What *Can't* Large Language Models Do? The Future of AI-Assisted Academic Writing

Raymond Fok
rayfok@cs.washington.edu
University of Washington
Seattle, WA, USA

Daniel S. Weld
danw@allenai.org
Allen Institute for AI &
University of Washington
Seattle, Washington, USA

ABSTRACT

Large language models have revolutionized the way we interact with the world around us, yet their relative nascency suggests its transformative potential on society is still underexplored. Applications built on these models have excelled at summarizing articles, engaging in realistic conversations, and writing creative stories. However, there remain open questions in how we can design tools that effectively leverage these models to support complex, cognitive demanding, and factual writing processes. In this position paper, we consider emergent paradigms in human-AI collaborative writing and their implications on future academic writing assistants.

1 INTRODUCTION

General-purpose generative models, and large language models (LLMs) in particular, have seen a prominent rise to fame, both within the academic research community and throughout mainstream media. These models have been harnessed countless applications, including some for addressing information overload and traversing the vast amounts of stored knowledge within the scientific literature. One infamous example is Galactica [29], an inaugural attempt at training large language models to learn, reason with, and generate scientific text. While Galactica's shortcomings were quickly revealed—from its tendency to generate authoritative-sounding yet inaccurate scientific text, to its suggestion of text marred by bias and toxicity [10]—the model also demonstrated state-of-the-art performance across many scientific and general-purpose benchmarks for question-answering and reasoning tasks. Against a backdrop of ethical concerns, but intrigued by the potential new interface into scientific knowledge, researchers have begun to wonder how these LLMs, and the applications they empower, could be developed responsibly to further science while reducing potential harms to the scientific community and society at large. In contrast to more cynical perspectives on the impending destabilization of established societal constructs at the hands of these language models, we consider how they may coexist with knowledge workers—with a focus on academic researchers in particular—complementing rather than replacing existing creative and problem-solving abilities.

2 LLM-POWERED WRITING ASSISTANTS

2.1 Interaction Paradigms of Writing Assistants

The application of large language models to the writing domain has garnered widespread attention through mainstream consumer applications such as Grammarly's writing assistant and Google's Smart Compose [7] and Smart Reply [20] features. Within the academic literature, researchers have developed LLM-powered writing assistants to study collaborative writing within domains such as creative fiction [3, 15, 28, 32], correspondence [16, 21], screenwriting [24], and scientific writing [14]. We first identify five human-AI collaboration paradigms of writing assistants in creative writing.

- **P1**. **Ideation.** Writing assistants may suggest creative directions for the text, particularly at the beginning of the writing process. Writers can also steer the ideation process by *seeding* the LLMs with specific prompts.
- **P2**. **Continuation.** Writing assistants may build upon existing text, adding to the beginning, end, or middle of passages.
- **P3**. **Elaboration.** Writing assistants may provide more details about a span of text specified by a writer. Similar to the ideation paradigm, elaboration could offer writers creative support, assisting in brainstorming, world building, or character development.
- **P4**. **Rewriting.** Writing assistants may rewrite text by suggesting alternatives for a selected span of text (i.e., infilling), or suggest alternatives that better reflect a desired property such as clarity, conciseness, or tone (i.e. style transfer).
- **P5**. **Questioning.** Writing assistants may generate questions about existing text, prompting writers to clarify or elaborate. Rather than continuing or infilling text, this form of questioning may help writers in reflecting upon and revising the text.

These co-writing paradigms also suggest natural interaction and interface design patterns. Available actions are often abstracted into lightweight affordances (e.g., buttons) within a writing assistant, and user requests are translated into pre-optimized prompts for the underlying LLM. In each paradigm, writing assistants can suggest multiple alternatives for writers, who in turn select aspects of the suggestions to incorporate into their own writing, iterate upon, or ignore. Some systems offer greater flexibility by allowing writers to interact with the LLMs directly via custom prompting.

2.2 Limitations of Writing Assistants

Studies observing writers collaborating with LLM-infused writing assistants have also revealed several limitations [8, 14, 15, 18, 24, 28, 31, 32]. Though emerging from studies largely involving creative writing, these limitations may also concern academic writing.

