

From Fabricated Citations to Unverified Journal Matches: Problems and Solutions in LLM Assisted Manuscript Generation

Indri Anjar Kartika Sari¹, Onno Widodo Purbo², Abi Julian Saputra Pratama³, Rahardian Dwi Saputra⁴

¹ Program Studi S1 Cyber Security, IDS Digital College, Surabaya, Indonesia; indri.cs2026@ids.ac.id

² Institut Teknologi Telkom Surabaya (ITTS), Surabaya, Indonesia; onno@itts.ac.id

³ Universitas Muhammadiyah Bengkulu, Bengkulu, Indonesia; abijulian@umb.ac.id

⁴ University of Muhammadiyah Malang, Malang, Indonesia; radiandwi2206@gmail.com

ARTICLE INFO

Keywords:

Keyword 1; Citation Hallucination
Keyword 2; Multi Agent Systems
Keyword 3; Design Science Research
Keyword 4; Retrieval Generation

Article history:

Received 2026-09-24

Revised 2026-09-25

Accepted 2026-12-01

ABSTRACT

The growing use of large language models (LLMs) in scholarly writing introduces risks of fabricated references, unsupported claims, and unreliable journal recommendations. Evidence gating addresses these risks by allowing outputs only when verifiable conditions are met and requiring explicit abstention otherwise. This study evaluates evidence gating through a design science research approach, examining scopus-jurnal, an open-source Python multi-agent pipeline that retrieves literature, drafts IMRaD sections, validates citations, reviews manuscripts, and recommends journals. Five evaluation episodes combined traceability audits and white-box experiments covering 46 in-text citations, 122 verification cases, 14 reviewer output formats, and 40 journal-matching probes. Results show that although the pipeline abstained when evidence was absent, weaknesses emerged between LLM outputs and deterministic code. Two of 17 specified phases lacked implementation, restricting the pipeline to drafting framing sections. Citation detection resolved only 31.6% of valid APA citations, while fabricated citations were sometimes accepted. Title matching accepted all 14 fabricated near-miss titles and rejected eight of nine legitimate short titles. Reviewer-score parsing failed on seven of ten plausible formats. A revised verification rule reduced false acceptance from 33.3% to 2.4% and false rejection from 43.1% to 0%. Journal-matching experiments showed that scope evidence without a score threshold enabled safe abstention, whereas introducing a threshold eliminated abstention. These findings emphasize tool-level evaluation and contribute an audit protocol, seven design principles, a reference architecture, eight integrity-preserving solutions, and five future research priorities.

This is an open access article under the [CC BY-NC-SA](https://creativecommons.org/licenses/by-nc-sa/4.0/) license.



Corresponding Author:

Indri Anjar Kartika Sari

Program Studi S1 Cyber Security, IDS Digital College, Surabaya, Indonesia; indri.cs2026@ids.ac.id

1. INTRODUCTION

Scholarly publishing is under strain. The volume of indexed articles has grown faster than the pool of editors and reviewers able to assess them (Bornmann & Mutz, 2015; Hanson et al., 2024). For researchers in systems that reward publication in Scopus- or Web of Science-indexed journals, the work of preparing a submission has also grown. They have to locate current literature, frame a defensible gap, write to the conventions of a target venue, choose a journal whose scope and standing fit the work, and assemble cover letters, highlights and declarations. Large language models (LLMs) have entered each of these steps.

Since late 2022, commentators, editors and researchers have documented both the promise and the hazards of generative AI in research writing (Dwivedi et al., 2023; Kasneci et al., 2023; van Dis et al., 2023). Publishers and integrity bodies have responded with policies that exclude AI systems from authorship and require disclosure of their use (COPE Council, 2023; Nature, 2023; Thorp, 2023). A distinct genre of software has grown around these steps: the AI submission assistant. Such tools combine some or all of four functions.

