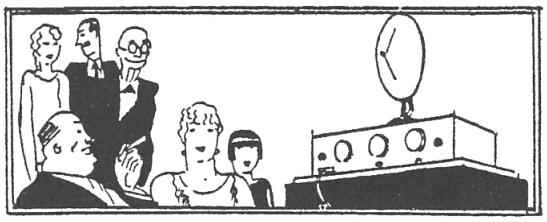
Generative AI for Scholarly Information Access



By **Lucy Lu Wang** (Information School, University of Washington)
<lucylw@uw.edu>

Many in the information seeking community are excited about the promise of large language models and Generative AI to improve scholarly information access. These models can quickly transform the content of scholarly works in ways that can make them more approachable, digestible, and suitably written for audiences for whom the works may not have been originally intended. However, the current technical implementation of Generative AI can limit their utility in these settings. Issues of hallucination (models generating false or misleading information) or bias propagation are still common, making it difficult to recommend these technologies for critical tasks. Dominant paradigms for addressing these issues and achieving alignment between AI and human values can also cause a reduction in the diversity of output, which can lead to information censorship for stigmatized topics, going against the goal of broad access to high-quality information. In this essay, I discuss the promises of AI for improving access to scholarly content, how current practices in Generative AI training may lead to undesirable and possibly unintended consequences, and how libraries and other community organizations could place themselves at the forefront of solutions for improving the individual and community relevance of these technologies.

The Promises of AI for Scholarly Content Creation and Understanding

Most scholarly content is written for other scholars. These texts make heavy use of technical jargon and assume a high level of background knowledge and domain expertise, raising barriers to reading and understanding. Engaging with these works can be difficult even from within their scholarly sub-communities, not to mention trying to do so from outside the academic sphere. Yet, the goal of many scholarly communities is to produce results and insights that can help improve individual and societal well-being. For example, while clinical trial reports for new medications are published in clinical journals, the people that stand to gain the most from promising results are patients and their caregivers.

Generative AI, such as large language models (LLMs) and text-to-vision models, have the ability to quickly transform the content of scholarly works, changing the language, tone, form, and presentation of these works to make them more approachable, understandable, and engaging for different audiences. In recent work, we have shown that simplifying text using language models can make difficult information more digestible (August et al., 2022); changing the form and presentation of material can reduce barriers to engagement (Shin et al., 2024); and synthesizing information across many papers can simplify the process of reviewing the literature (Giorgi et al., 2024). LLMs can also help with content creation, especially in cases where scholars lack the time and knowledge to easily perform this work without training. For example, they have been used successfully to help review papers (Zyska et al., 2023; D'Arcy et al., 2024) and describe complex figures to blind and low vision readers (Singh et al., 2024). When assisting users in writing, these models can reduce the time needed to produce high-quality content, while emphasizing the role of human authors in verifying that generated text accurately reflects their original intent.

Realizing the potential of Generative AI technologies, however, requires deeper understanding of individual and community-specific factors that influence scholarly information access, and the interplay between these factors and the design limitations of Generative AI systems. A specific tension is — we know that Generative AI can produce false or misleading output, which is a deal-breaker in scientific settings. They can also produce toxic or biased output, which may perpetuate social biases and would warn against their use for critical decision-making tasks. Current mitigation strategies for these negative behaviors, on the other hand, can also lead to homogenization of output, which may not support diverse community needs. This can manifest more when searching for information on stigmatized topics such as mental illness, reproductive health, or disability, where access to high-quality information is already difficult. So while these systems have the potential to improve information access, we must balance the various sides of this tension (accuracy, bias mitigation, and individual relevance) to achieve the promise of these technologies for end users.

Technical Limitations of Generative AI

Most popular LLMs are accessed through a chat interface, with users prompting the model in human language to provide a response or perform some task. Compared to classic search engines, this setup introduces new points of friction. First, LLMs compress information across many sources without attributing output to any specific source (or at least, without any guarantee of the correct source). At the same time, the tone of LLM-generated content tends to be official-sounding and confident. Together, these features reduce our ability to judge credibility based on previously reliable heuristics such as language quality or the trustworthiness of specific information sources. In a scholarly communication setting, false information and the inability to judge the veracity of information are unacceptable outcomes, since misleading content can pollute the scholarly record and reduce the value and trustworthiness of the entire scholarly enterprise.

Something else causing problems is unrepresentative training data. High quality data is essential to achieving good model performance, yet most LLMs are trained on mixtures of text and images scraped from the web, which is not representative of human society. Much of the toxicity and bias in AI output can be attributed to the presence of such issues within the training data (Zhao et al., 2017; Dodge et al., 2021; Buschek and Thorp, [n.d.]). Previous work has shown how marginalized groups tend to receive biased treatment from these models in terms of less equitable representation (Ghosh and Caliskan, 2023; Zack et al., 2023), higher rates of flagging in content moderation (Sap et al., 2019; Davidson et al., 2019), and more. While companies scramble to acquire better training data,[1] reinforcement

learning with human feedback (RLHF) has become the dominant paradigm for mitigating such issues from the modeling side (Glaese et al., 2022; Bai et al., 2022; Wu et al., 2023).

