RAGing Against the Literature: LLM-Powered Dataset Mention Extraction

Priyangshu Datta
Indian Institute of Science Education and Research, Kolkata, India
priyangshu.datta@outlook.com

Suchana Datta
University College Dublin, Ireland
suchana.datta@ucd.ie

Dwaipayan Roy
Indian Institute of Science Education and Research, Kolkata, India
dwaipayan.roy@iiserkol.ac.in

Abstract

Dataset Mention Extraction (DME) is a critical task in the field of scientific information extraction, aiming to identify references to datasets within research papers. In this paper, we explore two advanced methods for DME from research papers, utilizing the capabilities of Large Language Models (LLMs). The first method employs a language model with a prompt-based framework to extract dataset names from text chunks, utilizing patterns of dataset mentions as guidance. The second method integrates the Retrieval-Augmented Generation (RAG) framework, which enhances dataset extraction through a combination of keyword-based filtering, semantic retrieval, and iterative refinement. We observe that both of the proposed methods achieve more than a 25% improvement in recall compared to the baselines. Further, the RAG-based model achieves an extensive 26% improvement over the baselines. We also propose `exData`, a web-based tool for extracting dataset name mentions from a given article.

CCS Concepts

• **Information systems → Information extraction**.

Keywords

Bibliometrics, Dataset Mention Extraction, LLM, RAG.

ACM Reference Format:
Priyangshu Datta, Suchana Datta, and Dwaipayan Roy. 2024. RAGing Against the Literature: LLM-Powered Dataset Mention Extraction. In *The 2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL '24), December 16–20, 2024, Hong Kong, China.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3677389.3702523

1 Introduction

The rapid increase in publications across diverse research fields has significantly intensified the challenge of locating desired papers or references. This surge in scholarly output is well-documented in recent studies [6, 7, 39, 41], highlighting an overwhelming expansion of literature. As a result, researchers face increasing difficulty in navigating through a vast amount of data to identify pertinent and high-quality references. This growing complexity underscores the need for advanced search tools and methodologies to efficiently manage and retrieve relevant academic references in an ever-expanding knowledge landscape. In the rapidly expanding realm of research, a significant challenge arises in empirical domains where researchers must demonstrate the applicability and usefulness of their proposed models on real-life data. This is essential to showcase that their proposed method indeed works. In this context, the availability and accessibility of high-quality benchmark datasets play a pivotal role in advancing scientific inquiry and innovation.

However, a critical challenge exists in efficiently identifying and tracking these datasets within the vast corpus of scientific literature. Unlike traditional citations of papers, dataset mentions often lack a standardized format, can be phrased ambiguously, and may not be explicitly referenced. For example, if we consider the following excerpts from [55], the dataset name mentioned can be easily understood as *FairytaleQA*:

> *We train and evaluate mQG on the FairytaleQA dataset, a well-structured QA dataset based on storybooks, with narrative questions.*

In contrast, the mentions are not that explicit in the following excerpts from [50]:

> *RECESS builds on the CNC (Tan et al., 2022b), which is based on the GLOCON dataset of an Event Extraction Shared Task at CASE2021 (Hürriyetoglu et al., 2021).*

Here, the datasets mentioned are: *RECESS*, *CNC*, and *GLOCON*. Although they are all written in uppercase, it is not explicitly stated in the case of *RECESS* and *CNC*. This inconsistency hinders the discovery of relevant datasets, obstructs the assessment of their usage trends, and impedes the proper acknowledgements of their creators. Automatic dataset mention extraction has emerged as a promising solution to address this challenge. By leveraging natural language processing (NLP) techniques, DME aims to automatically identify and extract dataset mentions from scientific articles. This extracted information can then be used to build comprehensive data repositories, facilitate dataset discovery, and improve the overall transparency and reproducibility of scientific research.

While DME may appear similar to a standard Named Entity Recognition (NER) task [27] applied to academic papers, it presents a significantly greater challenge. NER typically focuses on well-defined categories like people, locations, or organizations, often relying on clear capitalization or specific word lists. Dataset mentions in academic papers, however, lack such standardization. They can be phrased ambiguously, use abbreviations or synonyms, and appear within complex sentences describing the dataset's characteristics. Furthermore, unlike named entities which are inherent to the subject matter, dataset mentions may be indirectly referenced or even embedded within methodological descriptions. These factors necessitate a more sophisticated approach that can not only

JCDL '24, December 16–20, 2024, Hong Kong, China
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1093-3/24/12
https://doi.org/10.1145/3677389.3702523

identify entities but also understand the context in which they are mentioned to accurately distinguish relevant dataset mentions from unrelated terms.

