Identifying Implicit Research Data References in Paper Citations

Koshi Motegi[1], **Koichiro Ito**[1], **Shigeki Matsubara**[1,2]
[1]Graduate School of Informatics, Nagoya University, Japan
[2]Information Technology Center, Nagoya University, Japan
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
motegi.koshi.h9@s.mail.nagoya-u.ac.jp,
{ito.koichiro.z5, matsubara.shigeki.z8}@f.mail.nagoya-u.ac.jp

Abstract

To encourage the public release of research data under open science, it is beneficial to establish mechanisms for evaluating research data based on metrics such as citation counts. In scholarly papers, authors sometimes cite papers that report the creation or release of research data instead of citing the research data themselves. In this paper, as a step toward computing citation counts of research data, we investigate the feasibility of identifying paper citations that refer to research data. We conducted an identification experiment using large language models and evaluated their performance.

Keywords: research data, paper citations, scholarly papers

1. Introduction

In recent years, open science has been promoted (UNESCO, 2021; G7, 2023), and the sharing of research artifacts has been encouraged. Research artifacts include not only scholarly papers but also research data such as datasets and tools. Although open access to scholarly papers has advanced, the openness of research data remains limited. One reason for this limited openness is that evaluation metrics for research data are less established. In contrast, citation counts are widely used as a representative evaluation metric for scholarly papers, and researchers tend to focus on increasing them.

To address this situation, citation counts of research data could be considered as an evaluation metric. However, as shown in Figure 1, research data are cited in various ways in scholarly papers. Specifically, the following citation ways are considered:

(a) **Formal citation**: Citing research data by describing relevant information about the research data in the reference list, in the same way as citing scholarly papers

(b) **URL citation**: Citing research data by describing the URLs of research data in the body text or in the footnote

(c) **Paper citation**: Citing research data implicitly by citing scholarly papers that report the creation or release of research data

For the formal citation and URL citation, several initiatives have focused on aggregating citation counts of research data (Clarivate, 2023; Rosonovski et al., 2023). In contrast, for paper citations, little effort has been made to develop mechanisms for aggregating citation counts of research data. This is because paper citations that refer to research data have generally been counted as citations to the papers themselves without distinguishing whether the citations refer to research data.

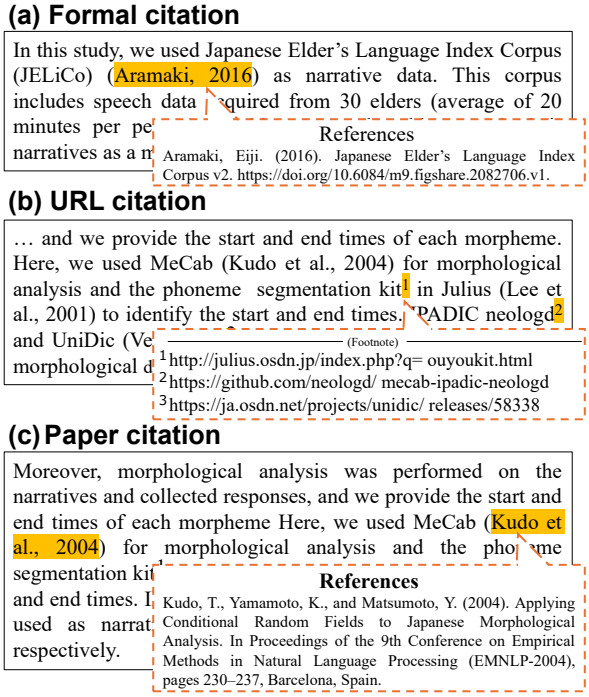

Figure 1: Ways to cite research data (examples are excerpted from Ito et al., 2022).

This study discusses the feasibility of identifying paper citations that refer to research data in scholarly papers to compute the citation counts of re-

Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026, pages 186–192
12 May 2026.

search data. To this end, we attempted to automatically classify paper citations in scholarly papers as referring to research data or not, using texts surrounding citations (i.e., the citation contexts) and bibliographic information. We conducted a preliminary experiment to classify paper citations using large language models (LLMs) and evaluated their performance.

