Engineering the Reproducible Literature Review Section for Scholarly Publications and Grant Applications

Yuanxi Fu[1], **Jodi Schneider**[1, 2]

[1]School of Information Sciences, University of Illinois Urbana-Champaign
[2]Harvard Radcliffe Institute for Advanced Study, Harvard University
fu5@illinois.edu, jodi@illinois.edu and jschneider@pobox.com

Abstract

Large language models (LLMs) have the potential to transform the synthesis of scientific knowledge. While literature review sections generated with the assistance of LLMs raise legitimate concerns due to limitations of the technology, researchers' interest in automation brings a rare opportunity to change scientific practice to increase the robustness and reproducibility of literature review sections. This position paper proposes a digital object called a reproducible literature review section containing a discourse graph and a bibliography in a computable format. By leveraging technologies including query-focused summarization with retrieval-augmented generation, discourse graphs, and scholarly big data infrastructure, the reproducible literature review section could address trust issues with human-generated literature review sections and LLM-generated text.

Introduction

Literature review is important because it underpins the progress of science: the synthesis of old facts helps prioritize which new facts should be generated. Literature review comes in different forms (Grant and Booth 2009). The most familiar form is the (narrative) literature review section (also called 'related work' or 'background') in scholarly publications and grant applications. Another well-known type of literature review is the systematic review, which synthesizes all available evidence to answer a given question. Researchers take extra methodological care when producing systematic reviews to ensure their quality and avoid biases (Cooper, Hedges, and Valentine 2019), enabling systematic reviews to help justify critical decisions, such as in medicine and public policy. Finally, scholarly big data infrastructure (e.g., citation databases, academic knowledge graphs) combined with modern data analytics have engendered new forms of research synthesis, such as scientometrics reviews (e.g., Tapeh and Naser 2023) and systematic maps (James, Randall, and Haddaway 2016).

Large language models (LLMs) have the potential to transform research synthesis once again. In this position paper, we focus on the literature review sections in scholarly publications and grant applications because they are prevalent and LLMs pose an attractive alternative for busy researchers preparing them. A recent large-scale survey found that researchers mention literature review as one of the "benefits/usefulness" of LLMs, although it did not report statistics on the actual degree of integration of LLMs in literature review writing (Liao et al. 2024).

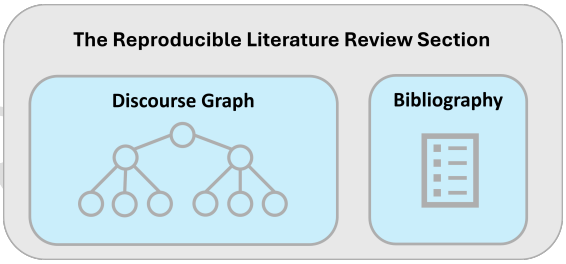

Figure 1: The reproducible literature review section

Literature review sections generated with the assistance of LLMs raise legitimate concerns due to limitations of the technology including fabricated references (Walters and Wilder 2023), inaccurate summarization (Tang et al. 2023) and socio-technical concerns such as proliferation of low quality research (Bail 2024) and erasure of intellectual individuality (Anderson, Shah, and Kreminski 2024). Yet researchers' interest in automation brings a rare opportunity to change scientific practice.

This position paper proposes a *reproducible literature review section* that could address trust issues with both human-generated literature review sections and LLM-generated text (Figure 1). We envision the reproducible literature review section as a digital object comprised of (1) a bibliography of the references cited in the literature review section in a computable format (e.g., CSV, BibTeX, RIS) and (2) a discourse graph (Chan et al. 2024). Authors would be free to use LLMs to craft their literature review section, provided that they submit a reproducible literature review section to accompany the manuscript for a gatekeeper (such as a peer reviewer, journal editor, or program officer) to inspect and verify the resulting literature review section.

Below we first describe trust issues with human-generated literature review sections and LLM-generated content in scholarly publications and grant applications. Second, we review existing technologies that enable a reproducible literature review section. Third, we consider nondeterminism

and drifting, two significant challenges for the reproducibility of research produced with LLMs, and their ramifications for our proposal. Finally, we describe how the reproducible literature review section could help build trust.

