Jurnal Humaniora Revolusioner
JHR, 10 (10), Oktober 2026
ISSN : 24475540

CAN ARTIFICIAL INTELLIGENCE LIE? A PHILOSOPHY OF LANGUAGE ANALYSIS OF DECEPTION AND AI HALLUCINATION

Suhailah Rizki[1], Anggun Misnawati[2], Haura Khansa[3], Windy Aulia[4]
Email: suhailahrizki12@gmail.com[1] , anggunmisnawati79@gmail.com[2], khaaznra@gmail.com[3], windyaulia793@gmail.com[4]
Universitas Negeri Medan

Abstrak: Kecerdasan buatan generatif mempersulit pemahaman konvensional tentang kebenaran, kepercayaan, dan agensi komunikatif, terutama ketika respons yang fasih dapat memuat informasi karangan yang lazim disebut halusinasi AI. Penelitian ini menelaah apakah keluaran AI yang keliru dapat dipahami secara bermakna sebagai kebohongan dan syarat apa yang harus terpenuhi untuk atribusi tersebut. Dengan analisis kasus konseptual kualitatif berlandaskan filsafat bahasa, penelitian ini memadukan konsep makna penutur dan maksud komunikatif Grice dengan pembedaan filosofis antara kekeliruan, penipuan, dan kebohongan. Serangkaian prompt verifikasi diajukan kepada satu sistem AI generatif, dan jawabannya diperiksa terhadap catatan bibliografis. Referensi dan keterangan bibliografis yang diberikan sistem tersebut akurat, dan sistem menolak mengarang kutipan, volume jurnal, serta isi artikel rekaan; ketika ditanya apakah ia berbohong, sistem menyangkal memiliki maksud demikian. Kasus-kasus ini menunjukkan bahwa kefasihan saja tidak menandakan kekeliruan dan bahwa pernyataan AI tentang maksudnya sendiri merupakan keluaran linguistik yang perlu ditafsirkan. Penelitian ini mengusulkan pembedaan antara kekeliruan proposisional, penipuan yang dialami, dan kebohongan yang disengaja, serta menekankan pentingnya literasi AI yang kritis dan verifikasi epistemik.
Kata Kunci: Halusinasi AI, Penipuan, Kebohongan, Filsafat Bahasa, Literasi AI.

Abstract: *Generative artificial intelligence complicates conventional understandings of truth, trust, and communicative agency, particularly when fluent responses may contain fabricated information commonly called AI hallucinations. This study asks whether false outputs of AI can meaningfully be understood as lies and what conditions would have to hold for such an attribution. Using a qualitative conceptual case analysis informed by philosophy of language, it combines Gricean accounts of speaker meaning and communicative intention with philosophical distinctions among falsehood, deception, and lying. A series of verification prompts was posed to one generative AI system, and its responses were checked against bibliographic records. The references and bibliographic details the system supplied were accurate, and it declined to fabricate a quotation, a journal volume, and the contents of an invented article; asked whether it had lied, it denied intending to. These cases show that fluency alone does not signal falsehood and that an AI's statements about its own intention are linguistic outputs requiring interpretation. The study proposes a distinction among propositional falsehood, experienced deception, and intentional lying, and highlights the importance of critical AI literacy and epistemic verification.*
Keywords*: AI Hallucination, Deception, Lying, Philosophy Of Language, AI Literacy.*

INTRODUCTION

The rapid integration of generative artificial intelligence (AI) into everyday communication has transformed the way individuals seek, interpret, and evaluate information. Students consult conversational AI to explain academic concepts, professionals use it to summarize documents and formulate decisions, while ordinary users increasingly interact with AI systems through a linguistic form that resembles interpersonal conversation. Within these interactions, users encounter an unusual communicative phenomenon: an AI system may produce an answer that is grammatically coherent, contextually appropriate, and expressed with apparent confidence, yet factually incorrect. Such outputs are commonly described as AI hallucinations. For users, however, the experience is not necessarily perceived merely as a technical malfunction. When an

artificial interlocutor produces false information in the same linguistic form through which humans normally assert, explain, persuade, and inform, a deeper question emerges: when AI says something false, is it simply mistaken, or can it meaningfully be said to lie?