RLHF methods use human-labeled preferences between different answer choices to further train AI models, with the goal of achieving closer alignment with human values and expectations. While effective at reducing some forms of toxic and biased output, these methods have limitations. RLHF as currently implemented assumes the existence of a homogeneous shared set of human values that models can learn and optimize for, while we know that communities are diverse in their beliefs and needs (Kirk et al., 2023). The diversity of model responses for models trained with RLHF has been observed to be lower (Padmakumar and He, 2023). This homogenization of output (Anderson et al., 2024) can cause a tendency to prioritize "safe" answers about "normalized" topics, which reduces a model's ability to provide accurate and actionable answers for longtail or stigmatized topics (Oliva et al., 2020; Gadiraju et al., 2023). The potential for harm from additional information censorship is high, since people already face difficulties accessing high-quality information about these topics. In a scholarly communication setting, this may extend to a likelihood for AI to perpetuate the status quo, recommending work and findings from only the most "canonical" scholars and institutions.

Improving the Community Relevance of Generative AI

Community-oriented Generative AI must be aware of the needs and challenges of individuals and adapt to better serve those needs. While current LLMs are hampered by the limitations I discussed previously, there are several promising developments, such as the increasing availability of open-source LLMs, more awareness around the need for representative training data, and techniques to adapt LLMs to different domains through retrieval augmentation or additional models (perhaps maintained and governed at the local level). I focus here on the retrieval-augmentation paradigm and how this may be a viable and appropriate way to adapt language models to the needs of individual communities. Specifically, it is a paradigm that works well with the existing nature of libraries as community repositories and curators of knowledge.

The Retrieval-augmentation Paradigm

To offset issues of hallucination, researchers have proposed retrieval augmentation as a way to encourage more faithful, attributable, and accurate model output (Lewis et al., 2020). Instead of relying on a model's parametric knowledge (what it learned during training), the retrieval augmented generation (RAG) paradigm acknowledges that training data is limited, and proposes to augment the model at the time of use with additional information. Models can then combine retrieved information with their parametric knowledge to produce more useful outputs. This can address cases where the information needed to provide an answer is missing from the model's training data or simply out of date. Additionally, because we know the origin of the retrieved information, we can attribute model-generated content directly to primary and/or secondary sources, which makes it easier to assess the credibility of the model's statements. As an example of how this might work: take the FDA's recent March approval of a new drug for treating resistant hypertension, aprocitentan.[2] A model trained on data before this time may be able to infer that approval is likely, based on prior publications documenting the drug's effectiveness in clinical trials (Schlaich et al., 2022), but it cannot be certain. In cases like these, retrieval augmentation would bridge the gap in knowledge and allow us to identify new publications — e.g., news articles, reports, press releases etc. — documenting the approval.

Using Retrieval Augmentation to Achieve Community Relevance

Libraries are community organizations with deep knowledge of their patrons and their needs, and have the power to acquire print, digital, and other resources to serve those needs. We should be leveraging the role of libraries as curators of information resources as a bridge between their communities and Generative AI technologies. AI and LLMs should be able to access content on behalf of the members of the communities they serve, through retrieval augmentation, in order to produce more relevant and useful outputs.

How to manage and implement this type of access is an open question. To make physical or subscription content usable for retrieval augmentation, these materials must be digitized and parsed into machine-readable units of text. If shouldered by individual libraries or community archives, the time and resource costs would be untenable, in addition to producing lots of redundant work. Crafting a centralized digitization repository is more straightforward for open access publications — though issues of funding and maintenance would still need to be addressed. But even with all the efforts made in that space in recent years, a majority of scholarly materials still require fee-based or subscription access. Publishers managing closed access materials would need to push such content to different institutions based on the terms of their subscription agreements. After access is obtained, local instantiations of LLMs could be hooked up to what is institutionally accessible as an additional and unique-to-that-community source of data for retrieval augmentation. Current standards for data exchange[3] would be sufficient for the transfer of scholarly works between centralized and local databases, but additional clauses would be necessary to communicate appropriate use of the content by language models (i.e., consumption, attribution, distribution, adaptation), the scope and duration of use, as well as data transformations to enable efficient and reliable retrieval.