Trust Issues with Human-generated Literature Review Sections

The literature review section in a scholarly publication or grant application meshes objectivity with subjectivity. Ideally, it should be a truthful narrative of what is known. Yet it must also provide the basis for the work reported or proposed, giving strong incentives for citation cherry-picking and citation distortions (Greenberg 2009).

The trustworthiness of any literature review depends heavily upon accurate citations. Yet inaccurate citations are common. Inaccurately representing the source cited (a phenomenon known as 'quotation error') affects 25.4% of medical citations, according to a systematic review and meta-analysis synthesizing 28 studies (Jergas and Baethge 2015). Inaccurate citations can have profound impact on science: a chain of inaccurate citations can convert a hypothesis to "fact" (Greenberg 2009). Inaccurate citations can impact society, too: inaccurate citations to a letter published in the *New England Journal of Medicine* in 1980 may have contributed to the opioid crisis (Leung et al. 2017). Therefore we need accurate citations.

Trust Issues with LLM-generated Content in Scholarly Publications and Grant Applications

LLM-generated content creates an unprecedented challenge for readers, reviewers, and editors, who are rightly suspicious because such content frequently contains factual errors (including fabricated references) and can make authors unknowingly infringe upon others' intellectual property (Brainard 2023). The *Science* journal family initially issued a ban on text or images generated by ChatGPT or other AI tools (Thorp 2023) but then reverted their strong stance to a disclosure-based policy.[1] Increasingly, journals, conferences, and funders have written policies requiring disclosure of generative AI usage. Yet disclosure alone cannot address mistrust toward LLM-generated content since it does not address root issues such as questions regarding the content's trustworthiness and the level of authors' active involvement in the review creation process.

Technologies Enabling a Reproducible Literature Review Section

Query-focused Summarization (QFS) with Retrieval-augmented Generation (RAG) LLMs made breakthroughs in the computational task of text summarization (Goyal, Li, and Durrett 2023; Pu, Gao, and Wan 2023; Edge et al. 2024). Query-focused summarization (QFS) over multiple documents (Roy and Kundu 2023) is akin to the human review process, in which human researchers read several publications, extract several pieces of relevant information, and synthesize these pieces of information into a text snippet. For instance, authors often need to mention other solutions to the problem addressed (P) in their literature review section (Teufel 2014). To automatically generate this snippet, they can prompt a LLM Retrieval-Augmented Generation (RAG) agent (Skarlinski et al. 2024) to answer: what are the solutions for P?, based on a folder of publications the authors select.

Query-focused summarization of scientific publications (in PDF format) has been realized in PaperQA2 (Skarlinski et al. 2024), which performs publication search, retrieval, and summarization. Literature search is not necessary for our envisioned reproducible literature review system, because we envision the authors supplying all references. To confine the answers to a set of publications selected by the authors, QFS operations need to be coupled with RAG. An adaption of PaperQA2 would be part of a system that can be used by (1) authors to generate text snippets from query-focused summarizations over several specified references and (2) gatekeepers to verify text snippets.

Discourse graph A discourse graph (Chan et al. 2024) can be used to explicitly represent the discourse structure of a literature review section. Figure 2, adapted from Chan et al. (2024), shows a toy example; a real literature review section will be more complex.

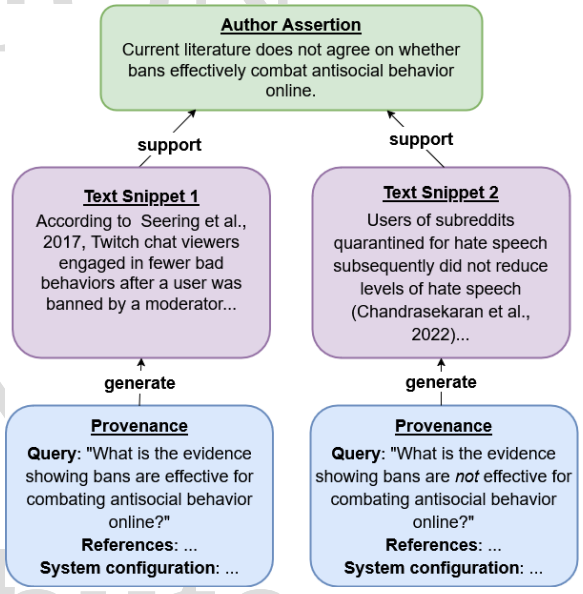

Figure 2: A discourse graph showing the discourse structure of a hypothetical literature review section adapted from Figure 1 of Chan et al. (2024).