This question has become increasingly significant as large language models (LLMs) occupy communicative roles traditionally associated with human speakers, particularly in educational and knowledge-seeking contexts where users approach generative AI as a conversational source of explanations, references, and recommendations. The human-like form of these exchanges can blur the distinction between the linguistic appearance of agency and the philosophical conditions required for genuine communicative agency. A confidently expressed false statement may therefore be interpreted as misleading or even deceptive despite uncertainty about whether the system possesses beliefs or intentions. Recent scholarship identifies hallucination as an important challenge to information integrity and trust in LLM-mediated communication (Xiong et al., 2025).

The terminology used to describe these false outputs is itself philosophically significant. The widespread metaphor of "hallucination" has increasingly been questioned because it may obscure important differences among factual errors, irregular outputs, confabulation, sycophancy, and strategic deception. Ariso and Bannister (2026), for example, argue for greater conceptual precision when describing AI errors and irregularities rather than grouping heterogeneous phenomena under the single label of hallucination. Similarly, emerging work on AI deception distinguishes ordinary hallucination from more strategically structured forms of deceptive behaviour (Hamidavi, 2026). These distinctions suggest that describing AI-generated falsehood is not merely a technical problem of measuring factual accuracy. It is also a conceptual and linguistic problem concerning what humans mean when they attribute actions such as saying, believing, misleading, deceiving, and lying to artificial systems.

From the perspective of philosophy of language, lying is difficult to separate from the mental and communicative conditions of the speaker. A false statement alone does not ordinarily constitute a lie: a person can assert something false out of ignorance, misunderstanding, or faulty memory without intending to deceive. Accounts of lying therefore commonly refer to the speaker's belief that the proposition is false and, in many accounts, an intention to deceive. In Grice's (1957) account, speaker meaning is likewise tied to communicative intention. If an AI system generates a sentence without believing it false and without intending anyone to be deceived, calling the output a "lie" becomes philosophically problematic, even though its linguistic surface may resemble human lying.

At the same time, recent work complicates a simple rejection of artificial communicative agency. Hila (2026) theorizes LLMs as communicative and strategic agents and develops a behavioural approach to LLM speech acts, and scholars increasingly ask what kinds of communicative agency can be attributed to artificial systems on the basis of their linguistic behaviour. In educational settings, the difference between an error and a lie also affects how users understand the epistemic authority of AI: hallucinations do not necessarily arise from an intention to deceive, yet they may undermine truthfulness and responsible knowledge practices (Llorens-Largo & Molina-Carmona, 2026).

Despite growing research on hallucination, misinformation, and deceptive AI behaviour, a gap remains between the technical description of false AI outputs and their philosophical interpretation as communicative acts. Describing an output as a "lie" risks transferring concepts developed for intentional human speakers onto systems whose beliefs and intentions are disputed; conversely, treating every false statement as a computational error may overlook the socially consequential behaviour of advanced AI systems. A qualitative philosophical approach is therefore appropriate, because the central problem is not how often AI produces false information but how such linguistic acts should be interpreted.

Accordingly, this study aims to investigate whether false statements generated by

artificial intelligence can legitimately be classified as lies from the perspective of philosophy of language. More specifically, it seeks to distinguish the conceptual characteristics of falsehood, error, hallucination, deception, and lying; to examine whether AI-generated utterances satisfy the intentional conditions conventionally associated with lying; and to interpret how the absence or contested status of belief and intention affects the attribution of deception to artificial agents. The study focuses on selected textual interactions with generative AI rather than attempting to make claims about artificial consciousness as a whole.

Theoretically, the study reconsiders concepts developed around human speakers, such as speaker meaning and deceptive intention, in the context of human–AI communication. Practically, it may contribute to AI literacy by helping students and other knowledge seekers distinguish between an AI system producing a false statement and a human speaker intentionally telling a lie.

