MOLE: Metadata Extraction and Validation in Scientific Papers Using LLMs

Zaid Alyafeai[1] **Maged S. Al-Shaibani**[2] **Bernard Ghanem**[1]
[1]KAUST [2]SDAIA-KFUPM Joint Research Center for AI, KFUPM

Abstract

Metadata extraction is essential for cataloging and preserving datasets, enabling effective research discovery and reproducibility, especially given the current exponential growth in scientific research. While Masader (Alyafeai et al., 2021b) laid the groundwork for extracting a wide range of metadata attributes from Arabic NLP datasets' scholarly articles, it relies heavily on manual annotation. In this paper, we present MOLE, a framework that leverages Large Language Models (LLMs) to automatically extract metadata attributes from scientific papers covering datasets of languages other than Arabic. Our schema-driven methodology processes entire documents across multiple input formats and incorporates robust validation mechanisms for consistent output. Additionally, we introduce a new benchmark to evaluate the research progress on this task. Through systematic analysis of context length, few-shot learning, and web browsing integration, we demonstrate that modern LLMs show promising results in automating this task, highlighting the need for further future work improvements to ensure consistent and reliable performance. We release the code[1] and dataset[2] for the research community.

1 Introduction

"*Metadata is data about data*"

The scientific community is experiencing an unprecedented data revolution, with researchers producing and sharing datasets at an extraordinary rate. However, the value of these datasets decreases when they are inadequately documented or difficult to discover. Extracting structured information that describes the characteristics, origins, and usage of datasets, is a critical challenge (Borgman, 2012; Wilkinson et al., 2016), especially

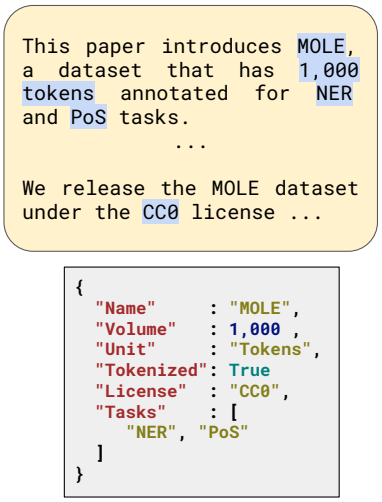

Figure 1: Sample metadata extracted from a dummy paper with highlighted attributes.

in vastly and rapidly growing domains such as natural language processing (NLP). These datasets vary widely in structure, size, format, purpose, and language. Without robust metadata extraction, valuable datasets remain underutilized, research efforts are duplicated, and reproducibility is compromised (Gebru et al., 2021; Dodge et al., 2019). With hundreds of thousands of new publications annually[3], automating metadata extraction is essential to maintain the scalability of the scientific ecosystem.

This work attempts to approach this problem utilizing LLMs to extract the metadata. We define metadata as a JSON object that holds many attributes, such as Year, License, and Paper and Dataset Links. Such attributes vary in constraints: some fixed options (License), some free form (Dataset Description), and some context-dependent (Paper & Dataset Links). While existing automated approaches typically extract around 5-10 attributes (Ahmad and Afzal, 2020; Tkaczyk et al., 2015),

[1] https://github.com/IVUL-KAUST/MOLE
[2] https://huggingface.co/datasets/IVUL-KAUST/MOLE

[3] As of April 2025, there are more than 2.7 million articles on arXiv.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12236–12264
November 4-9, 2025

Table 1: Comparison between MOLE and other methods in the literature.

Method	Attributes	Models	Schema	Benchmark	Formats	Browsing
MOLE	32	7 LLMs	✓	126	3	✓
(Watanabe et al., 2024)	8	2 LLMs	✗	✗	1	✗
(Giner-Miguelez et al., 2022)	23	2 LLMs	✗	12	1	✗
(Ahmad and Afzal, 2020)	9	SVC	✗	✗	1	✗
(Tkaczyk et al., 2015)	17	SVM	✗	✗	1	✗
(Constantin et al., 2013)	18	Rule-based	✗	✗	1	✗
(Councill et al., 2008)	23	CRF	✗	40	1	✗

our work automatically extracts around 30 different attributes per paper, providing a substantially more comprehensive metadata profile (see Table 1). Figure 1 overviews our work of extracting sample attributes from a paper as a simplified example.

Current metadata extraction approaches typically rely on rule-based systems, supervised machine learning, or combinations thereof (Rodríguez Méndez et al., 2021; Ahmad and Afzal, 2020; Tkaczyk et al., 2015; Constantin et al., 2013; Granitzer et al., 2012; Councill et al., 2008). While effective for structured documents, these methods struggle with the heterogeneity of scientific papers and require domain knowledge and maintenance to accommodate evolving document structures (Rizvi, 2020).

Recent advances in Large Language Models (LLMs) have opened new possibilities for information extraction (Team et al., 2023; OpenAI, 2023), with models showing promising results in extracting structured data, (Butcher et al., 2025; Li et al., 2024; Liu et al., 2024; Tam et al., 2024). A prominent advantage of modern LLMs is their ability to handle long contexts (Peng et al., 2023; Zhang et al., 2023; Team et al., 2023), allowing our methodology to process entire papers. We summarize our contributions as follows:

1. A generalized approach for dataset metadata extraction from scientific papers, capable of extracting more than 30 distinct attributes organized in a structured schema hierarchy.

2. A benchmark for metadata extraction involving datasets in multiple languages, covering Arabic, English, Russian, French, and Japanese, enabling systematic evaluation of performance across linguistic domains.

3. An examination of our approach on 7 LLMs, including proprietary and open-source models, analyzing the impact of long-context handling, few-shot learning, and constrained output generation.

2 Methodology

Figure 2 illustrates our MOLE framework. This framework processes scientific papers in either LaTeX source or PDF format, leveraging the power of LLMs' extraction capabilities to identify dataset metadata attributes. When processing LaTeX, the framework directly analyzes the source; for PDF, we study extracting text manually using the available PDF tools vs. prompting an LLM (with vision capabilities) to extract a structured output from the paper. Then, an LLM identifies and structures the metadata according to our schema, which is subsequently validated before producing the final JSON output. This pipeline enables an automated, efficient, and reliable extraction of a comprehensive dataset metadata from diverse document formats and across multiple languages.

2.1 Metdata

We build on the work of Masader (Alyafeai et al., 2021b) by modifying the following attributes:

- **HF Link** we add the Hugging Face[4] link to the dataset.

- **Access** we change the *Cost* attribute to show the actual cost of the dataset and add *Access* to highlight the accessibility of the dataset, including Free, Upon-Request, or Paid.

- **Derived From** we replace *Related datasets* by this attribute to mention all datasets that are used as a seed for a given dataset. This attribute is critical for evaluation as it can indicate any contamination issues.

[4]Currently, `https://huggingface.co/datasets` contains more than 300K datasets.

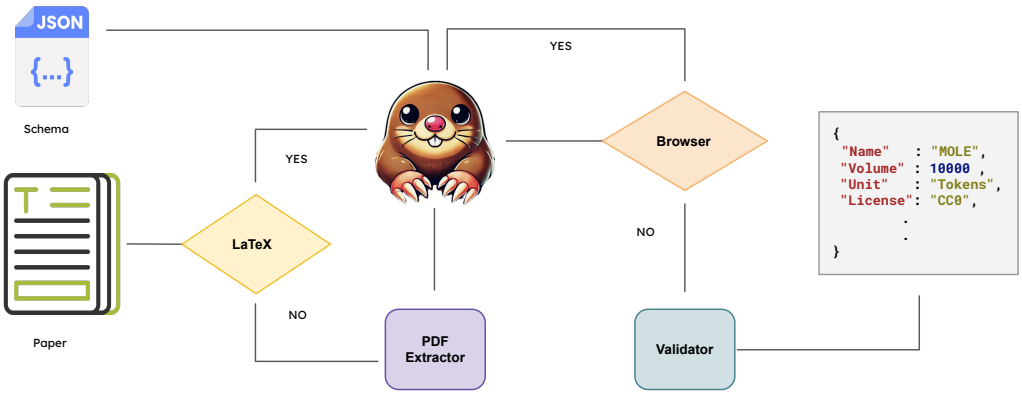

Figure 2: MOLE pipeline. The paper text and Schema are used as input, and the output is the extracted metadata content.

- **Domain** we use the following options to describe the attribute: social media, news articles, commentary, books, Wikipedia, web pages, public datasets, TV channels, captions, LLM, or other. The options are improved by including recent approaches for synthetic data generation using LLMs.
- **Collection Style** similar to *Domain* we use the following options to describe the attribute: crawling, human annotation, machine annotation, manual curation, LLM generated, or other.

In total, we have 32 attributes that can be used to annotate Arabic datasets (See Figure 11). We also extend our approach to datasets in other languages by considering the more general attributes. In this study, we consider the following language categories: Arabic (ar), English (en), French (fr), Japanese (jp), Russian (ru), and multilingual (multi). Subsequently, we extend the metadata attributes to such categories. For example, the *Script* attribute in Japanese could be Kanji, Hiragana, Katakana, or mixed. In other languages, it is fixed, as it is Latin in English and French and Cyrillic in Russian. Note that other languages don't have dialects, so we remove the *Dialect* and *Subset* attributes for monolingual datasets. For the multilingual category, we remove the *Script* attribute and use the *Subsets* to indicate the languages in the dataset and their corresponding size.

2.2 Schema

We define a schema as a JSON representing how an LLM should generate the metadata. Each metadata attribute is represented by a key in the schema. Our metadata schema mainly consists of five keys:

1. **question** specifies what metadata attribute to extract from the document. For example, we ask *What is the license of the dataset?* for the *License* attribute.
2. **options** a list of string values to choose from. The LLM must choose an answer from this list.
3. **option description** a dictionary that explains ambiguous options, for example, it might not be clear what low, medium, or high Ethical Risks mean.
4. **answer type** this field represents the output type for each metadata attribute. The complete list of output types is shown in Table 2.

Table 2: Permissible data types in our schema. The data types are provided in the schema to force the model to generate specific data types.

Type	Description
`str`	string
`url`	link
`date[year]`	year of the date
`List[str]`	list of strings
`float`	floating point number
`bool`	true or false
`List[Dict]`	list of dictionaries

5. **validation group** this field is used to collect similar attributes in a group. Mainly we use this for evaluation.

6. **answer min and max** this field specifies the length of an answer for a given question. As an example, take the *Tasks* attribute; then we assume that each dataset must have at least one task associated with it and at maximum 3 tasks. Hence we will have *answer_min = 1, answer_max = 3*. In general, if *answer_min = 0*, this attribute is optional. If the *answer_max* is not defined, then there are no constraints on the output max length.

In the following example, we show a schema for the *License* attribute. The answer min and max are set to 1 because a dataset must have only one license.

```
{
    "question": "What is the license of
    the dataset?",
    "options": [
        "Apache-2.0", "MIT", ...
    ],
    "answer_type": "str",
    "validation_group":"ACCESSIBILITY",
    "answer_min": 1,
    "answer_max": 1
}
```

Code 1: Example Schema for the License metadata attribute. Options are truncated for better visualization

Table 3: Number of annotated papers and annotated metadata attributes for each category in the collected test dataset.

Category	# papers	# fields	# annotations
ar	21	64	1,344
en	21	58	1,218
fr	21	58	1,218
jp	21	60	1,260
ru	21	58	1,218
multi	21	60	1,260
total	126	358	7,518

2.3 Validation

We mainly use three types of validations to make sure the output is consistent with our schema:

1. **Type Validation** if the output type for a given question is not correct, then we either cast it or use the default value. For example, the volume can be casted to `float` if it is given as `str`.