A bibliography of the references cited in the literature review section, in a computable format, checked by scholarly big data infrastructure A bibliography in a computable format (e.g., CSV, BibTeX, RIS) can be readily processed and checked against scholarly big data infrastructure (Waltman and Larivière 2020) to verify that the cited

[1] https://www.science.org/content/page/science-journals-editorial-policies

references exist. The bibliography file can also be characterized with descriptive statistics, particularly the bibliography's distribution of publication years, publication types, and topics (Ye 2023). A gatekeeper with data science skills can also query data sources (e.g., OpenAlex, Crossref) for citation networks, author networks, and funding data. A close-knit author network may reflect a lack of intellectual diversity, while a network's over-reliance on industry funding may imply sponsor bias.

Towards a Standard for Reproducibility

State-of-the-art LLMs are nondeterministic—meaning that the same instruction generates different outputs at different runs—because they use probabilistic random sampling to generate the next token and run on distributed systems where it is difficult to coordinate random seeds among different subsystems in order to produce a deterministic output (Blackwell, Barry, and Cohn 2024).[2] Nondeterminism threatens the reproducibility of LLM research (Ouyang et al. 2025; Aronson et al. 2024; Blackwell, Barry, and Cohn 2024; Staudinger et al. 2024). The ramification for our proposal is that even the authors (not to mention gatekeepers) cannot be guaranteed to regenerate identical text snippets locally using provenance information stored in the discourse graph.

Another issue is drifting, which refers to the changing behavior of LLMs over longer time periods (e.g., days to months) (Aronson et al. 2024). LLM providers continuously update and deliver their models to users (La Malfa et al. 2024). Continuous delivery, which is a standard practice in Software-as-a-Service, creates a serious challenge for reproducibility when applied to Language-Models-as-a-Service (La Malfa et al. 2024). The ramification is that even if authors achieve some sort of reproducibility at the time of submission, due to drifting, gatekeepers may still observe variation at the time of peer review.

How can the literature review section be reproducible when LLMs exhibit nondeterminism and drifting?

Our immediate response is that two text snippets generated by the LLMs with the same instructions at different times may not be identical, but they can still convey the same meaning. The reproducibility is judged based upon whether replacing one set of text snippets with another maintains the argument structure of the given discourse graph. For example, replacing *Text Snippet 1* in Figure 2 with a snippet such as "*no evidence showing bans are effective for combating antisocial behaviors online*" disrupts the argument structure since the statement no longer supports the author assertion that "*current literature does not agree on whether bans effectively combat antisocial behavior*," thus making the literature review section irreproducible.

However, our immediate response is likely insufficient. Reproducibility in general is a complex issue (McPhillips et al. 2019), and reproducibility of narrative literature reviews is understudied because the knowledge-intensive process of generating a narrative of existing research has been difficult to make transparent (Cram, Templier, and Paré 2020). We need to establish a standard for reproducibility for literature reviews generated with the assistance of LLMs, one that considers the inherent difficulty of achieving computational repeatability with LLMs. This standard will likely incorporate a taxonomy of reproducibility so that lower-level irreproducibility (e.g., variations in text snippets due to nondeterminism or drifting) will not affect higher-level reproducibility (e.g., overall argument structure). Cohen et al. (2018) have proposed such a taxonomy for natural language processing research, which can be instructive for us.

Building Trust with the Reproducible Literature Review Section

Figures 3 and 4 envision how the reproducible literature review section will function. In the author workflow (Figure 3), it is important for authors to achieve reproducibility locally and save proof of reproducibility (i.e., two additional discourse graphs). The effort paid to reproduce a piece of text will educate researchers about LLMs' nondeterministic nature, increasing their caution when using LLMs for other tasks. In the gatekeeper workflows (Figure 4), the reproducibility test aims not only at holding authors accountability but also at a continuous and distributed assessment of trustworthiness of LLMs as our research partners, as drifting will be noticed and recorded. The reproducible literature review section helps address trust issues in the following ways.