They search for literature, generate manuscript text, run pre-submission checks such as similarity, reviewer simulation and citation verification, and recommend journals. What sets them apart from a general-purpose chatbot is that they present assurance signals: a "93.7% match" to a journal, a similarity estimate "below 8%", a reviewer score of "95/100", or a citation labelled "verified". These signals are persuasive because they look like measurements. Research on automation bias shows that people over-rely on automated aids, especially when outputs are confident and the aid is usually right (Goddard et al., 2012; Parasuraman & Riley, 1997). Human-AI interaction guidelines therefore ask systems to make clear how well they can do what they claim to do (Amershi et al., 2019).

An assurance signal that does not measure what it claims to measure is more than a cosmetic defect. It can lead authors to submit to unsuitable venues and to include unverifiable claims. At scale, it can also feed the paper-mill and predatory-publishing dynamics that editors already struggle to contain (Grudniewicz et al., 2019; Kendall & Teixeira da Silva, 2024). The risks of LLM-generated scholarly content are well established at the level of the model. Fabricated and erroneous references are frequent in unaided LLM output (Bhattacharyya et al., 2023; Chelli et al., 2024; Walters & Wilder, 2023). Hallucination more generally is a structural property of current generation methods (Huang et al., 2025; Ji et al., 2023).

Generated abstracts can deceive expert readers (Else, 2023; Gao et al., 2023), and text-based AI detectors are unreliable (Weber-Wulff et al., 2023), which has prompted proposals to complement text analysis with evidence about how a text was produced (Sari et al., 2026). Much less is known at the level of the tool: the application code that wraps a model, orchestrates retrieval and verification, and turns intermediate results into the scores an author sees. Engineering choices at this layer decide whether a tool's outputs are grounded in evidence. They include retrieval sources, similarity functions, thresholds, parsers and fallback defaults. These choices are rarely documented, and they are almost never evaluated independently.

Algorithm-audit research has shown that the gap between what a system claims and what it does can be surfaced only by systematic external testing (Metaxa et al., 2021; Raji et al., 2020). This article addresses that gap through a design science evaluation of scopus-jurnal, an open-source Python multi-agent pipeline that covers the full "generator plus journal match" workflow. The pipeline audits a research code repository and retrieves literature from CrossRef, OpenAlex and Semantic Scholar. It builds a gap matrix, writes IMRaD sections with an LLM, validates citations against Crossref, runs a 15-dimension reviewer loop, recommends target journals and exports DOCX and LaTeX.

It is built on an explicit evidence-gating philosophy. Its governing prompt states that it must "never fabricate data, citations, or results", that it must "never claim Scopus/Q1 acceptance", and that missing evidence must be marked with an explicit [EVIDENCE_REQUIRED] tag. The pipeline is therefore a good test case for a question that matters beyond any single tool: when a submission assistant is designed to abstain rather than fabricate, do its gates actually hold? We pose three research questions: RQ1. To what extent does the implemented pipeline realise its evidence-gating specification, and where does it abstain, act or fail silently? RQ2. Under what realistic input conditions do the pipeline's evidence gates for citation extraction, citation verification and reviewer decisions fail, and in which direction (false acceptance or false rejection)? RQ3. Which design choices make journal matching and citation verification integrity-preserving, and do revised components measurably reduce the observed failures? The study makes four contributions. First, it provides a specification-to-implementation traceability audit of an evidence-gated pipeline.

The audit shows that abstention works as a safe failure mode, and it also locates silent failures. Second, it introduces the notion of gate leakage and distinguishes three forms: coverage, threshold and default leakage. Each is quantified with white-box experiments that execute the pipeline's own code. Third, it contributes and evaluates two revised components. One is a citation-verification rule that sharply reduces both false acceptance and false rejection. The other is a journal-matching design whose ablation shows that scope evidence and the absence of a score floor are both necessary for meaningful abstention. Fourth, it distils seven design principles, a reference architecture and eight concrete solutions for integrity-preserving submission assistants, sets out their benefits and a five-plan research agenda, and draws implications for authors, editors and research institutions.