DME goes beyond even keyphrase [35, 58] or metadata extraction [9] from academic papers. Scientific document processing tasks like keyphrase extraction (typically targeting 10-15 keywords per document) or named entity recognition in short excerpts (identifying 5-10 entities within 30-40 tokens) deal with much denser data compared to identifying mentions of datasets within full-text scientific documents [42]. While keyphrases can sometimes indicate the presence of a dataset, they often lack the specificity required to pinpoint the exact dataset being referred to. DME requires a deeper understanding of the relationships between words and the ability to differentiate between a general description of a dataset type and a mention of a specific dataset instance.

Recent advances in NLP, particularly in the field of large language models (LLMs) and retrieval-augmented generation (RAG), have opened up new avenues for tackling the dataset mention extraction task. LLMs, pre-trained on vast corpora of text, can capture rich linguistic patterns and contextual information, while RAG techniques leverage both retrievable knowledge sources and generative language modelling capabilities. In this paper, we present a novel approach that integrates LLMs and RAG for high-performance dataset mention extraction from academic papers. Our method utilizes the contextual understanding and generalization capabilities of LLMs, combined with the ability of RAG to selectively retrieve and incorporate relevant information from external knowledge sources. In summary, our contributions reported in this paper are threefold.

(1) We introduce a novel approach that systematically combines the strengths of the Retrieval Augmented Generation framework together with the Large Language Models for the challenging task of dataset mention extraction from academic papers.
(2) A comprehensive evaluation of the proposed method against state-of-the-art baselines on diverse datasets is conducted, providing insights into the performance gains achieved by the proposed methods as compared to the baselines. Further, the execution time is also compared to highlight the applicability of the methods in real-life scenarios.
(3) We release `exData`, a publicly accessible web-based tool that demonstrates the practical application of our proposed model and empowers researchers to efficiently extract dataset mentions from their own documents.

All the codes written for the project are freely available for reuse[1]. The remainder of this paper is organized as follows. Section 2 provides a comprehensive overview of related work in dataset mention extraction. Section 3 introduces our proposed methods: an LLM-based approach and a novel RAG-based approach. The experimental setup is detailed in Section 4, followed by the presentation and analysis of results in Section 5. Section 6 presents `exData`, a web-based tool built upon the proposed RAG-DME model. The paper concludes with a summary of contributions and potential future research directions in Section 7.

2 Related work

While information extraction, in general, has a well-established body of research [10, 52, 53, 60], the surge in interest towards automatic scientific information extraction has not yet translated into a broad research landscape. This is evident from the limited research scope compared to the general text domain, as shown in [33, 40]. This disparity stems from the inherent challenges associated with diverse scientific fields. A key obstacle lies in the specialized knowledge required to annotate training data. This expertise makes such data expensive and scarce, hindering the development of robust models for scientific information extraction.

The diversity of writing styles across academic disciplines presents a significant challenge in constructing benchmark datasets for DME. Existing datasets offer a range of coverage and granularity. Early efforts focused on specific domains like bioinformatics, with datasets containing around 60 papers Duck et al. [12]. Efforts have also been made to include papers from empirical studies in computer science and related fields [2, 14]. Additionally, benchmark datasets for the DME task have been created using papers from top AI venues [22, 34]. A multidisciplinary dataset has been proposed by D'Souza et al. [11], while a repository focused on natural language processing papers was introduced in Hou et al. [21]. Other datasets featuring scientific articles from the computer science domain are discussed in [19, 49]. However, these datasets have various limitations. For instance, most of them only include abstracts [11, 14, 34] or narrow snippets extracted from papers [19, 21, 49], while others consist of a relatively small number of papers (significantly fewer than $1,000$) [2, 12, 22]. The Coleridge Rich Context Competition (RCC)[2] has been a significant milestone in advancing the task of Dataset Mention Extraction (DME). However, the dataset utilized in this competition also suffers from limitations, including a relatively small corpus of fewer than $2,300$ papers. This constraint highlights the ongoing challenge of developing extensive and diverse benchmark datasets that can robustly support the evaluation and improvement of DME methodologies.

