Metadata Generation for Research Data from URL Citation Contexts in Scholarly Papers: Task Definition and Dataset Construction

Yu Watanabe[1], Koichiro Ito[1], Shigeki Matsubara[1,2],
[1]Graduate School of Informatics, Nagoya University,
[2]Information Technology Center, Nagoya University,
Correspondence: watanabe.yu.x3@s.mail.nagoya-u.ac.jp

Abstract

This paper proposes a new research task aimed at automatically generating metadata for research data, such as datasets and code, to accelerate open science. From the perspective of "Findable" in the FAIR data principles, research data is required to be assigned a global unique identifier and described with rich metadata. The proposed task is defined as extracting information about research data (specifically, *name*, *generic mention*, and *in-text citation*) from texts surrounding URLs that serve as identifiers for research data references in scholarly papers. To support this task, we constructed a dataset containing approximately 600 manually annotated citation contexts with URLs of research data from conference papers. To evaluate the task, we conducted a preliminary experiment using the constructed dataset, employing the In-Context Learning method with LLMs as a baseline. The results showed that the performance of LLMs matched that of humans in some cases, demonstrating the feasibility of the task.

1 Introduction

Open science is a movement to promote the utilization of research data by making them publicly available (G7 OSWG, 2023). To utilize research data, such as datasets and code, effectively, it is necessary to assign metadata. One solution to accelerate this process is to extract information on research data from texts referring to the data, such as scholarly papers.

The FAIR Guiding Principles (Wilkinson et al., 2016) outlines the criteria for achievement in open science. FAIR stands for "Findable," "Accessible," "Interoperable," and "Reusable." The most fundamental principle is "Findable," and the requirements for research data to be findable are that a unique identifier is assigned and that rich metadata are described. However, no previous study on extracting information about research data from scholarly papers has explicitly considered the above requirements.

This paper proposes a new research task of extracting information about research data from scholarly papers. We define the task based on DataCite (DataCite Metadata Working Group, 2024), a global standard metadata schema. Specifically, the task is defined as extracting information corresponding to *name*, *generic mention*, and *in-text citation* of research data. This information is extracted from the citation context, i.e., the paragraph containing URL citations.

To perform the proposed task, we manually annotated approximately 600 paragraphs of text (citation contexts) containing URLs citing research data from conference papers. We then conducted a preliminary experiment to evaluate our task. In the experiment, we adopted In-Context Learning (ICL) using LLMs as the baseline method and compared it with the performance of humans. The results demonstrated that the performance of LLMs for *generic mention* and *in-text citation* was comparable to that of humans.

2 Extraction of Information about Research Data

2.1 Metadata of Research Data

Research data are data collected or generated through research activities. In this study, data, such as datasets and code, were treated as research data. For research data to meet the most fundamental principle in the FAIR, i.e., "Findable," it is required to be assigned a unique identifier and described with rich metadata.

2.2 Utilization of Scholarly Papers for Metadata Generation

In scholarly papers, information about research data is provided when the data are created or used for a study. When mentioning the created research

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 72–79

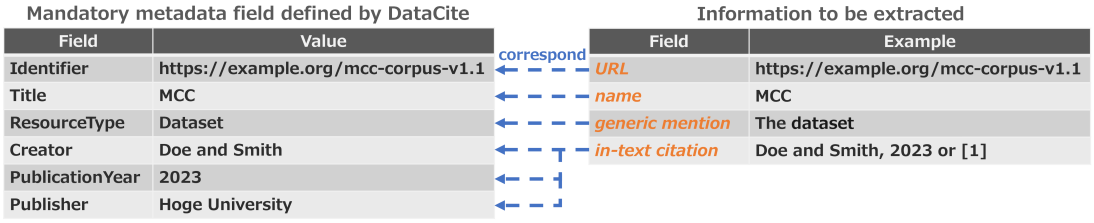

Figure 1: Correspondence between information fields to be extracted and DataCite mandatory metadata fields. Creator, PublicationYear, and Publisher can be obtained not only from the body text of the citing paper, but also from metadata of the cited paper, such as its authors and affiliations. This metadata is often available in the reference list or the header section of the cited paper. *In-text citation* provides access to the cited paper.

