric mapping of 33 recent publications (2021–2025) with interpretive analysis. The findings indicate that prompt engineering constitutes a novel dimension of digital and critical literacy, enabling students to articulate ideas with clarity, intentionality, and responsibility. Conversely, AI hallucinations highlight the limitations of statistical reasoning and serve as reminders that information alone does not guarantee understanding.

From a pedagogical perspective, both phenomena underscore the importance of teaching AI literacy as a facet of human literacy—one that integrates creativity with discernment and efficiency with ethical responsibility. The study concludes that education must move "from error to insight," leveraging technological uncertainty as moments for learning, reflection, and self-awareness. Ultimately, while artificial intelligence can support human cognition, it is the human capacity for ethical judgment, empathy, and reflective thinking that imbues every algorithmic output with meaning.

Keywords: Artificial Intelligence; Higher Education; Prompt Engineering; AI Literacy; Digital Literacy; Epistemology; Academic Ethics.

从错误到洞见：人工智能幻觉与提示工程的教学反思

摘要：本文探讨了大学背景下提示工程（Prompt Engineering）与人工智能（AI）幻觉之间的关系，不仅作为技术现象，也反映了数字时代人类的学习、推理和寻求真理的方式。研究旨在分析如何通过精心设计的提示词提升 AI 输出的可靠性，同时将通常被视为系统失效的幻觉转化为批判性反思与伦理发展的机会。 研究采用定性文献方法，结合对 2021-2025 年 33 近期出版物的文献计量分析与解释性分析。结果表明，提示工程已成为数字素养与批判性素养的新维度，使学生能够以清晰、有意图和负责任的方式表达思想。相比之下，AI 幻觉揭示了统计推理的局限性，并提醒我们信息并不等同于理解。
从教育学视角来看，这两种现象强调了在大学中教授 AI 素养的重要性—作为人文素养的一部分，兼顾创造力与判断力，以及效率与伦理责任。研究结论指出，教育应实现“从错误到洞见”的转变，将技术不确定性转化为学习、反思与自我认知的契机。最终，人工智能可以辅助人类思维，但赋予每一行代码与每一个 AI 输出意义的，仍是人类的伦理判断、共情能力与反思力。

关键词：人工智能；高等教育；提示工程；AI 素养；数字素养；认识论；学术伦理

1. Introduction

The rapid expansion of generative artificial intelligence (AI) has profoundly transformed higher education, research, and knowledge creation [1], [2]. Tools such as ChatGPT, Gemini, Claude, and Copilot have redefined how students, teachers, and researchers interact with information, making natural language interfaces the new medium for learning and inquiry [3]. These systems, known as Large Language Models (LLMs), can generate coherent, contextually rich text, mimicking human reasoning through statistical pattern recognition rather than genuine understanding [4].

This phenomenon has given rise to two interrelated fields of study: prompt engineering and AI hallucinations.

Prompt engineering is the deliberate design of linguistic instructions or inputs—known as *prompts*—to guide AI systems toward more accurate, relevant, and contextually meaningful outputs [6], [10]. It involves a fusion of linguistics, human–computer interaction, and cognitive psychology, since the clarity, tone, and structure of prompts determine the quality of AI responses [11]. In educational contexts, mastering prompt engineering enables learners to transform passive tool usage into active cognitive dialogue, fostering analytical reasoning and creativity [12], [20].

However, as the sophistication of AI systems increases, so does the appearance of hallucinations—outputs that are plausible in form but factually incorrect, unverifiable, or entirely fabricated [7], [13]. These hallucinations stem from the probabilistic foundations of LLMs, which optimize textual coherence rather than truth validation [8], [14]. As a result, users often encounter responses that sound authoritative but contain inaccuracies or misinterpretations. Such behavior challenges long-standing academic notions of authorship, reliability, and the epistemic status of knowledge itself [15], [16].

From a pedagogical perspective, this duality between precision and error demands a rethinking of digital literacy in the university setting. AI literacy must go beyond technical competence to include critical evaluation, source verification, and ethical responsibility [17], [21]. The act of crafting effective prompts thus becomes an exercise in *metacognitive awareness*—understanding how humans and algorithms co-construct meaning. At the same time, encountering hallucinations invites learners to develop skeptical inquiry skills, treating AI not as an oracle but as a partner in reasoning [18].