Recently, two more robust datasets have been proposed, containing a substantial number of papers to better evaluate the effectiveness of DME methods. The CoCon dataset [46] includes snippets from over $340,000$ papers with $7,592$ dataset mentions while, the DMDD dataset [40] provides the full content of $31,000$ papers with over $449,000$ dataset mentions. These larger and more comprehensive datasets represent a significant advancement in the resources available for DME research, allowing for more rigorous and extensive evaluation of extraction methods.

Research on dataset mention extraction (DME) has explored a diverse range of methods, spanning the spectrum of Named Entity Recognition (NER) techniques. Early work heavily relied on rule-based approaches, which offer the advantage of interpretability due to their reliance on handcrafted rules [8]. In another work, [16] focuses on a simpler approach using TF-IDF and cosine similarity for matching dataset mentions in social science articles within the da|ra registry[3]. However, the need for extensive linguistic expertise to develop these rules limits their scalability and adaptability

[1] https://github.com/priyangshu-datta/jcdl2024

[2] https://www.kaggle.com/c/coleridgeinitiative-show-us-the-data

[3] https://www.da-ra.de/

to new domains [38]. The emergence of machine learning techniques has significantly advanced the field of DME. Conditional Random Fields (CRFs) have established themselves as a powerful tool for NER tasks, achieving state-of-the-art performance across various languages [15, 29]. CRFs excel at modelling sequential data, making them well-suited for identifying and classifying dataset mentions within the text. Building on the strengths of both LSTMs and CRFs, the BiLSTM-CRF [45] architecture has emerged as a leading approach for DME. Several studies have reported the superior performance of BiLSTM-CRFs for DME tasks [57, 59].

While established methods like BiLSTM-CRFs have demonstrated strong performance in DME, recent research explores promising avenues beyond these techniques. Leveraging the strengths of language models in predicting the next words, Kumar et al. [24] proposes a simple BERT-based NER model enriched with part-of-speech (POS) information, demonstrating its effectiveness for DME. This approach reinforces the notion that DME can be effectively tackled as a specialized NER problem, as [19] argues. Their findings suggest that SciBERT [4], a transformer model pre-trained on scientific literature, achieves state-of-the-art performance when used for NER-based DME.

Moving beyond traditional NER approaches, Stavropoulos et al. [49] frame the DME task as an instruction-based question answering problem, aligning with the idea of generalized research artifact finding. This perspective suggests formulating the task as retrieving specific information based on a predefined question. Building on this concept, Younes and Scherp [56] delves deeper into transformer models, proposing more sophisticated methods that utilize a question-answering framework. Their work highlights the potential advantages of this approach, particularly for extracting previously unseen datasets, which can be a challenge for NER-based methods.

These recent advancements showcase the ongoing exploration of diverse DME methods, pushing the boundaries of traditional NER approaches and paving the way for even more robust and comprehensive dataset identification in scientific publications. It is worth noting that advancements in related tasks, like keyphrase extraction, are also informing DME research. For instance, the study reported in [36] explores the use of Graph Neural Networks (GNNs) to enhance keyphrase extraction from long documents. While this work focuses on a different task, it highlights the potential of leveraging GNNs to capture relationships between entities within text.

The recent development and ease of applications of large language models (LLMs) have revolutionized natural language processing and related fields of research. Trained on massive datasets of text, LLMs have become proficient at understanding and generating human languages. Their ability to process information, identify patterns, and respond comprehensively and coherently make them valuable tools across various applications. Improvements, such as, In-Context Learning (ICL) and Retrieval-Augmented Generation (RAG) push the boundaries of their capabilities further. In-Context Learning empowers LLMs to leverage information beyond their initial training data [54]. This allows them to adapt with specific tasks or situations, demonstrating promise in tasks like, question answering where the model can access and reference relevant documents to formulate a response [28]. On the other hand, RAG focuses on combining the strengths of both generative and retrieval techniques. By retrieving relevant passages from a vast knowledge base and then using those passages to inform its generation process, RAG can create more factually accurate and coherent texts. This approach shows promise in tasks like summarization [25], where the LLM can condense information while maintaining key details and context. These advancements, along with others like factual language understanding and commonsense reasoning, are transforming LLMs from text generators to powerful tools for a wide range of tasks, including machine translation [1], dialogue systems [47], and creative writing [18].