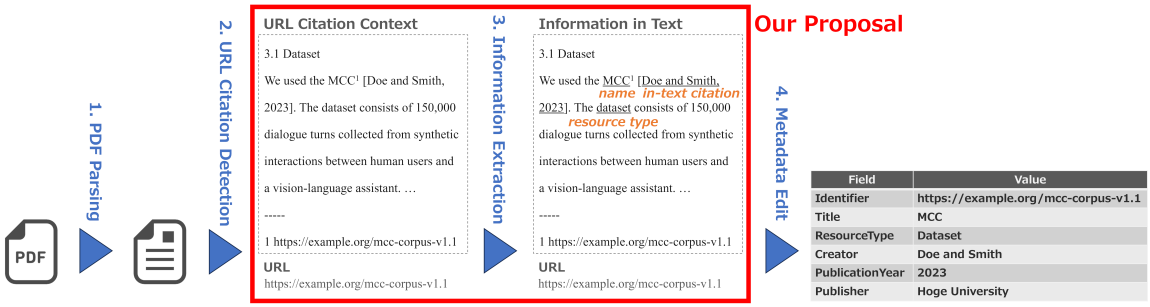

Figure 2: Flow of generating metadata on research data from scholarly papers.

data, the creator notes the name and description of the data and provides access to it, e.g., using a URL. When mentioning the used research data, the user provides the identifier, an overview of the research data, and its usage in the study. This information about research data is often included in the text surrounding mentions of research data in papers. Thus, the information extracted from the text can be used as a source to generate metadata.

2.3 Requirements for Information to be Extracted from Papers

In this study, we assumed metadata generation based on DataCite, which is a global standard and universal metadata schema. Furthermore, DataCite offers an additional advantage of interoperability, as its fields can be mapped to other metadata schemas such as the Dublin Core[1]. DataCite defines a metadata schema with six mandatory fields: "Identifier," "Creator," "Title," "Publisher," "PublicationYear," and "ResourceType." Thus, at minimum, it is necessary to extract the information corresponding to these metadata fields.

Based on the above, we define the following four conditions for the information to be extracted from scholarly papers.

1. Identifier for research data
2. Name of research data
3. Information about the type of research data
4. Information related to the creation of the research data

Figure 1 shows the correspondence between these conditions and mandatory fields in DataCite. The information in the above conditions 1 and 2 can be used for "Identifier" and "Title," respectively. The information in the condition 3 can be used to classify "ResourceType." From the information in the condition 4, it may be possible to generate the "Creator," "Publisher," and "PublicationYear."

2.4 Related Work

Previous studies have tackled the task of extracting information about research data from scholarly papers. Most of these studies extracted information by identifying the names and mentions of research data (Luan et al., 2018; Jain et al., 2020; Schindler et al., 2021; Hou et al., 2021; Pan et al., 2023; Otto et al., 2023; Stavropoulos et al., 2023; Pan et al., 2024; Watanabe et al., 2024). The name and mention detection realized comprehensive extraction of information from scholarly papers, satisfying the condition 2. However, this approach does not necessarily satisfy the condition 1 because it may not include identifiers such as URLs.

[1] http://purl.org/dc/elements/1.1/

Table 1: Fields of information to be extracted.

Field	Explanation
name	name given to research data
generic mention	generic reference to research data
in-text citation	in-text reference marker for research data

Table 2: Statistics of the dataset.

Annotation unit		Value
#(paragraph, URL)		601
# span	*name*	571
	generic mention	435
	in-text citation	202

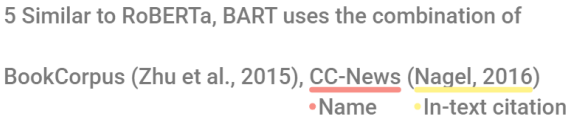

Figure 3: Annotation interface.