Preventing intentional misquotation Gatekeepers can generate text snippets (such as those shown in Figure 2) using the provenance provided in the discourse graph, and this possibility may deter authors from misrepresenting references to serve their own interests. However, since LLMs can still introduce quotation errors, a reproducible literature review section is still susceptible to quotation errors.

Preventing fabricated references Gatekeepers can verify the existence of the references using scholarly big data infrastructure. Moreover, authors are more likely to verify references themselves when they expect their bibliography to be inspected for fabricated references.

Assessing the author's level of active participation in the creation process The discourse graph reflects authors' participation in crafting the literature review section. The discourse graph can be used to evaluate the authors' rigor: A discourse graph showing evidence for and against author assertions and providing detailed rebuttals has better argument quality than a discourse graph that merely shows evidence supporting author assertions. The complexity of their discourse graph differentiates authors who carefully considered the structuring of the section from those who just prompted the LLM to write the literature review section for them.

Detecting cherry-picked literature The historical problem of detecting cherry-picked literature may finally be solvable. When the pertinent literature is small, query-focused summarization can provide a global answer over the entire corpus. When the pertinent literature is large and sampling is

[2] Setting temperature to 0 does not eliminate nondeterminism (Ouyang et al. 2025; Aronson et al. 2024).

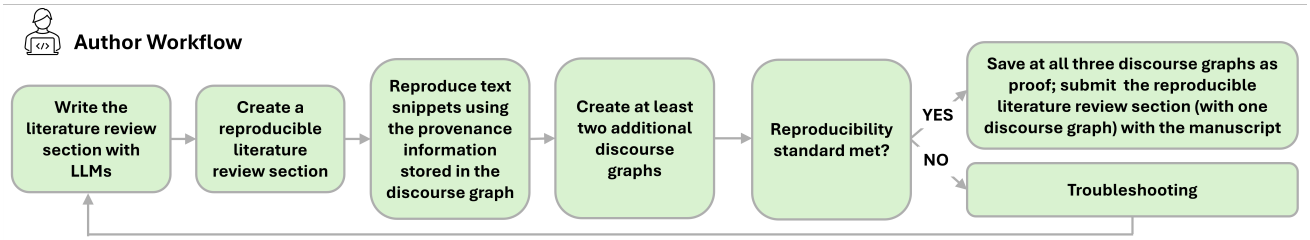

Figure 3: The author workflow

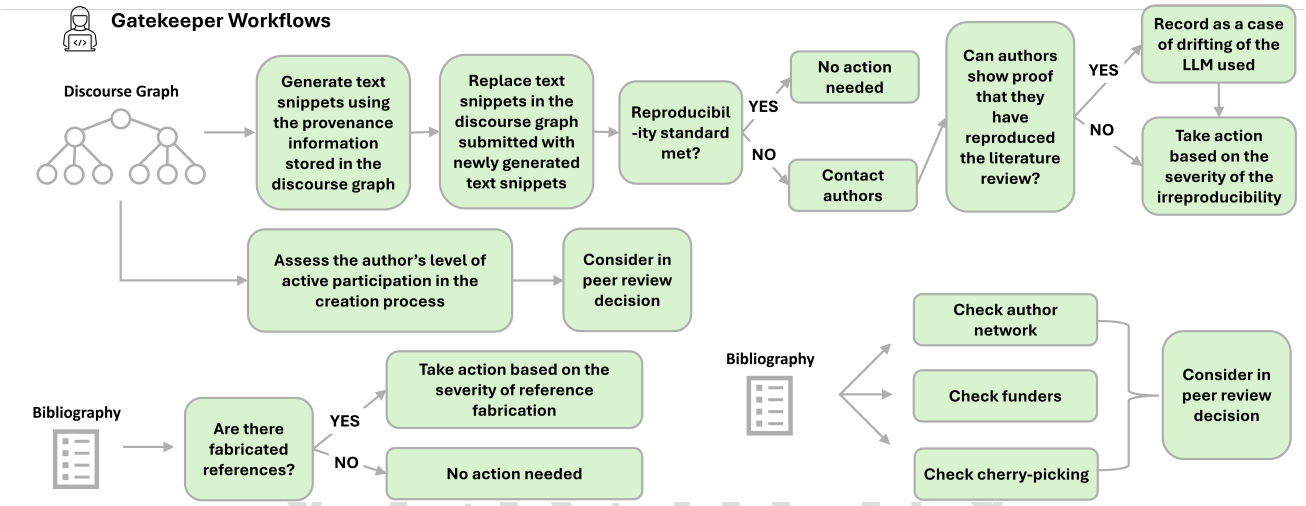

Figure 4: The gatekeeper workflows

required, a gatekeeper can apply query-focused summarization to several selected samples and verify whether the outcome is robust against different samples. We note that the search and retrieval in PaperQA2 is not needed for testing reproducibility, however, these functions may be very useful for checking cherry-picking since they allow a gatekeeper to explore a broader range of references beyond those the authors provided.

Conclusions

Rarely does a scholarly publication or grant proposal completely omit a review of the literature. Using LLMs to write these literature review sections is an attractive alternative for busy researchers and possibly already a practice among the tech-savvy. Literature review sections generated with the assistance of LLMs raise legitimate concerns, but we see a rare opportunity for improving literature reviews, leveraging researchers' interest in their