2. **Option Validation** if there are options to answer the question, then if the answer does not belong to one of the options, we use similarity matching to choose the most similar option.

3. **Length Validation** the output length must be within the range [answer_min, answer_max], otherwise the model will have a low score for length enforcing.

4. **JSON Validation** The generated JSON must be loadable using `json.loads(...)`. To fix unloadable strings, we apply some regex rules. As an example, we remove "'`json` prefixes in some generated JSONs.

2.4 Metrics

We use either exact match or list matching, depending on the answer type in the schema. For list matching, we use set intersection to compare the results. We use flexible matching where we allow at most one difference between the predicted and the ground truth. For dictionary matching, in addition to values, we also match the keys. We use F1 to calculate the score for a given metadata extraction. The precision calculates how much LLMs hallucinate values of non-existent attributes in the paper (lower precision), while recall calculates the accuracy of predicting metadata that exists in the paper.

3 Dataset

We manually annotated 126 papers covering datasets in different languages for testing. We additionally annotate 6 more papers for validation. Two authors participated in the annotation process [5]. We annotate each metadata with two values; the first one is the value of the metadata, and the second with a binary value of 1 if the attribute exists in the paper; otherwise, it is 0. The binary annotation will help us in measuring which metadata exists in the papers and which ones require browsing the Web. For example, the License of a given dataset might not exist in the paper itself, but most likely it can be accessed through the Link. The articles collected span six different categories: Arabic (ar), English (en), French (fr), Japanese (jp), Russian (ru), and Multilingual (multi) datasets. The annotation process included annotating a subset and calculating the agreement percentage of the two annotations. The percentage of agreement between the two authors was found to be 85.93. In Table 3,

[5] Guidelines: https://github.com/IVUL-KAUST/MOLE/blob/molev2_anonymized/GUIDLINES.md

Table 4: Results of all models on the main categories Arabic (ar), English (en), Japanese (jp), French (fr), Russian (ru), Multilingual (multi) datasets. Average shows the weighted average of all categories. Maximum across category is **bold** and the second maximum is underlined.

Model	ar	en	jp	fr	ru	multi	Average
Random	32.29	27.86	32.74	30.27	32.13	23.18	29.74
Keyword	45.37	43.89	44.72	45.81	46.21	37.36	43.89
Gemma 3 27B	59.08	69.35	70.93	69.29	67.88	60.81	66.23
Qwen 2.5 72B	66.26	67.37	70.05	71.94	65.78	65.88	67.88
Llama 4 Maverick	60.77	73.38	72.35	70.01	71.30	68.02	69.30
DeepSeek V3	66.64	73.78	71.26	72.13	71.48	64.88	70.03
Claude 3.5 Sonnet	65.50	71.14	71.63	**75.71**	75.62	68.45	71.34
GPT 4o	67.32	76.14	71.00	72.95	73.85	67.00	71.38
Gemini 2.5 Pro	**68.73**	**80.91**	**77.60**	75.06	**78.00**	**71.09**	**75.23**

we highlight the annotated papers in each category in addition to the number of annotated metadata attributes in each category. Additionally, we create six schemata for the different language categories.

Table 5: Details of models used in our evaluation. Context refers to the maximum context window size in tokens. Underlined models are closed source.

Model	Size (B)	Context
Gemini 2.5 Pro	-	1M
GPT-4o	-	128K
Claude 3.5 Sonnet	-	200K
DeepSeek V3	685	164K
Qwen 2.5	72	33K
Llama 4 Maverick	400	1M
Gemma 3	27	131K

4 Evaluation

We evaluate on 7 state-of-the-art models and 2 baselines. For the baselines, we evaluate a model based on a random choice of options and a model based on keyword extraction. For the other models, we use the OpenRouter[6] API to run all the experiments. The temperature is set to 0.0. We use the validation set to tune the system prompt. We repeat each inference a maximum of 6 times until there is no error. We evaluate a diverse set of proprietary and open-source LLMs—ranging from around 30 to over 600 billion parameters—to assess their capabilities on this task, as detailed in Table 5.

4.1 Categories

We evaluated all models in the different language categories in Table 4. Generally speaking, Gemini 2.5 Pro achieves the highest average score. Across different categories, Gemini 2.5 Pro also achieves the highest score in 5 out of the 6 categories. Smaller LLMs like Gemma 3, with only 27B parameters, achieves decent results on the benchmark. We note that Claude Sonnet 3.5 results in many errors, which caused its score to be lower for some categories.

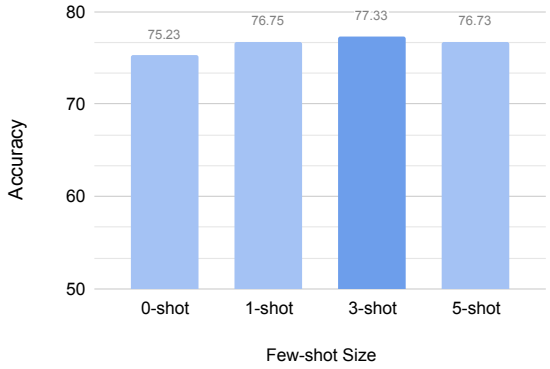

Figure 3: Few-shot results with 0, 1, 3, and 5 -shot examples using the Gemini 2.5 Pro model.

4.2 Few-shot

Since our benchmarks requires structured formatting, we test with different n-shot examples. Since processing multiple papers in a few-shot is expensive, we rely on synthetic examples creation and only evaluate the results using our top model, which is Gemini 2.5 Pro. We create the examples using a template and fill the attributes randomly (see Appendix F for more details). In Figure 3, we show that providing examples improve the results compared to zero-shot. In particular, 3-shot examples achieve the highest gain in results compared to zero-shot.

[6] https://openrouter.ai

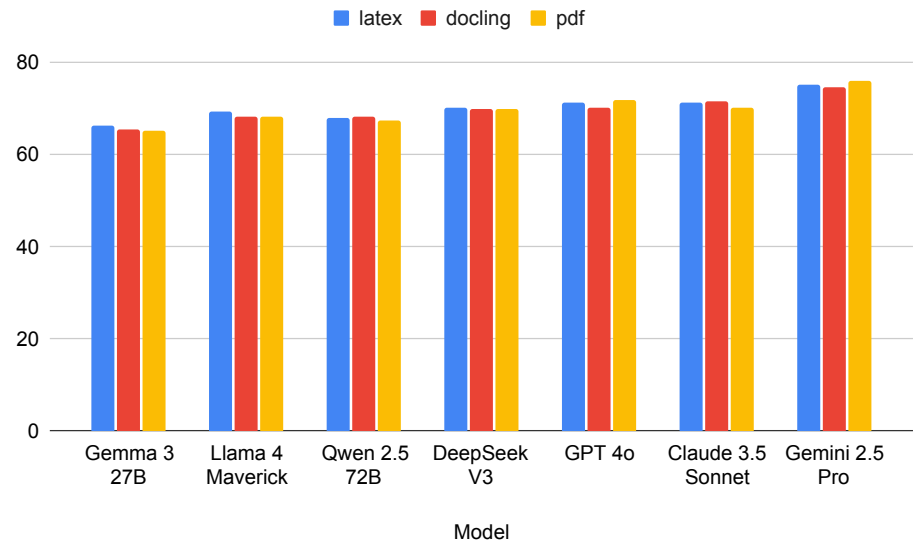

Figure 4: Latex vs. PDF vs. Docling input formats results across all models.

4.3 Input format

We experiment with three approaches for test input, using LaTeX, PDF text, and structured output using Docling (Auer et al., 2024). We are interested in validating the performance on other input formats, as, in many scenarios, the LaTeX source may not necessarily be available. To extract the text content of a PDF, we use Python pdfplumber[7]. We observe that for smaller models, the LaTeX format is slightly better compared to PDF and Docling. However, we don't observe a clear trend across all models.

4.4 Browsing

Some annotated attributes might not exist in the papers. For example, the License attribute is mostly extracted from the repository where the dataset is hosted. To allow all the models to browse, we use the extracted metadata from the non-browsing approach and the page where the dataset is hosted to predict the updated metadata attributes. For repositories that contain a README.md file, like GitHub and HuggingFace, we fetch the file directly from the repository. In Figure 5, we show the results across all the models. We observe a clear improvement when using browsing for all of the models. We note that the precision highlights more the effect of browsing, as it indicates whether models can reliably predict attributes that are not in the paper. Our experiences show that recall is affected by browsing, as the attributes that exist in the paper might get deflected by browsing data.

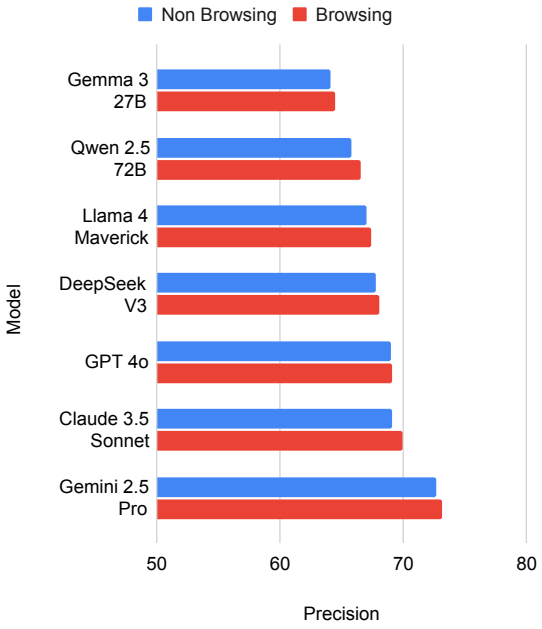

Figure 5: Browsing vs. no Browsing for all models in our evaluation benchmarks.

4.5 Length Enforcing

We use the min and max of the answer to check if the answer from a given LLM respects the required length from the schema. Additionally, to see the effect of increasing the length constraints, we define the following three granularities:

[7] https://pypi.org/project/pdfplumber/

Table 6: Model scores across three different constraints for the length of answer output (Low, Mid, High). The values are normalized by all fields in the metadata.

Model	Low	Mid	High
Gemma 3 27B	0.99	0.96	0.93
Qwen 2.5 72B	1.00	0.98	0.94
DeepSeek V3	1.00	0.99	0.95
GPT 4o	0.99	0.96	0.95
Claude 3.5 Sonnet	0.99	0.99	0.96
Llama 4 Maverick	1.00	0.98	0.96
Gemini 2.5 Pro	1.00	0.98	0.97

1. **Low** this is the standard type of constraint used in all previous experiments. It is considered more relaxed compared to others.

2. **Mid** a medium constraint used to decrease the range of the following attributes: Name, Description, Provider, Derived From, and Tasks.

3. **High** similar to Mid, we use the same attributes but with more stricter range.

As an example, the attribute *Description*, will have the *answer_max* as (50, 25, 12) for low, mid, and high, respectively. In Table 6, we highlight the results across all the models for low, mid, and high length constraints. Still, Gemini 2.5 Pro achieves the highest adherence to high constraints. We note that LLMs that cause many errors in structured output will adhere more to constraints as the output schema will be empty.

4.6 Context Length

In Figure 6, we show the effect of varying the context length on the results of all models. Interestingly, Gemini 2.5 Pro can still achieve competitive results by only selecting half or a quarter of the context. This shows that most of the metadata can be extracted at the upper part of the paper. A similar trend can also be seen for GPT-4o, Qwen, and DeepSeek. For other models, the results are affected significantly, especially for Llama and Claude Sonnet, where the results decrease dramatically. Also, we noticed the error frequency increased for such models when using a smaller context.