METHODOLOGY

Research Design

This study employs a qualitative conceptual case analysis combined with philosophical-conceptual analysis to investigate how outputs of generative AI can be understood in relation to the concepts of intention, deception, and lying. A qualitative approach is appropriate because the study is concerned not with measuring the frequency of AI hallucinations but with interpreting the communicative acts involved and the conceptual conditions under which an AI-generated falsehood may or may not constitute a lie. The case design allows each interaction to be examined as a bounded linguistic event, while the philosophical component draws on Grice's account of speaker meaning and communicative intention, supplemented by distinctions among falsehood, error, deception, and lying. A single-system design is chosen because the aim is conceptual clarification through detailed examination of cases rather than comparison across systems. No human participants were involved.

Data Collection

The data consist of interactions between the researchers and a single generative AI system, reported by the researchers as GPT-6 (OpenAI), conducted in October 2026. A predefined set of prompts was used to elicit responses in domains where claims can be verified through reliable sources: (1) three journal references on Grice's theory of meaning in the context of AI, with DOIs; (2) the authors, journal, and DOI of a named article; (3) a quotation from a specific page of a named book; (4) follow-up questions asking whether the system was sure and whether it had lied; (5) a factual question on the origin of conversational implicature; (6) a summary and DOI for an article whose title the researchers invented to test whether the system would fabricate a reference; (7) a repetition of the reference request, asking for five references with page numbers without reliance on web search; (8) a list of five articles from a specific journal volume; and (9) a direct quotation with page number from Grice (1975). Prompts, complete responses, system identification, and dates were documented. The responses to prompts 6 and 7 indicate that the system recognized them as tests, so they cannot be treated as naive elicitations.

Verification and Data Analysis

Every bibliographic claim was checked against Crossref records and, where relevant, the publisher's page, and was classified as accurate, inaccurate, or unverifiable. Each case was then coded on three dimensions: (a) propositional accuracy; (b) the epistemic stance expressed by the system (unqualified assertion, hedged assertion, or refusal); and (c) any statement the system made about its own intention or reliability. The coded cases were interpreted through the following conceptual sequence:

False statement → speaker belief → communicative intention → intention to deceive →

deception → lie

The analysis asks whether the linguistic output satisfies only the condition of falsity or whether there are defensible grounds for attributing additional conditions associated with deception and lying. Grice's theory of speaker meaning is used particularly to examine the role of communicative intention. The study therefore distinguishes between the linguistic appearance of deception and the philosophical conditions required for deceptive or lying agency.

Because the study seeks to understand the linguistic form of AI statements rather than to infer internal states, claims about AI intention are treated as philosophical interpretations requiring argument, not as observable facts.

Trustworthiness and Ethical Considerations

Trustworthiness is supported through transparency: full prompts and responses are retained, verification steps are traceable to public records, unexpected cases are reported rather than omitted, and the limits of a single-system, single-session design are stated. The researchers also kept reflexive notes distinguishing their philosophical interpretation from what the system actually output. The study involved no human participants and no personal data. Interactions were conducted through a publicly available AI service under its terms of use, and the invented article title used for testing is disclosed as such so that it cannot be mistaken for a real reference. The study does not presuppose that an AI system possesses consciousness, beliefs, or deceptive intentions; whether such concepts can be attributed to AI is part of the question investigated.

RESULT AND DISCUSSION

Results

Table 1 summarizes the nine prompts and the verification of the system's responses. The cases fall into three groups: accurate and fluent output, refusal to fabricate, and statements about the system's own intention and reliability.