These implications are especially relevant in contexts such as Indonesia, where publication targets in indexed journals are closely tied to career progression and research funding. The remainder of the article is organised as follows. Section 2 reviews related work and derives the integrity requirements used in the evaluation. Section 3 describes the design science method, the pipeline and the five evaluation episodes. Section 4 reports the results. Section 5 discusses theoretical and practical implications, design principles, proposed solutions and their benefits, limitations and a five-plan research agenda. Section 6 concludes.

2. METHODS

This study employed a qualitative library research approach using content analysis and comparative discourse analysis to examine and compare the educational thought of Syed Muhammad Naquib al-Attas and Ismail Raji al-Faruqi. The primary data were drawn from their seminal works, *Islam and Secularism* and *Islamization of Knowledge*, while secondary data consisted of relevant peer-reviewed journal articles and scholarly books on Islamic epistemology and the Islamization of knowledge. Data were collected through systematic identification, selection, and documentation of relevant textual passages and analyzed in three stages: reconstructing the epistemological foundations of each thinker, comparing their educational objectives, and interpreting the contemporary relevance of their ideas for Islamic education. The analysis focused on three predetermined dimensions: epistemology, educational objectives, and contemporary relevance, with particular attention to points of convergence, divergence, and potential integration. Data analysis involved reduction, thematic organization, comparative interpretation, and verification against primary texts and supporting literature to ensure interpretive accuracy. To enhance analytical rigor and minimize interpretive bias, the study established explicit comparative criteria, maintained traceability between interpretations and source texts, and presented convergences and divergences without presupposing conceptual compatibility. As a conceptual and

philosophical inquiry, the study aimed to develop an integrative framework for contemporary Islamic education rather than empirically test its implementation in educational institutions.

3. FINDINGS AND DISCUSSION

The evaluation shows that the pipeline is only partially aligned with its stated research-integrity specification. Of seventeen specified phases, ten are implemented as functional code, five are partial, and two data validation (Phase 3) and reproducible analysis (Phase 4) have no implementing code. Of fourteen slash commands, six are functional, six partial, and two `r` computation; it performs bibliographic retrieval from three APIs, DOI resolution through Crossref, and LLM-based writing and reviewing under anti-fabrication instructions.

The missing validation and analysis phases create a direct downstream limitation. The file `analysis/validated_results.md` is required by the results writer and the figure and table generators, yet no agent produces it. Consequently, the figure and table commands terminate with “no validated results found,” while the results writer receives an empty evidence block. In that situation, the writer emits [EVIDENCE_REQUIRED] rather than inventing results. This safe failure mode satisfies the intended evidence gate, but leaves the pipeline able to draft framing sections while unable to report empirical findings from the repository it audits. The `/scopus-check` function also remains only partially assured because it asks the LLM to nominate verifiable journals without using the available SCOPUS API_KEY or a retrieval-based score. In addition, the final audit’s “No HALLUCINATED citations” check is a permanent false alarm because the audit template itself contains the word “HALLUCINATED.”

Inline-citation testing further identifies a major integrity weakness. Of 38 legitimate citation instances, only 12 (31.6%) were detected and resolved to the correct DOI. Successful detection was limited to single-author parenthetical citations, parenthetical “et al.” citations, and year-suffixed citations. Two-author citations using an ampersand, narrative citations, semicolon-separated multiple citations, citations with page locators, surnames with lower-case particles, and citations introduced by signal phrases all produced 0% detection. Of eight fabricated instances, five were correctly flagged as unresolvable, one two-author fabrication was missed entirely, and two fabricated citations were falsely resolved to real papers because matching compared only the first five characters of the cited surname. The audit also deduplicates resolved inline citations against already retrieved papers, meaning it primarily verifies retrieved records rather than every citation actually placed in the manuscript.