4.7 Validation Groups

Our metadata attributes are divided into four validation groups, which are diversity, accessibility, content, and evaluation (see Appendix G and Figure 11). In Table 7, we compare the results of different models on the different validation groups. Gemini 2.5 Pro achieves the highest results in 3 out of the 4 groups. In general, we see all models struggle to achieve high scores for accessibility, which requires extracting the Link, License, and Host, etc. On the other hand, diversity is the easiest group to extract attributes for. Figure 7 shows the results of 6 attributes across different models. While Gemini 2.5 Pro achieves the highest average score, it doesn't achieve the highest scores across all attributes. Interestingly, some LLMs might achieve the highest scores on a single attribute like Llama on the *Link* attribute.

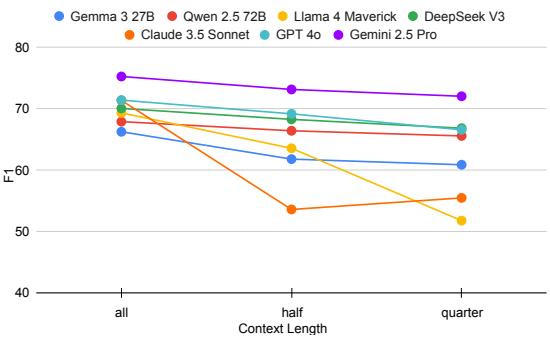

Figure 6: The effect of changing the context length on the results of all models.

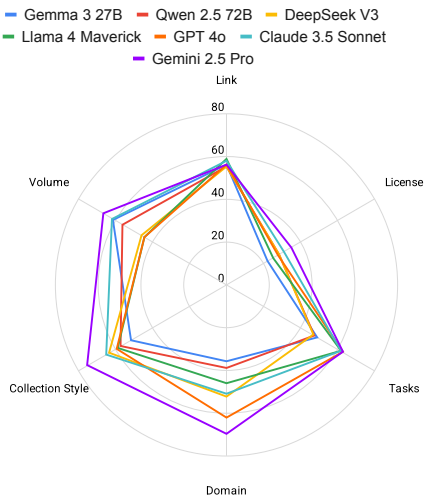

Figure 7: Precision across 6 different metadata attributes (Link, Volume, License, Collection Style, Domain, and Tasks).

Table 7: Model comparison across different validation groups using F1 score.

Model	Diversity	Accessibility	Content	Evaluation	Average
Random	46.83	24.95	27.64	34.32	33.43
Keyword	67.20	34.78	44.43	49.26	48.91
Qwen 2.5 72B	79.50	67.00	68.85	58.73	68.52
Gemma 3 27B	81.08	66.00	62.14	67.46	69.17
Llama 4 Maverick	81.08	68.04	70.88	62.28	70.57
DeepSeek V3	82.54	67.55	70.99	63.60	71.17
Claude 3.5 Sonnet	78.57	67.35	72.33	70.48	72.18
GPT 4o	84.13	67.78	70.64	**71.69**	73.56
Gemini 2.5 Pro	**87.17**	**69.24**	**78.75**	70.42	**76.40**

5 Related Work

The exponential growth of research data has made metadata extraction increasingly critical (Gebru et al., 2021; Mahadevkar et al., 2024; Yang et al., 2025). We examine the evolution and current state of metadata extraction research across three areas.

Evolution of Metadata Extraction Approaches Early systems relied on rule-based and traditional machine learning methods. CERMINE (Tkaczyk et al., 2015) employed modular architectures for bibliographic extraction, building on methods like PDFX (Constantin et al., 2013). FLAG-PDFe (Ahmad and Afzal, 2020) introduced feature-oriented frameworks using SVMs for scientific publications.

Deep learning marked a paradigm shift in this field. (An et al., 2017) introduced neural sequence labeling for citation metadata extraction. Multimodal approaches (Boukhers and Bouabdallah, 2022; Liu et al., 2018) integrated NLP with computer vision to handle layout diversity in PDF documents. Recent work includes domain-specific applications in chemistry (Schilling-Wilhelmi et al., 2024; Zhu and Cole, 2022), HPC (Schembera, 2021), and cybersecurity (Pizzolante et al., 2024), with some exploring LLMs for metadata extraction. Cross-lingual approaches have addressed language-specific challenges, including Korean complexities (Kong et al., 2022), Persian (Rahnama et al., 2020), and Arabic NLP resource cataloging through Masader (Alyafeai et al., 2021b; Altaher et al., 2022).

Standardization Efforts (Gebru et al., 2021) proposed standardized templates for ML dataset documentation, influencing practices across digital heritage (Alkemade et al., 2023), healthcare (Rostamzadeh et al., 2022), energy (Heintz, 2023), art (Srinivasan et al., 2021), and earth science (Connolly et al., 2025). DescribeML (Giner-Miguelez et al., 2022) provided a domain-specific language with IDE integration for practical implementation[8].

Evaluation Benchmarks Several benchmarks exist for metadata extraction evaluation. PARDA (Fan et al., 2019) provides annotated samples across domains and formats. The unarXive corpus (Saier and Färber, 2020) represents one of the largest scholarly datasets with full-text publications and metadata links. DocBank (Li et al., 2020b) and (Meuschke et al., 2023) offer additional evaluation frameworks. However, these focus on general paper attributes (title, authors, abstract) rather than detailed dataset characteristics like volume, license, and subsets that our work addresses.

6 Conclusion

This paper introduced a methodology for using LLMs to extract and validate metadata from scientific papers. Through our framework, which extracts around 30 distinct metadata attributes, we offer a more robust approach to automating metadata extraction. We experiment with multiple approaches, including different input formats, few-shot examples, and browsing. We also highlight the effect of length constraints and how LLMs still struggle to follow strict instructions. Throughout our experiments, we show recent advancements in processing long context, especially in flagship models like Gemini 2.5 Pro and GPT-4o that continue to achieve better results. We also release a benchmark of 126 papers manually annotated to facilitate research in metadata extraction. This work contributes not only to the advancement of metadata extraction techniques but also to the broader goal of making scientific research more transparent, accessible, and reusable.

[8] https://code.visualstudio.com/

Limitations

In this section, we provide some possible limitations and our proposed mitigation and possible future suggestions to improve our work:

1. **Cost** Processing thousands of tokens for a given paper increases the cost. Interestingly, We showed in this paper that most metadata can be extracted using a smaller context. One possible future direction is to make this process more cost-effective by reducing the context size. The context size can be reduced by using some kind of early skimming using a lighter/less expensive LLM.

2. **Length Enforcing** Length constraining is a difficult problem, and current LLMs are not capable of reliably predicting the exact number of tokens (Muennighoff et al., 2025). As a possible direction, we can use precise controlling methods like (Butcher et al., 2025) to extract structured data with better adherence to length constraints. If LLMs become optimal at length enforcing, they can become more cost-effective as we can generate as many tokens as we need.

3. **Source Code Availability** Our approach mostly depends on the availability of LaTeX source code. To mitigate this, we also compare the results for other input formats like PDF and structured format using Docling. However, using such an approach might be difficult to scale, especially due to time constraints. As a future direction, we will explore how to improve the skimming process and cleaning of PDF content. We can utilize this method to scale our approach to thousands of papers.

Acknowledgments

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST) - Center of Excellence for Generative AI, under award number 5940.

We want also to thank Ali Fadel and Amr Keleg for the useful discussions.

References

Muhammad Waqas Ahmad and Muhammad Tanvir Afzal. 2020. FLAG-PDFe: Features oriented metadata extraction framework for scientific publications. *IEEE Access*, 8:99458–99470.

Anna Aksenova, Ekaterina Gavrishina, Elisey Rykov, and Andrey Kutuzov. 2022. Rudsi: graph-based word sense induction dataset for russian. *arXiv preprint arXiv: 2209.13750*.

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2014. Polyglot-ner: Massive multilingual named entity recognition. *arXiv preprint arXiv: 1410.3791*.

Nadège Alavoine, Gaëlle Laperriere, Christophe Servan, Sahar Ghannay, and Sophie Rosset. 2024. New semantic task for the french spoken language understanding media benchmark. *arXiv preprint arXiv: 2403.19727*.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. *arXiv preprint arXiv: 1609.05625*.

Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell, and Steve Renals. 2015. Automatic dialect detection in arabic broadcast speech. *arXiv preprint arXiv: 1509.06928*.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic mgb-3. *arXiv preprint arXiv: 1709.07276*.

Henk Alkemade, Steven Claeyssens, Giovanni Colavizza, Nuno Freire, Jörg Lehmann, Clemens Neudeker, Giulia Osti, and 1 others. 2023. Datasheets for digital cultural heritage datasets. *Journal of open humanities data*, 9(17):1–11.

Manel Aloui, Hasna Chouikhi, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024. 101 billion arabic words dataset. *arXiv preprint arXiv: 2405.01590*.

Ali Alshehri, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2020. Understanding and detecting dangerous speech in social media. *arXiv preprint arXiv: 2005.06608*.

Yousef Altaher, Ali Fadel, Mazen Alotaibi, Mazen Alyazidi, Mishari Al-Mutairi, Mutlaq Aldhbuiub, Abdulrahman Mosaibah, Abdelrahman Rezk, Abdulrazzaq Alhendi, Mazen Abo Shal, and 1 others. 2022. Masader plus: A new interface for exploring+ 500 arabic nlp datasets. *arXiv preprint arXiv:2208.00932*.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv: 2004.06465*.

Zaid Alyafeai, Maged S. Al-shaibani, Mustafa Ghaleb, and Yousif Ahmed Al-Wajih. 2021a. Calliar: An online handwritten dataset for arabic calligraphy. *arXiv preprint arXiv: 2106.10745*.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran A. Q. Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged S. Al-Shaibani. 2024. Cidar: Culturally relevant instruction dataset for arabic. *arXiv preprint arXiv: 2402.03177*.

Zaid Alyafeai, Maraim Masoud, Mustafa Ghaleb, and Maged S Al-shaibani. 2021b. Masader: Metadata sourcing for arabic text and speech data resources. *arXiv preprint arXiv:2110.06744*.

Mohamed Seghir Hadj Ameur, Farid Meziane, and Ahmed Guessoum. 2019. Anetac: Arabic named entity transliteration and classification dataset. *arXiv preprint arXiv: 1907.03110*.

Dong An, Liangcai Gao, Zhuoren Jiang, Runtao Liu, and Zhi Tang. 2017. Citation metadata extraction via deep neural network-based segment sequence labeling. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1967–1970.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv: 1812.10464*.

Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. Xor qa: Cross-lingual open-retrieval question answering. *arXiv preprint arXiv: 2010.11856*.

Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikolaos Livathinos, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Fabian Lindlbauer, Kasper Dinkla, and 1 others. 2024. Docling technical report. *arXiv preprint arXiv:2408.09869*.

Sophie Balech, Christophe Benavent, and Mihai Calciu. 2020. The first french covid19 lockdown twitter dataset. *arXiv preprint arXiv: 2005.05075*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv: 2308.16884*.

David Beauchemin, Julien Laumonier, Yvan Le Ster, and Marouane Yassine. 2022. "fijo": a french insurance soft skill detection dataset. *arXiv preprint arXiv: 2204.05208*.

Yonatan Belinkov, Alexander Magidow, Alberto Barrón-Cedeño, Avi Shmidman, and Maxim Romanov. 2018. Studying the history of the arabic language: Language technology and a large-scale historical corpus. *arXiv preprint arXiv: 1809.03891*.

Yonatan Belinkov, Alexander Magidow, Maxim Romanov, Avi Shmidman, and Moshe Koppel. 2016. Shamela: A large-scale historical arabic corpus. *arXiv preprint arXiv: 1612.08989*.

