Citation Parsing and Analysis with Language Models

parth sarin[1], Juan Pablo Alperin[2]

[1] *Stanford University*

[2] *Simon Fraser University*

Purpose

Science is not an equal playing field. There is a large—though shrinking—inequality between knowledge production in the Global North and Global South. To properly understand this gap, national research capacities, and the broader power dynamics in global research contributions, it is necessary to have high quality and complete indexes of scholarly production. Accurate and complete metascientific information is also central to decision making at every level of the research system from the tenure and promotion decisions about individual faculty, to institutional resource allocation, and to national science policies. Such information is not indexed today: It has been observed for decades that research in the Global South is significantly underrepresented in supposedly global indexes, databases, and search engines for scholarly work (Cetto et al., 1998; Khanna et al., 2022; Mongeon & Paul-Hus, 2016). So, for over 50 years, decisions about funding, tenure, collaboration, and governance have been made without the same type of reliable quantitative data that exists to tabulate research in the Global North.

Fortunately, a number of projects are seeking to address global gaps in indexing and discoverability. New bibliographic databases like OpenAlex are taking a more inclusive indexing approach that has allowed them significantly outperform existing databases in terms of coverage (Alperin et al., 2024; Culbert et al., 2024; Jiao et al., 2023). However, despite the significantly greater coverage of works, including those from the Global South, there continues to be an enormous gap in the indexing of references (Alperin et al., 2024; Culbert et al., 2024). While more complete bibliographic records are useful for understanding knowledge production, references and citations allow us to better understand the connections and circulation of knowledge, and are thus an essential aspect of research information.

As such, our project seeks to contribute to the efforts of OpenCitations to develop parse-and-extract technologies for references and citation data. Our approach builds on developments in natural language processing to create and evaluate tools for the automated parsing of citations in published articles. We evaluated a wide variety of models and prompting techniques on a labeled dataset of citations. For reference, we compared these evaluations to existing state of the art tools. Then, we conducted distribution tests to assess whether small models—which could conceptually run in very low-resource environments—could be trained for citation parsing.

Methods

We began by assembling a dataset of citations in plain, formatted text, along with the same citations marked up in JATS format. We used two sources for these citations: PKP assembled an XML Markup Evaluation Corpus in 2016 with DOCX versions of 850 published articles and their JATS XML markup (Garnett, 2016); and the Open Research Europe (ORE) platform hosts PDF versions of 848 published articles and their JATS XML markup. We extracted the marked up version of each citation from the reference list in the JATS XML markup and matched it to a (markdown-formatted) plaintext citation extracted from the article. Extraction of plaintext citations was performed using Llama-3.1-8B-Instruct, using the prompt format below, and we programmatically verified that each of the citations appeared in the article to prevent hallucinated citations. Matching was performed by computing pairwise edit distances between each plaintext citation and text-only versions of each marked up citation, then matching in order of increasing distance. In the end, we had a dataset of citations formatted in markdown along with their marked up JATS versions. We sampled 1,000 data points from each of the two corpora and did our testing on a final dataset with 2,000 citations.

Prompt 1. Prompt for citation extraction

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are an expert annotator that specializes in reading academic articles and
isolating their bibliography. You will see the text of an academic article and
you should write out a list of the citations in the article, one on each line.
Copy the citations from the article
verbatim.<|eot_id|><|start_header_id|>user<|end_header_id|>

{article_text}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

We tested fourteen popular edge language models: deepseek-ai/DeepSeek-R1-Distill-Qwen-14B, deepseek-ai/DeepSeek-R1-Distill-Qwen-7B, Qwen/Qwen2.5-7B-Instruct, Qwen/Qwen2.5-3B-Instruct, Qwen/Qwen3-4B, Qwen/Qwen3-1.7B, Qwen/Qwen3-0.6B, microsoft/phi-4, microsoft/Phi4-mini-instruct, meta-llama/Llama-3.1-8B-Instruct, and meta-llama/Llama-3.2-3B-Instruct. For each model we tested its performance with and without chain-of-thought prompting.

Results

This project is currently underway, with results expected by May 2025. At the workshop, and in the updated version of this manuscript, we will report the results of our analysis for the three aforementioned language models as well as Grobid, AnyStyle, and other state-of-the-art citation parsing tools.

Accuracy information will include details on eight citation fields designated by the National Library of Medicine as the most important components of a <mixed-citation> element. These sub-fields are generally required for inclusion in indexing in both Crossref and PubMed Central. A summary of the reported fields is shown in Table 1.

Table 1. Fields for markup accuracy evaluation

Evaluation specification

Field	Description	Edit distance threshold	Type
Model name	The name of the model being evaluated	N/A	Text
Average validation accuracy	The average edit distance that the model's achieved during the adaptation step on the validation set	N/A	Number
Coverage	The percentage of citations in the evaluation set for which the model returned valid markup	N/A	Percentage
`<source>` accuracy	The average accuracy for the `<source>` field prediction on the valid responses	5	Percentage
`<volume>` accuracy	The average accuracy for the `<volume>` field prediction on the valid responses	0	Percentage
`<issue>` accuracy	The average accuracy for the `<issue>` field prediction on the valid responses	0	Percentage
`<fpage>` accuracy	The average accuracy for the `<fpage>` field prediction on the valid responses	0	Percentage
`<surname>` accuracy	The average accuracy for the first `<surname>` field prediction on the valid responses	5	Percentage
`<year>` accuracy	The average accuracy for the `<year>` field prediction on the valid responses	0	Percentage

Value

Our intent is to use the most accurate version of this model to extract and parse the citations of millions of research articles published by journals using OJS and to release this corpus in the OpenCitations format for further citation analysis. Similarly, we expect that the model—and accompanying extraction and markup pipeline—can be used to extract and parse citations from future articles. Finally, we will explore ways of including such models as part of the article production tools in OJS. Such a deployment would better enable automated JATS markup of citations and lower the cost of the kind of high-fidelity XML production for tens of thousands of journals around the world.

References

Alperin, J. P., & Fischman, G. (Eds.). (2015). *Made in Latin America: Open access, scholarly journals, and regional innovations*. CLACSO. https://biblioteca-repositorio.clacso.edu.ar/handle/CLACSO/16332

Alperin, J. P., Portenoy, J., Demes, K., Larivière, V., & Haustein, S. (2024). *An analysis of the suitability of OpenAlex for bibliometric analyses* (arXiv:2404.17663). arXiv. https://doi.org/10.48550/arXiv.2404.17663

Cetto, A. M., Alonso-Gamboa, O., Altbach, P., & Teferra, D. (1998). Scientific and scholarly journals in Latin America and the Caribbean. *Knowledge Dissemination in Africa: The Role of the Scholarly Journal, Ed. Philip G. Altbach and Damtew Teferra*, 99–126.

Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2024). *Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus* (arXiv:2401.16359). arXiv. http://arxiv.org/abs/2401.16359

Garnett, A. (2016, April 18). *The XML Markup Evaluation Corpus*. Public Knowledge Project. https://pkp.sfu.ca/2016/04/18/the-xml-markup-evaluation-corpus/

Jiao, C., Li, K., & Fang, Z. (2023). How are exclusively data journals indexed in major scholarly databases? An examination of four databases. *Scientific Data*, *10*(1), 737. https://doi.org/10.1038/s41597-023-02625-x

Khanna, S., Ball, J., Alperin, J. P., & Willinsky, J. (2022). Recalibrating the scope of scholarly publishing: A modest step in a vast decolonization process. *Quantitative Science Studies*, *3*(4), 912–930. https://doi.org/10.1162/qss_a_00228

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5