Organizations already positioned as community repositories of knowledge should lead the charge. These are suitable places for the deployment of Generative AI systems adapted to the local community, as well as for educating those communities around appropriate uses and limitations of AI. While major contenders in Generative AI development race to make their models more powerful and performant, libraries and community organizations should focus on creating adaptations and connections to make these systems usable for their patrons. If taking the approach of adapting a general-purpose AI model, an organization would need to create the data infrastructure to make information resources available to these models, while maintaining the ability to define the scope of use for each resource.

An alternate approach might be to develop models of one's own, trained for specific tasks and goals using the available data. While this seems like a weighty ask, the difficulty of training and maintaining such a model is dropping as these technologies mature; and this option will become much more feasible over time. As for why this is an attractive approach, in many cases researchers have demonstrated that training smaller models using curated training data can lead to better or comparable performance when compared to adapting a larger general-purpose model for the same tasks (Li et al., 2023; Jiang

et al., 2023). Taking such an approach allows an organization to maintain control over the goals of these technologies, define the scope of abilities based on community needs, and address shortcomings quickly, while freeing the organization from the whims and decisions of distant and unresponsive tech giants.

References

Barrett R. Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization Effects of Large Language Models on Human Creative Ideation. *ArXiv* abs/2402.01536 (2024).

Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2022. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Transactions on Computer-Human Interaction* 30 (2022), 1 – 38.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, et al. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *ArXiv* abs/2204.05862 (2022).

Christo Buschek and Jer Thorp. [n.d.]. Models All the Way Down. https://knowingmachines.org/models-allthe-way. ([n. d.]).

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. MARG: Multi-Agent Review Generation for Scientific Papers. *ArXiv* abs/2401.04259 (2024).

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. *ArXiv* abs/1905.12516 (2019).

Jesse Dodge, Ana Marasovic, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Conference on Empirical Methods in Natural Language Processing*.

Vinitha Gadiraju, Shaun K. Kane, Sunipa Dev, Alex S Taylor, Ding Wang, Emily Denton, and Robin N. Brewer. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (2023).

Sourojit Ghosh and Aylin Caliskan. 2023. ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (2023).

John Giorgi, Amanpreet Singh, Doug Downey, Sergey Feldman, and Lucy Lu Wang. 2024. TOPICAL: Topic Pages Automagically. *NAACL System Demonstrations* (2024).

Amelia Glaese, Nathan McAleese, Maja Trkebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *ArXiv* abs/2209.14375 (2022).

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothee Lacroix, and´ William El Sayed. 2023. Mistral 7B. *ArXiv* abs/2310.06825 (2023).

Hannah Rose Kirk, Andrew M. Bean, Bertie Vidgen, Paul Rottger, and Scott A. Hale. 2023. The Past, Present¨ and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values. *ArXiv* abs/2310.07629 (2023).

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NeurIPS* abs/2005.11401 (2020).

Yuan-Fang Li, Sebastien Bubeck, Ronen Eldan, Allison Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks Are All You Need II: phi-1.5 technical report. *ArXiv* abs/2309.05463 (2023).

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2020. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture* 25 (2020), 700 – 732.

Vishakh Padmakumar and He He. 2023. Does Writing with Language Models Reduce Content Diversity? *ArXiv* abs/2309.05196 (2023).

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Annual Meeting of the Association for Computational Linguistics*.

Markus P. Schlaich, Marc Bellet, Michael A. Weber, Parisa Danaietash, George L. Bakris, John M. Flack, et al. 2022. Dual endothelin antagonist aprocitentan for resistant hypertension (PRECISION): a multicentre, blinded, randomised, parallel-group, phase 3 trial. *The Lancet* 400 (2022), 1927–1937.

Donghoon Shin, Lucy Lu Wang, and Gary Hsieh. 2024. From Paper to Card: Transforming Design Implications with Generative AI. In *Proceedings of the 2024 ACM CHI conference on Human Factors in Computing Systems*.

Nikhil Singh, Lucy Lu Wang, and Jonathan Bragg. 2024. FigurA11y: AI Assistance for Scientific Alt Text Writing. In *Proceedings of the 2024 ACM International Conference on Intelligent User Interfaces (IUI)*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. *ArXiv* abs/2306.01693 (2023).

Travis Zack, Eric P. Lehman, Mirac Suzgun, Jorge Alberto Rodriguez, Leo Anthony Celi, Judy Gichoya, Daniel Jurafsky, Peter Szolovits, D. Bates, E. Raja-Elie, Abdulnour, Atul Janardhan Butte, and Emily Alsentzer. 2023. Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare. In *medRxiv*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Conference on Empirical Methods in Natural Language Processing*.

Dennis Zyska, Nils Dycke, Jan Buchmann, Ilia Kuznetsov, and Iryna Gurevych. 2023. CARE: Collaborative AI-Assisted Reading Environment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Danushka Bollegala, Ruihong Huang, and Alan Ritter (Eds.). Association for Computational Linguistics, Toronto, Canada, 291–303.

endnotes on page 24