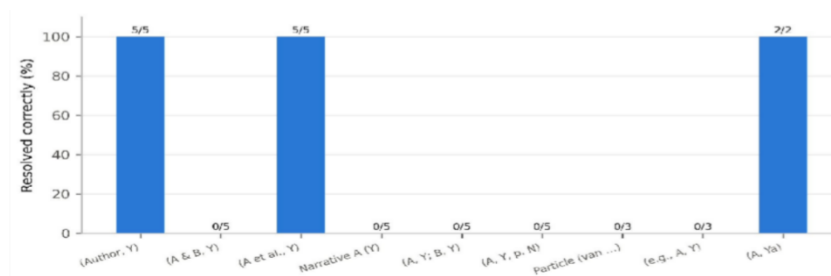


Figure 1. Share of legitimate in-text citations detected and resolved to the correct DOI by the pipeline's citation extractor, by APA 7 citation pattern (n = 38). Labels above bars give counts.

DOI and title verification produced similarly mixed results. Across 122 test cases, the original word-overlap rule correctly verified exact titles and case/punctuation variants, rejected titles swapped onto other DOIs, rejected non-resolving DOIs, and returned UNVERIFIED on network timeouts. However, all fourteen near-miss fabricated titles attached to real DOIs were accepted,

eight of nine legitimate short-form titles were rejected, every empty claimed title was certified as VERIFIED, and every genuine work without a DOI was labelled HALLUCINATED. Across genuine cases ($n = 51$), the false-rejection rate was 43.1%, including 27.5% labelled HALLUCINATED; across fabricated cases ($n = 42$), the false-acceptance rate was 33.3%. A revised rule based on normalized equality, controlled main-title matching, UNVERIFIED outcomes for empty titles and DOI-less works, and MISMATCH otherwise reduced false acceptance from 33.3% to 2.4% and false rejection from 43.1% to 0%. One false acceptance remained, showing that title evidence alone is insufficient and should be supplemented by registry metadata such as first-author surname and publication year.

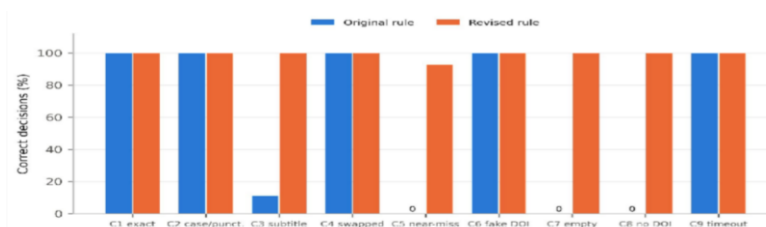


Figure 2. Share of correct decisions per verification case type for the original word-overlap rule and the revised rule ($n = 122$ cases from 14 Crossref records). C7 is scored as correct when the rule refuses to certify (UNVERIFIED)

The score parser recovered all fifteen dimensions in only three of ten plausible LLM output formats: the canonical table, decimal scores, and a table with an additional trailing column. Bold scores, spaced fractions, bare integers, scores out of 100, bolded dimension codes, and list-form outputs were not parsed; a table without the leading pipe recovered only eight of fifteen dimensions and missed dimension N, the citation-integrity score used by the pass condition. When parsing fails, the citation score defaults to zero and the manuscript is repeatedly revised until the maximum number of rounds is reached, even when the reviewer has judged it acceptable. The issue extractor is permissive: it recognizes dash bullets under level-three headings but can miss critical issues expressed as numbered lists or bold pseudo-headings, allowing a manuscript to pass despite unresolved concerns. The section-selection heuristic may also rewrite uncriticized sections because it matches keywords anywhere in the review.

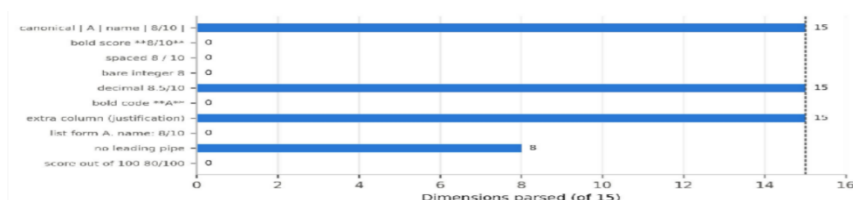


Figure 3. Number of the fifteen reviewer dimensions recovered by the pipeline's score parser for ten plausible LLM output formats. The dashed line marks complete recovery.