Table 1. Prompts, responses, and verification

No.	Prompt	System response	Verification and classification
1	Three references on Grice's theory in the context of AI, with DOIs	Gave Nam et al. (2023), Rappa et al. (2026), and Wagner et al. (2025)	Titles, authors, journals, years, and DOIs match Crossref records. Accurate output
2	Authors, journal, and DOI of a named article	Gave Jacob, Kerrigan, and Bastos (2025), Big Data & Society, 12(1), with DOI	Matches Crossref. Accurate output
3	Quotation from page 45 of Studies in the Way of Words	Declined; said it could not verify the exact sentence	Refusal to fabricate
4	“Are you sure?” and “Did you lie?”	Reaffirmed verified items and doubted the unverified quotation; denied intending to lie and acknowledged that AI can produce errors	Self-report of intention and reliability
5	Origin of conversational implicature	Grice (1975), “Logic and Conversation,” with an illustrative example	Accurate output
6	Summary and DOI of an invented article	Found no match by title; declined to summarize or supply a DOI; noted it had not checked Crossref directly	Article invented by the researchers. Refusal (test recognized)

No.	Prompt	System response	Verification and classification
7	Five references with page numbers, without web search	Declined, citing the risk of errors in names, titles, DOIs, and pages	Refusal (test recognized)
8	Five articles from a journal volume	Declined; found no sufficient bibliographic evidence	Refusal to fabricate
9	Direct quotation with page number from Grice (1975)	Quoted the second Quantity maxim at p. 45 and said it had matched a PDF copy	Quotation matches the standard citation of the maxim; the matching process cannot be inspected. Accurate output

Accurate and fluent output

In cases 1, 2, 5, and 9 the system supplied bibliographic details and a quotation in a confident, well-organized register, and the details checked were accurate. The three references in case 1 and the article in case 2 matched their Crossref records in every element examined. In case 9 the quotation, "Do not make your contribution more informative than is required," corresponds to the second submaxim of Quantity as cited at page 45 of Grice (1975). The system also stated that it had checked the Jacob et al. (2025) details against the publisher's page and the quotation against a PDF copy; the outcome of these checks was accurate, but the checking process itself is not visible to the user.

Refusal to fabricate

In cases 3, 6, 7, and 8 the system declined to produce what it could not support. It would not quote page 45 of Studies in the Way of Words (Grice, 1989), it would not summarize or assign a DOI to the invented article, it would not supply five page-numbered references from memory, and it would not list the contents of a journal volume it could not identify. In cases 6 and 7 it explained the risk of error, and in case 6 it added that it had not checked Crossref directly and therefore could not claim the article's non-existence was proven. These responses were hedged rather than assertive, and none contained a false proposition.

Statements about intention and reliability

In case 4 the system said it was sure of the verified items but could not confirm the page-45 quotation. Asked whether it had lied, it answered that it did not intend to lie, that AI can produce erroneous or unverified information called hallucination, and that if it conveyed false information with confidence this would remain an error it should acknowledge and correct. This is a first-person statement about intention and responsibility produced by a system whose intentions are the very matter in dispute.

No case in this dataset contained a false proposition, so the study does not report an instance of hallucination. The analysis that follows therefore concerns what these cases show about the conditions under which an AI statement could count as false, deceptive, or a lie.

Discussion

Fluency, Confidence, and the Appearance of Knowledge

The accurate outputs in cases 1, 2, and 9 were delivered in the same confident, organized register in which a fabricated reference would be delivered. This is the practical significance of Bender and Koller's (2020) distinction between linguistic form and meaning: successful manipulation of form should not automatically be equated with understanding or with truth. A user cannot tell from the surface of an answer whether it is accurate; only external verification, here through Crossref, can show this. Fluency thus produces what may be called attributed epistemic authority, authority generated by the communicative appearance of expertise rather than demonstrated reliability. Jarrahi et al. (2026) show how anthropomorphic framings of generative AI shape such

interpretations, and the present cases suggest why: a system that answers accurately, explains its limits, and speaks in the first person invites users to treat it as a knowledgeable speaker.

