The Case of the Mysterious Citations

Amanda Bienz, *University of New Mexico, Albuquerque, NM, USA*

Carl Pearson, *Sandia National Laboratories, Albuquerque, NM, USA*

Simon Garcia de Gonzalo, *Sandia National Laboratories, Albuquerque, NM, USA*

Abstract—Mysterious citations are routinely appearing in peer-reviewed publications throughout the scientific community. In this paper, we developed an automated pipeline and examine the proceedings of four major high-performance computing conferences, comparing the accuracy of citations between the 2021 and 2025 proceedings. While none of the 2021 papers contained mysterious citations, every 2025 proceeding did, impacting 2-6% of published papers. In addition, we observe a sharp rise in paper title and authorship errors, motivating the need for stronger citation-verification practice. No author within our dataset acknowledged using AI to generate citations even though all four conference policies required it, indicating current policies are insufficient.

Large language models (LLMs) have transitioned in just a few years from research prototypes to widely available tools. Systems such as ChatGPT [1], Claude [2], and Gemini [3] are now embedded in everyday workflows, offering fluent text generation, summarization, and synthesis at low cost. Accessibility, ease of use, and apparent competence have made them attractive assistants for writing tasks in various domains, including academic research [4].

This shift has begun to surface explicitly in the academic publication ecosystem. Major venues and professional societies, including ACM and IEEE, now publish guidelines on the use of generative AI in paper preparation[5], [6], ranging from disclosure to authorship attribution. These policies acknowledge a reality: LLMs are already being used in conference and journal submissions, particularly for drafting, editing, and literature review assistance [4], and their presence is no longer exceptional.

At the same time, the use of LLMs in publications is often not transparent. Many venues do not require explicit disclosure, and even when policies exist, compliance is uneven. Anecdotal reports from program committees and reviewers increasingly describe a specific failure mode: "hallucinated" or outright fabricated citations inserted by LLMs for related work sections or to substantiate claims. These references may appear superficially plausible with complete author lists, venues, and years, however correspond to no real publication.

This phenomenon poses a burden on the peer-review process. Reviewers cannot reliably detect such errors by inspection alone. Verifying a suspicious reference may require checking reference data (e.g. titles, author names, venues, DOIs) against external publication records (libraries, metadata aggregators), an effort that scales poorly when combined with an already time-consuming review workflow. As a result, fabricated or incorrect citations can pass unnoticed, undermining the integrity of the evaluation process.

This paper outlines an effort to make this emerging issue visible. By analyzing papers published in computer science conferences, we quantify the prevalence and characteristics of incorrect and fabricated citations associated with LLM-assisted writing. Rather than treating these incidents as isolated anecdotes, we provide empirical evidence of their scope, offering a foundation for informed discussion and policy in an era where generative models are a common part of academic authorship.

USE AND MISUSE OF LLMS IN ACADEMIC WRITING

Not all uses of generative models in writing are intrinsically problematic. There is a broad consensus that regular tasks such as grammar correction, formatting, and stylistic polishing do not, in themselves, threaten the foundations of academic communication. These uses can be considered similar to spellcheckers, reference managers, or automated typesetting tools, operating at the level of presentation rather than substance.

The scientific method, by contrast, relies on transparent interaction with existing knowledge. Claims must be justified by evidence, and citations serve as verifiable trace links in a chain of progress. Generating invented sources or misrepresenting the literature breaks this chain. In other words, the unverified insertion of references that do not correspond to real works amounts to a form of citation fabrication, an operation indistinguishable from falsification, one of IEEE's and ACM's recognized categories of scientific misconduct [7], [8].

This distinction matters due to LLMs no longer being a fringe tool. Large-scale empirical work has shown that LLMs are already embedded in mainstream scientific writing practices. Liang et al. [4] analyzed more than one million pre-prints and published articles and found a significant and rapidly increasing presence of LLM-modified text across disciplines with particular double digit adoption numbers in computer science. Their results provide clear evidence that generative models are not just experimental aids but are actively reshaping how scientific work is produced.