automation. We propose a digital object called a reproducible literature review section containing (1) a bibliography of the references cited in the literature review section in a computable format and (2) a discourse graph. Authors should be free to use LLMs to craft their literature review section provided that they submit a digital object that helps gatekeepers inspect and verify the resulting literature review section. Future work should develop a reproducibility standard for literature review sections generated with the assistance of LLMs; rather than computational repeatability the standard will consider argument-level reproducibility. The reproducible literature review section leverages existing technologies including query-focused summarization with retrieval-augmented generation, discourse graphs, and scholarly big data infrastructure. It could address trust issues with human-generated literature reviews and LLM-generated text.

Acknowledgments

This research was supported by NSF 2046454 CAREER: Using network analysis to assess confidence in research synthesis. Jodi Schneider was supported in part as the 2024–2025 Perrin Moorhead Grayson and Bruns Grayson Fellow, Harvard Radcliffe Institute for Advanced Study. Thanks to Timothy McPhillips and Lars Vilhuber for discussions on reproducibility. Thanks to Tim Clark, Peter Darch, Tobias Kuhn, Cyril Labbé, Joe Menke, Corinne McCumber, Theodore Ledford, Madelyn Sanfilippo, Maria Janina Sarol, Aaron Tay, Anita de Waard, and Heng Zheng for feedback.

References

Anderson, B. R.; Shah, J. H.; and Kreminski, M. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*, 413–425.

Aronson, S. J.; Machini, K.; Shin, J.; Sriraman, P.; Hamill, S.; Henricks, E. R.; Mailly, C. J.; Nottage, A. J.; Amr, S. S.; Oates, M.; and Lebo, M. S. 2024. GPT-4 Performance, nondeterminism, and drift in genetic literature review. *NEJM AI*, 1(9).

Bail, C. A. 2024. Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21): e2314021121.

Blackwell, R. E.; Barry, J.; and Cohn, A. G. 2024. Towards reproducible LLM evaluation: quantifying uncertainty in LLM benchmark scores. arXiv:2410.03492.

Brainard, J. 2023. As scientists explore AI-written text, journals hammer out policies. *Science*, 379(6634): 740–741.

Chan, J.; Akamatsu, M.; Vargas, D.; Kawerau, L.; and Gartner, M. 2024. Steps towards an infrastructure for scholarly synthesis. arXiv:2407.20666.

Cohen, K. B.; Xia, J.; Zweigenbaum, P.; Callahan, T.; Hargraves, O.; Goss, F.; Ide, N.; Névéol, A.; Grouin, C.; and Hunter, L. E. 2018. Three dimensions of reproducibility in natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Cooper, H.; Hedges, L. V.; and Valentine, J. C., eds. 2019. *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation, 3rd edition.