Previous research explored web-based approaches for dataset extraction [48] or identification and retrieval [32]. However, these methods relied on external services that might be unreliable or incur computational costs. Unfortunately, both web tools are no longer functional (last accessed October, 2024). This highlights a key limitation: the current landscape lacks readily available, live systems that can automatically extract dataset mentions directly from academic papers provided as input. Our work in this paper aims to mitigate this gap by proposing a novel DME approach that operates cost-effective, prompt-based LLMs.

3 LLM-based dataset mention extraction

In this section, we propose two approaches for extracting dataset mentions from academic papers: one leveraging a prompt-based method with Large Language Models (LLMs), and the other utilizing a Retrieval-Augmented Generation (RAG) framework. Both methods aim to accurately identify and extract names of datasets within textual documents, utilizing the advanced capabilities of LLMs but differing in their specific implementations and enhancements.

The *first* method, based on LLM prompts, employs pre-trained language models to generate contextually relevant responses that highlight dataset mentions. This approach works on the advanced understanding and contextualization capabilities of LLMs to select and extract relevant information from the text. The *second* method, using Retrieval-Augmented Generation, combines the strengths of both retrieval systems and generative models. By incorporating a retrieval component, RAG enhances the ability of the generative model to initially select a potential paragraph containing dataset names before performing the final selection thereby, improving the precision and recall of dataset mention extraction.

Extracting dataset mentions from research papers often begins with converting the documents from PDF, HTML, simple text or a URL, into a suitable format for further processing. For PDF documents, tools like GROBID [31], PDFExtract [5], Icecite [3] etc. can be used for conversion into plain textual format. Similarly, XML or HTML formatted papers can leverage Document Object Model (DOM) parsers like SAX, HtmlMonkey parser, `HtmlMonkey`[4], or `htmlparser2`[5] for text extraction. Once converted to plain text, the data undergoes preprocessing to remove unnecessary characters, and eliminate extraneous information, such as URLs and email addresses.

Our primary objective at this stage is to prepare the textual content for subsequent processing by large language models. Processing an entire research paper often exceeds several thousand words. Fitting a large text within a single LLM context window can

[4] https://github.com/fb55/htmlparser2
[5] https://github.com/SoftCircuits/HtmlMonkey

increase the risk of hallucinations. This limitation necessitates techniques, like chunking, to mitigate the issue. Hence, the input paper needs to be divided into smaller chunks before being processed by the LLMs. Before chunking, sentence segmentation is essential to preserve contextual information, and to preserve contextual information within text chunks, accurate sentence segmentation is essential. Sentence segmentation, the task of identifying sentence boundaries, is a critical preprocessing step for many NLP tasks and systems, as models often require individual sentences as input [30, 43]. Errors in segmentation can adversely impact the performance of downstream tasks, such as machine translation [37]. While various approaches exist, including rule-based methods (e.g., spaCy, PySBD [20, 44]), unsupervised methods (e.g., [23, 51]), and supervised methods (e.g., [17, 51]), our work utilizes the recently developed SaT model [13] for sentence segmentation.

Following sentence segmentation, the entire text is further divided into passages applying a sliding window approach. Specifically, we slide the window over the input text in such a way that each passage comprises nearly 200 tokens and overlaps with its adjacent passages by two sentences. This 'sliding window' approach, facilitated by our custom chunking function, ensures that context is preserved across chunks. Such preservation is crucial for LLMs to understand the relationships between sentences while extracting information effectively.

Undergoing these common steps, our proposed LLM-based and RAG-based methods, namely LLM-DME and RAG-DME respectively, further diverge in their approaches towards dataset mention extraction which we detail in the following two sections.

3.1 LLM-DME

The LLM-DME method leverages the capabilities of large language models to extract dataset names mentioned within research papers. This approach is designed to handle the complexities of academic language and the variability of how datasets are referenced. By using LLMs with carefully crafted prompts, LLM-DME aims to provide a comprehensive and precise extraction of dataset mentions.

Once the input snippet information is chunked, as discussed in Section 3, each chunk is then passed through a large language model (LLM) along with a carefully crafted prompt. The prompt is designed to instruct the LLM to identify and extract the dataset names mentioned within each chunk. This process precisely involves the following steps:

(1) **Crafting the prompt:** The prompt is formulated to clearly direct the attention of the LLM to the task of extracting dataset names. In-context learning is utilized, where the prompt includes specific patterns and examples of how dataset names are typically mentioned in research papers. This method leverages the LLM's ability to learn from the provided context within the prompt itself. The exact prompt used in our study is outlined in Appendix A.1.