Along with this broad adoption, however, a specific and consequential failure mode has begun to surface: the fabrication or corruption of references. This pattern mirrors earlier high-profile incident outside of academia, where a legal brief drafted with LLM assistance was submitted containing judicial opinions and citations to cases that did not exist [9]. The error was not stylistic, it produced plausible authority where none existed.

Comparable cases are now emerging in scientific publishing. Investigations of submissions to the 2026 International Conference on Learning Representations (ICLR) reported that 20% (60 out of 300 papers sampled from 20,000 total submissions), contained at least one AI hallucination [10]. GPTZero [11] "Hallucination Check" scan found that 50 peer-reviewed submissions to ICLR contained at least one hallucinated citation that peer reviewers had not flagged. These fake references often include made-up authors, incorrect venues, or entirely invented citations [12]. At the time of writing, however, this citation-verification tool is not publicly available to authors or reviewers.

By situating LLM-assisted writing within both its legitimate uses and its emerging issues, we hope to underscore a central tension: generative models can responsibly assist with how academic writing is done, but if they are allowed to invent what they cite, they undermine the very mechanism by which scientific knowledge is validated and extended.

METHODOLOGY

Our experiments examine a single question: how has the prevalence of *substantially erroneous* bibliographic entries changed since LLMs became publicly available.

This is distinct from the task of determining how prevalent LLM use is in the academic writing process *more broadly*. In general, deciding whether a given piece of text is generated by an LLM used by an adversarial author is an unsolved problem, and may even be impossible. Even if it were possible, it would not be directly useful to determining the scientific validity of that text. For example, a researcher acting in good faith could carry out a valid experiment, generate scientifically-grounded bullet points describing their work, use an LLM to produce a paragraph of prose, checked the resulting text for accuracy, and submit it for publication. The mere fact that an LLM emitted the text is unrelated to the scientific validity.

Unlike other prose, scholarly writing feature structured references to preexisting related publications. To a much greater extent than other parts of prose, the references themselves can be checked for correctness. This is not a judgment of the merit of the referenced work or it's relationship to the text, but rather a straightforward check of whether the reference points to an extant, preexisting work.

These structured references provide an opportunity to examine the prevalence of a specific pattern of LLM misuse: generating related bibliography entries without actually consulting the related work in question. It is difficult to defend this use of LLMs as being part of a valid scholarly process, separate from other potential uses of LLMs.

To validate references within a published paper, its PDF file was passed through a software tool that produced a summary report (❶-❺ in Figure). The results of the final report were manually verified by the authors. Figure 1 summarizes this process.

❶ **Extract PDF text**: PyPDF's [13] `ParsePDF` functionality is used to convert the input PDF file into plain text.

❷ **Isolate bibliography**: String search heuristics are used to isolate the paper's bibliography; specifically, the text between the last occurrence of "bibliography" or "references" and the following occurrence of "appendix" (or the end of the document).

❸ **Split bibliography into entries**: String search heuristics are used split the bibliography into separate entries. For the conferences analyzed in this work, all bibliography entries started with a number in square brackets (e.g. "[17]").

❹ **Parse entry**: String search is used to find DOIs and ArXiv identifiers. Regular expressions are used to identify title, authors, and venue from 12 empirically-identified reference formats.