The journal-matching ablation demonstrates that meaningful recommendation depends on evidence-rich journal representations and abstention. With scope profiles and no score floor (M1), the matcher abstained for all ten nonsense probes and nine of ten degenerate probes at $\tau = 0.10$. For the fifteen probes with an a priori acceptable venue, hit@1 was 93.3% at $\tau = 0.10$, 100% at $\tau = 0.05$, and 73.3% at $\tau = 0.15$. Replacing scope profiles with journal names only (M2) reduced mean top scores for AI-topic probes from 0.178 to 0.041 and hit@1 to 33.3%. Adding a 0.70 floor eliminated abstention entirely: under M3 and M4 every probe received a top score of at least 0.70. M3 reached 100% hit@1 but compressed the distinction between relevant and nonsense probes, while M4

produced 66.7% hit@1 with very weak separation. These results show that score floors manufacture confidence rather than improve evidence.

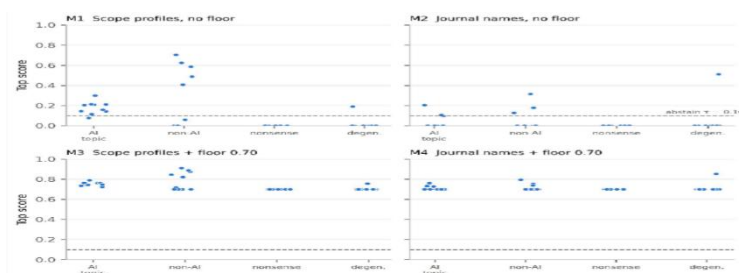


Figure 4. Top journal-match score per probe under four matcher designs (E5). M1: TF-IDF cosine over scope profiles written by the authors, no floor. M2: cosine over journal names only. M3 and M4: as M1 and M2 with scores mapped onto a 0.70 floor. The dashed line marks the abstention threshold $\tau = 0.10$. Each dot is one probe; horizontal jitter is added for legibility.

Overall, the evaluated pipeline fully satisfies integrity requirements R5 (abstention) and R6 (disclosure and human accountability), while R1–R4 are only partially satisfied. With the revised citation-verification rule and a scope-based, floor-free journal matcher, R1 and R3 would be largely satisfied. R2 still requires a numeric-claim gate and completion of Phases 3–4, while R4 requires schema-constrained reviewer output. For manuscripts prepared with the pipeline, citations accepted only by the original title check should be re-verified, journal suggestions should be confirmed on index source page, and all EVIDENCE_REQUIRED markers should be resolved before submission.

DISCUSSION

The discussion situates the broader educational implications in the thought of Syed Muhammad Naquib al-Attas and Ismail Raji al-Faruqi. Both perspectives begin from the proposition that knowledge is value-laden and that modern education cannot be treated as epistemologically neutral. Al-Attas argues that knowledge is embedded in a worldview; therefore, Islamic education should not merely import modern curricula without critically examining their ontological and methodological assumptions. This requires purification of meaning (*tashfiyah al-ma'na*) so that key concepts of knowledge are harmonized with *tawhid* (Sunarti & Rahman, 2025). Al-Faruqi similarly rejects a purely instrumental and utilitarian educational orientation, emphasizing social justice and the benefit of the ummah as central purposes of education (Sunarti & Rahman, 2025). Their positions are therefore complementary in reconstructing Islamic education in response to secular epistemological dominance.

This shared foundation also reshapes the normative aims of education. Al-Attas places the formation of the civilized person (*insan adabi*) at the center of education. Education is therefore not limited to social mobility or technical competence, but should integrate spiritual, intellectual, and moral dimensions into a balanced personality (Ammar, 2023). Al-Faruqi extends this orientation toward social transformation by emphasizing the formation of Muslims who master modern knowledge and apply it ethically (Khikmatun, 2025). Together, these dimensions provide an educational purpose relevant to spiritual formation and social responsibility (Sunarti & Rahman, 2025).

