Breaking Boundaries in Citation Parsing: A Comparative Study of Generative LLMs and Traditional Out-of-the-box Citation Parsers

Iana Atanassova[1,2,*,†], Marc Bertin[3,†]

[1]*Université de Franche-Comté, CRIT, France*
[2]*Institut Universitaire de France (IUF)*
[3]*ELICO, Université Claude Bernard Lyon 1, France*

Abstract

The task of citation string parsing has been the focus of many efforts. Traditional tools explicitly designed to parse bibliographic information, such as Bilbo, Grobid, and Parscit, have long been established in the academic landscape. Recently, with the emergence of general conversational LLMs (Large Language Models) such as OpenAI's ChatGPT and Llama, an interesting question arises: can such language models, originally developed for natural language understanding (NLU), be employed to efficiently process bibliographies, and how would their performance for this task compare to that of dedicated bibliographic parsing tools? In this article, we propose an experiment to measure the ability of LLMs to analyse citation strings in different citation styles. We use a synthetic dataset with 12 different citation styles. We evaluate the output of two generative LLMs, ChatGPT 3.5 and Llama 2 7B, and two out-of-the-box citation parsers, CERMINE and Neural ParsCit. The results show that the LLMs tend to outperform the citation parsers for all citation styles and labels.

Keywords

Generative LLMs, Citation string parsing, Reference parsing, BibTeX, ChatGPT, Neural ParsCit, CERMINE, Llama, References

1. Introduction

A remaining challenge in Bibliometrics and scholarly publishing is the parsing of bibliographic references, which is an important step in processing full-text articles and linking them to bibliographic databases. The Open Science movement has contributed to this field by making large corpora of publications available. However, in order to make the scientific literature more accessible and easier to navigate, it is necessary to develop efficient tools for linking full-text articles and their corresponding references, with the aim of creating corpora. This issue is not only of concern to bibliometric research, but is related to wider real-world needs, especially in

Joint Proceedings of BIR 2024: 14th International Workshop on Bibliometric-enhanced Information Retrieval and IR4U2 2024: 1st Workshop on Information Retrieval for Understudied Users
*Corresponding author.
†These authors contributed equally.
iana.atanassova@univ-fcomte.fr (I. Atanassova); marc.bertin@univ-lyon1.fr (M. Bertin)
https://iana-atanassova.github.io/ (I. Atanassova); https://elico-recherche.msh-lse.fr/membres/marc-bertin (M. Bertin)
0000-0002-0877-7063 (I. Atanassova); 0000-0001-7116-9338 (M. Bertin)

times of crisis. For example, the COVID-19 Open Research Dataset (CORD-19) Database [1] is a free resource[1] of tens of thousands of scholarly articles about COVID-19, SARS-CoV-2, and related coronaviruses for use by the global research community.

PDF is currently the most widely used format for publishing scientific articles, although some publishers offer HTML access to their articles. However, obtaining structured text and bibliographic data from PDFs is a complex and error-prone process. The XML format offers specific tagsets for representing journal articles, the JATS (Journal Article Tag Suite) and NLM DTD. They are used for example by PubMed[2] and PLOS[3] who provide direct access to the articles in XML. The LaTeX format is also widely used in scientific publishing by many journals and preprint databases, such as arXiv. ArXiv hosts over two million scientific articles in eight fields, mostly in the Natural and Applied Sciences. The UnarXiv corpus [2] was constructed using the arXiv data in LaTeX format, using a method that avoids the distortions introduced by PDF processing. Processing peer-reviewed publications, beyond Pubmed and PLOS datasets, poses a considerable challenge, particularly in parsing citation strings from PDF files. This issue is also confronting the scientific publishing sector [3, 4].

Conversational LLMs (Large Language Models) have recently had a significant impact in many domains, particularly in coding. Thus, with the emergence of general chatbots such as OpenAI's GPT-3.5, which were initially developed for natural language understanding (NLU), an intriguing question arises:

Conversational Large Language Models (LLMs), such as OpenAI's GPT-3.5, have had a significant impact in various fields, , particularly in coding. While they are originally designed for Natural Language Understanding (NLU), an intriguing question arises:

> Can generative LLMs be employed to efficiently process bibliographies, and how would their performance for this particular task compare to that of dedicated citation string parsing tools?

This question is relevant for two reasons. First, the existence and easy access to conversational LLMs might render task-specific tools obsolete in the near future. Are we approaching this point? Second, the accessibility of conversational LLMs to a broad audience, including non-technical users, impacts academics and information science. Researchers, bibliometricians, librarians, and students could leverage these models' advanced parsing abilities through simple natural language prompts, thus democratizing the access to sophisticated bibliographic data management.