2. Citations and Evaluation of Research Data

Typically, research data are cited in one of the following three ways: formal citation, URL citation, or paper citation. In this section, we discuss existing research and efforts to compute citation counts of research data across different citation citation ways.

2.1. Formal Citation

Formal citation is a way to cite research data by providing its information, such as its title, creators, and persistent identifiers (e.g., DOI) in the reference list, in the same way as citing scholarly papers. Figure 1 (a) shows an example of a formal citation. This way has been recommended for citing research data in guidelines such as the Joint Declaration of Data Citation Principles (Group, 2014) and the Tromsø Recommendations (Andreassen et al., 2019). In addition, several conferences and journals treat formal citation as the standard to cite research data, including LREC[1], which is an international conference on language resources, and Scientific Data[2], an open-access journal that focuses on the creation and reuse of research data.

Furthermore, several systems have been developed to compute the citation counts of research data based on formal citations, such as Make Data Count[3], the Data Citation Index (Clarivate, 2023), and the Europe PMC API (Rosonovski et al., 2023). Using the citation counts obtained through these systems, research data and their creators have been evaluated in several indices, including the OmicsDI (Perez-Riverol et al., 2019) and the Data-Index (Hood and Sutherland, 2021).

2.2. URL Citation

URL citation is a way to cite research data by describing the URLs of research data in the body text or a footnote of a scholarly paper. Figure 1 (b) shows an example of URL citation. Computing URL citation counts requires identifying URLs of research data within scholarly papers.

Previous studies extracted URLs of research data from scholarly papers using regular expressions and pretrained language models (Yamamoto and Takagi, 2007; Tsunokake and Matsubara, 2022; Otto et al., 2023). In addition, some studies attempted to classify resource types referenced by URLs in scholarly papers (Zhao et al., 2019; Tsunokake and Matsubara, 2022; Wada et al., 2025).

(a) Research data

The FINECITE dataset was built from a subset of ACL Anthology Network Corpus (Radev et al., 2009). The ACL Anthology Network contains over 80K papers from several ACL conferences and other venues in computational …

(b) Concept

By embedding scientific progress and argumentation, citations serve a critical function that has been extensively examined—a research field known as citation context analysis (CCA) (Kunnath et al., 2022; Swales, 1986).

(c) Method

We considered four baselines for the classification task: (i) the scaffolding approach presented in Cohan et al.,(2019), (ii) the best-performing citation classification model from the …

: Citation tag, : Cited content

Figure 2: Examples of paper citations and cited contents (examples are excerpted from Jantsch et al., 2025).

2.3. Paper Citation

Paper citation of research data is a way to cite research data implicitly by citing a scholarly paper that reports the creation or release of research data. Figure 1 (c) shows an example of paper citation. Generally, as shown in Figure 2, paper citations are made to refer to the concepts, methods, or findings presented in the cited paper. That is, citations that refer to research data are not necessarily frequent. Therefore, to compute the citation counts of research data accurately, it is necessary to identify paper citations that refer to research data.

A previous study introduced metrics to evaluate research data that consider the citation counts of papers reporting the creation or release of the data (Callahan et al., 2018). However, in that study, the papers were identified using indexing terms from PubMed's thesaurus, MeSH (Medical Subject Headings)[4], which makes it challenging to apply the approach beyond the medical domain. In

[1] https://lrec2026.info/authors-kit/

[2] https://www.nature.com/sdata/submission-guidelines

[3] https://makedatacount.org

[4] https://www.nlm.nih.gov/mesh/meshhome.html

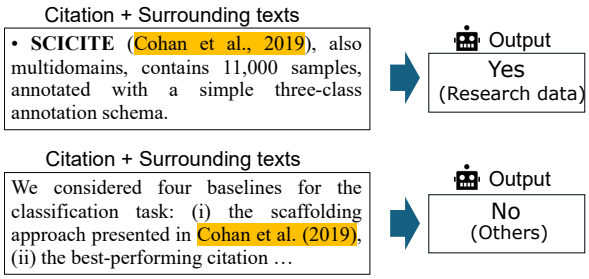

Figure 3: Overview of task setting (examples are excerpted from Jantsch et al., 2025).

addition, the method proposed in that study does not consider whether each paper is actually cited as a reference to the research data.