Hallucination. The tendency for large language models to hallucinate, or generate factually inaccurate text, is not a new problem. In many cases, the generated text is further conveyed in authoritative or confident tone. Hallucinated content that is subtly inaccurate may not only be dangerous for novices, but also underwhelming for experienced writers in more technical domains who expect these

assistants to generate semantically meaningful text but receive instead a mimicry of the writing form.

Inconsistent content and style. Current writing assistants have limited working memory and generated text may often appear to lack contextual awareness. Though assistants can offer multiple fluent, well-written continuations of a passage, the suggestions may be inconsistent with the existing content. They may also be ignorant of style, failing to offer suggestions that reflect aspects of the text such as voice, word choice, tone, or tense.

Repetition. Writing assistants can descend into repetition, particularly in prompted scenarios less represented in the model's training data. The generated text may also suggest biased narratives, rehashed tropes, or clichés, evidencing the models' susceptibility to inconsistency and lack of nuanced understanding.

Mediocrity. Under the guise of creativity, writing assistants can require writers to wade through dozens of similar suggestions before finding a "good" one [18]. Equipped with these writing assistants, writers may still struggle to efficiently pick through mediocre, let alone repetitive and biased, suggestions.

Ethical concerns. Writers may be concerned with the origin of the generated text, for instance how biases or misrepresentations encoded within the training data could be reflected in the model's output. Issues with plagiarism could also arise; for instance, writers might become concerned regarding ownership of the co-created writing given the inclusion of model generated text into their own writing, and whether the generated text itself plagiarizes from its learned sources.

3 SUPPORTING ACADEMIC WRITING

Since writing assistants have mostly been studied in the context of leisure writing, we are interested in understanding the opportunities and challenges in applying these assistants to academic writing, a more constrained and knowledge-driven task. How do the previously described interactions and limitations manifest for academic writing, and what are the implications on the design of future academic writing assistants? Moreover, beyond writing itself, how can these writing assistants support other challenging or tedious aspects of conducting research? As a starting point, we present an overview of five stages in the research process, envisioning how future intelligent writing assistants, empowered by LLMs, may support researchers in each stage.

Ideation. Academic writing provides a medium for documentation and communication of scientific knowledge. Before writing, researchers first accrue knowledge by identifying and executing upon a research idea. The ideation process can be challenging however, since researchers are required to identify an unsolved and relevant problem, justify why such a problem significant enough to solve, and determine feasibility given the available methods. Few intelligent tools have been successfully supported researchers with ideation, and it is unlikely current LLMs will change this status quo. Unlike in creative writing where absurdist tendencies of LLMs can be useful, the same creativity for academic writing ideation may generate nonsensical ideas agnostic to significance or feasibility. That said, throughout their training, these models may have learned to identify and then generate what it deems as the more significant, interesting, or unsolved research trends. And while many of the suggestions will be trite or nonsensical, researchers may still find utility in riffing off ideas from their newfound collaborator.

Engaging with the literature. Academic writing requires researchers to position their work within a continually expanding body of prior work. Some intelligent tools have recently been developed to assist in the navigation and consumption of academic literature (e.g., [1, 11, 12, 17, 19, 22]). With the addition of LLMs, assistants will not only assist in writing, but also the retrieval, comprehension, and synthesis of relevant papers. Designing question answering systems over single papers and even large corpora of documents is now more tractable and effective with recent LLMs than ever before, empowering research assistants such as Scholarcy[1], Elicit[2], and Scite[3]. Moreover, the vastly improved information retrieval and summarization capabilities of LLMs facilitate need-driven exploration and sensemaking over the academic literature, enabling researchers to focus more on creative tasks such as research ideation and clearly communicating their novel ideas through writing.