In this paper, we evaluate the effectiveness of conversational generative LLMs, specifically ChatGPT 3.5 and Llama 2 7B, in citation string parsing, by comparing their performance against traditional tools like CERMINE and Neural ParsCit. These tools were employed directly, with no additional training, to ensure accessibility and ease of use for non-specialists.

Our objective is to assess these parsers across a large variety of citation styles which reflect different academic disciplines. Existing datasets typically cover one or two disciplines with limited variation in citation styles, and with the Humanities notably underrepresented. To

[1] https://github.com/allenai/cord19
[2] https://www.ncbi.nlm.nih.gov/pmc/pmcdoc/tagging-guidelines/article/style.html
[3] https://plos.org/text-and-data-mining/

address this, we have developed a synthetic dataset utilizing the BibTeX format and the LaTeX biber package, allowing for a comprehensive representation of citation styles.

2. Citation String Parsing: State of the Art and Limitations

Over the last decade, many tools have been developed to carry out the task of citation string parsing, i.e. to produce structured bibliographic metadata from character strings that represent bibliographic references. The two main categories of approaches, as described in [5], are *Non-machine Learning based* and *Machine Learned (ML) based* Approaches. Non-machine Learning based Approaches include rule-based approaches, knowledge-based approaches, and template matching. Machine Learned based Approaches include Support Vector Machines (SVMs), Hidden Markov Models (HMM), Conditional Random Fields (CRF), and Deep Learning based approaches. The work of [6] proposes a state of the art and a study to compare *out-of-the-box* and *re-trained* ML and rule-based approaches. The results showed that ML approaches tend to outperform non-ML approaches. However, the study was limited to a specific set of metadata and did not include an in-depth evaluation of essential fields of the bibliographic references, such as title or authors.

There are several datasets available for training and evaluating citation parsers, but they are often limited to specific disciplines (see [5] for a complete analysis). For instance, Cora [7], CiteSeer [8], and Flux-CIM [9] are designed for use in Computer Science and Artificial Intelligence, while CS-SW[10] is intended for use in Semantic Web. GROTOAP2 [11] is based on articles from PubMed Central Open Access Subset, and was used for training the CERMINE citation parser [12].

There are two multi-domain datasets available: GROBID [13] and GIANT [14, 15]. GROBID was developed using the datasets cited above, but its evaluation is essentially based on life sciences and prepublications[4]. On the other hand, the GIANT dataset is a synthetic corpus of generated citation strings, designed to cover a wide range of citation styles[5].

The task of citation string parsing is an integral part of building large full-text annotated corpora of publications, such as The Semantic Scholar Open Research Corpus (S2ORC) [16] or ISTEX [17, 18]. S2ORC is a large corpus that contains 81.1 million English language academic papers from a wide range of disciplines. ISTEX is the largest repository of standardized scientific archives in France, serving the research community for documentary and TDM use. It contains over 27 million scientific publications spanning 700 years in all disciplines and in several languages. GROBID is a key component in both ISTEX and S2ORC's processing pipelines.

The diversity of scientific fields and citation practices plays an important role in citation string parsing. Current ML methods require large annotated corpora for model training. The tools perform well when trained on corpora adapted to their task. However, as noted by [5], the IEEE and ACM citation styles differ significantly from MLA, which is primarily used in the Humanities. The existence of numerous citation styles across various disciplines makes it difficult to identify and parse citation strings independently of the styles. At the same time, it appears that the datasets may not be large enough to encompass all styles required for the

[4] https://grobid.readthedocs.io/en/latest/Principles/.
[5] https://github.com/BeelGroup/

efficient training of the models. To address this limitation, [19] conducted a study comparing the performance of tools for citation parsing using synthetic and real citation strings. The study found that training models with synthetic data did not result in decreased performance compared to real data, confirming that synthetic citation strings can be generated as an alternative to corpus-based training.

3. Method

Figure 1 shows the main steps of the processing pipeline for our experiment.