(2) **Contextual understanding:** Each chunk of text, containing around 200 tokens with an overlap of 2 sentences with adjacent chunks, is paired with the prompt. The overlapping sentences ensure that the LLM retains context and continuity between chunks, improving the accuracy of dataset extraction.

(3) **Individual processing:** The LLM processes each chunk individually, applying the prompt to identify relevant dataset names. This step leverages the LLM's ability to comprehend complex language and recognize variations in how datasets might be mentioned.

(4) **Aggregation of results:** After processing all chunks, the extracted dataset names are aggregated. This step involves combining the results from each chunk, ensuring that all mentions are captured and duplications are minimized.

(5) **Final output:** The final output is a comprehensive list of dataset names extracted from the research paper, providing a detailed and accurate representation of the datasets mentioned throughout the document.

3.2 RAG-DME

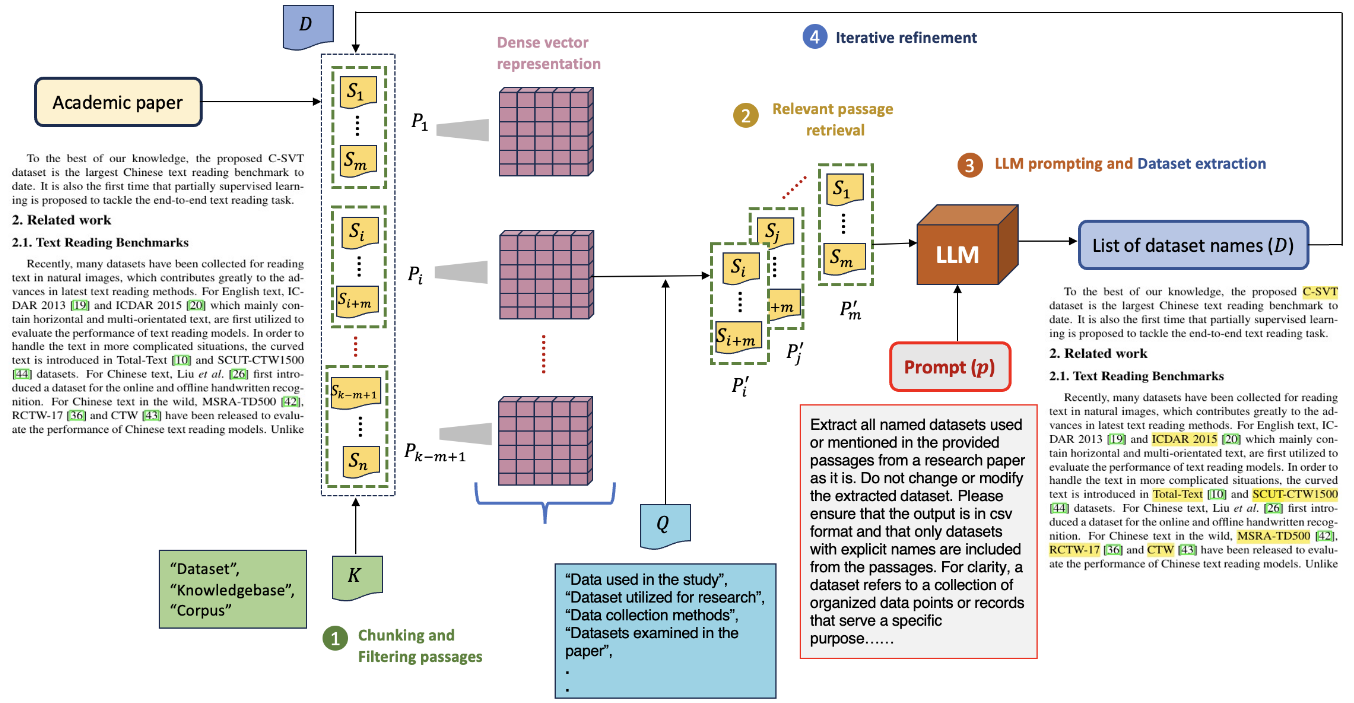
The RAG-DME method utilizes the Retrieval-Augmented Generation (RAG) framework for dataset mention extractions from research papers. A graphical representation of the model is presented in Figure 1. The preprocessing and chunking steps, denoted as 'Chunking and filtering passages' in Figure 1, for this method are the same as presented earlier in the Section 3.1. These steps include converting the research paper into textual format, preprocessing the text, splitting it into sentences, and then grouping the sentences into passages following the sliding window protocol. With the text prepared in this manner, the specific steps involved in the RAG-DME method are as follows.