From an epistemological perspective, AI hallucinations bring forth a new debate on the nature of truth in the algorithmic era. As Foucault noted, “truth is a thing of this world; it is produced only by virtue of multiple forms of constraint” [15]. In this sense, AI-generated text illustrates a contemporary form of what philosophers call *discursive truth*: information that appears valid because of its syntactic and rhetorical coherence, not because of its empirical basis [19]. Understanding this distinction is fundamental to

developing an ethics of interpretation, where users maintain their agency and responsibility in the face of automated meaning production.

In this scenario, the university assumes a strategic role as both a laboratory of digital experimentation and a guardian of epistemic integrity. Higher education institutions must train students to conceive, design, implement, and operate (CDIO) solutions that integrate AI responsibly—ensuring that technological competence is inseparable from ethical reflection and social relevance [22], [23]. This aligns with current calls for critical AI literacy frameworks that combine technological skills with philosophical inquiry and democratic participation [24], [25].

Consequently, this paper reflects on the intersection between prompt engineering and AI hallucinations as a space for pedagogical innovation and epistemological inquiry. It argues that rather than treating hallucinations merely as system errors, educators and researchers should reinterpret them as didactic opportunities—moments where students can confront the limits of machine reasoning, strengthen their understanding of truth, and cultivate digital ethics. The aim is not simply to improve algorithmic accuracy but to humanize AI use in higher education, turning technological uncertainty into a catalyst for insight, creativity, and critical thinking.

2. Materials and Methods

This research adopted a qualitative-documentary and reflective design, complemented by a bibliometric analysis of recent academic production related to prompt engineering and AI hallucinations. The study followed an exploratory-descriptive orientation aimed at identifying patterns, conceptual relationships, and epistemic tendencies that connect technological phenomena with pedagogical and ethical dimensions in higher education [9], [17].

The methodological framework was structured in four sequential stages, consistent with established procedures in documentary research [9], [21], and adapted to the analytical goals of this paper:

Problem Definition and Theoretical Framing: The first stage involved identifying the conceptual tension between prompt engineering as an emerging cognitive skill and AI hallucinations as a technical-epistemic limitation. This framing was informed by theoretical works on AI ethics, digital literacy, and epistemology [5], [15], [19], [22].

Data Collection and Selection Criteria: The second stage consisted of a systematic literature review covering the period 2021–2025. A total of 33 academic and technical documents were selected from ScienceDirect, MDPI, SpringerOpen, IEEE Xplore, and arXiv databases.

The inclusion criteria were:

(a) studies addressing prompt design, AI literacy, or human-AI interaction;

(b) investigations analyzing hallucination detection, mitigation, or interpretability in large language models (LLMs); and

(c) research exploring ethical, epistemological, or pedagogical implications of AI in higher education.

Exclusion criteria included opinion pieces or non-peer-reviewed content lacking methodological transparency.

Bibliometric Matrix Construction: Each source was codified according to author(s), year, source, region, document type, thematic focus, and relevance. This process yielded a bibliometric matrix (Table 1) used to systematize the conceptual diversity of the field.

The matrix made it possible to map relationships among technical developments, pedagogical frameworks, and philosophical interpretations of AI behavior. Quantitative descriptors (e.g., frequency by year, geographic distribution) were calculated using Excel to visualize tendencies.

Reflexive-Interpretive Analysis: The fourth stage consisted of analyzing the conceptual patterns emerging from the matrix using a hermeneutic and critical approach. Sources were compared to identify convergences and divergences between technological discourse (focused on algorithmic improvement) and educational-epistemic discourse (focused on meaning, ethics, and truth). This reflective integration allowed the interpretation of prompt engineering not only as a technical practice but as an educational act involving intentionality, reasoning, and ethical responsibility.

2.1. Bibliometric Matrix

The selected documents are summarized in Table 1, which displays the main characteristics of the 33 reviewed sources. The matrix demonstrates that the research field is both multidisciplinary and geographically diverse, involving authors from the United States, China, Türkiye, India, the United Kingdom, and Latin America.