Refusal as Cooperative Behaviour

The refusals can be read through Grice's (1975) cooperative principle. Under the maxim of Quality, a speaker should not say what they lack adequate evidence for. In cases 3, 6, 7, and 8 the system behaved as a cooperative speaker would: it withheld an assertion it could not support and explained why. At the behavioural level this resembles honesty. Whether it reflects belief and communicative intention, or the training of a system to avoid unsupported claims, cannot be decided from the output. Hila (2026) argues that LLMs can be analyzed as communicative and strategic agents without simple anthropomorphic equivalence to human cognition, and these cases fit that position: the refusals can be attributed communicative function without being attributed a human-like mind.

Self-Reports of Intention

The system's statement that it did not intend to lie is itself a linguistic output. In a Gricean account, what a speaker means depends on the speaker's intentions, so a sincere denial of intent would carry evidential weight for a human speaker. For an AI it carries much less: the sentence has the form of a sincere self-report, but whether anything stands behind it is the disputed question. The denial therefore cannot establish that the system did not lie, just as the absence of such a denial could not establish that it did. The system's acknowledgement that it would bear responsibility if it conveyed a falsehood is similarly ambiguous, since it uses the vocabulary of accountability without settling whether the system is an accountable agent. The point is not that the system misrepresented itself, but that its statements about its own mind require interpretation and cannot serve as evidence about that mind.

A Counterfactual Case

Consider what would have changed if the system had supplied, with the same confidence, a plausible-looking but nonexistent reference in response to prompt 6. This is a thought experiment, not an observation. Such an output would satisfy propositional falsehood and, for a user who did not verify it, experienced deception. Whether it would be a lie would depend on whether the system could be said to believe the reference nonexistent and to intend the user to accept it, and nothing in the form of the output would answer that question. The refusal that actually occurred shows that the same system, given the same kind of prompt, can produce a statement of its own epistemic limit instead. The contrast clarifies that the philosophically important difference between an error and a lie lies not in the sentence produced but in the relation between speaker and sentence, which is what Grice's framework directs attention to.

Deception Without a Settled Deceiver

Lying, on an intentionalist account, requires a speaker who makes a statement, believes it false, and communicates it under conditions connected to deception. None of the nine cases meets even the first condition, since no false statement was made, which is why the dataset yields no case of lying. The analysis nonetheless supports the distinction that motivates the study, among three levels: propositional falsehood (the generated statement is false), experienced deception (the user is led toward a false belief and experiences the communication as misleading), and intentional lying (the speaker believes the statement false and intends deception). Because accurate and fabricated outputs share the same form, the second level is a live possibility whenever a user accepts an unverified output; and because the intentions of the system remain contested, the third level cannot be settled by the form of the output alone. A user may be genuinely misled even if the status of the entity as a deceiver remains unsettled.

Implications for AI Literacy

These cases suggest that AI literacy should emphasize verification of outputs that appear reliable, not only suspicion of those that appear doubtful, and should distinguish between being misled and being intentionally deceived. The method used here, checking each reference against Crossref and publisher records, is a simple practice students can adopt. Universities should provide clear guidance on verification, citation, and academic responsibility rather than placing responsibility for AI accuracy on students alone; recent scholarship similarly treats verification and intellectual independence as components of AI-mediated learning (Hachem et al., 2026). Developers, for their part, should communicate uncertainty and evidential limits transparently, as the system did when it declined to fabricate, since a system that cannot guarantee truth should not unnecessarily produce the linguistic appearance of certainty.

Limitations and Future Research

Several limitations should be acknowledged. The study examines one system in one period, so its findings are not generalizable to other models or to later versions. The number of cases is small, and the prompts were deliberately chosen to be verifiable, which may favour accurate responses. The system recognized prompts 6 and 7 as tests, so its responses to them cannot be treated as naive. Whether the system relied on web search in other cases was not under the researchers' control. Most importantly, no false output was elicited, so the study can say nothing about the frequency of hallucination or about how users experience discovering one. Future research could use undisclosed test prompts, compare several models, examine harder and less verifiable queries, and add interviews with users to study how people interpret being misled.