(1) **Keyword-based passage filtering**: In this step, a set of predefined keywords, denoted as K, is used to filter the passages P from a research paper (refer to ① in the Figure 1). The motivation here is to identify passages that are likely to contain dataset name mentions. Examples of such predefined keywords are '*dataset*', '*knowledgebase*', '*corpus*' etc. A comprehensive list of keywords is curated through an in-depth analysis of randomly selected papers from the DMDD [40] dataset, supported by the expertise of co-authors of this article in research paper writing. The complete keyword list is provided later in the Appendix A.2. Specifically, a simple boolean retrieval is performed with the terms from the keyword list and a union of selected passages are finally taken into account.

(2) **Relevant passage retrieval**: The filtered list of passages P is then further tailored via dense retrieval using cosine similarity semantic search. This retrieval process identifies the most relevant passages, denoted as P', that are likely to contain dataset mentions as shown in ② in Figure 1. A predefined list of queries are used for retrieval, such as "*data used in the study*". This approach narrows down to the most pertinent passages. The entire list of queries used for this step is provided in the Appendix A.3.

(3) **Dataset extraction**: The relevant passages P' are then passed through a large language model with the same prompt used in the LLM-DME method, denoted as p. This prompt instructs the LLM to extract mentions of datasets from the input passages (see ③ in the Figure 1). By using a consistent prompt, we ensure that the extraction process is uniform across different methods.

(4) **Iterative refinement**: Once an initial set of datasets D is extracted, this set is treated as a new set of keywords, effectively


D
Academic paper
S_1
S_m
S_i
S_{i+m}
S_{k-m+1}
S_n
P_1
P_i
P_{k-m+1}
Dense vector representation
K
"Dataset", "Knowledgebase", "Corpus"
1 Chunking and Filtering passages
Q
2 Relevant passage retrieval
S_j
P_i'
P_j'
P_m'
3 LLM prompting and Dataset extraction
LLM
Prompt (p)
List of dataset names (D)
4 Iterative refinement

To the best of our knowledge, the proposed C-SVT dataset is the largest Chinese text reading benchmark to date. It is also the first time that partially supervised learning is proposed to tackle the end-to-end text reading task.

2. Related work

2.1. Text Reading Benchmarks

Recently, many datasets have been collected for reading text in natural images, which contributes greatly to the advances in latest text reading methods. For English text, ICDAR 2013 [19] and ICDAR 2015 [20] which mainly contain horizontal and multi-orientated text, are first utilized to evaluate the performance of text reading models. In order to handle the text in more complicated situations, the curved text is introduced in Total-Text [10] and SCUT-CTW1500 [44] datasets. For Chinese text, Liu *et al.* [26] first introduced a dataset for the online and offline handwritten recognition. For Chinese text in the wild, MSRA-TD500 [42], RCTW-17 [36] and CTW [43] have been released to evaluate the performance of Chinese text reading models. Unlike

"Data used in the study",
"Dataset utilized for research",
"Data collection methods",
"Datasets examined in the paper",
.
.

Extract all named datasets used or mentioned in the provided passages from a research paper as it is. Do not change or modify the extracted dataset. Please ensure that the output is in csv format and that only datasets with explicit names are included from the passages. For clarity, a dataset refers to a collection of organized data points or records that serve a specific purpose……

To the best of our knowledge, the proposed C-SVT dataset is the largest Chinese text reading benchmark to date. It is also the first time that partially supervised learning is proposed to tackle the end-to-end text reading task.