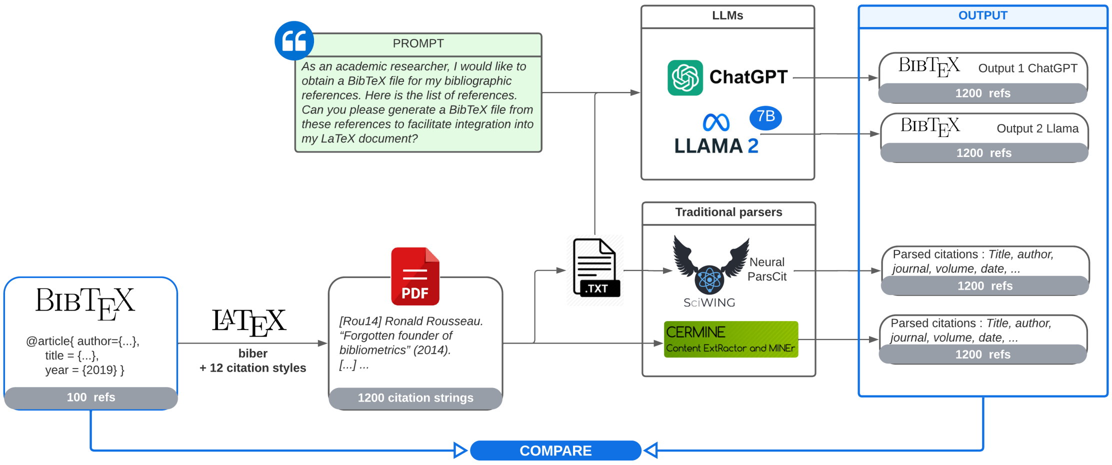

Figure 1: Overview of the processing pipeline

3.1. Building a synthetic dataset of citation strings in various styles

The benchmark dataset required for our task consists of citation strings and their corresponding parsed structures. To obtain a high-quality dataset that covers the most common citation styles, we followed these steps:

1. We processed the obtained BibTeX database using LaTeX with the `biber` package, applying 12 different citation styles. The list of the citation styles that we used is: `apa`, `mla`, `chem-acs`, `phys`, `nature`, `science`, `ieee`, `chicago-authordate`, `numeric`, `alphabetic`, `authoryear`, `authortitle`[6].
2. To ensure that the PDF-to-text conversion does not intervene with the quality of the citation strings, they were manually extracted from the produced PDF articles and stored in text files, with one citation string per line. The extracted citation strings were then used as input for both Neural ParsCit and the LLMs.

[6] BibLaTeX allows some variants of these styles, e.g. `alphabetic-verb`, `authoryear-comp`, `authortitle-ibid`, but they did not produce any modification in the generated citation strings.

Table 1
Fields present in the original BibTeX database

ENTRYTYPE	author	title	journal	year	volume	number	pages	series	booktitle	doi
@article	93	93	93	93	90	88	90	0	0	62
@inproceedings	7	7	0	7	2	0	6	7	7	3
Total	100	100	93	100	92	88	96	7	7	65

Using the above steps, we obtained 1,200 citation strings in 12 different styles that correspond to the BibTeX entries in our database of 100 references. Since this dataset only includes strings produced by the LaTeX `biber` package, we consider that they do not contain any formatting or punctuation errors. As this procedure follows the typical method for producing a bibliography in a paper, we believe that this type of dataset accurately reflects citation string structures that are commonly found in real articles, while also encompassing a wide range of citation styles used by various disciplines and journals.

3.2. Test protocol for the generative LLMs

Our task was to test two readily available generative LLMs:

1. OpenAI's ChatGPT: free online version 3.5, January 2024;
2. Llama 2 7B, that we loaded locally using the LM Studio server.

We divided the dataset of 1,200 citation strings into sets and submitted them to the two models preceded by the following prompt:

> "As an academic researcher, I would like to obtain a BibTeX file for my bibliographic references. Here is the list of references. Can you please generate a BibTeX file from these references to facilitate integration into my LaTeX document?"

The sets submitted to ChatGPT were of 10 citation strings, while for Llama we had to reduce this size to 5 because we found that the quality of the output for this model deteriorated rapidly after the first 6 or 7 BibTeX entries that were generated. Also, for both models we cleared the conversation history after every 10 sets of citations, so as to prevent too long a conversation history from affecting the quality of the responses.

Both models produce BibTeX entries in response to the prompt, most of which follow the correct BibTeX syntax. Llama's responses contained, in addition to the BibTeX entries, several introductory and concluding sentences that we had to remove, e.g. *"Of course, I can help you generate a BibTeX file for your references. Here is the output for each reference: [...] This will output the reference in the standard BibTeX format. [...]"*.

3.3. Test protocol for CERMINE and Neural ParsCit

CERMINE (Content ExtRactor and MINEr) [12] extracts metadata and content from scientific articles in PDF format. It's output includes the metadata, the structured content of the article, and the parsed bibliographic references in an NLM XML record. For our experiment, we used the online version available from http://cermine.ceon.pl/.

As CERMINE relies on the structure of the paper to identify the bibliography section, we have provided it with full PDF papers generated using the `llncs` LaTeX template for articles. Each article contains a title, authors and affiliations, an abstract and keywords. The body of the text follows the IMRaD structure, with several paragraphs per section and references to all 100 citations in our dataset. The last section is the References section. We generated 12 such articles, one for each citation style. The articles are identical except for the citation style that is used.