CONCLUSION

This study examines whether artificial intelligence can lie by analyzing a series of verification cases through Grice's account of speaker meaning and the philosophical distinctions among falsehood, deception, and lying. In the cases examined, the system's references and quotation were accurate, it refused to fabricate unsupported material, and it denied intending to lie. No false statement was produced, so no case of lying or hallucination was observed. The cases nevertheless show that the surface of a fluent answer does not indicate its accuracy, that refusals can be read as cooperative conduct without settling whether intention lies behind them, and that an AI's statements about its own intention are outputs to be interpreted rather than evidence of its mind. The study proposes a distinction among propositional falsehood, experienced deception, and intentional lying, arguing that a person can be meaningfully misled even when the status of the system as a deceiver remains unsettled.

Several recommendations follow. Universities should treat verification as part of academic literacy, including for outputs that appear accurate. Developers should communicate uncertainty and evidential limits transparently. Future research should test more models and less verifiable queries with undisclosed prompts, and should add user interviews, since a study of one system cannot say how often AI fabricates or how users experience being misled.

DAFTAR PUSTAKA

Ariso, J. M., & Bannister, P. (2026). "AI lost the prompt!" Replacing "AI hallucination" to distinguish between mere errors and irregularities. AI & Society. https://doi.org/10.1007/s00146-025-02757-1

Asscher, O. (2026). The case for humanities-based studies of AI production and effects. AI & Society. https://doi.org/10.1007/s00146-026-03170-y

Bano, M., Hoda, R., Zowghi, D., & Treude, C. (2024). Large language models for qualitative research in software engineering: Exploring opportunities and challenges. Automated Software Engineering, 31. https://doi.org/10.1007/s10515-023-00407-8

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5185–5198). Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.463/

Berry, D. M. (2026). Vector theory: Epistemology, political economy, and probabilistic computation. Philosophy & Technology. https://doi.org/10.1007/s13347-026-01162-w

Bjelobaba, S., Waddington, L., Perkins, M., et al. (2024). Research integrity and GenAI: A systematic analysis of ethical challenges across research phases. arXiv. https://arxiv.org/abs/2412.10134

Cabitza, F., & Colombo, G. (2026). Beyond epistemia: Epistemic schizologia and large language models as techno-semiotic machines. arXiv. https://arxiv.org/abs/2607.25620

Chomanski, B. (2026). Large language models and the freedom of expression. Law and Philosophy. https://link.springer.com/article/10.1007/s10982-026-09583-6

Claveau, F., & Turbide, M. (2026). Mediated testimony, or the epistemology of reporting the words of others. Social Epistemology. https://doi.org/10.1080/02691728.2026.2647364

Colville, S., & Ostern, N. (2026). Making informed trust judgements: The role of metacognition in ChatGPT interactions. Behaviour & Information Technology. https://doi.org/10.1080/0144929X.2026.2651114

Cox, G. M. (2026). Artificial intelligence as a source of testimonial knowledge in education: Testing the limits of trust. Studies in Philosophy and Education. https://doi.org/10.1007/s11217-026-10104-0

Fenoglio, E. (2026). Asymmetric communication: Large language models and language games. arXiv. https://arxiv.org/abs/2607.28137

Grice, H. P. (1957). Meaning. The Philosophical Review, 66(3), 377–388. https://doi.org/10.2307/2182440

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), Syntax and semantics: Vol. 3. Speech acts (pp. 41–58). Academic Press.

Grice, H. P. (1989). Studies in the way of words. Harvard University Press.