Writing. Most of the collaboration paradigms identified for creative writing map closely to academic writing. For instance, assistants can help writers continue upon their existing text, elaborate upon a topic sentence, or suggest alternative phrasings of a sentence. Researchers could also instead sketch an outline of their paper in informal, high-level claims, with the assistant generating detailed prose around the outline and researchers verifying the generated content. On the other hand, the assistant could generate an outline for a paper, with the researcher filling in details around the provided structure. These workflows resemble how developers use generative models such as Github Copilot [6] to translate natural language comments of code functionality into suggested code implementations. Studies on user interactions with these models (e.g., [2, 25, 30]) could help inform the design of future academic writing assistants. Transferable insights might include how users prefer to communicate intent with their task-based assistants, or how affordances should be provided to support validation and recovery from errors in the generated content.

One key challenge lies in developing writing assistants that can emphasize factuality and consistency, or otherwise provide affordances to facilitate researcher oversight of the generated content. Models far less sophisticated than current LLMs have previously shown that they can generate full conference papers that *appear* scientific, yet the generations are not actually factual, verifiable, or scientific. Efforts to understand and mitigate hallucination of LLMs are well underway within the NLP community, though the insights have yet to be adopted and evaluated within human-centered applications. Future academic writing assistants may resemble evidentiary QA systems (e.g., WebGPT [26], Perplexity AI[4], GopherCite [23]), designed to incorporate external knowledge and cite relevant sources within their generated text.

Revision. Like those for creative writing, academic writing assistants can assist researchers through revision. These models are more than capable of emulating human copywriters, suggesting

[1] https://www.scholarcy.com/
[2] https://elicit.org/
[3] https://scite.ai/
[4] https://www.perplexity.ai/

grammatical fixes, rephrasing text, rewriting passages for clarity, ensuring consistency of jargon, verifying logical soundness, checking for controversial or biased statements [27], identifying additional supporting references [4], and more. We imagine most researchers will retain agency throughout the revision process, with writing assistants playing the role of an eager copy-writing assistant, or perhaps a critical reviewer suggesting direct revisions and questioning the researcher to elaborate or clarify further.

Disseminating knowledge. In addition to helping write papers, writing assistants can support more efficient creation of accompanying knowledge artifacts, such as presentations [13], text summaries [5, 9], online blogs, or social media threads [14]. While some systems have been explored previously to automate the creation of such artifacts, the general-purpose capacities of LLMs to organize, summarize, and simplify scientific text will improve the utility of these artifact generation tools. By facilitating the communication of knowledge, these assistants can further the impact and accessibility of novel academic research.

4 CLOSING THOUGHTS

Reflecting on current LLM-powered creative writing assistants, we see opportunities in HCI and NLP research to improve the design of future academic writing assistants. Open challenges remain in ensuring factuality and providing provenance for text generated by LLMs, and designing tools that effectively harness the strengths of LLMs to augment researchers' existing capabilities. We argue that while these writing assistants may help scaffold or automate rote aspects of academic writing, they are inadequate without their co-creators. In this paper, we identified design paradigms and limitations of academic writing assistants in the hopes of fostering insightful discussion about the envisioned opportunities and challenges of leveraging such tools, for instance throughout the five stages of the research process. Researchers are necessary for navigating the complex landscape of academic research, and now more than ever, are responsible for designing the future in which these models will support, and not displace, researchers.

REFERENCES

[1] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Trans. Comput.-Hum. Interact.* (apr 2023). https://doi.org/10.1145/3589955

[2] Shraddha Barke, Michael B. James, and Nadia Polikarpova. 2022. Grounded Copilot: How Programmers Interact with Code-Generating Models. *ArXiv* abs/2206.15000 (2022).

[3] Eden Bensaid, Mauro Martino, Benjamin Hoover, Jacob Andreas, and Hendrik Strobelt. 2021. FairyTailor: A Multimodal Generative Framework for Storytelling. *ArXiv* abs/2108.04324 (2021).

[4] Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-Based Citation Recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 238–251.

[5] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme Summarization of Scientific Documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 4766–4777.

[6] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[7] Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail Smart Compose: Real-Time Assisted Writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2287–2295.

[8] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 329–340.

[9] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 615–621.