Neural ParsCit [20] uses a deep learning model, Long Short Term Memory (LSTM), to perform sequence-to-sequence labeling. It parses reference strings into their component tags such as Author, Journal, Location, Date, etc. The output is a string in which each token is followed by its label. For our experiment we used the implementation of Neural ParsCit which is part of the Scientific Document Processing Toolkit (SciWing) and uses Bi-LSTM-CRF + GloVe + Elmo + Char-LSTM[7].

3.4. Evaluation of the output

The evaluation of each parser's output was based on a predefined list of fields and labels. This was necessary due to the varying formats and labels produced by the different parsers, despite their intended retrieval of the same level of detail and number of fields. Table 2 displays the specific lists that we used for each parser.

The input for our processing is a BibTeX database, but only the two LLMs provide output in the BibTeX format. CERMINE and Neural ParsCit use their own annotation labels to render the structure of the citation strings. Table 3 displays the correspondence between the labels in the three types of outputs: the sub-tags of the NML XML `ref` element that are used in CERMINE, the labels produced by Neural ParsCit, and the BibTeX fields. Each parser was evaluated solely on the fields it was intended to provide, considering this correspondence.

Table 2
Fields/labels used for the evaluation of the output

Parser	List of fields/labels
ChatGPT & LLama	"ENTRYTYPE", "author", "title", "journal", "year", "volume", "number", "pages", "series", "booktitle", "doi"
CERMINE	"author", "title", "journal", "year", "pages", "volume", "number"
Neural ParsCit	"author", "title", "journal", "year", "pages", "volume", "booktitle"

The BibTeX format that we use for our input inherently allows for certain variations in the data, that should be taken into account when we need to compare the original data with the output of the parsers. To do this, we normalised all white spaces and converted all titles to titlecase. Some of the punctuation had to be normalized, e.g. the different types of hyphens (-) that can appear in the `pages` field. Non-Unicode characters have been removed, and punctuation signs

[7] https://sciwing.io/, https://pypi.org/project/sciwing/.

Table 3
Correspondance between fields/labels produced by CERMINE, Neural ParsCit and BibTeX

CERMINE (NML XML ref) element	Neural ParsCit label	BibTeX field
<string-name>, <given-name>, <surname>	AUTHOR	author
<article-title>	TITLE	title
<source>	JOURNAL	journal
<source>	BOOKTITLE	booktitle
<volume>	VOLUME	volume
<issue>	VOLUME	number
<fpage>, <lpage>	PAGES	pages
<year>	DATE	year

were stripped from titles, which allows to eliminate trailing commas and points that are present in Neural ParsCit's output.

Author names in BibTeX require some specific processing. Figure 2 shows an example of a BibTeX entry and its citation strings in two different citation styles, with the output produced by the four citation parsers. Author names can be presented in a BibTeX field with one of the following syntaxes: "`First-name Surname`" or "`Surname, First-name`" or "`Surname, F.`". The generated citation string can follow one of these syntaxes depending on the citation style. In addition, long author lists are often abbreviated in the citation strings and replaced by the expression "et al".

In the example in figure 2, the author lists produced by ChatGPT, CERMINE and Neural ParsCit are correct for both `ieee` and `science` styles, although they do not contain all the author names of the original entry. In fact, the parsers rely only on the citation strings, which contain partial information for the authors, to produce the correct output. On the other hand, Llama missed several authors for the `ieee` style, and hallucinated several other authors for the `science` style. Furthermore, its output for the `ieee` style is not syntactically correct as a BibTeX entry, in which case we consider all the fields to be wrong.

Following these considerations, we applied the following algorithmic solution to correctly compare the output of the parsers for the author names:

1. Convert all author names to the "`First-name Surname`" syntax.
2. If the citation string contains "et al", then keep only the first author.
3. If the citation string contains only initials for the first names of authors, then assume that only initials are present in the original BibTeX entry for this style.
4. Remove all points after the initials, and convert all names to lowercase to avoid problems of capitalisation of names such as "McKein" to lead to incorrect output, etc.

The BibTeX database we use contains two types of entries, `@artile` and `@inproceedings`, which differ in that the `@artile` entries have a `journal` field and the `@inproceedings` entries have a `booktitle` field. As CERMINE and Neural ParsCit do not distinguish between these types of entries, and CERMINE does not provide a `booktitle` label, we considered that for these two parsers the `journal` label is equivalent to `booktitle` in cases where the original BibTeX entry contains a `booktitle` field.

When evaluating the output for optional fields, such as `doi`, we need to take into account that this output is only expected for those citation styles where the information is present in the citation string. For example, in figure 2, the reference in the `science` style does not contain any information about `doi` although the `doi` is present in the original BibTeX entry. ChatGPT correctly returned an entry without `doi`, as did CERMINE and Neural ParsCit. Llama hallucinated a `doi` and a `url`.