Abhyuday Bhartiya, Kartikeya Badola, and Mausam. 2021. Dis-rex: A multilingual dataset for distantly supervised relation extraction. *arXiv preprint arXiv: 2104.08655*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *arXiv preprint arXiv: 1911.11641*.

Sergey Bondarkov, Victor Ledenev, and Dmitriy Skougarevskiy. 2025. Russian financial statements database: A firm-level collection of the universe of financial statements. *arXiv preprint arXiv: 2501.05841*.

Christine L Borgman. 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6):1059–1078.

Zeyd Boukhers and Azeddine Bouabdallah. 2022. Vision and natural language for metadata extraction from scientific pdf documents: a multimodal approach. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–5.

Caroline Brun and Vassilina Nikoulina. 2024. Frenchtoxicityprompts: a large benchmark for evaluating and mitigating toxicity in french texts. *arXiv preprint arXiv: 2406.17566*.

Bradley Butcher, Michael O'Keefe, and James Titchener. 2025. Precise length control for large language models. *Natural Language Processing Journal*, page 100143.

Mauro Cettolo. 2016. An arabic-hebrew parallel corpus of ted talks. *arXiv preprint arXiv: 1610.00572*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv: 1905.10044*.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur P. Parikh. 2023. Seahorse: A multilingual, multifaceted dataset for summarization evaluation. *arXiv preprint arXiv: 2305.13194*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman.

2021. Training verifiers to solve math word problems. *arXiv preprint arXiv: 2110.14168*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv: 1809.05053*.

Charlotte J Connolly, Daniel M Hueholt, and Melissa A Burt. 2025. Datasheets for earth science datasets. *Bulletin of the American Meteorological Society*.

Alexandru Constantin, Steve Pettifer, and Andrei Voronkov. 2013. PDFX: Fully-automated PDF-to-XML conversion of scientific literature. In *Proceedings of the 2013 ACM symposium on Document engineering*, pages 177–180.

Isaac G Councill, C Lee Giles, and Min-Yen Kan. 2008. ParsCit: An open-source CRF reference string parsing package. In *Proceedings of LREC*, volume 8, pages 661–667.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv: 2005.00547*.

Martin d'Hoffschmidt, Wacim Belblidia, Tom Brendlé, Quentin Heinrich, and Maxime Vidal. 2020. Fquad: French question answering dataset. *arXiv preprint arXiv: 2002.06071*.

Shizhe Diao, Yu Yang, Yonggan Fu, Xin Dong, Dan Su, Markus Kliegl, Zijia Chen, Peter Belcak, Yoshi Suhara, Hongxu Yin, Mostofa Patwary, Yingyan, Lin, Jan Kautz, and Pavlo Molchanov. 2025. Climb: Clustering-based iterative data mixture bootstrapping for language model pre-training. *arXiv preprint arXiv: 2504.13161*.

Shahd Dibas, Christian Khairallah, Nizar Habash, Omar Fayez Sadi, Tariq Sairafy, Karmel Sarabta, and Abrar Ardah. 2022. Maknuune: A large open palestinian arabic lexicon. *arXiv preprint arXiv: 2210.12985*.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. 2019. Show your work: Improved reporting of experimental results. 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)*, pages 2185–2194.

Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2019. Sberquad – russian reading comprehension dataset: Description and analysis. *arXiv preprint arXiv: 1912.09723*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv: 2305.07759*.

Tiantian Fan, Junming Liu, Yeliang Qiu, Congfeng Jiang, Jilin Zhang, Wei Zhang, and Jian Wan. 2019. Parda: a dataset for scholarly pdf document metadata extraction evaluation. In *Collaborative Computing: Networking, Applications and Worksharing: 14th EAI International Conference, CollaborateCom 2018, Shanghai, China, December 1-3, 2018, Proceedings 14*, pages 417–431. Springer.

Chayma Fourati, Abir Messaoudi, and Hatem Haddad. 2020. Tunizi: a tunisian arabizi sentiment analysis dataset. *arXiv preprint arXiv: 2004.14303*.

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. From freem to d'alembert: a large corpus and a language model for early modern french. *arXiv preprint arXiv: 2202.09452*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Joan Giner-Miguelez, Abel Gómez, and Jordi Cabot. 2022. Describeml: a tool for describing machine learning datasets. In *Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings*, pages 22–26.

Taisia Glushkova, Alexey Machnev, Alena Fenogenova, Tatiana Shavrina, Ekaterina Artemova, and Dmitry I. Ignatov. 2020. Danetqa: a yes/no question answering dataset for the russian language. *arXiv preprint arXiv: 2010.02605*.

Rahul Goel, Waleed Ammar, Aditya Gupta, Siddharth Vashishtha, Motoki Sano, Faiz Surani, Max Chang, HyunJeong Choe, David Greene, Kyle He, Rattima Nitisaroj, Anna Trukhina, Shachi Paul, Pararth Shah, Rushin Shah, and Zhou Yu. 2023. Presto: A multilingual dataset for parsing realistic task-oriented dialogs. *arXiv preprint arXiv: 2303.08954*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv: 2106.03193*.

Michael Granitzer, Maya Hristakeva, Rhiannon Knight, Kris Jack, and Roman Kern. 2012. A comparison of layout based bibliographic metadata extraction techniques. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, pages 1–8.

Veronika Grigoreva, Anastasiia Ivanova, Ilseyar Alimova, and Ekaterina Artemova. 2024. Rubia: A russian language bias detection dataset. *arXiv preprint arXiv: 2403.17553*.

Ilya Gusev. 2020. Dataset for automatic summarization of russian news. *arXiv preprint arXiv: 2006.11063*.

Ilya Gusev and Alexey Tikhonov. 2021. Headlinecause: A dataset of news headlines for detecting causalities. *arXiv preprint arXiv: 2108.12626*.

Julien Hauret, Malo Olivier, Thomas Joubaud, Christophe Langrenne, Sarah Poirée, Véronique Zimpfer, and Éric Bavu. 2024. Vibravox: A dataset of french speech captured with body-conduction audio sensors. *arXiv preprint arXiv: 2407.11828*.

Ahmed Heakl, Youssef Mohamed, and Ahmed B. Zaky. 2024. Araspider: Democratizing arabic-to-sql. *arXiv preprint arXiv: 2402.07448*.

Quentin Heinrich, Gautier Viaud, and Wacim Belblidia. 2021. Fquad2.0: French question answering and knowing that you know nothing. *arXiv preprint arXiv: 2109.13209*.

Ilana Heintz. 2023. Datasheets for energy datasets: An ethically-minded approach to documentation. In *Companion Proceedings of the 14th ACM International Conference on Future Energy Systems*, pages 40–51.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv: 2009.03300*.

Kushan Hewapathirana, Nisansa de Silva, and C. D. Athuraliya. 2024. M2ds: Multilingual dataset for multi-document summarisation. *arXiv preprint arXiv: 2407.12336*.

Masanori Hirano, Masahiro Suzuki, and Hiroki Sakaji. 2023. llm-japanese-dataset v0: Construction of japanese chat dataset for large language models and its methodology. *arXiv preprint arXiv: 2305.12720*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2023. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv: 2309.12053*.

Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. Multi2woz: A robust multilingual dataset and conversational pretraining for task-oriented dialog. *arXiv preprint arXiv: 2205.10400*.

Julie Hunter, Jérôme Louradour, Virgile Rennard, Ismaïl Harrando, Guokan Shang, and Jean-Pierre Lorré. 2023. The claire french dialogue dataset. *arXiv preprint arXiv: 2311.16840*.

Benjamin Icard, François Maine, Morgane Casanova, Géraud Faye, Julien Chanson, Guillaume Gadek, Ghislain Atemezing, François Bancilhon, and Paul Égré. 2024. A multi-label dataset of french fake news: Human and machine insights. *arXiv preprint arXiv: 2403.16099*.

Yusuke Ide, Masato Mita, Adam Nohejl, Hiroki Ouchi, and Taro Watanabe. 2023. Japanese lexical complexity for non-native readers: A new dataset. *arXiv preprint arXiv: 2306.17399*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv: 1705.03551*.

Fanny Jourdan, Yannick Chevalier, and Cécile Favre. 2025. Fairtranslate: An english-french dataset for gender bias evaluation in machine translation by overcoming gender binarity. *arXiv preprint arXiv: 2504.15941*.

Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, and Elizaveta Petrova. 2023. Slovo: Russian sign language dataset. *arXiv preprint arXiv: 2305.14527*.

Nikolay Karpov, Alexander Denisenko, and Fedor Minkin. 2021. Golos: Russian dataset for speech research. *arXiv preprint arXiv: 2106.10161*.

Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano. 2020. Project piaf: Building a native french question-answering dataset. *arXiv preprint arXiv: 2007.00968*.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. *arXiv preprint arXiv: 2010.02573*.

Mohammed Khalil and Mohammed Sabry. 2024. Athar: A high-quality and diverse dataset for classical arabic to english translation. *arXiv preprint arXiv: 2407.19835*.

Ali Can Kocabiyikoglu, François Portet, Prudence Gibert, Hervé Blanchon, Jean-Marc Babouchkine, and Gaëtan Gavazzi. 2022. A spoken drug prescription dataset in french for spoken language understanding. *arXiv preprint arXiv: 2207.08292*.

Hyesoo Kong, Hwamook Yoon, Jaewook Seol, Mihwan Hyun, Hyejin Lee, Soonyoung Kim, and Wonjun Choi. 2022. Annotated open corpus construction and bert-based approach for automatic metadata extraction from korean academic papers. *IEEE Access*, 11:825–838.

Vladislav Korablinov and Pavel Braslavski. 2020. Rubq: A russian dataset for question answering over wikidata. *arXiv preprint arXiv: 2005.10659*.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy

Baldwin. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. *arXiv preprint arXiv: 2402.12840*.

Abdullatif Köksal and Arzucan Özgür. 2020. The relx dataset and matching the multilingual blanks for cross-lingual relation classification. *arXiv preprint arXiv: 2010.09381*.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Frenchmedmcqa: A french multiple-choice question answering dataset for medical domain. *arXiv preprint arXiv: 2304.04280*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv: 1704.04683*.

Antoine Lefebvre-Brossard, Stephane Gazaille, and Michel C. Desmarais. 2023. Alloprof: a new french question-answer education dataset and its use in an information retrieval case study. *arXiv preprint arXiv: 2302.07738*.

Thibaud Leteno, Irina Proskurina, Antoine Gourru, Julien Velcin, Charlotte Laclau, Guillaume Metzler, and Christophe Gravier. 2025. Histoires morales: A french dataset for assessing moral alignment. *arXiv preprint arXiv: 2501.17117*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv: 1910.07475*.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2020a. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. *arXiv preprint arXiv: 2008.09335*.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020b. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*.

Yinghao Li, Rampi Ramprasad, and Chao Zhang. 2024. A simple but effective approach to improve structured language model output for information extraction. *arXiv preprint arXiv:2402.13364*.

Michael Xieyang Liu, Frederick Liu, Alexander J Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J Cai. 2024. " we need structured output": Towards user-centered constraints on large language model output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9.

Runtao Liu, Liangcai Gao, Dong An, Zhuoren Jiang, and Zhi Tang. 2018. Automatic document metadata extraction based on deep networks. In *Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6*, pages 305–317. Springer.

Antoine Louis and Gerasimos Spanakis. 2021. A statutory article retrieval dataset in french. *arXiv preprint arXiv: 2108.11792*.

Natalia Loukachevitch, Ekaterina Artemova, Tatiana Batura, Pavel Braslavski, Ilia Denisov, Vladimir Ivanov, Suresh Manandhar, Alexander Pugachev, and Elena Tutubalina. 2021. Nerel: A russian dataset with nested named entities, relations and events. *arXiv preprint arXiv: 2108.13112*.