In contrast, other studies have obtained information on research data from URL citations in the text of scholarly papers. For example, Tsunokake and Matsubara classified whether URLs in scholarly papers cite research data or not (Tsunokake and Matsubara, 2021). Zhao et al., Tsunokake and Matsubara, and Wada et al. classified types of research data cited by URL using the text surrounding URL citations (Zhao et al., 2019; Tsunokake and Matsubara, 2022; Wada et al., 2024). These studies satisfy the condition 1 because URL is regarded as an identifier. They also satisfy the condition 3 by classifying the type of research data. However, the conditions 2 and 4 are not satisfied because they did not target information excluding URL and type.

3 Extraction from Citation Contexts

3.1 Prerequisites for the Task

The flow of generating metadata on research data from scholarly papers is shown in Figure 2. The procedure is summarized as follows.

1. Parse the paper in PDF format and convert it to semi-structured text.
2. Detect URLs that refer to research data among all URLs in the text and extract segments containing the URLs as body texts.
3. Extract the information about research data from the body text.
4. Edit the extracted strings and generate metadata on the research data.

In the above procedure, step 3 represents the task proposed in this study. A detailed definition of the proposed task is given in Section 3.2. For step 1, several tools have been developed to parse and convert scholarly papers in PDF format to text format (Lopez, 2009; The Apache Software Foundation, 2009; Abekawa and Aizawa, 2016; Mistral AI Team, 2025). Regarding step 2, some URL citations refer to related web pages or scholarly papers rather than the research data. To address this issue, we will adapt a previously proposed URL citation classification method (Tsunokake and Matsubara, 2021). Editing in step 4 is left for future work because it requires advanced techniques, e.g., integrating information extracted from multiple papers.

3.2 Task Settings

We define step 3, the proposed task, as follows.

Input: a pair of a URL citing research data and a URL citation context. If the URL appears in a footnote or the bibliography, its text is concatenated to the body text as the input URL citation context.

Output: strings included in the input text corresponding to *name*, *generic mention*, and *in-text citation* of research data. Table 1 explains these three fields of information.

This task takes text with URL citation as input; thus, it satisfies the condition 1. In addition, the information *name*, *generic mention*, and *in-text citation* satisfy the conditions 2 to 4, respectively.

4 Dataset Construction

For the annotation, we used a dataset constructed in a previous study (Tsunokake and Matsubara, 2022) that targets URL citations of research data. This dataset contains URLs citing research data and their corresponding paragraph texts (i.e., citation contexts), extracted from papers published in notable natural language processing conferences[2]. If URLs appeared in footnotes or bibliographies, the corresponding paragraphs were extracted. In this study, we used a total of 601 URL-paragraph pairs, where the URLs refer to datasets or code[3].

[2] https://aclanthology.org/

[3] Whether URLs refers to research data was determined manually in the previous study.

We asked an expert in corpus annotation in NLP to assign information *name*, *generic mention*, and *in-text citation* to paragraphs (if any). Assigning information of URL-cited research data was done by annotating spans and labels. Table 2 shows the statistics of the dataset. We used the doccano (Nakayama et al., 2018) annotation tool, where the worker annotated the text, as shown in Figure 3.

5 Preliminary Experiment

To verify the feasibility of the proposed task, we conducted a preliminary experiment.

5.1 Experimental Data

The constructed dataset was split into training, development, and test data based on the papers' publication years. The test data included paragraphs from papers published in 2021, the latest year in the dataset. The development data were split such that the proportion of publication years was uniform (excluding the test data). The training data were obtained by excluding the test and development data. Finally, the ratio of the training, development, and test data was 397:107:97.

5.2 Extraction Methods

In this experiment, we compared the extraction performance of LLMs with that of humans. For the human extraction, we asked another worker who majored in NLP to extract information.

We adopted ICL (Brown et al., 2020) with LLMs as the baseline method. We set few-shot settings because the performance of LLMs is affected by the given demo samples. The demo samples were selected based on the similarity between the test input text and the candidate texts.

The prompts comprised an instruction, a demonstration, and a test input[4]. The instruction provided the task definition and label in Section 3 and Table 1, respectively. The demonstration included samples retrieved from the training data.