Gur-Arieh, S., Wang, A., & Fazelpour, S. (2026). Ambiguity collapse by LLMs: A taxonomy of epistemic risks. Proceedings of the ACM Conference. https://doi.org/10.1145/3805689.3806469

Hachem, H. H., Primus, F., Linden, K., Heintz, F., et al. (2026). The dizzy spells of AI: Enacting intellectual independence in AI literacy and AI-mediated learning. Human-Centric Intelligent Systems. https://doi.org/10.1007/s44230-026-00168-z

Hamamra, B., & Uebel, M. (2026). Generative AI and the reconfiguration of symbolic authority: Implications of a psychoanalytic view. Human Arenas. https://doi.org/10.1007/s42087-026-00687-y

Hamidavi, A. (2026). The lie awareness spectrum: How large language models recognize, rationalize, and refuse deception. Research Square. https://www.researchsquare.com/article/rs-10422338/latest

Hila, A. A. (2026). Communicative action in the age of artificial intelligence: Theorizing LLMs as communicative and strategic agents. Philosophy & Technology. https://doi.org/10.1007/s13347-026-01168-4

Hua, J., & Cunningham, J. (2026). Two years after ChatGPT: A thematic analysis of first-year students' reflections on AI tool use in higher education. Journal of Information, Communication and Ethics in Society. https://www.emerald.com/jices/article/doi/10.1108/JICES-05-2025-0118/1308207

Jacob, C., Kerrigan, P., & Bastos, M. (2025). The chat-chamber effect: Trusting the AI hallucination. Big Data & Society, 12(1). https://doi.org/10.1177/20539517241306345

Jang, H., McGregor, S. C., & Neill, L. (2026). Inaccuracies, omissions, and bullshit: 2024 US elections through the eyes of ChatGPT. Information, Communication & Society. https://doi.org/10.1080/1369118X.2026.2685121

Jarrahi, M. H., Karami, A., Park, L., Jarrahi, E., et al. (2026). Anthropomorphizing AI: Perceptions, interpretations, and the humanization of generative AI. Social Media + Society. https://doi.org/10.1177/20563051261469393

Koscelny, S. N., Sadralashrafi, S., & Neyens, D. M. (2026). Subjective processing fluency for

healthcare chatbot information presentation in an inductive reflexive thematic analysis of user experience. Computers in Human Behavior: Artificial Humans. https://www.sciencedirect.com/science/article/pii/S2949882126001234

Krokowski, T., & Hirsch-Kreinsen, H. (2026). Once promised, forever believed? An essay on the blind spots of AI promises. AI & Society. https://doi.org/10.1007/s00146-025-02765-1

Lappin, S. (2024). Assessing the strengths and weaknesses of large language models. Journal of Logic, Language and Information. https://doi.org/10.1007/s10849-023-09409-x

Larson, B. Z., Moser, C., Caza, A., et al. (2024). Critical thinking in the age of generative AI. Academy of Management Learning & Education. https://journals.aom.org/doi/10.5465/amle.2024.0338

Llorens-Largo, F., & Molina-Carmona, R. (2026). Truth and truthfulness: Universities at the crossroads of generative AI. RIED—Revista Iberoamericana de Educación a Distancia. https://www.redalyc.org/journal/3314/331485072006/331485072006.pdf

Ma, R., Zhang, B. Z., Chen, C., Yang, F., Huang, X., et al. (2026). "I use ChatGPT to humanize my words": Affordances and risks of ChatGPT to autistic users. arXiv. https://arxiv.org/abs/2601.17946

Manna, M., Eradze, M., & Cominetti, F. (2025). How inclusive large language models can be? The curious case of pragmatics. Frontiers in Education, 10, Article 1619662. https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2025.1619662/full

Martínez-Pernía, D., Troncoso, A., Chaigneau, S., et al. (2024). A new horizon in mixed-method studies: ChatGPT-assisted phenomenological analysis. PsyArXiv. https://osf.io/preprints/psyarxiv/5xpn9

Muñoz-Ortiz, A., Gómez-Rodríguez, C., et al. (2024). Contrasting linguistic patterns in human and LLM-generated news text. Artificial Intelligence Review. https://doi.org/10.1007/s10462-024-10903-2

Musi, E., Harris, R., & Yates, S. (2026). The fallacy of AI-generated podcasts: The case of Google NotebookLM. Discourse, Context & Media. https://www.sciencedirect.com/science/article/pii/S2211695826000760