❺ **Search external aggregators for matching publications**: Depending on the extracted metadata, a variety of external sources are search for matching

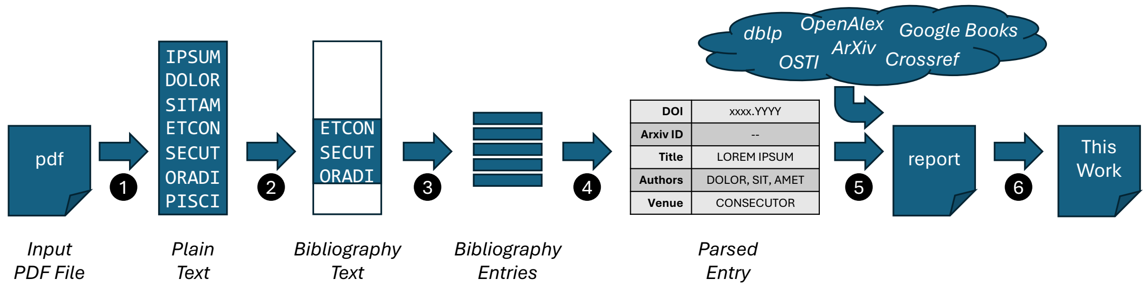

FIGURE 1. Summary of the analysis methodology. ❶ extract text via PyPDF's `ParsePDF`. ❷ isolate bibliography. ❸ split bibliography into entries. ❹ parse bibliography entry. ❺ search external aggregators for matching publications. ❻ manually validate machine results.

publications. Those sources are ArXiv, Crossref, dblp, Google Books, OpenAlex, and the U.S. Department of Energy Office of Science and Technical Information (OSTI).

❻ **Manually validate machine results**: If a citation was not found automatically, manually search for the article. If a web search does not validate its existence or returns only ResearchGate articles, open the cited location, such as the journal volume or proceedings, and check the cited pages. For ArXiv pre-prints, check against previous versions for title and author discrepancies.

MYSTERIOUS CITATIONS

We have chosen to anonymize all presented results to avoid implicating any specific individuals, institutions, or venues. Rather, we hope to raise awareness of this systemic challenge to the scholarly writing process.

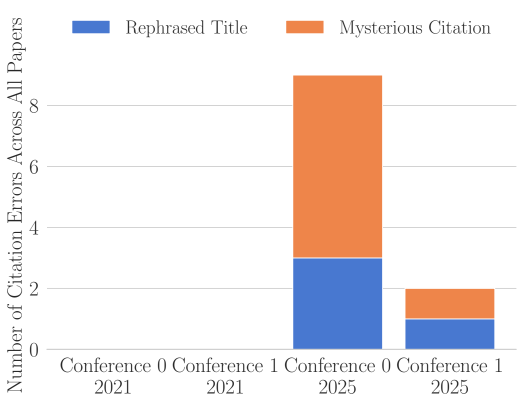

FIGURE 2. Baseline comparison: number of papers, per conference, with at least one citation issue, colored by highest level of severity across all citations within the paper.

We analyzed the accuracy of citations in the proceedings of four high-performance computing conferences. All examined conferences are well-known and respected within the community and were selected due to our familiarity as previous authors or program committee members. These conferences, labeled as Conferences 0 through 3 throughout the remainder of this section, have ERA conference rankings of A, C, A, and unranked, respectively [14]. The 2021 proceedings were analyzed as a baseline while the 2025 proceedings of the same conferences were examined for current trends. We analyzed all papers within each proceeding for correctness. Note that the proceedings vary in size, from less than 50 to over 100 publications.

Citation errors naturally can occur for a number of reasons, including typos by the author and discrepancies among subtitle inclusion. Therefore, we have categorized citation errors by severity, as follows:

1) **Minor Error:** A minor error within the title, such as a misspelled word or a small number of missing words. The presumptive reference is found at the listed location.

2) **Rephrased Title:** The title has been rephrased but retains the meaning of the true title. The publication is found at the listed location.

3) **Mysterious Citation:** No paper a similar enough title exists. The cited location either does not exist or holds an unrelated paper with different authors.