5.3 Implementation and Evaluation

We used Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen3-8B (Qwen Team, 2025) as open LLMs, and GPT-4.1 (OpenAI, 2025) as a closed LLM. In all LLMs, the decoding method was greedy, and the output format was the JSON schema. In the few-shot setting, e5-Mistral-7b-instruct (Wang et al., 2023) retrieved five samples for each input from the training data.

The evaluation was performed for each information field shown in Table 1. We used entity-based F1 as the evaluation metric. Note that we used the edit distance to determine the match between the output and the ground truth because the format of the model output is not span[5].

Table 3: Comparison of extraction performance between LLMs and humans.

	Llama	Qwen	GPT	Human
name	39.33	41.94	45.98	59.86
generic mention	18.63	29.11	42.60	40.40
in-text citation	38.78	39.58	61.33	63.77
Macro average	32.25	36.88	49.97	54.68

5.4 Results

Table 3 shows the performance of all models, alongside the human performance. Overall, the best-performing LLM, GPT-4.1, still underperformed humans by approximately 5 points. Focusing on each information field, both LLMs and the human achieved high performance for *name* and *in-text citation*. Notably, the extraction of *name* showed a gap of approximately 10 points, suggesting that this aspect remains challenging for LLMs. In contrast, for *in-text citation*, GPT-4.1 achieved an F1 of 61.33%, demonstrating extraction performance comparable to that of humans. For *generic mention*, while the open LLMs' performance was lower than the human performance, GPT-4.1 outperformed humans with an F1 of 42.60%.

6 Error Analysis

To reveal the challenges of information extraction, we conducted an error analysis on the outputs of the best-performing model, GPT-4.1, as well as the human. To analyze errors in detail, we introduced the four categories defined in the Message Understanding Conference (Chinchor and Sundheim, 1993).

Correct (COR): both span and label are perfectly matched.

Partial (PAR): span is partially matched, and labels are matched.

Missing (MIS): ground truth is missed by a system.

[4]The prompt is provided in Appendix A.

[5]The Levenshtein distance was used, and a similarity greater than 0.8 was considered a match.

URL	http://opus.nlpl.eu/
Citation Context	... This data is derived from two main sources: (1) open-source repository of parallel corpora, OPUS [Cite_Footnote_3] (Tiedemann, 2012) and (2) ParaCrawl (Esplà et al., 2019). From OPUS, we use the JW300 corpus (Agić and Vulić, 2019), OpenSubtitles (Lison and Tiedemann, 2016), XhosaNavy, Memat, and QED (Abdelali et al., 2014). Despite the existence of this paral-lel data, these text datasets were often collected from large, relatively unclean multilingual corpora, ... *footnote*: 3 http://opus.nlpl.eu/
ground truth	{*name*: "OPUS", *generic mention*: "this paral-lel data", *in-text citation*: "Tiedemann, 2012"}
GPT	{*name*: ["OPUS", "JW300", "OpenSubtitles", "XhosaNavy", "QED", "Memat"], *generic mention*: "paral-lel data", *in-text citation*: "N/A"}
Human	{*name*: "OPUS", *generic mention*: "N/A", *in-text citation*: "Tiedemann, 2012"}

Figure 4: Representative example of the observed error cases. "[Cite_Footnote_3]" denotes a footnote citation tag (which would normally be the number 3).

Table 4: Number of error categories for each information field.

	name		*generic mention*		*in-text citation*	
	GPT	Human	GPT	Human	GPT	Human
COR	37	43	22	17	21	22
PAR	3	1	14	3	2	0
SPU	67	36	68	14	16	11
MIS	27	23	29	45	13	14

Spurious (SPU): a system produces a response that doesn't exist in the ground truth.

As in the experiment in Section 5, we used the edit distance for span matching.