Several other studies attempted to classify the intent of citing scholarly papers (Teufel et al., 2006; Abu-Jbara et al., 2013; Jurgens et al., 2018; Cohan et al., 2019; Nambanoor Kunnath et al., 2022). These studies defined citation intent types, such as citations used to provide background for the study (Background type) and those used to describe existing methods employed in the study (Method type). Based on these definitions, previous studies constructed annotated corpora and developed methods to classify the citation intent. Representative methods include classifiers based on pretrained language models, e.g., SciBERT (Beltagy et al., 2019), (Shui et al., 2024; Paolini et al., 2025) and LLM-based classifiers (Jantsch et al., 2025; Koloveas et al., 2025). However, the citation intent types defined by these previous studies were not based on whether the citations refer to research data.

3. Task Setting

In this study, we investigated the feasibility of identifying paper citations that implicitly refer to research data. Here, we formulate the target problem as a binary classification task: given a citation, its citation context, and the bibliographic information of the cited paper as input, classifying whether it refers to research data (Yes) or not (No). Figure 3 shows an overview of the task setting.

In this study, research data refer to tangible research data created, generated, or collected through scholarly activities, e.g., datasets, software, tools, and machine learning models. In contrast, intangible research data, e.g., algorithms or methods, machine learning architectures, and task definitions, are excluded.

We treat paper citations as referring to research data in cases where the citation contexts contain references to the research data produced in the cited study. Even when the names of the research data are not explicitly mentioned, we treat citations as referring to research data if the context makes it clear that the cited works are being cited for their research data.

In the upper example in Figure 3, the paper citation refers to the "SCICITE" dataset, and thus is classified as referring to research data. In the lower example, the citation is made to refer to a method (i.e., "the scaffolding approach"); thus, it is not classified as a paper citation that implicitly refers to research data.

4. Preliminary Experiment

We conducted a preliminary experiment to investigate the feasibility of identifying paper citations that implicitly refer to research data.

4.1. Experimental Data

We conducted an experiment on the GSAP-NER dataset (Otto et al., 2023), which contains 100 machine learning papers annotated with 10 entity types, with a focus on machine learning models and datasets. In this dataset, citation tags are also annotated as an entity type.

In this study, we used 10 papers from the GSAP-NER dataset [5]. First, based on the GSAP-NER annotations, we extracted 783 citation tags and the paragraphs containing them. Then, for each extracted citation tag, we linked it to the corresponding bibliographic information of the cited paper, relying on information such as author names. We then identified the paper citations that implicitly refer to research data based on the citation tags, the paragraphs containing citation tags, and the bibliographic information of the cited papers. As a result, among the 783 citations in the 10 papers, 191 (24.4%) were identified as referring to research data.

4.2. Experimental Settings

We used Llama3.1 8B instruct (Grattafiori et al., 2024), Ministral-3-{3B, 8B, 14B}-Instruct-2512 (AI, 2025), GPT-4.1 series (GPT-4.1-nano, GPT-4.1-mini, and GPT-4.1) (OpenAI, 2025a), GPT-5 series (GPT-5-nano, GPT-5-mini, and GPT-5) (OpenAI, 2025b), GPT-5.2 (OpenAI, 2025c) as instruction-tuned LLMs. For models other than GPT series, we used vLLM[6] (Kwon et al., 2023). We limited the model responses to either "Yes" (paper citations implicitly referring to research data) or "No" (otherwise) using Structured Outputs. We set the temperature to 0.0 and used default parameters

[5] Specifically, we used six of the 10 folds created by Otto et al. (2023) for cross-validation.