The values of precision, recall and F-measure were calculated taking into account all the fields/labels that were produced by the parsers according to the table 2. Only fields for which the parsers produced values identical to those of the original BibTeX entries were considered correct. Fields for which the values differed from those of the original BibTeX entries, after applying all of the above considerations, were considered incorrect. Other types of error include fields added by the parser that were not present in the original BibTeX record, or fields missing from the parser's output.

4. Results

The table 4 shows the results for the precision (P), recall (R) and F-measure (F) of the output of the parsers for the different citation styles, and the table 5 shows the results by field/label. Figure 3 shows the F-measures that were obtained for the parsers, for each citation style and field/label. The values were calculated per field/label, in order to account for partially-correct parsing. ChatGPT obtains the best scores, and this result is consistent for all citation styles and for all fields, with F-scores between 0.751 and 0.996. Llama is particularly good for the `mla` style (F-score of 0.776) and for retrieving titles, journals, volumes and pages. However, Llama is relatively poor at retrieving years, while all other parsers perform better for the years.

CERMINE and Neural ParsCit both perform well for `apa` (F-scores of 0.755 and 0.5 respectively) and both are relatively good for `chem-acs`. CERMINE performs well also for `ieee`, but its scores for the other citation styles are rather low. CERMINE was trained on data from the PMC OA subset, which shows little variability, and this may explain why CERMINE fails to deal with the different citation styles. The scores for Neural ParsCit are generally lower, with the best F-score being 0.5 for the `apa` style. It performs relatively well in retrieving years and journals.

Some of the errors of CERMINE and Llama are due to the fact that their output could not be fully processed. For some bibliography entries CERMINE produced no output, and in other cases Llama produced code with incorrect BibTeX syntax that could not be parsed. The table 6 shows the number of entries for which the parsers produced syntactically incorrect output, or no output, and their percentage relative to all 1,200 citation strings in the dataset.

The data that was produced for this experiment and the output of the parsers are available [21][8].

5. Discussion and Limitations

The present study is designed to conduct a comparative analysis between existing NLP tools and LLMs with regard to their performance on the task of citation parsing. It is important to

[8]See also the github repository https://github.com/iana-atanassova/citation-parsers-bir2024.git

Table 4
Precision, recall and F-measure for the citation string parsing, by citation style. The best values for each parser are in bold. All parsers are used out-of-the-box without any pretraining.

Citation style		ChatGPT	Llama 2 7B	CERMINE	Neural ParsCit
alphabetic	P	0.995	0.826	0.188	0.239
	R	0.986	0.628	0.142	0.233
	F	0.991	0.713	0.162	0.236
apa	P	0.998	0.766	**0.757**	**0.504**
	R	0.962	0.623	**0.753**	**0.495**
	F	0.980	0.687	**0.755**	**0.500**
authortitle	P	0.999	0.849	0.204	0.326
	R	**0.994**	0.653	0.178	0.321
	F	**0.996**	0.738	0.190	0.324
authoryear	P	0.996	0.837	0.341	0.332
	R	0.991	0.635	0.294	0.327
	F	0.993	0.722	0.316	0.329
chem-acs	P	0.997	0.540	0.661	0.474
	R	0.746	0.411	0.469	0.422
	F	0.853	0.467	0.549	0.446
chicago-authordate	P	0.998	0.815	0.374	0.276
	R	0.993	0.607	0.313	0.250
	F	0.995	0.696	0.341	0.262
ieee	P	**1.000**	0.829	0.644	0.171
	R	0.988	0.660	0.598	0.163
	F	0.994	0.735	0.620	0.167
mla	P	**1.000**	**0.856**	0.528	0.370
	R	0.986	**0.711**	0.488	0.364
	F	0.993	**0.776**	0.507	0.367
nature	P	0.996	0.655	0.454	0.290
	R	0.870	0.493	0.344	0.286
	F	0.928	0.562	0.391	0.288
numeric	P	0.979	0.806	0.208	0.240
	R	0.845	0.596	0.180	0.233
	F	0.907	0.685	0.193	0.236
phys	P	0.999	0.759	0.428	0.390
	R	0.861	0.615	0.293	0.361
	F	0.925	0.680	0.348	0.375
science	P	0.998	0.650	0.514	0.211
	R	0.608	0.494	0.320	0.192
	F	0.756	0.561	0.394	0.201

Table 5
Precision, recall and F-measure for the citation string parsing, by field/label. All parsers are used out-of-the-box without any pretraining.