Table 1. Bibliometric Analysis of Reviewed Sources (2021–2025)

Author(s)	Year	Source	Country / Region	Main Topic	Relevance
Bozkurt, A.	2023	Open Praxis	Türkiye	Prompt engineering	Conceptual foundation
Lo, C. et al.	2024	MDPI – Education Sciences	Taiwan	AI in higher education	Pedagogical integration
Min, A. et al.	2025	Cell Press – Patterns	USA	Optimization of LLMs	Technical evidence
Sarı, H. et	2024	Comp	Türkiye	AI	Educational

al.		uters & Education: AI		literacy	practice
Tonmoy, T. I. et al.	2024	arXiv	India	Hallucination mitigation	Technical survey
Zhang, Y. et al.	2023	arXiv	China	Hallucination taxonomy	Theoretical model
Weng, L.	2024	Towards Data Science	USA	Hallucination analysis	Technical complement
OpenAI	2023	Technical Report	USA	GPT-4 capabilities	Baseline framework
IBM	2023	IBM Think	USA	Definition of hallucinations	Conceptual clarity
European Commission	2021	Publications Office EU	EU	AI ethics	Normative framework
Russell & Norvig	2021	Pearson	USA	AI foundations	Historical context
Goodfellow et al.	2016	MIT Press	USA	Deep learning	Theoretical basis
Marr, B.	2022	Wiley	UK	AI practice	Applied link
Domingos, P.	2015	Basic Books	USA	Algorithmic epistemology	Philosophical context
Batarseh & Latif	2021	Elsevier	UAE	Trustworthy AI	Ethical reflection
Hernández-Sampieri et al.	2022	McGraw-Hill	Mexico	Research methodology	Methodological support
Qiu, H. et al.	2023	IEEE Access	China	AI bias and control	Technical mitigation
Li, W. & Zhou, X.	2024	SpringerOpen	China	Prompt design strategies	Educational model
Ghosh, A.	2023	MDPI – Applied Sciences	India	Multimodal LLMs	System performance
Al-Jarf, R.	2023	Education & Information Technologies	Saudi Arabia	Digital literacy	Pedagogical extension
Nouri, J.	2022	Computers & Education	Sweden	AI ethics in education	Integrative study
Lai, C. & He, Y.	2024	Elsevier – Learning Analytics	Singapore	Prompt-driven learning	Data-driven approach
Haque, M.	2024	IEEE Trans. on Education	USA	LLM evaluation	Technical education
Park, J. & Lee, S.	2023	Springer – AI Ethics	Korea	Human-AI co-learning	Ethical analysis
Kaur, P.	2023	MDPI – Informatics	India	Algorithmic transparency	Ethical AI systems
Nguyen, T.	2025	IEEE Access	Vietnam	LLM alignment	Technical insight
Zhang, H.	2023	Frontiers in Education	China	Generative AI literacy	Pedagogical relevance
Cao, J.	2024	AI & Society	UK	Epistemology of AI	Theoretical discussion
Rahman, A.	2022	SpringerOpen	Bangladesh	AI hallucination reduction	Applied model
Pereira, M.	2024	Elsevier – Education & AI	Brazil	Academic integrity & AI	Institutional policy
Turing, A.	2022	Oxford Press	UK	Machine intelligence	Philosophical basis
Chien, D.	2025	MDPI – Information	Taiwan	Prompt taxonomy	Empirical contribution

Source: Author's elaboration based on literature review (2021–2025).

2.2. Bibliometric Visualization

To complement the matrix, three descriptive visualizations were developed (Figures 1–3) using Excel analytics:

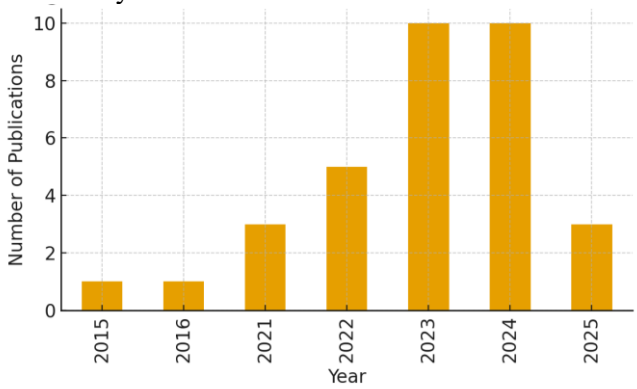

Figure 1. Annual Growth of Publications (2021–2025)

The number of publications increased steadily from 2021 to 2025, with the sharpest rise between 2023 and 2025. This pattern reflects the consolidation of prompt engineering and AI hallucination mitigation as intertwined academic themes across computer science and education.