Michael J. Lyons. 2021. "excavating ai" re-excavated: Debunking a fallacious account of the jaffe dataset. *arXiv preprint arXiv: 2107.13998*.

Supriya V Mahadevkar, Shruti Patil, Ketan Kotecha, Lim Way Soong, and Tanupriya Choudhury. 2024. Exploring ai-driven approaches for unstructured document analysis and future horizons. *Journal of Big Data*, 11(1):92.

Igor Markov, Sergey Nesteruk, Andrey Kuznetsov, and Denis Dimitrov. 2023. Rustitw: Russian language text dataset for visual text in-the-wild recognition. *arXiv preprint arXiv: 2303.16531*.

Norman Meuschke, Apurva Jagdale, Timo Spinde, Jelena Mitrović, and Bela Gipp. 2023. A benchmark of pdf information extraction tools using a multi-task and multi-domain evaluation framework for academic documents. In *International Conference on Information*, pages 383–405. Springer.

Paul Michel and Graham Neubig. 2018. Mtnt: A testbed for machine translation of noisy text. *arXiv preprint arXiv: 1809.00388*.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. Rucola: Russian corpus of linguistic acceptability. *arXiv preprint arXiv: 2210.12814*.

Elena Mikhalkova. 2021. A russian jeopardy! data set for question-answering systems. *arXiv preprint arXiv: 2112.02325*.

Mehrad Moradshahi, Tianhao Shen, Kalika Bali, Monojit Choudhury, Gaël de Chalendar, Anmol Goel, Sungkyun Kim, Prashant Kodali, Ponnurangam Kumaraguru, Nasredine Semmar, Sina J. Semnani, Jiwon Seo, Vivek Seshadri, Manish Shrivastava, Michael Sun, Aditya Yadavalli, Chaobin You, Deyi Xiong, and Monica S. Lam. 2023. X-risawoz: High-quality end-to-end multilingual dialogue datasets and few-shot agents. *arXiv preprint arXiv: 2306.17674*.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2019. Jparacrawl: A large scale web-based english-japanese parallel corpus. *arXiv preprint arXiv: 1911.10668*.

Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. 2022. Emojis as anchors to detect arabic offensive language and hate speech. *arXiv preprint arXiv: 2201.06723*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Taichi Murayama, Shohei Hisada, Makoto Uehara, Shoko Wakamiya, and Eiji Aramaki. 2022. Annotation-scheme reconstruction for "fake news" and japanese fake news dataset. *arXiv preprint arXiv: 2204.02718*.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2014. Labr: A large scale arabic sentiment analysis benchmark. *arXiv preprint arXiv: 1411.6718*.

Atsumoto Ohashi, Ryu Hirai, Shinya Iizuka, and Ryuichiro Higashinaka. 2024. Jmultiwoz: A large-scale japanese multi-domain task-oriented dialogue dataset. *arXiv preprint arXiv: 2403.17319*.

Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. 2024. Building a large japanese web corpus for large language models. *arXiv preprint arXiv: 2404.17733*.

Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. 2024. Jdocqa: Japanese document question answering dataset for generative language models. *arXiv preprint arXiv: 2403.19454*.

OpenAI. 2023. GPT-4 technical report. arXiv:2303.08774.

Hiroki Ouchi, Hiroyuki Shindo, Shoko Wakamiya, Yuki Matsuda, Naoya Inoue, Shohei Higashiyama, Satoshi Nakamura, and Taro Watanabe. 2023. Arukikata travelogue dataset. *arXiv preprint arXiv: 2305.11444*.

Aissam Outchakoucht and Hamza Es-Samaali. 2021. Moroccan dialect -darija- open dataset. *arXiv preprint arXiv: 2103.09687*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv: 1606.06031*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv: 2306.01116*.

Bowen Peng, Sami Xie, Hao Yao, Karthik Gu, Skyler Reynolds, Tri Shen, and Denny Zhou. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, and 1090 others. 2025. Humanity's last exam. *arXiv preprint arXiv: 2501.14249*.

Dina Pisarevskaya and Tatiana Shavrina. 2022. Wikiomnia: generative qa corpus on the whole russian wikipedia. *arXiv preprint arXiv: 2204.08009*.

Raffaele Pizzolante, Arcangelo Castiglione, and Francesco Palmieri. 2024. Unlocking insights: An extensible framework for automated metadata extraction from online documents. In *2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 1512–1521. IEEE.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv: 2005.00333*.

Reid Pryzant, Yongjoo Chung, Dan Jurafsky, and Denny Britz. 2017. Jesc: Japanese-english subtitle corpus. *arXiv preprint arXiv: 1710.10639*.

Alexander Pugachev, Alena Fenogenova, Vladislav Mikhailov, and Ekaterina Artemova. 2025. Repa: Russian error types annotation for evaluating text generation and judgment capabilities. *arXiv preprint arXiv: 2503.13102*.

Mohadese Rahnama, Seyed Mohammad Hossein Hasheminejad, and Jalal A Nasiri. 2020. Automatic metadata extraction from iranian theses and dissertations. In *2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–5. IEEE.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv: 1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv: 1606.05250*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv: 2311.12022*.

Virgile Rennard, Guokan Shang, Damien Grari, Julie Hunter, and Michalis Vazirgiannis. 2023. Fredsum: A dialogue summarization corpus for french political debates. *arXiv preprint arXiv: 2312.04843*.

Syed Toufeeq Ahmad Rizvi. 2020. *A hybrid approach for information extraction from biomedical text*. Ph.D. thesis, University of Bedfordshire.

Julia Rodina and Andrey Kutuzov. 2020. Rusemshift: a dataset of historical lexical semantic change in russian. *arXiv preprint arXiv: 2010.06436*.

Sergio J Rodríguez Méndez, Pouya G Omran, Armin Haller, and Kerry Taylor. 2021. MEL: Metadata extractor & loader. In *European Semantic Web Conference*, pages 159–163. Springer.

Mihaela Rosca and Thomas Breuel. 2016. Sequence-to-sequence neural network models for transliteration. *arXiv preprint arXiv: 1610.09565*.

Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. 2022. Healthsheet: development of a transparency artifact for health datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1943–1961.

Tarek Saier and Michael Färber. 2020. unarxive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics*, 125(3):3085–3108.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv: 1907.10641*.

Yuya Sakaizawa and Mamoru Komachi. 2017. Construction of a japanese word similarity dataset. *arXiv preprint arXiv: 1703.05916*.

Björn Schembera. 2021. Like a rainbow in the dark: metadata annotation for hpc applications in the age of dark data. *The Journal of Supercomputing*, 77(8):8946–8966.

Marcel Schilling-Wilhelmi, Melissa Ríos-García, Sameena Shabih, Marcos V. Gil, Santiago Miret, Christoph T. Koch, José A. Márquez, and Kevin Maik Jablonka. 2024. From text to insight: large language models for materials science data extraction. *arXiv preprint arXiv:2407.16867*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. Mlsum: The multilingual summarization corpus. *arXiv preprint arXiv: 2004.14900*.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. *arXiv preprint arXiv: 2210.01613*.

ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. 2022. Jaquad: Japanese question answering dataset for machine reading comprehension. *arXiv preprint arXiv: 2202.01764*.

Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2023. Jcola: Japanese corpus of linguistic acceptability. *arXiv preprint arXiv: 2309.12676*.

Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2017. Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. *arXiv preprint arXiv: 1711.00354*.

Ramya Srinivasan, Emily Denton, Jordan Famularo, Negar Rostamzadeh, Fernando Diaz, and Beth Coleman. 2021. Artsheets for art datasets. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *arXiv preprint arXiv: 1906.00591*.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O. Arik, Danqi Chen, and Tao Yu. 2024. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv: 2407.12883*.

Tomoki Sugimoto, Yasumasa Onoe, and Hitomi Yanaka. 2023. Jamp: Controlled japanese temporal inference dataset for evaluating generalization capacity of language models. *arXiv preprint arXiv: 2306.10727*.

Daisuke Suzuki, Yujin Takahashi, Ikumi Yamashita, Taichi Aida, Tosho Hirasawa, Michitaka Nakatsuji, Masato Mita, and Mamoru Komachi. 2022. Construction of a quality estimation dataset for automatic evaluation of japanese grammatical error correction. *arXiv preprint arXiv: 2201.08038*.

Shinnosuke Takamichi, Ludwig Kürzinger, Takaaki Saeki, Sayaka Shiota, and Shinji Watanabe. 2021. Jtubespeech: corpus of japanese speech collected from youtube for speech recognition and speaker verification. *arXiv preprint arXiv: 2112.09323*.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. Rublimp: Russian benchmark of linguistic minimal pairs. *arXiv preprint arXiv: 2406.19232*.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on performance of large language models. *arXiv preprint arXiv:2408.02442*.

Kota Tanabe, Masahiro Suzuki, Hiroki Sakaji, and Itsuki Noda. 2024. Jafin: Japanese financial instruction dataset. *arXiv preprint arXiv: 2404.09260*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yingtao Tian, Chikahiko Suzuki, Tarin Clanuwat, Mikel Bober-Irizar, Alex Lamb, and Asanobu Kitamoto. 2020. Kaokore: A pre-modern japanese art facial expression dataset. *arXiv preprint arXiv: 2002.08595*.

Dominika Tkaczyk, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Łukasz Bolikowski. 2015. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(4):317–335.

Alena Tsanda and Elena Bruches. 2024. Russian-language multimodal dataset for automatic summarization of scientific papers. *arXiv preprint arXiv: 2405.07886*.

Nobuhiro Ueda, Hideko Habe, Yoko Matsui, Akishige Yuguchi, Seiya Kawano, Yasutomo Kawanishi, Sadao Kurohashi, and Koichiro Yoshino. 2024. J-cre3: A japanese conversation dataset for real-world reference resolution. *arXiv preprint arXiv: 2403.19259*.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. *arXiv preprint arXiv: 2003.08385*.

Viacheslav Vasilev, Julia Agafonova, Nikolai Gerasimenko, Alexander Kapitanov, Polina Mikhailova, Evelina Mironova, and Denis Dimitrov. 2025a. Ruscode: Russian cultural code benchmark for text-to-image generation. *arXiv preprint arXiv: 2502.07455*.

Viacheslav Vasilev, Vladimir Arkhipkin, Julia Agafonova, Tatiana Nikulina, Evelina Mironova, Alisa Shichanina, Nikolai Gerasimenko, Mikhail Shoytov, and Denis Dimitrov. 2025b. Craft: Cultural russian-oriented dataset adaptation for focused text-to-image generation. *arXiv preprint arXiv: 2505.04851*.

Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Huu Nguyen. 2022a. Mee: A novel multilingual event extraction dataset. *arXiv preprint arXiv: 2211.05955*.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2022b. Minion: a large-scale and diverse dataset for multilingual event detection. *arXiv preprint arXiv: 2211.05958*.

Josiah Wang, Pranava Madhyastha, Josiel Figueiredo, Chiraag Lala, and Lucia Specia. 2021. Multisubs: A large-scale multimodal and multilingual dataset. *arXiv preprint arXiv: 2103.01910*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv: 2406.01574*.

Michelle Wastl, Jannis Vamvas, Selena Calleri, and Rico Sennrich. 2025. 20min-xd: A comparable corpus of swiss news articles. *arXiv preprint arXiv: 2504.21677*.