Figure 4 shows a representative example of the observed error cases. In this example, the correct data name is "OPUS" (COR case), but GPT additionally extracted unrelated names such as "JW300" and "OpenSubtitles" (SPU case). For *generic mention*, the human failed to extract "this parallel data" (MIS case), while GPT produced a partial extraction by outputting "parallel data" (PAR case). Regarding *in-text citation*, GPT failed to extract the citation in this example, again resulting in a MIS error.

Table 4 shows the number of error categories for each information field. For *name*, the human produced approximately six more COR cases than GPT, indicating more accurate extraction. In contrast, GPT produced a substantially larger number of SPU cases, suggesting that it is more likely to extracting incorrect information and that its extraction precision remains challenging. For *generic mention*, both GPT and the human yielded far more SPU and MIS cases than COR, demonstrating that extracting this field is generally challenging. Moreover, GPT tends to generate a huge number of incorrect mentions (higher SPU), whereas the human more frequently fail to extract valid mentions (higher MIS). For *in-text citation*, both GPT and the human produced a high number of COR cases, indicating that this information can be extracted reliably by both humans and models.

7 Conclusion

This paper proposed the task of extracting information about research data from URL citation contexts in scholarly papers, and constructed a dataset thorough text annotations according to the DataCite schema. The result of the preliminary experiment demonstrated that the performance of LLMs matched that of humans in some cases, indicating the feasibility of the proposed task.

8 Limitations

Task We defined the output of the task as the information to generate the mandatory metadata fields in the DataCite schema. However, the schema also includes recommended and optional fields, such as "subject" and "size," which could potentially be extracted from scholarly papers. To generate richer metadata, we should expand the scope of the task to cover a wider range of metadata fields.

Dataset The dataset constructed in this study is limited to conference papers in the field of natural language processing and their associated research data. However, research data, such as datasets or code, are also frequently mentioned in papers from diverse domains, specifically digital libraries and medical research. To improve the domain adaptability of the proposed task, we should extend the dataset to cover a broader range of domains.

Experiment The evaluation in this study was designed as a preliminary investigation, and consequently, the reported performance should be considered exploratory. To perform a more comprehensive evaluation of the proposed task, future experiments should be conducted, including evaluations of several supervised approaches.

9 Ethical Considerations

In this project, annotation workers were employed by a staffing agency in Japan. The workers annotated a total of 601 paragraph-URL pairs. Workers were paid approximately 800 yen ($5) per pair.

Acknowledgments

This work was partially supported by the Grant-in-Aid for Scientific Research (B) (No. 23K21844) of JSPS.

References

Takeshi Abekawa and Akiko Aizawa. 2016. SideNoter: Scholarly Paper Browsing System based on PDF Restructuring and Text Annotation. In *Proceedings of the 26th International Conference on Computational Linguistics: System Demonstrations (COLING 2016)*, pages 136–140.

Tom Brown, Benjamin Mann, Nick Ryder, and 1 others. 2020. Language Models are Few-Shot Learners. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 1877–1901.

Nancy Chinchor and Beth Sundheim. 1993. MUC-5 evaluation metrics. In *Proceedings of the 5th Message Understanding Conference (MUC-5)*.

DataCite Metadata Working Group. 2024. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.6.

G7 OSWG. 2023. Annex 1: G7 Open Science Working Group (OSWG).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXive:2407.21783*.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2021. TDMSci: A Specialized Corpus for Scientific Literature Entity Tagging of Tasks Datasets and Metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*, pages 707–714.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A Challenge Dataset for Document-Level Information Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 7506–7516.

Patrice Lopez. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *Proceedings of the 13th European Conference on Digital Libraries (ECDL 2009)*, pages 473–474.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 3219–3232.

Mistral AI Team. 2025. Mistral OCR.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. Doccano: Text Annotation Tool for Human.

OpenAI. 2025. Introducing GPT-4.1 in the API.

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 *Proceedings of the Findings of the Association for Computational Linguistics (EMNLP 2023)*, pages 8166–8176.

Huitong Pan, Qi Zhang, Cornelia Caragea, Eduard Dragut, and Longin Jan Latecki. 2024. SciDMT: A Large-Scale Corpus for Detecting Scientific Mentions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14407–14417.