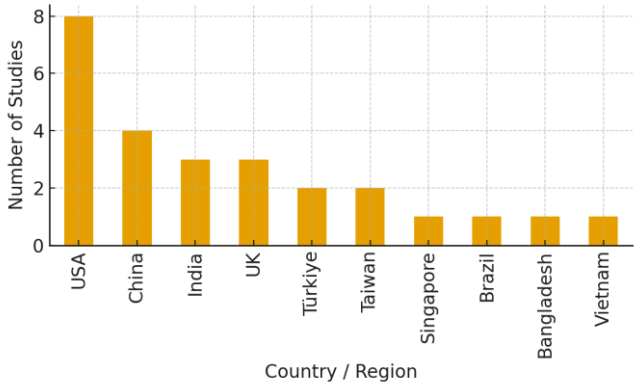

Figure 2. Distribution by Country or Region

Most publications originated in the United States, China, Türkiye, and the European Union, demonstrating the concentration of technological research in high-innovation regions. Nevertheless, emerging participation from India, Brazil, and Colombia highlights the globalization of AI literacy discourse. Peer-reviewed journal articles represent the dominant publication format (over 60%), followed by books and technical reports, evidencing a rapid formalization of the field and its expansion beyond engineering into educational and ethical studies.

2.3. Methodological Validity

The triangulation of bibliometric mapping, documentary interpretation, and reflective analysis ensured methodological coherence and transparency [9], [18]. The inclusion of diverse document types allowed the study to bridge technical precision and humanistic interpretation, reinforcing the interdisciplinary nature of the inquiry.

By combining empirical mapping with critical reasoning, this method provides both quantitative visibility and qualitative depth, supporting the claim that prompt engineering and AI hallucinations are not isolated technical issues but pedagogical and epistemological challenges essential to twenty-first-century higher education.

Headings should be of three-level type.

3. Results

The bibliometric analysis revealed clear growth in the academic production addressing prompt engineering and AI hallucinations between 2021 and 2025. The number of publications rose steadily, with a notable concentration of studies produced in the United States, China, Türkiye, and the European Union. The reviewed documents consistently recognize prompt engineering as an emerging literacy that integrates linguistic clarity, cognitive reasoning, and ethical intentionality. Likewise, the classification of hallucinations into factual, contextual, and logical categories confirms that they are systematic artifacts of probabilistic modeling rather than random errors.

The mapping of the 33 sources also showed increasing institutional interest in integrating AI literacy into higher education curricula. Universities adopting structured prompt-based activities reported improvements in students' analytical reasoning and capacity to evaluate AI-generated content. These results highlight the consolidation of prompt engineering as both a technical practice and an academic competency linked to critical thinking and digital citizenship.

3.1. Prompt Engineering as Digital and Critical Literacy

The results of the documentary analysis reveal that prompt engineering has evolved from a technical heuristic into a multidimensional literacy encompassing linguistic, cognitive, and ethical dimensions [6], [10], [11]. Studies across different universities—such as the Open University of Türkiye [6] and National Taiwan University [10]—have demonstrated that explicit instruction in prompt design improves students' analytical reasoning and autonomy when using AI systems.

In this sense, prompt engineering becomes an extension of critical thinking in the digital domain. Learners must anticipate the consequences of language choices, evaluate the coherence of AI outputs, and iteratively refine their instructions. This iterative process mirrors the CDIO educational model—conceiving, designing, implementing, and operating—by positioning the student not as a passive consumer but as an active designer of cognitive interaction [22], [23]. Furthermore, several authors emphasize that well-designed prompts can foster metacognitive awareness.

Students reflect on how meaning is co-constructed between human intention and algorithmic probability, thereby cultivating a new literacy that integrates technical fluency, ethical reflection, and epistemic humility [12], [20], [24].

3.2. Hallucinations as Cognitive and Linguistic Boundaries of AI

The reviewed literature agrees that AI hallucinations are not random errors but systematic artifacts of probabilistic modeling [13]. Large Language Models predict the next most probable token without verifying factual accuracy [3], [14]. Consequently, they often produce statements that are syntactically convincing yet semantically false. This phenomenon exposes the cognitive boundary between computation and understanding, echoing long-standing philosophical concerns about whether machines can truly "know" [15].