[6] https://github.com/vllm-project/vllm

Model	Con + Cite				Con + Cite + Bib			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Random	0.500	0.247	0.499	0.330	0.500	0.247	0.499	0.330
Llama3.1 8B	0.785	0.557	0.637	0.594	0.711	0.444	0.684	0.539
Ministral3 3B	0.799	0.821	0.238	0.369	0.780	**0.920**	0.119	0.211
Ministral3 8B	0.815	0.610	0.689	0.647	0.806	0.611	0.585	0.598
Ministral3 14B	0.839	0.694	0.622	0.656	0.826	0.665	0.596	0.628
GPT-4.1-nano	0.757	0.538	0.109	0.181	0.745	0.434	0.119	0.187
GPT-4.1-mini	0.854	0.661	0.767	0.710	0.844	0.673	0.715	0.693
GPT-4.1	0.880	0.777	0.720	0.747	**0.888**	0.797	0.731	**0.762**
GPT-5-nano	0.861	0.698	0.767	0.731	0.854	0.682	0.767	0.721
GPT-5-mini	0.874	0.712	**0.819**	**0.761**	0.874	0.724	**0.788**	0.754
GPT-5	0.885	0.841	0.658	0.738	0.881	0.813	0.674	0.737
GPT-5.2	**0.891**	**0.886**	0.642	0.745	0.884	0.892	0.601	0.718

Table 1: Classification performance for identifying paper citations that implicitly refer to research data.

for all other settings[7]. We considered two input settings:

- Con + Cite: citation context + citation tag
- Con + Cite + Bib: citation context + citation tag + bibliographic information of the cited paper

We also employed a random baseline that outputs "Yes" or "No" at random. The extraction performance of the models was evaluated using accuracy, precision, recall, and F1-score.

4.3. Results

Table 1 shows the classification performance of each model. Under the Con + Cite setting, GPT-5.2 yielded the best accuracy and precision, with values of 0.891 and 0.886, respectively, and GPT-5-mini achieved the highest recall and F1-score, with values of 0.819 and 0.761, respectively.

Focusing on models within the same serie, we found that performance tended to improve as model size increased. In particular, the GPT series consistently achieved high scores: all models except GPT-4.1-nano achieved an F1 score of approximately 0.7. Overall, the experimental results indicated that while performance varied across models, using relatively larger models in the GPT series enabled more accurate identification of paper citations that implicitly refer to research data.

We also compared the performance of each model with and without bibliographic information. Although the F1-scores improved for GPT-4.1-nano and GPT-4.1, they decreased for the other models. Thus, we did not find that the use of bibliographic information yielded a clear performance benefit.

[7]For Ministral-3, we used the vLLM configuration recommended for Mistral models. We also set tensor type to BF16 to unify precision across all models.

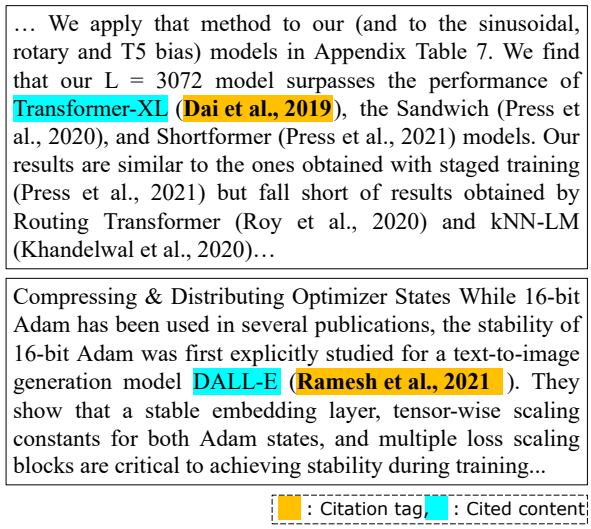

Figure 4: Misclassification cases of GPT-4.1 (input setting: Con + Cite + Bib).

4.4. Discussion

We conducted an error analysis of GPT-4.1 under the Con + Cite + Bib setting, which achieved the best F1 score. Although this model achieved strong overall performance, it still produced several misclassifications. Figure 4 shows representative examples of the misclassification cases. In these cases, the citations refer to machine learning models ("Transformer-XL" and "DALL-E") and should be classified as "Yes" (i.e., paper citations implicitly referring to research data), yet the model outputs "No." This suggests that GPT-4.1 tended to miss citations referring to machine learning models more frequently than to other types of research data. To mitigate such errors, it may be beneficial to provide additional information about the research data, such as explicitly highlighting the names of the machine learning models appear-

ing near the citation tag.