Yu Watanabe, Koichiro Ito, and Shigeki Matsubara. 2024. Capabilities and challenges of llms in metadata extraction from scholarly papers. In *International Conference on Asian Digital Libraries*, pages 280–287. Springer.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv: 1707.06209*.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, and 1 others. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Ankit Yadav, Shubham Chandel, Sushant Chatufale, and Anil Bandhakavi. 2023. Lahm : Large annotated dataset for multi-domain and multilingual hate speech identification. *arXiv preprint arXiv: 2304.00913*.

Wenli Yang, Rui Fu, Muhammad Bilal Amin, and Byeong Kang. 2025. Impact and influence of modern ai in metadata management. *arXiv preprint arXiv:2501.16605*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv: 1809.09600*.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv: 1705.00823*.

Jamil Zaghir, Mina Bjelogrlic, Jean-Philippe Goldman, Soukaïna Aananou, Christophe Gaudet-Blavignac, and Christian Lovis. 2023. Frasimed: a clinical french annotated resource produced through crosslingual bert-based annotation projection. *arXiv preprint arXiv: 2309.10770*.

Wajdi Zaghouani and Anis Charfi. 2018. Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. *arXiv preprint arXiv: 1808.07674*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv: 1905.07830*.

Shouyuan Zhang, Xiaohan Ding, and Alexander M Rush. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Miao Zhu and Jacqueline M Cole. 2022. Pdfdataextractor: A tool for reading scientific text and interpreting metadata from the typeset literature in the portable document format. *Journal of Chemical Information and Modeling*, 62(7):1633–1643.

A Datasets

In Total, we annotated 132 papers including validation and testing. Here is a list of the paper datasets: RuBQ (Korablinov and Braslavski, 2020), RuCoLA (Mikhailov et al., 2022), Slovo (Kapitanov et al., 2023), WikiOmnia (Pisarevskaya and Shavrina, 2022), DaNetQA (Glushkova et al., 2020), RuSemShift (Rodina and Kutuzov, 2020), REPA (Pugachev et al., 2025), SberQuAD (Efimov et al., 2019), RuBLiMP (Taktasheva et al., 2024), Golos (Karpov et al., 2021), RFSD (Bondarkov et al., 2025), RuDSI (Aksenova et al., 2022), HeadlineCause (Gusev and Tikhonov, 2021), RusTitW (Markov et al., 2023), CRAFT (Vasilev et al., 2025b), Gazeta (Gusev, 2020), RuBia (Grigoreva et al., 2024), RusCode (Vasilev et al., 2025a), Russian Jeopardy (Mikhalkova, 2021), Russian Multimodal Summarization (Tsanda and Bruches, 2024), NEREL (Loukachevitch et al., 2021), HISTOIRES-MORALES (Leteno et al., 2025), PxSLU (Kocabiyikoglu et al., 2022), French COVID19 Lockdown Twitter Dataset (Balech et al., 2020), MEDIA with Intents (Alavoine et al., 2024), FrenchMedMCQA (Labrak et al., 2023), FRASIMED (Zaghir et al., 2023), FQuAD1.1 (d'Hoffschmidt et al., 2020), FIJO (Beauchemin et al., 2022), BSARD (Louis and Spanakis, 2021), FairTranslate (Jourdan et al., 2025), 20min-XD (Wastl et al., 2025), Alloprof (Lefebvre-Brossard et al., 2023), FRED-Sum (Rennard et al., 2023), Vibravox (Hauret et al., 2024), MTNT (Michel and Neubig, 2018), PIAF (Keraron et al., 2020), FrenchToxicityPrompts (Brun and Nikoulina, 2024), OBSINFOX (Icard et al., 2024), CFDD (Hunter et al., 2023), FREEM-max (Gabay et al., 2022), FQuAD2.0 (Heinrich et al., 2021), HellaSwag (Zellers et al., 2019), GPQA (Rein et al., 2023), GoEmotions (Demszky et al., 2020), SQuAD 2.0 (Rajpurkar et al., 2018), LAMBADA (Paperno et al., 2016), ClimbMix (Diao et al., 2025), RACE (Lai et al., 2017), MMLU-Pro (Wang et al., 2024), BoolQ (Clark et al., 2019), GSM8K (Cobbe et al., 2021), HotpotQA (Yang et al., 2018), SQuAD (Rajpurkar et al., 2016), RefinedWeb (Penedo et al., 2023), MMLU (Hendrycks et al., 2020), PIQA (Bisk et al., 2019), BRIGHT (Su et al., 2024), HLE (Phan et al., 2025), TinyStories (Eldan and Li, 2023), WinoGrande (Sakaguchi et al., 2019), SciQ (Welbl et al., 2017), TriviaQA (Joshi et al., 2017), JEC (Suzuki et al., 2022), JParaCrawl (Morishita et al., 2019), KaoKore (Tian et al., 2020), llm-japanese-dataset v0 (Hirano et al., 2023), JaLeCoN (Ide et al., 2023), JaQuAD (So et al., 2022), JAFFE (Lyons, 2021), JaFIn (Tanabe et al., 2024), Japanese Fake News Dataset (Murayama et al., 2022), JMultiWOZ (Ohashi et al., 2024), Japanese Web Corpus (Okazaki et al., 2024), J-CRe3 (Ueda et al., 2024), JSUT (Sonobe et al., 2017), Japanese Word Similarity Dataset (Sakaizawa and Komachi, 2017), STAIR Captions (Yoshikawa et al., 2017), Arukikata Travelogue (Ouchi et al., 2023), JTube-Speech (Takamichi et al., 2021), JCoLA (Someya et al., 2023), JESC (Pryzant et al., 2017), JDocQA (Onami et al., 2024), Jamp (Sugimoto et al., 2023), 101 Billion Arabic Words Dataset (Aloui et al., 2024), WinoMT (Stanovsky et al., 2019), ArabicMMLU (Koto et al., 2024), CIDAR (Alyafeai et al., 2024), Belebele (Bandarkar et al., 2023), MGB-2 (Ali et al., 2016), ANETAC (Ameur et al., 2019), TUNIZI (Fourati et al., 2020), Shamela (Belinkov et al., 2016), POLYGLOT-NER (Al-Rfou et al., 2014), DODa (Outchakoucht and Es-Samaali, 2021), LASER (Artetxe and Schwenk, 2018), MGB-3 (Ali et al., 2017), Arap-Tweet (Zaghouani and Charfi, 2018), FLORES-101 (Goyal et al., 2021), Transliteration (Rosca and Breuel, 2016), ADI-5 (Ali et al., 2015), Maknuune (Dibas et al., 2022), EmojisAnchors (Mubarak et al., 2022), Calliar (Alyafeai et al., 2021a), LABR (Nabil et al., 2014), ACVA (Huang et al., 2023), ATHAR (Khalil and Sabry, 2024), OpenITI-proc (Belinkov et al., 2018), AraDangspeech (Alshehri et al., 2020), Arabic-Hebrew TED Talks Parallel Corpus (Cettolo, 2016), ARASPIDER (Heakl et al., 2024), XNLI (Conneau et al., 2018), X-stance (Vamvas and Sennrich, 2020), DiS-ReX (Bhartiya et al., 2021), RELX (Köksal and Özgür, 2020), MultiSubs (Wang et al., 2021), MEE (Veyseh et al., 2022a), XCOPA (Ponti et al., 2020), MLQA (Lewis et al., 2019), M2DS (Hewapathirana et al., 2024), XOR-TyDi (Asai et al., 2020), Multilingual Hate Speech Detection Dataset (Aluru et al., 2020), MINION (Veyseh et al., 2022b), SEA-HORSE (Clark et al., 2023), Mintaka (Sen et al., 2022), Multi2WOZ (Hung et al., 2022), MTOP (Li et al., 2020a), X-RiSAWOZ (Moradshahi et al., 2023), PRESTO (Goel et al., 2023), LAHM (Yadav et al., 2023), MARC (Keung et al., 2020), and

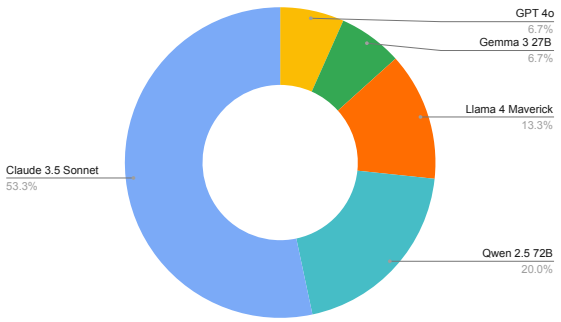

Figure 8: Distribution of the number of errors per model for one experiment. This graph only shows the models with at least one error.

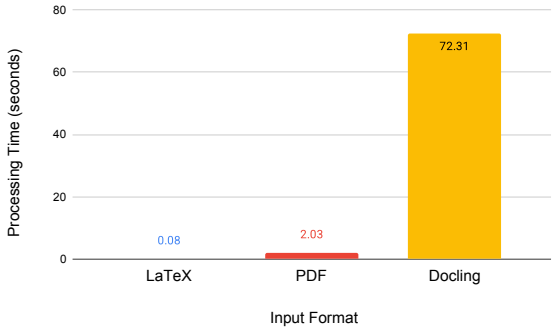

Figure 9: Average processing time comparison between different input formats. LaTeX source processing (0.08s) is most efficient, followed by pdfplumber-based PDF extraction (2.03s), while Docling structured parsing (72.31s) requires substantially more computation time.

MLSUM (Scialom et al., 2020).

B Costs

In Table 9, we show the costs for each model in terms of input tokens, output tokens, and cost in one single experiment. In general, we observe that Claude is more costly while processing the least number of tokens. Gemini 2.5 Pro achieves the best cost vs performance results as it is cheaper than other models but achieves better results across different experiments. In total, it costs around 20 USD to run the evaluation and extract the metadata for all the papers.

C Errors

In Figure 8, we show the number of errors for each model in a single experiment. The major proportion of errors comes from Claude Sonnet 3.5. In general, we run more than 1K inference requests to the OpenRouter API, and around 16 errors happened. Most of the errors occur due to failing to read the generated JSON file.

D Input Format Processing Time

We analyze the processing time of each input format. Figure 9 shows the average of processing time where we applied each of these methods on the subset of the testset. The figure shows the varying differences in preprocessing time across these methods. LaTeX source processing is remarkably efficient (0.08s), while PDF extraction via pdfplumber is reasonably fast (2.03s). In contrast, Docling processing requires significantly more compute time (72.31s), highlighting important efficiency trade-offs when scaling metadata extraction to large document collections.

```
{name}: A {task} dataset for {schema}
{authors}
{affs}
{name}, is a {schema} {task} dataset, that contains
{volume} {unit}. {language_table} {provider_stmt}
The dataset was collected from {collection_style} of
{domain} in {year}. The dataset is publicly available
through this link {link}. {license_stmt}. {hf_stmt}.
```

Figure 10: Template-based few-shot example creation.

E Model's Access

In Table 8, we highlight the models used for evaluation. We use a collection of closed and open models. We use the OpenRouter API to perform inference on all models.

F Synthetic Template Generation

For few-shot experiments, we create synthetic examples, which are short paper templates that can be filled using the different attributes in each metadata schema. In Figure 10, we show an example of a template. The language table is used for the multilingual schema, where we create a table for the language subsets of the datasets. To add variance in the results, we sample different options for attributes like unit, task, collection style, and domain. We also sample a different number each time for the volume. For each schema category in our dataset we generate five examples.

G Validation Groups

We have four groups that gather similar metadata attributes. Figure 11 shows the different validation

Table 8: Models used for evaluation and their respective links in OpenRouter.