At the curricular and pedagogical levels, al-Attas emphasizes a hierarchy of knowledge grounded in the Islamic worldview, with revealed knowledge occupying a central and orienting position. Integration between religious and general sciences should therefore be substantive rather

than merely administrative, accompanied by the purification of key concepts in accordance with tawhid (Sassi, 2022). Al-Faruqi offers a more operational framework through the Islamization of knowledge, which calls for restructuring curricula, research methodologies, and teaching materials so that scientific disciplines are reconstructed on monotheistic principles. The task is to reconsider the theoretical and methodological assumptions underlying each discipline. Such integrative curriculum design seeks synergy across fields while strengthening epistemological foundations (Sumiran, 2025) and developing graduates who combine technical competence with ethical judgment (Muaz et al., 2023).

The same principles extend to institutional management. Islamic educational institutions should function not only as sites of knowledge transmission but also as environments for cultivating a monotheistic worldview, scientific adab, academic ethos, and moral responsibility. Al-Faruqi further emphasizes interdisciplinary knowledge production directed toward solving the problems of the ummah and the wider global community. In this sense, Islamic education becomes a dynamic institution capable of responding to contemporary challenges rather than a static repository of inherited knowledge (Coil & Aprison, 2023).

In the Indonesian context, these ideas are relevant because separation between religious and general sciences can produce graduates insufficiently prepared for global challenges. The integrative framework of tawhid and adab offered by al-Attas and al-Faruqi can support curriculum reform, strengthen educators' epistemological competence, and reinforce institutional visions grounded in Islamic values. Its application can orient Islamic education toward a more holistic model that connects scientific rationality with spirituality while maintaining global competitiveness without abandoning its religious identity (Nisok, 2024).

Taken together, both thinkers provide a foundation for Islamic education that is intellectually integrative, morally grounded, socially responsive, and institutionally transformative. Across curriculum, pedagogy, institutional management, and contextual relevance, their ideas emphasize not only cognitive and practical competence but also character and moral formation. The synthesis therefore positions Islamic education as a moral force capable of contributing to human progress while addressing the challenges of modernity (Khairunnisa et al., 2025).

4. CONCLUSION

AI submission assistants promise to take a manuscript from literature search to a matched journal. The signals they display along the way (verification labels, reviewer scores, journal matches) look like measurements, and authors are likely to treat them as such. This study evaluated an open-source multi-agent pipeline built on evidence gating. The pipeline retrieves literature from bibliographic APIs, resolves DOIs, and abstains with explicit markers when evidence is missing. At its central boundary, the gating worked: the pipeline did not fabricate the results it lacked. Its gates still leaked at predictable points where the LLM meets deterministic code. The citation extractor could not see most APA citation forms. A title threshold admitted every near-miss fabrication, and defaults turned missing titles into verification. Reviewer-score parsers let formatting, rather than judgement, decide whether a manuscript passed. Small, principled revisions closed most of these gaps. A stricter title rule reduced false acceptance of fabricated citations from 33.3% to 2.4% without rejecting legitimate ones. In journal matching, an ablation showed that scope evidence and a floor-free decision rule are both needed if a matcher is to decline when nothing fits. The broader conclusion is that the integrity of AI-assisted publishing depends on the tool layer as much as on the model. A tool that abstains honestly is more useful to a researcher than one that always answers confidently, but only if its gates are tested as rigorously as its model. The eight solutions, seven design principles, reference architecture and audit protocol offered here give developers, authors,

editors and institutions a practical way to build and check such tools. The five-plan research agenda sets out how to test them at scale, from a revised release of the pipeline to open benchmarks, a user study and a field audit of the tools researchers already use.