Figure 2 compares citation errors within the 2021 and 2025 proceedings of two conferences. Minor errors are not included in the plot as they occur consistently across all years. Both rephrased titles and mysterious citations were only found in recent proceedings, corre-

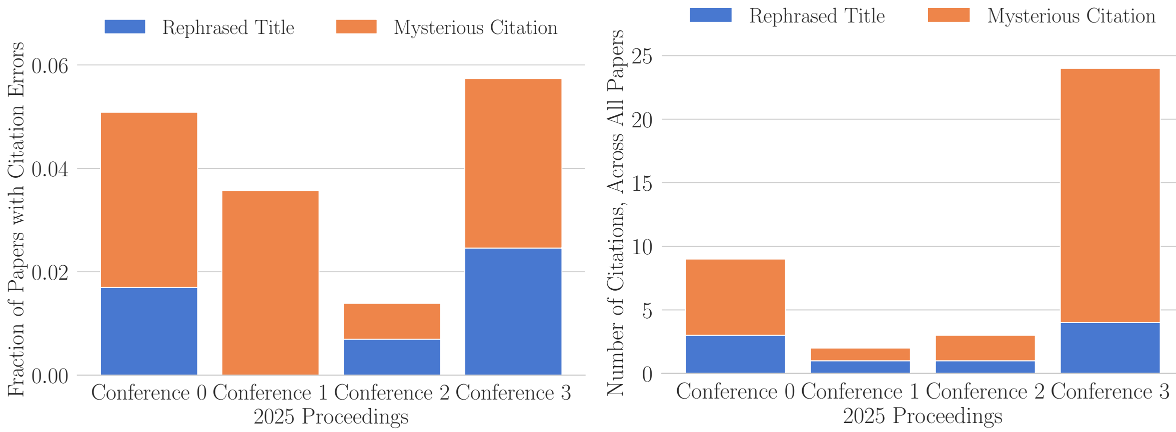

FIGURE 3. Citation errors in 2025 proceedings: The fraction of papers with citation errors, colored by highest severity across all citations within the paper (left) and the total number of erroneous citation across all papers, colored by severity (right).

lating with the advent of generative AI.

Figure 3 shows citation errors found within the 2025 proceedings of the four analyzed conferences. Both rephrased titles and mysterious citations were found in all four conference proceedings, with the fraction of papers in which they were found ranging from under 2% of papers to nearly 6%. **In every proceedings that we examined, there are *published* papers that include citations to papers for which we were able to find no evidence of existence.** In most of the papers in which a mysterious citation was found, one or a small number of citations had severe errors. However, in one paper, over half of the citations did not exist at the cited location and we were unable to locate the cited titles through Google searches.

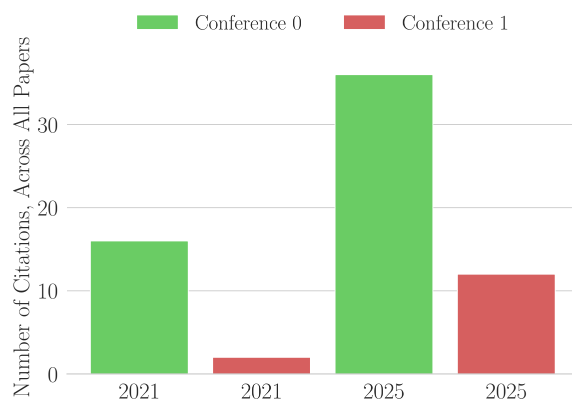

FIGURE 4. Baseline comparison: number of citations, per conference, with missing authors, extra authors, or both.

Author and DOI Errors

In addition to the mysterious citations, we also analyzed errors within author lists. Author name misspellings were frequent across all proceedings, and as such, we ignored these within our analysis. We also ignored minor errors such as reversal of first and last names, ‘et al‘ when all authors are listed, and missing special characters. We examined recent proceedings only for instances of missing authors, extra authors, or both. In the case of mysterious citations, authors were classified as incorrect. For all other citations, including those without title errors, we compared author lists to the true publication.