Model	Link
GPT 4o	https://openrouter.ai/openai/gpt-4o
Claude 3.5 Sonnet	https://openrouter.ai/anthropic/claude-3.5-sonnet
Gemini 2.5 Pro	https://openrouter.ai/google/gemini-2.5-pro-preview-03-25
DeepSeek V3	https://openrouter.ai/deepseek/deepseek-chat-v3-0324
Llama 4 Maverick	https://openrouter.ai/meta-llama/llama-4-maverick
Gemma 3 27B	https://openrouter.ai/google/gemma-3-27b-it
Qwen2.5 72B	https://openrouter.ai/qwen/qwen-2.5-72b-instruct

Table 9: Input, output, and total tokens and cost in USD for running one experiment on all test sets using LaTeX as input. We use OpenRouter to estimate the number of input and output tokens.

Model	Input Tokens	Output Tokens	Total Tokens	Cost (USD)
Gemma 3 27B	2162756	85738	2248494	0.23
Qwen 2.5 72B	2163466	77643	2241109	0.29
Llama 4 Maverick	2159318	71856	2231174	0.41
DeepSeek V3	2163754	77280	2241034	0.72
Gemini 2.5 Pro	2163823	160811	2324634	4.31
GPT 4o	2163823	80514	2244337	6.21
Claude 3.5 Sonnet	2163934	74944	2238878	**7.62**

groups in the Arabic subset of the dataset. The general group is not used for validation, as it shows easily extractable attributes. Each group covers attributes that are similar in terms of the grouping function.

H Full metrics

As we explained in our data annotation procedure, we also add a binary value to indicate whether the attribute is actually extracted from the paper or elsewhere. To test the model's ability in only attributes that exist in papers, we show the results in Table 12. As expected, considering only the attributes in the paper, we observe a significantly higher recall compared to precision.

I System Prompt

In Figure 12, we show the system prompt to generate the metadata given the paper and the input schema.

J Individual Attribute Evaluation

In Table 10, we highlight the results on individual attributes of metadata. We observe that most models struggle with Links and License in general. This could be attributed to missing such attributes in most papers. To observe the effect of browsing, we plot the difference between browsing and non-browsing models in Table 11. Generally, we see the most gain in extracting the license attribute, which is mostly reported in repositories. However, we see a decrease in some other attributes. For example, the Volume is decreased because some papers report different numbers than what is actually released in the data. Additionally, the Link attribute is worse because it seems browsing forces the model to change the link based on what is in the repository. Interestingly, Gemma's results increase across multiple attributes without any decrease.

K Contamination

To consider contamination, we also evaluate the LLMs using papers after the model release. The results show no significant difference that might indicate the presence of some contamination.

L Schema

In Code 2, we show the full schema used for the Arabic datasets. Note in some experiments, we change the answer_max to reflect to different constraints.

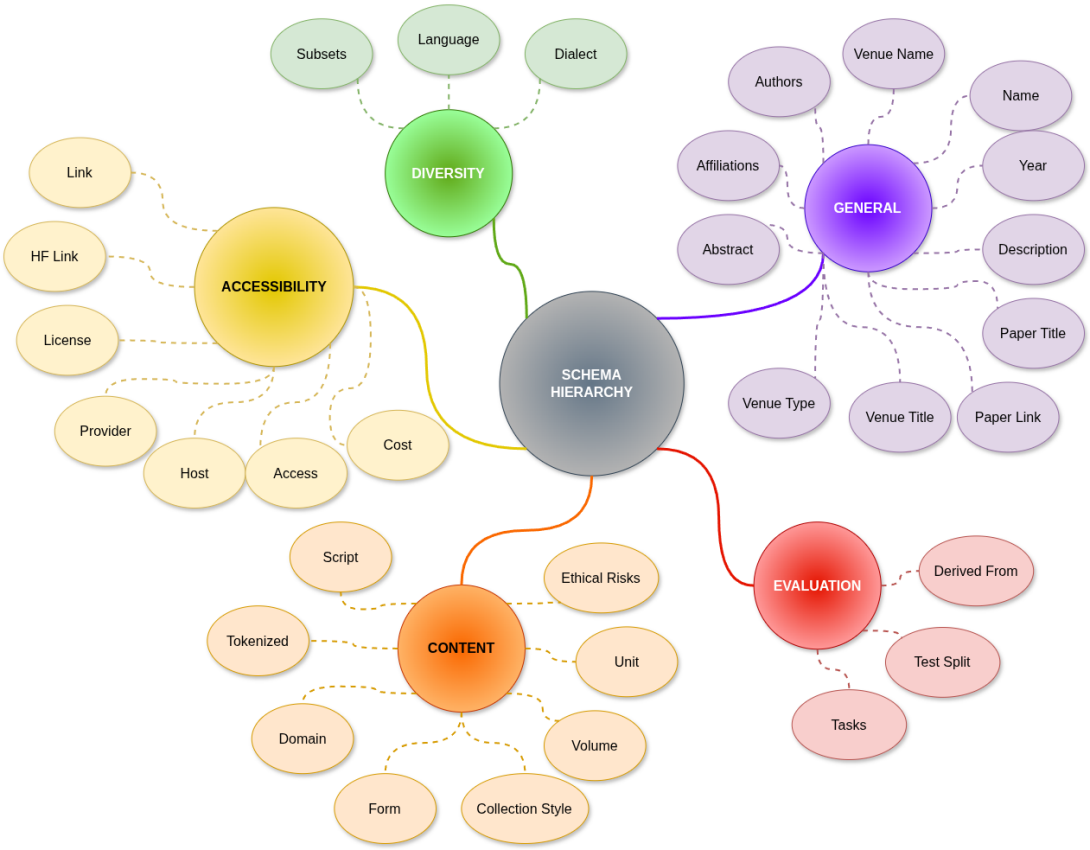

Figure 11: Schema validation groups and their associated attributes for the Arabic metadata.

Table 10: Results across different attributes that exist in all categories. For better visualization, we remove the versions and naming conventions of each model.

Model	Llama	Gemini	Qwen	DeepSeek	GPT	Gemma	Claude
Link	58.73	56.35	55.56	55.56	56.35	56.35	57.94
HF Link	47.62	47.62	44.44	47.62	48.41	45.24	48.41
License	25.40	34.92	28.57	28.57	27.78	22.22	30.95
Language	89.68	93.65	88.10	90.48	92.86	88.89	84.13
Domain	46.03	69.84	38.89	52.38	61.90	35.71	50.79
Form	94.44	97.62	95.24	96.03	96.03	89.68	88.89
Collection Style	58.73	75.40	57.14	63.49	59.52	51.59	65.08
Volume	44.44	66.67	56.35	46.03	44.44	61.11	61.90
Unit	69.05	70.63	53.17	59.52	54.76	58.73	59.52
Ethical Risks	83.33	80.16	82.54	81.75	80.16	34.13	84.13
Provider	50.00	53.17	50.00	52.38	50.00	48.41	47.62
Derived From	53.17	61.90	53.97	66.67	66.67	73.81	65.08
Tokenized	92.86	85.71	91.27	92.86	91.27	95.24	90.48
Host	68.25	69.84	68.25	69.05	67.46	65.87	69.84
Access	97.62	92.06	97.62	96.03	96.83	94.44	89.68
Cost	96.83	100.00	100.00	96.83	100.00	99.21	100.00
Test Split	71.43	86.51	74.60	76.98	85.71	79.37	84.13
Tasks	61.90	62.70	47.62	46.83	62.70	49.21	61.90

Table 11: Difference between Browsing and non-browsing models in all attributes. Model names are shortened for better visualization.

Model	Llama	Gemini	Qwen	DeepSeek	GPT	Gemma	Claude
Link	-0.79	-4.76	-5.56	0.79	-6.35	-1.59	-2.38
HF Link	0.79	0.79	1.59	0.00	-0.79	-0.79	0.00
License	15.08	15.08	19.05	7.14	15.08	13.49	21.43
Language	-0.79	0.79	0.00	0.00	0.00	-0.79	0.79
Domain	0.79	0.79	0.00	0.00	0.00	1.59	1.59
Form	0.00	0.00	0.79	0.00	0.00	0.00	0.79
Collection Style	0.00	0.00	0.00	0.00	0.00	1.59	0.00
Volume	-1.59	-4.76	-2.38	-0.79	-4.76	-1.59	-3.17
Unit	0.00	0.79	0.79	0.00	0.00	-0.79	-0.79
Ethical Risks	-0.79	0.79	0.00	0.00	0.00	0.00	1.59
Provider	-1.59	0.00	0.00	0.00	0.00	-0.79	-0.79
Derived From	-2.38	-0.79	0.79	-0.79	0.00	-1.59	-2.38
Tokenized	1.59	0.79	0.00	0.79	0.00	0.00	-1.59
Host	0.00	-0.79	0.00	0.00	0.00	-1.59	1.59
Access	-0.79	-0.79	-0.79	-0.79	0.00	-1.59	-0.79
Cost	-0.79	-0.79	0.00	0.00	0.00	0.00	0.00
Test Split	-0.79	0.79	0.00	0.00	0.79	0.79	-0.79
Tasks	0.00	2.38	0.79	0.00	0.00	2.38	1.59

Figure 12: System prompt for generating the metadata.

You are a professional research paper reader. You will be provided 'Input schema' and 'Paper Text' and you must respond with an 'Output JSON'. The 'Output Schema' is a JSON with the following format key:answer where the answer represents an answer to the question. The 'Input Schema' has the following main fields for each key: 'question': A question that needs to be answered. 'options' : If the 'question' has 'options' then the question can be answered by choosing one or more options depending on 'answer_min' and 'answer_max' 'options_description': A description of the 'options' that might be unclear. Use the descriptions to understand the options. 'answer_type': the type of the answer to the 'question'. The answer must follow the type of the answer. 'answer_min' : If the 'answer_type' is List[str], then it defines the minimum number of list items in the answer. Otherwise it defines the minimum number of words in the answer. 'answer_max' : If the 'answer_type' is List[str], then it defines the maximum number of list items in the answer. Otherwise it defines the maximum number of words in the answer. The answer must be the same type as 'answer_type' and its length must be in the range ['answer_min', 'answer_max']. If answer_min = answer_max then the length of answer MUST be answer_min. The 'Output JSON' is a JSON that can be parsed using Python 'json.load()'. USE double quotes "" not single quotes '' for the keys and values. The 'Output JSON' has ONLY the keys: 'columns'. The value for each key is the answer to the 'question' that represents the same key in the 'Input Schema'.

Table 12: Comparison between models in terms of precision, recall, and f1 scores.

Model	Precision	Recall	F1
Random	29.46	30.27	29.74
Keyword	42.39	45.65	43.89
Gemma 3 27B	64.08	68.72	66.23
Qwen 2.5 72B	65.78	70.28	67.88
Llama 4 Maverick	67.04	71.90	69.30
DeepSeek V3	67.72	72.70	70.03
Claude 3.5 Sonnet	69.08	73.94	71.34
GPT 4o	69.01	74.09	71.38
Gemini 2.5 Pro	**72.67**	**78.17**	**75.23**

Table 13: Model performance by evaluating on papers published after the model release.