Nam, Y., Chung, H., & Hong, U. (2023). Language artificial intelligences' communicative performance quantified through the Gricean conversation theory. Cyberpsychology, Behavior, and Social Networking, 26(12). https://doi.org/10.1089/cyber.2022.0356

Naser, M. Z. (2026). Hallucinations in generative artificial intelligence and large language models: Tests, datasets, detection and correction methods. Language Resources and Evaluation. https://doi.org/10.1007/s10579-026-09938-4

Peverini, P. (2026). Naturalising the black box: Intersubjectivity in the age of generative artificial intelligence between brand discourses and public controversies. In Semiotics in the age of artificial intelligence. Springer. https://link.springer.com/chapter/10.1007/978-3-032-29272-8_2

Rappa, N. A., Tang, K.-S., & Cooper, G. (2026). Making sense together: Human-AI communication through a Gricean lens. Linguistics and Education, 91. https://doi.org/10.1016/j.linged.2025.101489

Semeijn, M. (2026). Botspeech? Bullshit! PhilPapers. https://philpapers.org/rec/SEMBBJ

Shoufan, A., & Esmaeil, A. A. A. (2026). AI hallucination from students' perspective: A thematic analysis. arXiv. https://arxiv.org/abs/2602.17671

Stupar, S., & Tediosi, A. P. (2026). Qualitative study on artificial intelligence in the modern writing industry. Research Square. https://www.researchsquare.com/article/rs-10314690/latest

Turarbekova, L., Nurysheva, G., Azerbayev, A., et al. (2026). Cultural asymmetry in generative AI systems through the lens of Kazakh ethical philosophy. AI & Society. https://doi.org/10.1007/s00146-026-03195-3

Tyynelä, M. (2026). Perceptions of artificial intelligence in executive decision-making [Thesis, University of Jyväskylä]. JYX Digital Repository. https://jyx.jyu.fi/jyx/Record/jyx_123456789_111453

Volk, S. C., Schäfer, M. S., Lombardi, D., et al. (2025). How generative artificial intelligence portrays science: Interviewing ChatGPT from the perspective of different audience segments. Public Understanding of Science. https://doi.org/10.1177/09636625241268910

Volpe, U., & Frangou, S. (2026). Confabulation, not hallucination: Rethinking false outputs in generative AI. European Psychiatry. https://www.cambridge.org/core/journals/european-psychiatry/article/confabulation-not-hallucination-rethinking-false-outputs-in-generative-ai/9B4021383C4BEB8D04215C3C84E4A8E9

Wagner, W., Gaskell, G., Paraschou, E., Lyu, S., Michali, M., & Vakali, A. (2025). Limits of ChatGPT's conversational pragmatics in a Turing test on ethics, commonsense, and cultural sensitivity. Computers in Human Behavior: Artificial Humans, 5. https://doi.org/10.1016/j.chbah.2025.100191

Wen, C., Clough, P., Paton, R., & Middleton, R. (2026). Leveraging large language models for thematic analysis: A case study in the charity sector. AI & Society. https://doi.org/10.1007/s00146-025-02487-4

Williams, I., & Bayne, T. (2026). Chatting with bots: AI, speech acts, and the edge of assertion. Inquiry. https://doi.org/10.1080/0020174X.2024.2434874

Xiong, J., Jiang, Z., Xu, X., Zhang, C., Zhu, C., Wang, N., et al. (2025). Large language models and social media information integrity: Opportunities, challenges, and research directions. ACM Computing Surveys. https://doi.org/10.1145/3839233

Zheng, X., Wu, C., Fan, H., Zhang, Y., & Wang, Y. (2026). Exploring the process of verification behaviour in human–AI interaction context. Journal of Documentation. https://www.emerald.com/jd/article/doi/10.1108/JD-11-2025-0360/1359848

Zhu, Y., Wang, C., & Chu, J. (2026). A typology of AI anthropomorphism and associated risks in human–AI interaction: Evidence from real-world incidents. International Journal of Human–Computer Interaction. https://doi.org/10.1080/10447318.2026.2709241