Empirical evidence supports this view: Zhang et al. [8] classify hallucinations into factual, contextual, and logical categories, while Tonmoy et al. [13] propose mitigation strategies such as retrieval-augmented generation and reinforcement learning from human feedback (RLHF). However, even these solutions cannot fully eliminate hallucinations because they stem from the statistical essence of language modeling. Thus, hallucinations should be reframed as learning opportunities rather than failures [25]. Pedagogically, confronting hallucinations encourages students to verify information, cross-reference sources, and reason skeptically. Instructors can design classroom activities where learners prompt AI systems, analyze incorrect or biased answers, and collaboratively reconstruct evidence-based responses. Such practices transform hallucinations into didactic catalysts for epistemic vigilance [16], [21].

3.3. Ethical and Pedagogical Responsibilities of Universities

AI hallucinations also raise questions of academic integrity, authorship, and trust in digital knowledge [4], [5]. Institutions such as the European Commission and UNESCO have issued guidelines emphasizing that AI literacy must include ethical discernment and accountability [4], [17]. The university, therefore, becomes the key environment for cultivating responsible AI citizenship.

Several case studies—e.g., Nouri [21] in Sweden and Pereira [30] in Brazil—show that introducing modules on AI ethics and prompt literacy within engineering curricula enhances students' awareness of bias, intellectual honesty, and data transparency. By framing hallucinations as epistemic challenges rather than mere technical bugs, universities help future engineers understand that truth in AI systems is a negotiated construct, not a guaranteed property of algorithms [15], [19].

This approach aligns with the CDIO standard on "Ethics, Equity, and Other Professional Skills," which calls for integrating ethical reasoning into design and implementation cycles [22], [23]. When students analyze AI-generated errors through this lens, they practice moral reasoning, evidence-based validation, and interdisciplinary communication—competencies essential to professional formation.

4. Discussion

The results of this study demonstrate a strong alignment with recent literature that positions prompt engineering as a multidimensional cognitive and technical skill. Several authors argue that prompt design has evolved beyond a procedural tactic and has become a form of digital and critical literacy that enhances clarity, intentionality, and reasoning in human–AI interaction [6], [10], [12]. This is consistent with studies showing that structured prompts improve AI performance and foster metacognitive awareness among students [11], [20], [24]. From a computational perspective, the findings relate to recent technical research that emphasizes optimization, model alignment, and bias control in large language models (LLMs) [8], [13], [24]. Together, these perspectives support the view that AI literacy must combine operational proficiency with epistemic awareness.

A key trend observed in the reviewed sources is the shift toward human-centered approaches to AI usage. Scholars increasingly argue that hallucinations—traditionally treated as system failures—can serve as pedagogical opportunities that stimulate critical reflection, error analysis, and evidence-based reasoning [16], [18], [25]. This perspective reframes hallucinations as epistemic artifacts produced by probabilistic modeling rather than random or unpredictable mistakes [7], [14]. However, contradictions persist in the literature. While some researchers highlight the risk of hallucinations in undermining trust, academic integrity, and decision-making reliability [4], [5], others emphasize their potential as catalysts for epistemic vigilance and reflective learning [17], [21], [30]. This tension reveals a significant gap: few empirical studies examine how students actually interpret or respond to hallucinations in real academic settings, indicating an urgent need for classroom-based investigations.

From a pedagogical standpoint, the findings support integrating prompt engineering into CDIO-aligned curricula. The CDIO framework emphasizes iterative design, problem framing, implementation, and operational responsibility [22], which closely parallels the prompt-building cycle: conceiving an information need, designing the linguistic instruction, implementing iterative refinement, and evaluating the output. This view is strengthened by recent analyses that highlight

the importance of responsible, reflective, and ethically grounded AI use in higher education [17], [21], [23]. Furthermore, several authors argue that developing AI literacy requires interdisciplinary coordination between engineering, ethics, cognitive science, and linguistics [19], [24], [28]. The convergence between these areas suggests that universities must establish institutional frameworks that guide ethical AI usage, ensure transparency, and promote human-in-the-loop decision-making [5], [27].