Model	Average
Llama 4 Maveric	54.17
DeepSeek V3	60.00
Qwen 2.5 72B	61.11
Claude 3.5 Sonnet	61.73
Gemma 3 27B	61.90
GPT 4o	66.05
Gemini 2.5 Pro	**73.33**

```
{
    "Name": {
        "question": "What is the name of the dataset?",
        "answer_type": "str",
        "answer_min": 1,
        "answer_max": 5
    },
    "Subsets": {
        "question": "What are the dialect subsets of this dataset? The keys are '
    Name', 'Volume', 'Unit', 'Dialect'. 'Dialect' must be one of the country name
    from the dialect options",
        "answer_type": "List[Dict[Name, Volume, Unit, Dialect]]",
        "validation_group":"DIVERSITY",
        "answer_min": 0,
        "answer_max": 29
    },
    "Link": {
        "question": "What is the link to access the dataset? The link must contain
    the dataset. If the dataset is hosted on HuggingFace, use the HF Link.",
        "answer_type": "url",
        "validation_group":"ACCESSIBILITY",
        "answer_min": 1,
        "answer_max": 1
    },
    "HF Link": {
        "question": "What is the Huggingface link of the dataset?",
        "answer_type": "url",
        "validation_group":"ACCESSIBILITY",
        "answer_min": 0,
        "answer_max": 1
    },
    "License": {
        "question": "What is the license of the dataset?",
        "options": [
            "Apache-1.0",
            "Apache-2.0",
            "Non Commercial Use - ELRA END USER",
            "BSD",
            "CC BY 1.0",
            "CC BY 2.0",
            "CC BY 3.0",
            "CC BY 4.0",
            "CC BY-NC 1.0",
            "CC BY-NC 2.0",
            "CC BY-NC 3.0",
            "CC BY-NC 4.0",
            "CC BY-NC-ND 1.0",
            "CC BY-NC-ND 2.0",
            "CC BY-NC-ND 3.0",
            "CC BY-NC-ND 4.0",
            "CC BY-SA 1.0",
            "CC BY-SA 2.0",
            "CC BY-SA 3.0",
            "CC BY-SA 4.0",
            "CC BY-NC 1.0",
            "CC BY-NC 2.0",
            "CC BY-NC 3.0",
            "CC BY-NC 4.0",
            "CC BY-NC-SA 1.0",
            "CC BY-NC-SA 2.0",
            "CC BY-NC-SA 3.0",
            "CC BY-NC-SA 4.0",
            "CC0",
            "CDLA-Permissive-1.0",
            "CDLA-Permissive-2.0",
            "GPL-1.0",
            "GPL-2.0",
            "GPL-3.0",
            "LDC User Agreement",
            "LGPL-2.0",
```

```
            "LGPL-3.0",
            "MIT License",
            "ODbl-1.0",
            "MPL-1.0",
            "MPL-2.0",
            "ODC-By",
            "unknown",
            "custom"
        ],
        "answer_type": "str",
        "validation_group":"ACCESSIBILITY",
        "answer_min": 1,
        "answer_max": 1
    },
    "Year": {
        "question": "What year was the dataset published?",
        "answer_type": "date[year]",
        "answer_min": 1,
        "answer_max": 1
    },
    "Language": {
        "question": "What languages are in the dataset?",
        "options": ["ar", "multilingual"],
        "option_description": {
            "ar": "the dataset is purely in Arabic, there are no other languages
    involved",
            "multilingual": "the dataset contains samples in other languages"
        },
        "answer_type": "str",
        "validation_group":"DIVERSITY",
        "answer_min": 1,
        "answer_max": 1
    },
    "Dialect": {
        "question": "What is the dialect of the dataset?",
        "options": [
            "Classical Arabic",
            "Modern Standard Arabic",
            "United Arab Emirates",
            "Bahrain",
            "Djibouti",
            "Algeria",
            "Egypt",
            "Iraq",
            "Jordan",
            "Comoros",
            "Kuwait",
            "Lebanon",
            "Libya",
            "Morocco",
            "Mauritania",
            "Oman",
            "Palestine",
            "Qatar",
            "Saudi Arabia",
            "Sudan",
            "Somalia",
            "South Sudan",
            "Syria",
            "Tunisia",
            "Yemen",
            "Levant",
            "North Africa",
            "Gulf",
            "mixed"
        ],
        "option_description": {
            "mixed": "the dataset contains samples in multiple dialects i.e. social
    media. Assume Modern Standard Arabic if not specified."
        },
```

```
        "answer_type": "str",
        "validation_group":"DIVERSITY",
        "answer_min": 1,
        "answer_max": 1
    },
    "Domain": {
        "question": "What is the source of the dataset?",
        "options": [
            "social media",
            "news articles",
            "reviews",
            "commentary",
            "books",
            "wikipedia",
            "web pages",
            "public datasets",
            "TV Channels",
            "captions",
            "LLM",
            "other"
        ],
        "answer_type": "List[str]",
        "validation_group":"CONTENT",
        "answer_min": 1,
        "answer_max": 11
    },
    "Form": {
        "question":"What is the form of the data?",
        "options": ["text", "spoken", "images"],
        "answer_type": "str",
        "validation_group":"CONTENT",
        "answer_min": 1,
        "answer_max": 1
    },
    "Collection Style": {
        "question": "How was this dataset collected?",
        "options": [
            "crawling",
            "human annotation",
            "machine annotation",
            "manual curation",
            "LLM generated",
            "other"
        ],
        "option_description": {
            "crawling": "the dataset was collected by crawling the web",
            "human annotation": "the dataset was labelled by human annotators",
            "machine annotation": "the dataset was collected/labelled by machine
    programs",
            "manual curation": "the dataset was collected manually by human curators
    ",
            "LLM generated": "the dataset was generated by an LLM",
            "other": "the dataset was collected in a different way"
        },
        "answer_type": "List[str]",
        "validation_group":"CONTENT",
        "answer_min": 1,
        "answer_max": 5
    },
    "Description": {
        "question": "Write a brief description about the dataset",
        "answer_type": "str",
        "answer_min": 0,
        "answer_max": 50
    },
    "Volume": {
        "question": "What is the size of the dataset?. If the dataset is
    multilingual only use the size of the Arabic dataset",
        "answer_type": "float",
        "validation_group":"CONTENT",
```

```
        "answer_min": 1
    },
    "Unit": {
        "question": "What kind of examples does the dataset include?",
        "options": ["tokens", "sentences", "documents", "hours", "images"],
        "option_description": {
            "tokens": "the dataset contains individual tokens/words",
            "sentences": "the samples are sentences or short paragraphs",
            "documents": "the samples are long documents i.e. web pages or books",
            "hours": "the samples are audio files",
            "images": "the samples are images"
        },
        "answer_type": "str",
        "validation_group":"CONTENT",
        "answer_min": 1,
        "answer_max": 1
    },
    "Ethical Risks": {
        "question": "What is the level of the ethical risks of the dataset?",
        "options": ["Low", "Medium", "High"],
        "option_description": {
            "Low": "most likely no ethical risks associated with this dataset",
            "Medium": "social media datasets",
            "High": "hate/offensive datasets from social media, or web pages"
        },
        "answer_type": "str",
        "validation_group":"CONTENT",
        "answer_min": 1,
        "answer_max": 1
    },
    "Provider": {
        "question": "What entity is the provider of the dataset? Don't use Team.",
        "answer_type": "List[str]",
        "validation_group":"ACCESSIBILITY",
        "answer_min": 0,
        "answer_max": 10
    },
    "Derived From": {
        "question": "What datasets were used to create the dataset?",
        "answer_type": "List[str]",
        "validation_group":"EVALUATION",
        "answer_min": 0
    },
    "Paper Title": {
        "question": "What is the title of the paper?",
        "answer_type": "str",
        "answer_min" : 3
    },
    "Paper Link": {
        "question": "What is the link to the paper?",
        "answer_type": "str",
        "answer_min": 1
    },
    "Script": {
        "question": "What is the script of this dataset?",
        "options": ["Arab", "Latin", "Arab-Latin"],
        "option_description": {
            "Arab": "The script used is only in Arabic",
            "Latin": "The script used is only in Latin i.e. it has samples written
in Latin like Arabizi or transliteration",
            "Arab-Latin": "The script used is a mix of Arabic and Latin"
        },
        "answer_type": "str",
        "validation_group":"CONTENT",
        "answer_min": 1,
        "answer_max": 1
    },
    "Tokenized": {
        "question": "Is the dataset tokenized?",
        "options": [true, false],
```

```
        "option_description": {
            "true": "The dataset is tokenized. Tokenized means the words are split
    using a morphological analyzer",
            "false": "The dataset is not tokenized"
        },
        "answer_type": "bool",
        "validation_group":"CONTENT",
        "answer_min": 1,
        "answer_max": 1
    },
    "Host": {
        "question": "What is name of the repository that hosts the dataset?",
        "options": [
            "CAMeL Resources",
            "CodaLab",
            "data.world",
            "Dropbox",
            "Gdrive",
            "GitHub",
            "GitLab",
            "kaggle",
            "LDC",
            "MPDI",
            "Mendeley Data",
            "Mozilla",
            "OneDrive",
            "QCRI Resources",
            "ResearchGate",
            "sourceforge",
            "zenodo",
            "HuggingFace",
            "ELRA",
            "other"
        ],
        "answer_type": "str",
        "validation_group":"ACCESSIBILITY",
        "answer_min": 1,
        "answer_max": 1
    },
    "Access": {
        "question": "What is the accessibility of the dataset?",
        "options": ["Free", "Upon-Request", "With-Fee"],
        "option_description": {
            "Free": "the dataset is public and free to access",
            "Upon-Request": "the dataset is free to access but requires a submitting
     a request or filling out a form",
            "With-Fee": "the dataset is not free to access"
        },
        "answer_type": "str",
        "validation_group":"ACCESSIBILITY",
        "answer_min": 1,
        "answer_max": 1
    },
    "Cost": {
        "question": "If the dataset is not free, what is the cost?",
        "answer_type": "str",
        "validation_group":"ACCESSIBILITY",
        "answer_min": 0
    },
    "Test Split": {
        "question": "Does the dataset contain a train/valid and test split?",
        "options": [true, false],
        "option_description": {
            "true": "The dataset contains a train/valid and test split",
            "false": "The dataset does not contain a train/valid or test split"
        },
        "answer_type": "bool",
        "validation_group":"EVALUATION",
        "answer_min": 1,
        "answer_max": 1
```

```
    },
    "Tasks": {
        "question": "What NLP tasks is this dataset intended for?",
        "options": [
            "machine translation",
            "speech recognition",
            "sentiment analysis",
            "language modeling",
            "topic classification",
            "dialect identification",
            "text generation",
            "cross-lingual information retrieval",
            "named entity recognition",
            "question answering",
            "multiple choice question answering",
            "information retrieval",
            "part of speech tagging",
            "language identification",
            "summarization",
            "speaker identification",
            "transliteration",
            "morphological analysis",
            "offensive language detection",
            "review classification",
            "gender identification",
            "fake news detection",
            "dependency parsing",
            "irony detection",
            "meter classification",
            "natural language inference",
            "instruction tuning",
            "Linguistic acceptability",
            "other"
        ],
        "answer_type": "List[str]",
        "validation_group":"EVALUATION",
        "answer_min": 1,
        "answer_max": 5

    },
    "Venue Title": {
        "question": "What is the venue title of the published paper?",
        "answer_type": "str",
        "answer_min": 1
    },
    "Venue Type": {
        "question": "What is the venue type?",
        "options": ["conference", "workshop", "journal", "preprint"],
        "answer_type": "str",
        "answer_min": 1,
        "answer_max": 1
    },
    "Venue Name": {
        "question": "What is the full name of the venue that published the paper?",
        "answer_type": "str",
        "answer_min" : 0
    },
    "Authors": {
        "question": "Who are the authors of the paper?",
        "answer_type": "List[str]",
        "answer_min": 1,
        "answer_max": 20
    },
    "Affiliations": {
        "question": "What are the affiliations of the authors?",
        "answer_type": "List[str]",
        "answer_min": 0,
        "answer_max": 20
    },
    "Abstract": {
```

```
        "question": "What is the abstract of the paper? replace any double quotes in
    the abstract by single quotes '",
        "answer_type": "str",
        "answer_min": 5
    }
}
```

Code 2: Example Schema for Arabic datasets' metadata