Unlike title errors which were rampant only in recent proceedings, missing and extra authors were also found in 2021 proceedings,as shown in Figure 4. However, the number of citations with author list errors has drastically increased from 2021 to 2025.

Finally, we analyzed the correctness of provided DOIs and Arxiv IDs. The vast majority of citations did not include a DOI or Arxiv ID. While there were incorrect DOIs and ArXiv links in multiple of the examined proceedings, both in 2021 and 2025, trends are difficult to analyze due to the small sample size.

AI-Generated Citations

There is a correlation between the introduction of generative AI and the occurrence of mysterious citations within publications. However, we are unable to verify that AI was used to generate the incorrect citations found in this work. **All analyzed conferences required authors to acknowledge any generative AI**

usage, yet none of the papers acknowledged using AI to generate citations. During our data collection, we did find one citation that included a URL ending in `utm_source=chatgpt.com`, strongly indicating the citation was generated by ChatGPT. However, the authors only acknowledged using generative AI for grammatical edits.

While we limited our dataset to proceedings of HPC conferences with which we are familiar, AI-generated citations are appearing across research domains. For instance, the Google Scholar search displayed in Figure 5 shows that **at the time of writing this article, nearly 17,000 articles contain a link generated by ChatGPT**.

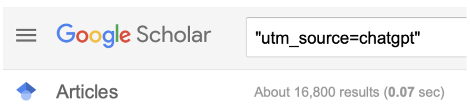

FIGURE 5. A Google Scholar search for articles containing URLs generated by ChatGPT, showing 17,000 results.

AI-Generated Papers

Without author acknowledgment, there is no way to verify which portions of a paper have been generated by AI. However, in recent years, non-peer-reviewed papers have been published online by unverified researchers at a rapid rate. This is particularly problematic on ResearchGate [15], where no verification process is required to create a profile and upload a paper, enabling fake online personas such as Larry, the world's most cited cat [16].

One of the mysterious citations from Conference 3 was verified as inaccurate, with no related article existing in the cited location. However, the exact title did appear on ResearchGate. The full text was uploaded, but it was not formatted for formal publication, contained bulleted lists instead of paragraphs, and had empty figures as shown in Figure 6. **This incomplete paper was uploaded to ResearchGate on the same date as the citing paper's submission deadline, as listed in Conference 3's Call for Papers.** The ResearchGate profile to which this paper was uploaded has since added 40 similar papers.

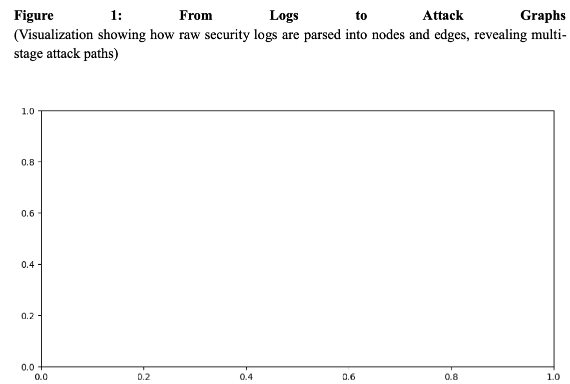

FIGURE 6. A screenshot of the data presented in the ResearchGate upload cited by a paper in Conference 3.

DISCUSSION

Inappropriate and unacknowledged use of generative AI poses a growing risk to the integrity of the scientific community. When a paper contains unacknowledged generated content, such as embedded `utm_source=chatgpt`, the community is left to infer that generative AI may have been used in other unacknowledged ways. More concerning, when a paper includes verifiably incorrect citations, readers are left to assume the authors either intentionally fabricated references or included unverified generated content throughout the publication. Such issues can undermine trust in the work as a whole and can negatively impact the reputation of all co-authors, even those who were unaware generated content was included.

Crucially, every mysterious citation identified in this study appears in published proceedings. These papers successfully passed peer review and editorial checks, yet still contain fabricated or corrupted references. This demonstrates that the current review process which is already under severe time and workload constraints, is not sufficient to reliably detect these classes of errors.