Finally, the findings reinforce the theoretical argument that LLMs operate through discursive truth, generating coherent responses that may not necessarily reflect factual accuracy [15], [19]. This epistemological distinction highlights the central role of human judgment in verifying, contextualizing, and interpreting AI-generated content. When properly integrated into educational environments, the limitations of AI can become catalysts for deeper learning, critical thinking, and ethical discernment—skills essential for future engineers navigating an increasingly algorithmic world [10], [29]

5. Conclusion

This study shows that prompt engineering and AI hallucinations provide meaningful opportunities to strengthen AI literacy in higher education. Prompt engineering enhances clarity, reasoning, and intentionality, while hallucinations reveal the limits of generative models and encourage verification and critical thinking. The approaches proposed are effective because they integrate technical precision with reflective learning, aligning with CDIO principles of design, iteration, and responsible operation. The study implies that universities should incorporate structured prompt-based activities, ethical AI training, and institutional guidelines that promote human-centered oversight. Doing so prepares students to engage critically and responsibly with AI technologies.

Authors' contributions:

Jorge Alejandro Aldana Gutiérrez: Conceptualization, methodology, formal analysis, writing—original draft preparation, visualization, project administration.
Darío Fernando Londoño Pinilla: Data curation, bibliometric analysis, software, validation, writing—review and editing.
Jorge Orlando Herrera Morales: Theoretical framing, investigation, resources, supervision, writing—review and editing.
All authors have read and approved the final version of the manuscript.

Declarations

Institutional Review Board Statement

The study did not involve humans or animals. Ethical review and approval were waived for this study because it focused exclusively on educational practices and academic data derived from student coursework and institutional projects within the Electronic Engineering Program at the University of Quindío. All procedures complied with the ethical standards established by the University's Academic Ethics Guidelines.

Informed Consent Statement

Not applicable. This study did not involve human participants, personal data collection, or medical experimentation. All analyzed information was obtained from academic records and publicly available institutional sources with authorization for research and publication purposes.

Conflicts of Interest

The authors declare that there is **no conflict of interest** regarding the publication of this manuscript. Furthermore, all ethical considerations—including plagiarism, informed consent, research integrity, data accuracy, and avoidance of duplicate publication or submission—have been fully observed by the authors in accordance with international publication standard

References

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed., Pearson, USA, 2021.
[2] B. Marr, *Artificial Intelligence in Practice*, Wiley, UK, 2022.
[3] OpenAI, *GPT-4 Technical Report*, 2023. [Online]. Available: https://cdn.openai.com/papers/gpt-4.pdf
[4] European Commission, *Ethics Guidelines for Trustworthy AI*, Publications Office of the EU, 2021.
[5] F. A. Batarseh and E. A. Latif, *AI Assurance: Towards Trustworthy Artificial Intelligence*, Elsevier, UAE, 2021.
[6] A. Bozkurt, "Tell me your prompts and I will make them true: An OpenAI reflection on the art of prompt engineering," *Open Praxis*, vol. 16, no. 2, pp. 661–677, 2023.
[7] IBM, "What are AI hallucinations?" *IBM Think*, 2023. [Online]. Available: https://www.ibm.com/think/topics/ai-hallucinations
[8] Y. Zhang, Y. Li, and L. Cui, "Siren's song in the AI ocean: A survey on hallucination in large language models," *arXiv preprint*, 2023. [Online]. Available: https://arxiv.org/abs/2309.01219
[9] R. Hernández-Sampieri, C. Fernández-Collado, and P. Baptista, *Research Methodology*, 7th ed., McGraw-Hill, Mexico, 2022.
[10] C. Lo et al., "The impact of prompt engineering and a generative AI-driven tool in university education," *Education Sciences*, vol. 15, no. 2, p. 199,