Field/label		ChatGPT	Llama 2 7B	CERMINE	Neural ParsCit
author	P	0.951	0.724	0.059	0.242
	R	0.951	0.716	0.054	0.237
	F	0.951	0.720	0.056	0.239
title	P	0.817	0.775	0.168	0.088
	R	0.817	0.770	0.128	0.078
	F	0.817	0.772	0.145	0.083
journal	P	0.994	0.890	0.248	0.398
	R	0.902	0.685	0.240	0.393
	F	0.946	0.774	0.244	0.396
year	P	0.995	0.907	0.913	0.833
	R	0.994	0.317	0.775	0.785
	F	0.995	0.470	0.838	0.808
volume	P	0.946	0.899	0.609	0.141
	R	0.948	0.705	0.408	0.146
	F	0.947	0.791	0.489	0.143
pages	P	0.995	0.743	0.783	0.199
	R	0.990	0.738	0.779	0.195
	F	0.992	0.740	0.781	0.197
number	P	0.982	0.833	0.508	
	R	0.609	0.023	0.277	
	F	0.751	0.045	0.358	
series	P	0.974	0.000		
	R	0.974	0.000		
	F	0.974	0.000		
booktitle	P	0.708	0.152		
	R	0.810	0.246		
	F	0.756	0.188		
doi	P	0.964	0.619		
	R	0.967	0.705		
	F	0.966	0.659		
ENTRYTYPE	F	0.987	0.731		
Macro F-score		0.929	0.670	0.419	0.313

Table 6
Citation strings producing no output or syntactically incorrect output, by parser

ChatGPT	Llama 2 7B	CERMINE	Neural ParsCit
1 (0,08 %)	219 (18,25 %)	160 (13,33 %)	0 (0,00 %)

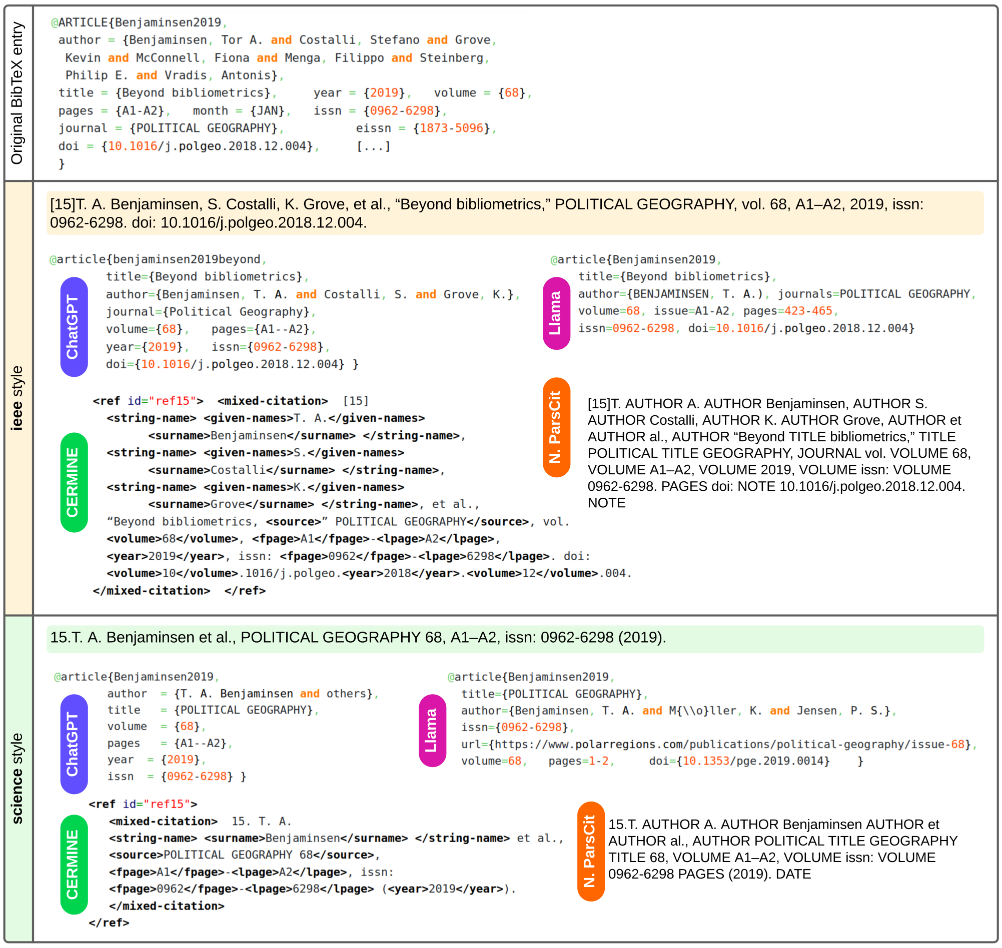

Figure 2: Example of an original BibTeX entry, with citation strings generated in `ieee` style and in `science` style, and the outputs of the citation parsers

clarify that this research does not introduce new methodologies within the field; instead, it aims to offer a novel perspective on citation parsing by assessing the results of different types of models. The choice to use a synthetic dataset, which might not be representative of real world situations, is motivated by the objective to cover as many different citation styles as possible. In particular, styles used in the Humanities are notably absent from existing datasets, hence our approach aims to bridge this gap and provide an analysis of citation parsing across different disciplines.