The scientific community could take steps to better prevent these mysterious citations from occurring within our peer-reviewed publications. If an author uses generative AI during a literature review, hallucinations can be avoided by manually finding and reading the publication, using only official resources to cite it. Co-authors can prevent mysterious citations by checking the full bibliography for correctness before submission.

During the review process, all citations for each article can be verified prior to acceptance. Conference organizing committees can further develop policies, such as desk rejecting papers with mysterious citations or/and treat fake citations as a form of scientific misconduct on par with fabrication. Major publication venues could adopt an explicit policy that defines unacceptable citation practices and associated penalties. For difficult to find references, authors could be required to supply verifiable evidence such as DOI or PDF at submission time.

Fully hallucinated citations are likely to disappear as generative AI continues to improve. To date, while citations generated by ChatGPT 5 consistently include extra words, incorrect authors, and invalid DOIs, all of our attempts to prompt the new model into generating a fully fabricated citation have been unsuccessful. However, hallucinated citations are only a tangible symptom of a larger underlying problem. Researchers are increasingly using AI to generate content and, in some cases, publishing the generated text without verifying it for correctness. This trend raises questions about whether research papers can remain a reliable method for sharing scientific progress.

CONCLUSION

Our study provides the first systemic evidence that unverified, AI-hallucinated citations have begun to appear in peer-reviewed conferences proceedings. By comparing four major venues from 2021 and 2025, we show that every 2025 proceeding we examined contains at least one mysterious citation. These errors, which affect up to 6% of *published papers* in some venues, pose a serious issue for the integrity of academic communication and are unlikely to be caught by today's review processes. In order to initiate community discussion on this topic, we briefly mention some actions that could be taken to mitigate mysterious citation errors. Combining individual diligence with explicit community-wide policy on this subject and developing automated tools could help improve traceability and maintain trust in the scientific process.

LIMITATIONS

It is not possible to prove that a citation error was generated through an LLM hallucination. While we are not aware of any confounding factors that would cause the incidence of human-derived bibliography errors to arise, we do not explicitly account for them either.

ACKNOWLEDGMENTS

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

This work made use of free, public API access to ArXiv, Crossref, dblp, OpenAlex, and the U.S. Department of Energy Office of Science and Technical Information (OSTI). The authors would like to thank those organizations for providing those resources. Generative AI was used to generate string parsing as API search keys when developing the citation verification tool.

This work was performed with partial support from the National Science Foundation under Grant Nos. CCF-2338077 and CNS-2450092 and the U.S. Department of Energy's National Nuclear Security Administration (NNSA) under the Predictive Science Academic Alliance Program Awards DE-NA0003966, DE-NA0004267.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation and the U.S. Department of Energy's National Nuclear Security Administration.

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC (NTESS), a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration (DOE/NNSA) under contract DE-NA0003525. This written work is authored by an employee of NTESS. The employee, not NTESS, owns the right, title and interest in and to the written work and is responsible for its contents. Any subjective views or opinions that might be expressed in the written work do not necessarily represent the views of the U.S. Government. The publisher acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this written work or allow others to do so, for U.S. Government purposes. The DOE will provide public access to results of federally sponsored research in accordance with the DOE Public Access Plan.

REFERENCES

1. OpenAI, "ChatGPT." https://openai.com/blog/chatgpt, 2022. Accessed: 2026-01-13.
2. Anthropic, "Claude." https://www.anthropic.com/claude, 2023. Accessed: 2026-01-13.
3. Google, "Gemini." https://deepmind.google/technologies/gemini/, 2023. Accessed: 2026-01-13.
4. W. Liang, Y. Zhang, Z. Wu, H. Lepp, W. Ji, X. Zhao, H. Cao, S. Liu, S. He, Z. Huang, *et al.*, "Quantifying large language model usage in scientific papers," *Nature Human Behaviour*, pp. 1–11, 2025.
5. Association for Computing Machinery, "ACM policy on authorship." https://www.acm.org/publications/policies/new-acm-policy-on-authorship, 2023. In-

cludes guidance on the use of generative AI in manuscript preparation. Accessed: 2026-01-13.