Cram, W. A.; Templier, M.; and Paré, G. 2020. (Re)considering the concept of literature review reproducibility. *Journal of the Association for Information Systems*, 21(5): 1103–1114.

Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; and Larson, J. 2024. From local to global: a graph RAG approach to query-focused summarization. arXiv:2404.16130.

Goyal, T.; Li, J. J.; and Durrett, G. 2023. News summarization and evaluation in the era of GPT-3. arXiv:2209.12356.

Grant, M. J.; and Booth, A. 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2): 91–108.

Greenberg, S. A. 2009. How citation distortions create unfounded authority: analysis of a citation network. *BMJ*, 339: b2680.

James, K. L.; Randall, N. P.; and Haddaway, N. R. 2016. A methodology for systematic mapping in environmental sciences. *Environmental Evidence*, 5(1): 7.

Jergas, H.; and Baethge, C. 2015. Quotation accuracy in medical journal articles—a systematic review and meta-analysis. *PeerJ*, 3: e1364.

La Malfa, E.; Petrov, A.; Frieder, S.; Weinhuber, C.; Burnell, R.; Nazar, R.; Cohn, A.; Shadbolt, N.; and Wooldridge, M. 2024. Language-Models-as-a-Service: overview of a new paradigm and its challenges. *Journal of Artificial Intelligence Research*, 80: 1497–1523.

Leung, P. T.; Macdonald, E. M.; Stanbrook, M. B.; Dhalla, I. A.; and Juurlink, D. N. 2017. A 1980 letter on the risk of opioid addiction. *New England Journal of Medicine*, 376(22): 2194–2195.

Liao, Z.; Antoniak, M.; Cheong, I.; Cheng, E. Y.-Y.; Lee, A.-H.; Lo, K.; Chang, J. C.; and Zhang, A. X. 2024. LLMs as research tools: a large scale survey of researchers' usage and perceptions. arXiv:2411.05025.

McPhillips, T.; Willis, C.; Gryk, M. R.; Nuñez-Corrales, S.; and Ludäscher, B. 2019. Reproducibility by other means: transparent research objects. In *2019 15th International Conference on eScience*, 502–509.

Ouyang, S.; Zhang, J. M.; Harman, M.; and Wang, M. 2025. An empirical study of the non-determinism of ChatGPT in code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2).

Pu, X.; Gao, M.; and Wan, X. 2023. Summarization is (almost) dead. arXiv:2309.09558.

Roy, P.; and Kundu, S. 2023. Review on query-focused multi-document summarization (QMDS) with comparative analysis. *ACM Computing Surveys*, 56(1): 1–38.

Skarlinski, M. D.; Cox, S.; Laurent, J. M.; Braza, J. D.; Hinks, M.; Hammerling, M. J.; Ponnapati, M.; Rodriques, S. G.; and White, A. D. 2024. Language agents achieve superhuman synthesis of scientific knowledge. arXiv:2409.13740.

Staudinger, M.; Kusa, W.; Piroi, F.; Lipani, A.; and Hanbury, A. 2024. A reproducibility and generalizability study of large language models for query generation. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 186–196.

Tang, L.; Sun, Z.; Idnay, B.; Nestor, J. G.; Soroush, A.; Elias, P. A.; Xu, Z.; Ding, Y.; Durrett, G.; Rousseau, J. F.; Weng, C.; and Peng, Y. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6: 158.

Tapeh, A. T. G.; and Naser, M. Z. 2023. Artificial intelligence, machine learning, and deep learning in structural engineering: a scientometrics review of trends and best practices. *Archives of Computational Methods in Engineering*, 30: 115–159.

Teufel, S. 2014. Scientific argumentation detection as limited-domain intention recognition. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, volume 1341. CEUR Workshop Proceedings.

Thorp, H. H. 2023. ChatGPT is fun, but not an author. *Science*, 379(6630): 313.

Walters, W. H.; and Wilder, E. I. 2023. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13(1): 14045.

Waltman, L.; and Larivière, V. 2020. Special issue on bibliographic data sources. *Quantitative Science Studies*, 1(1): 360–362.

Ye, Z. 2023. Using data visualization to characterize bibliographies. University of Illinois Undergraduate Research Symposium, https://hdl.handle.net/2142/117485.