We tested the smallest version of Llama 2, with 7B parameters, to demonstrate the impact of model size on performance. In contrast, ChatGPT has 175B parameters. Despite this, it is clear that even a small LLM can outperform traditional out-of-the-box tools based on ML approaches, in particular when the data covers various disciplines and citation styles. However, LLMs introduce new types of errors like hallucinations, demanding specific experimental adjustments. In our experiments, we limited ChatGPT to 10 citation strings and Llama to 5

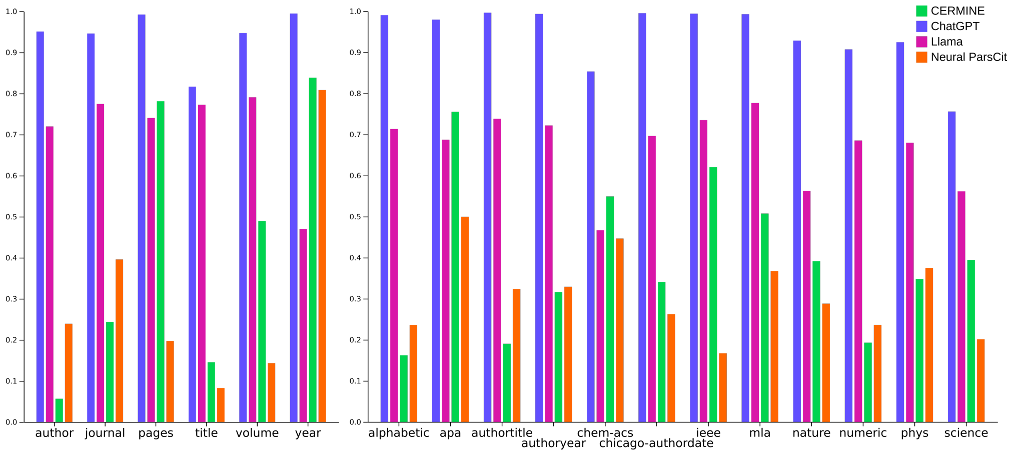

Figure 3: F-measures for the different parsers, by citation style and by field/label. All parsers are used out-of-the-box without any pretraining.

to mitigate hallucinations, also clearing conversation histories regularly. Despite this, Llama frequently hallucinated, while ChatGPT showed better performance. However, systematic hallucination control is essential before these models can effectively be used in real-case scenarios. Additionally, LLMs' use involves other considerations, such as prompt-specific responses and unnecessary text additions, observed with ChatGPT and Llama respectively. These issues highlight the need for improved prompts and post-processing in future LLM applications.

While LLMs have been trained on huge amounts of data, this is not the case for classical models. Consequently, the generalisation power of LLMs and classical models across different citation styles and datasets varies significantly. Direct comparisons between these two categories of models should be approached with caution, and any results derived from such comparisons must be interpreted within the context of these foundational differences.

6. Conclusion and Future Work

We proposed an experiment to measure the ability of LLMs to analyse citation strings in different citation styles and compare them to two out-of-the-box citation parsers, CERMINE and Neural ParsCit. We used a synthetic dataset of citation strings that allowes us to cover 12 different citation styles. The results indicate that the LLMs tend to outperform the citation parsers for all citation styles and labels, with ChatGPT 3.5 producing the best results.

Our next step is to develop an approach for testing more LLMs using Crossref data. Crossref is a DOI registration agency[9], that supports various metadata content types, making it possible to generate synthetic reference strings in both BibTeX and JSON formats. We also need to test other traditional tools, such as Grobid. Additionally, we must compare the performance of larger Open Source LLMs, such as the upcoming versions of Llama [22] and Mistral [23]. Prompt engineering may be a viable strategy for improving results. Another way to improve the output

[9]https://citation.crosscite.org/docs.html

of LLMs is to address the problem of hallucinations by establishing a framework to reduce this phenomenon.

Acknowledgments

This work was supported by French ANR grant number ANR-20-CE38-0003-01 and grant number ANR-21-CE38-0003-01.

Declaration of Generative AI

During the preparation of this work, the authors used ChatGPT in order to: grammar and spelling check. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, S. Kohlmeier, CORD-19: The COVID-19 open research dataset, in: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Association for Computational Linguistics, Online, 2020. URL: https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1.