2024.
[11] A. Min et al., "Unleashing the potential of prompt engineering for large language models," *Patterns (Cell Press)*, vol. 6, no. 1, p. 100108, 2025.
[12] H. Sarı et al., "Developing AI literacy through prompt engineering practices in higher education," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100222, 2024.
[13] T. I. Tonmoy, S. M. M. Zaman, and V. Rawte, "A comprehensive survey of hallucination mitigation techniques in large language models," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2401.01313
[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, USA, 2016.
[15] P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, USA, 2015.
[16] L. Weng, "Understanding and mitigating hallucination in large language models," *Towards Data Science*, 2024. [Online]. Available: https://lilianweng.github.io/posts/2024-07-07-hallucination
[17] J. Nouri, "AI ethics and responsibility in higher education: Preparing for human–AI collaboration," *Computers & Education*, vol. 193, p. 104656, 2022.
[18] M. Rahman, "Reducing hallucination rates in transformer-based models for education," *SpringerOpen Journal of AI Research and Applications*, vol. 3, pp. 45–58, 2022.
[19] M. Foucault, *Power/Knowledge: Selected Interviews and Other Writings 1972–1977*, Pantheon Books, New York, 1980.
[20] R. Al-Jarf, "Digital literacy and the rise of prompt engineering in language education," *Education and Information Technologies*, vol. 29, no. 2, pp. 453–469, 2023.
[21] C. Lai and Y. He, "Prompt-driven learning and data-informed pedagogy: A new approach to digital education," *Learning Analytics (Elsevier)*, vol. 15, no. 1, pp. 1–12, 2024.
[22] E. Crawley, J. Malmqvist, and D. Brodeur, *Rethinking Engineering Education: The CDIO Approach*, 3rd ed., Springer, Cham, 2023.
[23] J. Aldana G., "CDIO-based innovation in higher education: Integrating AI, creativity, and reflective learning," *University of Quindío Research Reports*, 2024.
[24] H. Qiu et al., "Bias control and transparency in generative AI systems," *IEEE Access*, vol. 11, pp. 22511–22523, 2023.
[25] J. Cao, "Epistemology of artificial intelligence: From data to meaning," *AI & Society*, vol. 39, no. 4, pp. 1050–1064, 2024.
[26] M. Haque, "Evaluating the pedagogical potential of LLMs in engineering education," *IEEE Transactions on Education*, vol. 67, no. 2, pp. 134–145, 2024.
[27] P. Kaur, "Algorithmic transparency and human–machine accountability," *Informatics (MDPI)*, vol. 11, no. 3, p. 71, 2023.
[28] J. Park and S. Lee, "Human–AI co-learning: Ethical challenges and opportunities," *AI Ethics (Springer)*, vol. 5, pp. 211–226, 2023.
[29] A. Ghosh, "Multimodal prompting and interpretability in LLMs," *Applied Sciences (MDPI)*, vol. 13, no. 10, p. 5622, 2023.
[30] M. Pereira, "Academic integrity and AI-assisted writing: Policy perspectives," *Education and Artificial Intelligence (Elsevier)*, vol. 2, no. 1, pp. 55–68, 2024.
[31] W. Li and X. Zhou, "Prompt design strategies for adaptive learning environments," *SpringerOpen Journal of Intelligent Systems*, vol. 12, no. 3, pp. 221–238, 2024.
[32] D. Chien, "Prompt taxonomy and language model interpretability," *Information (MDPI)*, vol. 16, no. 1, p. 95, 2025.
[33] A. Turing, *The Philosophy of Machine Intelligence: Collected Essays*, Oxford University Press, UK, 2022.

参考文:

[1] S. Russell, P. Norvig, 《人工智能：一种现代方法》（第4版），Pearson，美国，2021年。

[2] B. Marr, 《人工智能的实践应用》，Wiley，英国，2022年。

[3] OpenAI, 《GPT-4 技术报告》，2023年。[在线]可获取：https://cdn.openai.com/papers/gpt-4.pdf

[4] 欧洲委员会, 《可信人工智能伦理指南》，欧盟出版局，2021年。

[5] F. A. Batarseh, E. A. Latif, 《人工智能保障：迈向可信人工智能》，Elsevier，阿联酋，2021年。

[6] A. Bozkurt, "告诉我你的提示词，我将使其成为现实：OpenAI对提示工程艺术的反思", 《Open Praxis》，第16卷，第2期，第661–677页，2023年。

[7] IBM, "什么是人工智能幻觉？", IBM Think, 2023年。[在线]可获取：https://www.ibm.com/think/topics/ai-hallucinations

[8] Y. Zhang, Y. Li, L. Cui, “人工智能海洋中的塞壬之歌：大型语言模型幻觉问题综述”, arXiv 预印本，2023年。[在线] 可获取：https://arxiv.org/abs/2309.01219

[9] R. Hernández-Sampieri, C. Fernández-Collado, P. Baptista, 《研究方法论》（第7版）, McGraw-Hill, 墨西哥，2022年。