6. Institute of Electrical and Electronics Engineers, "Submission and peer review policies: Guidelines on AI-generated text." https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/guidelines-and-policies/submission-and-peer-review-policies/, 2023. Defines disclosure requirements for AI-generated content in IEEE publications. Accessed: 2026-01-13.
7. Association for Computing Machinery, "Acm policy on plagiarism, misrepresentation, and content falsification." https://www.acm.org/publications/policies/plagiarism-overview, 2025. ACM defines content falsification as intentional misrepresentation of results, supporting materials, or references, including manufactured or irrelevant citations that do not legitimately support the work.
8. Institute of Electrical and Electronics Engineers, "Ieee ethical requirements for authors: Avoiding fabrication, falsification, and citation misconduct." https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/ethical-requirements/, 2025. IEEE author ethics policies prohibit fabrication (inventing data/results), falsification (altering or omitting data/results), and inappropriate uses of citation such as citing irrelevant sources to inflate metrics or mislead readers.
9. B. Weiser and N. Schweber, "The ChatGPT lawyer explains himself." https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html, 6 2023. Accessed: 2026-01-14.
10. P. Esau, N. Shmatko, and A. Adam, "GPTZero finds over 50 new hallucinations in ICLE2026 submissions." https://gptzero.me/news/iclr-2026/, 2025. Accessed: 2026-01-14.
11. E. Tian and A. Cui, "Gptzero: Detection of ai-generated text," 2026. Accessed: 2026-01-15.
12. S. Martin, "Startup investigation reveals 50 peer-reviewed papers contained ai-hallucinated citations." https://betakit.com/start-up-investigation-reveals-50-peer-reviewed-papers-contained-hallucinated-citations/?utm_source=chatgpt.com, 2025. Published December 16, 2025; accessed 2026-01-15.
13. PyPDF Contributors, "Pypdf." https://github.com/py-pdf/pypdf, 2026. Accessed: 2026-01-15.
14. "ERA conference rankings." http://www.conferenceranks.com. Accessed: 2026-01-21.
15. H. S. Al-Sinani and C. J. Mitchell, "From content creation to citation inflation: A genai case study." https://arxiv.org/abs/2503.23414, 2025. Accessed: 2026-01-16.
16. C. Wilcox, "How easy is it to fudge your scientific rank? meet larry, the world's most cited cat." https://www.doi.org/10.1126/science.zl99qni, 2024. Accessed: 2026-01-16.

Amanda Bienz is an Assistant Professor in the Computer Science Department at the University of New Mexico. Her research interests include improving the performance and scalability of parallel solvers, applications, and MPI collective operations for emerging architectures. Bienz received her Ph.D in Computer Science from the University of Illinois Urbana–Champaign. Contact her at bienz@unm.edu.

Carl Pearson is a Senior Member of Technical Staff in the Scalable Algorithms group at Sandia National Laboratories. His research interests include future architectures for scientific computing, GPU communication for distributed linear algebra, and GPU acceleration of irregular operations. He received his Ph.D in Electrical and Computer Engineering from the University of Illinois Urbana–Champaign. Contact him at cwpears@sandia.gov.

Simon Garcia de Gonzalo is a Senior Member of Technical Staff in the Scalable Computer Architecture group at Sandia National Laboratories. His research interests include high-performance computing architectures, GPU communication and benchmarking, and performance analysis for emerging heterogeneous systems. He received his Ph.D. in Computer Science from the University of Illinois Urbana–Champaign. Contact him at simgarc@sandia.gov.