[2] T. Saier, M. Färber, unarXive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata, Scientometrics 125 (2020) 3085–3108. doi:10.1007/s11192-020-03382-z.

[3] Z. Boukhers, S. Ambhore, S. Staab, An end-to-end approach for extracting and segmenting high-variance references from pdf documents, in: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2019, pp. 186–195. doi:10.1109/JCDL.2019.00035.

[4] D. Tkaczyk, A. Collins, P. Sheridan, J. Beel, Evaluation and comparison of open source bibliographic reference parsers: a business use case, arXiv preprint arXiv:1802.01168 (2018).

[5] V. Jain, N. Baliyan, S. Kumar, Machine learning approaches for entity extraction from citation strings, in: International Conference on Information Technology, Springer, 2023, pp. 287–297.

[6] D. Tkaczyk, A. Collins, P. Sheridan, J. Beel, Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers, in: Proceedings of the 18th ACM/IEEE on joint conference on digital libraries, JCDL '18, ACM, 2018, pp. 99–108. doi:10.1145/3197026.3197048.

[7] K. Seymore, R. Rosenfeld, Learning hidden markov model structure for information extraction (1999).

[8] C. Caragea, J. Wu, A. Ciobanu, K. Williams, J. Fernández-Ramírez, H.-H. Chen, Z. Wu, L. Giles, Citeseer x: A scholarly big dataset, in: Advances in Information Retrieval: 36th

European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings 36, Springer, 2014, pp. 311–322.
[9] E. Cortez, A. S. da Silva, M. A. Gonçalves, F. Mesquita, E. S. de Moura, FLUX-CIM: flexible unsupervised extraction of citation metadata, in: Proceedings JCDL 2007, 2007, pp. 215–224.
[10] T. Groza, A. Grimnes, S. Handschuh, Reference information extraction and processing using random conditional fields, Information Technology and Libraries 31 (2012) 6–20.
[11] D. Tkaczyk, P. Szostek, L. Bolikowski, Grotoap2-the methodology of creating a large ground truth dataset of scientific articles, D-Lib Magazine 20 (2014).
[12] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, L. Bolikowski, Cermine: automatic extraction of structured metadata from scientific literature, International Journal on Document Analysis and Recognition (IJDAR) 18 (2015) 317–335. doi:10.1007/s10032-015-0249-8.
[13] P. Lopez, GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications, in: Proceedings ECDL 2009, Springer, Springer Berlin Heidelberg, 2009, pp. 473–474. doi:10.1007/978-3-642-04346-8_62.
[14] M. Grennan, M. Schibel, A. Collins, J. Beel, Giant: The 1-billion annotated synthetic bibliographic-reference-string dataset for deep citation parsing, in: 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, 2019, pp. 101–112.
[15] M. Grennan, M. Schibel, A. Collins, J. Beel, GIANT: The 1-Billion Annotated Synthetic Bibliographic-Reference-String Dataset for Deep Citation Parsing [Data] (2019). URL: https://doi.org/10.7910/DVN/LXQXAO. doi:10.7910/DVN/LXQXAO.
[16] K. Lo, L. L. Wang, M. Neumann, R. Kinney, D. Weld, S2ORC: The semantic scholar open research corpus, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4969–4983. URL: https://aclanthology.org/2020.acl-main.447. doi:10.18653/v1/2020.acl-main.447.
[17] P. Cuxac, A. Collignon, Istex, un projet national d'archives documentaires: au-delà de l'accès au texte intégral, l'enrichissement des données par méthodes de fouille de textes., in: Analyser la science: les bibliothèques numériques comme objet de recherche in 85ème Congrès ACFAS, 2017.
[18] P. Cuxac, N. Thouvenin, Archives numériques et fouille de textes: le projet istex, Atelier TextMine, EGC 2017 (Extraction et Gestion des Connaissances), Grenoble, France, January 24 27 (2017) 2017.
[19] M. Grennan, J. Beel, Synthetic vs. real reference strings for citation parsing, and the importance of re-training and out-of-sample data for meaningful evaluations: experiments with grobid, giant and cora, arXiv preprint arXiv:2004.10410 (2020). arXiv:2004.10410.
[20] A. Prasad, M. Kaur, M.-Y. Kan, Neural parscit: A deep learning based reference string parser, International Journal on Digital Libraries 19 (2018) 323–337. doi:10.1007/s00799-018-0242-1.
[21] M. Bertin, I. Atanassova, Synthetic dataset of citation strings in 12 styles, 2024. URL: https://doi.org/10.5281/zenodo.10839503. doi:10.5281/zenodo.10839503.
[22] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models,

arXiv preprint arXiv:2307.09288 (2023).
[23] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).