[10] Benj Edwards. 2022. New Meta AI demo writes racist and inaccurate scientific literature, gets pulled. https://arstechnica.com/information-technology/2022/11/after-controversy-meta-pulls-demo-of-ai-model-that-writes-scientific-papers/

[11] Joseph Chee Chang et al. 2022. CiteSee: Augmenting Citations in Scientific Papers with Persistent and Personalized Historical Context. *ArXiv* abs/2022.99999 (2022).

[12] Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Andrew Head, Marti A. Hearst, and Daniel S. Weld. 2023. Scim: Intelligent Skimming Support for Scientific Papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, Sydney, Australia, 23 pages.

[13] Tsu-Jui Fu, William Yang Wang, Daniel McDuff, and Yale Song. 2022. DOC2PPT: Automatic Presentation Slides Generation from Scientific Documents. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 1 (June 2022), 634–642. https://doi.org/10.1609/aaai.v36i1.19943

[14] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing Using Language Models. In *Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1002–1019.

[15] Maliheh Ghajargar, Jeffrey Bardzell, and Love Lagerkvist. 2022. A Redhead Walks into a Bar: Experiences of Writing Fiction with Artificial Intelligence *(Academic Mindtrek '22)*. Association for Computing Machinery, New York, NY, USA, 230–241.

[16] Steven M. Goodman, Erin Buehler, Patrick Clary, Andy Coenen, Aaron Donsbach, Tiffanie N. Horne, Michal Lahav, Robert MacDonald, Rain Breaw Michaels, Ajit Narayanan, Mahima Pushkarna, Joel Riley, Alex Santana, Lei Shi, Rachel Sweeney, Phil Weaver, Ann Yuan, and Meredith Ringel Morris. 2022. LaMPost: Design and Evaluation of an AI-Assisted Email Writing Prototype for Adults with Dyslexia. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 24, 18 pages.

[17] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2021. Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 413, 18 pages.

[18] Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. 2022. Creative Writing with an AI-Powered Writing Assistant: Perspectives from Professional Writers. *ArXiv* abs/2211.05030 (2022).

[19] Hyeonsu Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An Interactive System for Personalized Thread-Based Exploration and Organization of Scientific Literature. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 94, 15 pages.

[20] Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart Reply: Automated Response Suggestion for Email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 955–964.

[21] Florian Lehmann, Niklas Markert, Hai Dang, and Daniel Buschek. 2022. Suggestion Lists vs. Continuous Generation: Interaction Design for Writing with Generative Models on Mobile Devices Affect Text Length, Wording and Perceived Authorship. In *Proceedings of Mensch Und Computer 2022* (Darmstadt, Germany) *(MuC '22)*. Association for Computing Machinery, New York, NY, USA, 192–208.

[22] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu Kang, Egor Klevak, Bailey Kuehl, Michael Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol

Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2023. The Semantic Reader Project: Augmenting Scholarly Documents through AI-Powered Interactive Reading Interfaces. arXiv:2303.14334 [cs.DL]

[23] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147* (2022).

[24] Piotr Wojciech Mirowski, Kory Wallace Mathewson, Jaylen J. Pittman, and Richard Evans. 2022. Co-Writing Screenplays and Theatre Scripts with Language Models: An Evaluation by Industry Professionals. *ArXiv* abs/2209.14958 (2022).

[25] Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2022. Reading Between the Lines: Modeling User Behavior and Costs in AI-Assisted Programming. *ArXiv* abs/2210.14306 (2022).

[26] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* (2021).

[27] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 1650–1659.

[28] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2022. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Trans. Comput.-Hum. Interact.* (Feb. 2022).

[29] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony S. Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A Large Language Model for Science. *ArXiv* abs/2211.09085 (2022).

[30] Priyan Vaithilingam, Tianyi Zhang, and Elena L. Glassman. 2022. Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 332, 7 pages.

[31] David James Woo, Yanzhi Wang, and Hengky Susanto. 2022. Student-AI Creative Writing: Pedagogical Strategies for Applying Natural Language Generation in Schools. *ArXiv* abs/2207.01484 (2022).

[32] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852.