REFERENCES

- Agrawal, A., Suzgun, M., Mackey, L., & Kalai, A. T. (2024). Do language models know when they're hallucinating references? In Findings of the Association for Computational Linguistics: EACL 2024. Association for Computational Linguistics. <https://doi.org/10.48550/arXiv.2305.18248>
- Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15(2), Article e35179. <https://doi.org/10.7759/cureus.35179>
- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Paper 3). ACM. <https://doi.org/10.1145/3290605.3300233>
- Beel, J., Gipp, B., Langer, S., & Breiteringer, C. (2016). Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries*, 17(4), 305–338. <https://doi.org/10.1007/s00799-015-0156-0>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). ACM. <https://doi.org/10.1145/3442188.3445922>
- Bergstrom, C. T. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News*, 68(5), 314–316. <https://doi.org/10.5860/crln.68.5.7804>
- Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., & Miller, L. E. (2023). High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus*, 15(5), Article e39238. <https://doi.org/10.7759/cureus.39238>
- Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, 1(1), 363–376. https://doi.org/10.1162/qss_a_00018
- Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11), 2215–2222. <https://doi.org/10.1002/asi.23329>
- Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications*, 8, Article 25. <https://doi.org/10.1057/s41599-020-00703-8>
- Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26, Article e53164. <https://doi.org/10.2196/53164>
- Conroy, G. (2023). Scientific sleuths spot dishonest ChatGPT use in papers. *Nature*. <https://doi.org/10.1038/d41586-023-02477-w>
- COPE Council. (2023). COPE position statement: Authorship and AI tools. Committee on Publication Ethics. <https://publicationethics.org/cope-position-statements/ai-author>

- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., et al. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, Article 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
- Else, H. (2023). Abstracts written by ChatGPT fool scientists. *Nature*, 613(7944), 423. <https://doi.org/10.1038/d41586-023-00056-7>
- Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: A systematic literature review. *ACM Computing Surveys*, 52(6), <https://doi.org/10.1145/3345317>
- Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. <https://doi.org/10.1038/s41746-023-00819-6>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey (arXiv:2312.10997). arXiv. <https://doi.org/10.48550/arXiv.2312.10997>
- Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127. <https://doi.org/10.1136/amiajnl-2011-000089>
- Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2), 337–355. <https://doi.org/10.25300/MISQ/2013/37.2.01>
- Grudniewicz, A., Moher, D., Cobey, K. D., et al. (2019). Predatory journals: No definition, no defence. *Nature*, 576(7786), 210–212. <https://doi.org/10.1038/d41586-019-03759-y>
- Guerrero-Bote, V. P., & Moya-Anegón, F. (2012). A further step forward in measuring journals' scientific prestige: The SJR2 indicator. *Journal of Informetrics*, 6(4), 674–688. <https://doi.org/10.1016/j.joi.2012.07.001>
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330). PMLR. <https://doi.org/10.48550/arXiv.1706.04599>
- Hanson, M. A., Gómez Barreiro, P., Crosetto, P., & Brockington, D. (2024). The strain on scientific publishing. *Quantitative Science Studies*, 5(4), 823–843. https://doi.org/10.1162/qss_a_00327
- Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1), 414–427. https://doi.org/10.1162/qss_a_00022
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–106. <https://doi.org/10.2307/25148625>
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569–16572. <https://doi.org/10.1073/pnas.0507655102>
- S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., & Schmidhuber, J. (2024). MetaGPT: Meta programming for a multi-agent collaborative framework. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*. <https://doi.org/10.48550/arXiv.2308.00352>
- Hosseini, M., & Horbach, S. P. J. M. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research Integrity and Peer Review*, 8, Article 4.

- <https://doi.org/10.1186/s41073-023-00133-5>
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), Article 42. <https://doi.org/10.1145/3703155>
- 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), Article 248. <https://doi.org/10.1145/3571730>
- Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- Kasneci, E., Sessler, K., Küchemann, S., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kendall, G., & Teixeira da Silva, J. A. (2024). Risks of abuse of large language models, like ChatGPT, in scientific publishing: Authorship, predatory publishing, and paper mills. *Learned Publishing*, 37(1), 55–62. <https://doi.org/10.1002/leap.1578>
- Kinney, R., Anastasiades, C., Authur, R., et al. (2023). The Semantic Scholar Open Data Platform (arXiv:2301.10140). arXiv. <https://doi.org/10.48550/arXiv.2301.10140>
- E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track)*. <https://doi.org/10.48550/arXiv.2306.05685>