5. Conclusion

In this paper, toward automating the computation of citation counts of research data, we examined the feasibility of identifying paper citations that implicitly refer to research data. In this experiment, the classification was performed by providing citation context and bibliographic information as inputs to LLMs. The experimental results showed that LLMs could identify paper citations that implicitly refer to research data with high performance.

In scholarly papers, contribution of the cited paper to citing papers is not uniform. Previous studies explored metrics by classifying paper citations into fine-grained categories based on citation intent and importance, and then weighting citations. As future work, we plan to classify citations of research data based on their types and citation intent as a step toward weighted evaluation of citations of research data.

6. Acknowledgements

This research was partially supported by "Developing a Research Data Ecosystem for the Promotion of Data-Driven Science" of MEXT and the Grant-in-Aid for Challenging Research (Exploratory) (No. 23K18506) of JSPS.

7. Bibliographical References

Helene N. Andreassen, Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell, and Research Data Alliance Linguistic Data Interest Group. 2019. Tromsø recommendations for citation of research data in linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Alison Callahan, Rainer Winnenburg, and Nigam H. Shah. 2018. U-index, a dataset and an impact metric for informatics tools and databases. *Scientific Data*, 5(1):180043.

Clarivate. 2023. Recommended practices to promote scholarly data citation and tracking: The role of the data citation index. White paper. Version 4.

G7. 2023. Annex 1: G7 open science working group (OSWG). (Last Accessed: 2025-12-22).

FORCE11 Data Citation Synthesis Group. 2014. Joint declaration of data citation principles.

Amelia S. C. Hood and William J. Sutherland. 2021. The data-index: An author-level metric that values impactful data and incentivizes data sharing. *Ecology and Evolution*, 11(21):14344–14350.

Koichiro Ito, Masaki Murata, Tomohiro Ohno, and Shigeki Matsubara. 2022. Construction of responsive utterance corpus for attentive listening response production. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 7244–7252, Marseille, France. European Language Resources Association.

Lasse M. Jantsch, Dong-Jae Koh, Seonghwan Yoon, Jisu Lee, Anne Lauscher, and Young-Kyoon Suh. 2025. FineCite: A novel approach for fine-grained citation context analysis. In *Findings of the Association for Computational Linguistics*, pages 24525–24542, Vienna, Austria. Association for Computational Linguistics.

Paris Koloveas, Serafeim Chatzopoulos, Thanasis Vergoulis, and Christos Tryfonopoulos. 2025. Can LLMs predict citation intent? An experimental analysis of in-context learning and fine-tuning on open LLMs. In *Linking Theory and Practice of Digital Libraries: 29th International Conference on Theory and Practice of Digital Libraries*, pages 207–224. Springer.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, pages 611–626, New York, USA. Association for Computing Machinery.

Lorenzo Paolini, Sahar Vahdati, Angelo Di Iorio, Robert Wardenga, Ivan Heibi, and Silvio Peroni. 2025. CiteFusion: An ensemble framework for citation intent classification harnessing dual-model binary couples and shap analyses. *Scientometrics*, 130(11):5911–5981.

Yasset Perez-Riverol, Andrey Zorin, Gaurhari Dass, Manh-Tu Vu, Pan Xu, Mihai Glont, Juan Antonio Vizcaíno, Andrew F. Jarnuczak, Robert Petryszak, Peipei Ping, and Henning Hermjakob. 2019. Quantifying the impact of

public omics data. *Nature Communications*, 10(1):3512.

Summer Rosonovski, Maria Levchenko, Rajat Bhatnagar, Umamageswari Chandrasekaran, Lynne Faulk, Islam Hassan, Matt Jeffryes, Syed Irtaza Mubashar, Maaly Nassar, Madhumiethaa Jayaprabha Palanisamy, Michael Parkin, Jagadeeswararao Poluru, Frances Rogers, Shyamasree Saha, Mohamed Selim, Zunaira Shafique, Michele Ide-Smith, David Stephenson, Santosh Tirunagari, Aravind Venkatesan, Lijun Xing, and Melissa Harrison. 2023. Europe PMC in 2023. *Nucleic Acids Research*, 52(D1):D1668–D1676.