Huitong Pan, Qi Zhang, Eduard Dragut, Cornelia Caragea, and Longin Jan Latecki. 2023. DMDD: A Large-Scale Dataset for Dataset Mentions Detection. *Transactions of the Association for Computational Linguistics*, 11:1132–1146.

Qwen Team. 2025. Qwen3 Technical Report. *arXiv preprint arXive:2505.09388*.

David Schindler, Felix Bensmann, Stefan Dietze, and Frank Krüger. 2021. SoMeSci- A 5 Star Open Data Gold Standard Knowledge Graph of Software Mentions in Scientific Articles. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM 2021)*, page 4574–4583.

Petros Stavropoulos, Ioannis Lyris, Natalia Manola, Ioanna Grypari, and Haris Papageorgiou. 2023. Empowering Knowledge Discovery from Scientific Literature: A novel approach to Research Artifact

Analysis. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 37–53.

The Apache Software Foundation. 2009. Apache PDFBox.

Masaya Tsunokake and Shigeki Matsubara. 2021. Classification of URLs Citing Research Artifacts in Scholarly Documents based on Distributed Representations. In *Proceedings of 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE 2021)*, pages 20–25.

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 (WIESP 2022)*, pages 8–19.

Kazuhiro Wada, Masaya Tsunokake, and Shigeki Matsubara. 2024. On an Intermediate Task for Classifying URL Citations on Scholarly Papers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics and Language Resources and Evaluation (LREC-COLING 2024)*, pages 12359–12369.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving Text Embeddings with Large Language Models. *arXiv preprint arXiv:2401.00368*.

Yu Watanabe, Koichiro Ito, and Shigeki Matsubara. 2024. Capabilities and Challenges of LLMs in Metadata Extraction from Scholarly Papers. In *Proceedings of the 26th International Conference on Asia-Pacific Digital Libraries (ICADL 2024)*, volume 1, page 280–287.

Mark Wilkinson, Michel Dumontier, IJsbrand Aalbersberg, and 1 others. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3(160018).

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 (EMNLP-IJCNLP 2019)*, pages 5206–5215.

A Prompt

In the experiment, we employed the chat templates defined by each LLM's tokenizer. Figure 5 shows the prompt constructed using the Llama chat template. For Llama and Qwen, we embedded the target information fields into the prompt using function calling. For GPT-4.1, we incorporated them using the response format.

```
#system
Your task is to extract information about research data cited by the given URL
    from section title, body text and footnote/reference.

You have access to the following functions. To call a function, please respond
    with JSON for a function call.
Respond in the format
{
  "name": "information_extraction",
  "description": "Your task is to extract information about research data cited
      by the given URL from section title, body text and footnote/reference.",
  "parameters": {
    "title": "InfoSchema","type": "object",
    "properties": {
      "name": {"title": "Name",
        "type": "array","items": { "type": "string" },
        "description": "A name or title by which the research data is known. May
             be the title of a dataset or the name of a piece of software or an
            instrument. If no names are given, return N/A"
      },
      "genericmention": {"title": "Genericmention",
        "type": "array","items": { "type": "string" },
        "description": "Generic mention refers to a common noun phrase that
            references the research data. If no generic mentions are given,
            return N/A"
      },
      "citationtag": {"title": "Citationtag",
        "type": "array","items": { "type": "string" },
        "description": "Citation tag is a tag that indicates the citation of a
            scholarly paper related to research data. If no citation tags are
            given, return N/A."
      }
    },
    "required": ["name","genericmention","citationtag"]
  }
}

#demonstration
{pairs of demo input and demo output}

#user
Given URL: https://example.org/mcc-corpus-v1.1
Section Title: 3.1 Dataset
Body Text: We used the MCC[Cite_Footnote_1] (Doe and Smith, 2023). The dataset
    consists of 150,000 dialogue turns collected from synthetic interactions
    between human users and a vision-language assistant. ···
Footnote or Reference Text: 1 https://example.org/mcc-corpus-v1.1
```

Figure 5: A simplified version of the used prompt.