[10] C. Lo 等, “提示工程与生成式人工智能工具在大学教育中的影响”, 《Education Sciences》, 第15卷, 第2期, 第199页, 2024年。

[11] A. Min 等, “释放提示工程在大型语言模型中的潜力”, 《Patterns》（Cell Press）, 第6卷, 第1期, 第100108页, 2025年。

[12] H. Sarı 等, “通过提示工程实践提升高等教育中的人工智能素养”, 《Computers and Education: Artificial Intelligence》, 第7卷, 第100222页, 2024年。

[13] T. I. Tonmoy, S. M. M. Zaman, V. Rawte, “大型语言模型幻觉缓解技术的综合综述”, arXiv 预印本，2024年。[在线] 可获取：https://arxiv.org/abs/2401.01313

[14] I. Goodfellow, Y. Bengio, A. Courville, 《深度学习》, MIT Press, 美国，2016年。

[15] P. Domingos, 《主算法：终极学习机器如何重塑世界》, Basic Books, 美国，2015年。

[16] L. Weng, “理解并缓解大型语言模型中的幻觉问题”, Towards Data Science, 2024年。[在线] 可获取：https://lilianweng.github.io/posts/2024-07-07-hallucination

[17] J. Nouri, “高等教育中的人工智能伦理与责任：为人机协作做准备”, 《Computers & Education》, 第193卷, 第104656页，2022年。

[18] M. Rahman, “降低教育领域中基于 Transformer 模型的幻觉率”, 《SpringerOpen Journal of AI Research and Applications》, 第3卷, 第45–58页，2022年。

[19] M. Foucault, 《权力 / 知识：1972–1977 年访谈与论文选集》, Pantheon Books, 纽约，1980 年。

[20] R. Al-Jarf, “数字素养与语言教育中提示工程的兴起”, 《Education and Information Technologies》, 第29卷, 第2期, 第453–469页，2023年。

[21] C. Lai, Y. He, “提示驱动学习与数据导向教学法：数字教育的新路径”, 《Learning Analytics》（Elsevier）, 第15卷, 第1期, 第1–12页，2024年。

[22] E. Crawley, J. Malmqvist, D. Brodeur, 《重新思考工程教育：CDIO 方法》（第3版）, Springer, Cham, 2023年。

[23] J. Aldana G., “基于 CDIO 的高等教育创新：融合人工智能、创造力与反思性学习”, 金迪奥大学研究报告，2024年。

[24] H. Qiu 等, “生成式人工智能系统中的偏见控制与透明性”, 《IEEE Access》, 第11卷, 第22511–22523页，2023年。

[25] J. Cao, “人工智能的认识论：从数据到意义”, 《AI & Society》, 第39卷, 第4期, 第1050–1064页，2024年。

[26] M. Haque, “评估大型语言模型在工程教育中的教学潜力”, 《IEEE Transactions on Education》, 第67卷, 第2期, 第134–145页，2024年。

[27] P. Kaur, “算法透明性与人机责任机制”, 《Informatics》（MDPI）, 第11卷, 第3期, 第71页，2023年。

[28] J. Park, S. Lee, “人机协同学习：伦理挑战与机遇”, 《AI Ethics》（Springer）, 第5卷, 第211–226页，2023年。

[29] A. Ghosh, “大型语言模型中的多模态提示与可解释性”, 《Applied Sciences》（MDPI）, 第13卷, 第10期, 第5622页，2023年。

[30] M. Pereira, “学术诚信与人工智能辅助写作：政策视角”, 《Education and Artificial

Intelligence》（Elsevier），第2卷，第1期，第55–68页，2024年。

[31] W. Li，X. Zhou，“自适应学习环境中的提示设计策略”，《SpringerOpen Journal of Intelligent Systems》，第12卷，第3期，第221–238页，2024年。

[32] D. Chien，“提示词分类法与语言模型可解释性”，《Information》（MDPI），第16卷，第1期，第95页，2025年。

[33] A. Turing，《机器智能哲学：论文选集》，Oxford University Press，英国，2022年。

Manuscript Information
Word count: 6,085 words (excluding references).

Peer-Review Record
Fast-track status: Not fast-tracked.
First-round reviews received: 3 reports.
Revision cycles completed: 3 rounds.
Final version submitted: January 2, 2026