Zeren Shui, Petros Karypis, Daniel S. Karls, Mingjian Wen, Saurav Manchanda, Ellad B. Tadmor, and George Karypis. 2024. Fine-tuning language models on multiple datasets for citation intention classification. pages 16718–16732, Miami, Florida, USA. Association for Computational Linguistics.

Masaya Tsunokake and Shigeki Matsubara. 2022. Classification of URL citations in scholarly papers for promoting utilization of research artifacts. In *Proceedings of the 1st Workshop on Information Extraction from Scientific Publications*, pages 8–19, Online. Association for Computational Linguistics.

UNESCO. 2021. UNESCO recommendation on open science.

Kazuhiro Wada, Masaya Tsunokake, and Shigeki Matsubara. 2025. Classification of URL citations on scholarly papers using intermediate task training. *IEICE Transactions on Information and Systems*, E108-D(10):1183–1193.

Yasunori Yamamoto and Toshihisa Takagi. 2007. OReFiL: An online resource finder for life sciences. *BMC Bioinformatics*, 8:287.

He Zhao, Zhunchen Luo, Chong Feng, Anqing Zheng, and Xiaopeng Liu. 2019. A context-based framework for modeling the role and function of on-line resource citations in scientific literature. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5206–5215, Hong Kong, China. Association for Computational Linguistics.

8. Language Resource References

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 596–606, Atlanta, Georgia. Association for Computational Linguistics.

Mistral AI. 2025. Ministral3. (Last Accessed: 2025-12-14).

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3586–3596, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The Llama 3 herd of models.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Suchetha Nambanoor Kunnath, Valentin Stauber, Ronin Wu, David Pride, Viktor Botev, and Petr Knoth. 2022. ACT2: A multi-disciplinary semi-structured dataset for importance and purpose classification of citations. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 3398–3406, Marseille, France. European Language Resources Association.

OpenAI. 2025a. Introducing GPT-4.1 in the api. (Last Accessed: 2025-12-14).

OpenAI. 2025b. Introducing GPT-5. (Last Accessed: 2025-12-22).

OpenAI. 2025c. Introducing GPT-5.2. (Last Accessed: 2025-12-22).

Wolfgang Otto, Matthäus Zloch, Lu Gan, Saurav Karmakar, and Stefan Dietze. 2023. GSAP-NER: A novel task, corpus, and baseline for scholarly entity extraction focused on machine learning models and datasets. In *Findings of the Association for Computational Linguistics*,

pages 8166–8176, Singapore. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia. Association for Computational Linguistics.

A. LLMs Prompt

The system prompt is shown in Figure 5, and the user prompts for Con + Cite and Con + Cite + Bib are shown in Figure 6 and Figure 7, respectively.

System Prompt

You are a helpful assistant.

Your task is to determine whether, in the given context, the citation is made for referring to research data.

Here, "research data" includes:
– datasets (raw or processed data)
– software tools or code (usable tools or libraries)
– trained models (released weights, checkpoints, or model files)

Important:
The decision must be based on the citation intent in the provided Context, not on what the cited work provides in general.
A citation counts as referencing research data (answer "Yes") only if:
– the Context indicates that the cited work's dataset, tool, or trained model is being used, accessed, downloaded, analyzed, applied, or employed in the study; or the cited work's research data is directly referenced as a resource.

A citation should be labeled "No" if:
– the Context cites the work only for its method, algorithm, model architecture, theoretical idea, concept, or background explanation;
– the Context does not indicate the use or reference of any dataset, tool, or trained model provided by the cited work;
– the citation intent is ambiguous or unclear.

Respond with exactly one word:
Yes
No

Figure 5: System prompt.

User Prompt (Con + Cite)

Context:
[context]

Citation Tag:
[citation_tag]

Figure 6: User prompt (Con + Cite).

User Prompt (Con + Cite + Bib)

Context:
[context]

Citation Tag:
[citation_tag]

Bibliographic Information:
[bibliographic_information]

Figure 7: User prompt (Con + Cite + Bib).