2. Related work

2.1. Text Reading Benchmarks

Recently, many datasets have been collected for reading text in natural images, which contributes greatly to the advances in latest text reading methods. For English text, ICDAR 2013 [19] and ICDAR 2015 [20] which mainly contain horizontal and multi-orientated text, are first utilized to evaluate the performance of text reading models. In order to handle the text in more complicated situations, the curved text is introduced in Total-Text [10] and SCUT-CTW1500 [44] datasets. For Chinese text, Liu *et al.* [26] first introduced a dataset for the online and offline handwritten recognition. For Chinese text in the wild, MSRA-TD500 [42], RCTW-17 [36] and CTW [43] have been released to evaluate the performance of Chinese text reading models. Unlike

Figure 1: Schematic of our proposed iterative RAG-based dataset mention extraction model RAG-DME. A scientific paper is first (1) split into sentences $\{S_1, ..., S_n\}$ and then grouped into passages $\{P_1, ..., P_k\}$, followed by a filtration with the help of a list of predefined keywords, K. Next, (2) a set of queries Q is used to retrieve a relevant set of passages $\{P_1', ..., P_m'\}$ which are (3) passed through a LLM with a prompt p to extract the set of dataset mentions, D at the first iteration. From next iteration onwards until the model outcome converges, (4) the RAG-DME repeats the work flow by replacing the keyword list, K with the obtained dataset mention list, D at the fist step and (5) the LLM prompt is updated by concatenating D with it.

updating K to D. The method then loops back to step (1), utilizing the updated keyword set to refine the dataset extraction process as depicted by (4) in Figure 1. This recursive retrieval technique helps to iteratively enhance the accuracy and comprehensiveness of the dataset mentions extracted from the research papers.

(5) **Prompt updating**: Along with the iterative refinement of keywords, the prompt p is also updated to include the extracted datasets D as examples. The modified prompt is created by concatenating the older prompt with the selected dataset names. This updated prompt is then used in subsequent iterations to further improve the extraction process. This step ensures that the LLM has more context and examples to work with, thereby increasing its ability to accurately identify dataset mentions.

(6) **Termination condition**: The steps of keyword-based passage filtering, relevant passage retrieval, dataset extraction, iterative refinement, and prompt updating are repeated until no new datasets are found. This termination condition ensures that the method converges and that all possible dataset mentions are extracted.

From the architecture, the first step of keyword-based filtering (performed using a boolean retrieval) may seem redundant with the following semantic retrieval step. However, there is a difference between the two retrievals: while the first retrieval works based on a boolean condition check (presence or absence), the dense retrieval considers a high-level concept-based matching that goes beyond a term overlap. The aspect of this difference can be understood clearly from the list of queries as well: the queries are keywords for the filtering (step 1) while a set of conceptual topics is used as queries in the dense retrieval step. This step is important for the iterative refinement part of the model. First of all, this filtering serves to drastically reduce the number of passages needed to be processed in the subsequent step using semantic retrieval. This improves the efficiency and accuracy of the extraction process. Also, the filtering establishes a base to improve upon by including passages containing obvious keywords that may indicate the mention of dataset names. Another reason for keyword-based filtering prior to semantic retrieval is that dataset names are often mentioned in close proximity within the text. By initially filtering the passages that contain specific keywords related to datasets, we increase the likelihood of capturing relevant context around dataset mentions. This approach takes advantage of how research papers are written, where important terms and their explanations are often close together. By using keyword-based filtering first, we make sure that the next step, semantic retrieval, focuses on passages that are both relevant and full of useful context.

Note that, although we are referring to this model as 'RAG' based, it does not strictly adhere to the traditional Retrieval-Augmented

Table 1: Datasets used in our study.

Dataset	# papers	content	# dataset mentions	# queries
DMDD [40]	31,219	whole paper	over 449000	450
CoCon [46]	340,965	abstract	7,592	500

Table 2: Characteristic features of both the baselines and proposed DME models presented in this paper.

Model	Type	Embedding	NER	Input
SciBERT$_{sentence}$	Baseline	✓	✓	Sentence
SciBERT$_{sample}$	Baseline	✓	✓	Sentence
SciBERT$_{full}$	Baseline	✓	✓	Paper
BiLSTM-CRF	Baseline	✓	✓	Paper
LLM-DME	Proposed	✗	✗	Paper
RAG-DME	Proposed	✓	✗	Paper

Generation framework, where an external resource is utilized for retrieval, and the resultant list is subsequently fed into the Language Model (specifically an LLM). Instead, in this approach, the passages from which the dataset names are to be extracted are ranked based on semantic retrieval, similar to RAG, but the retrieval is performed on the internal passages rather than external resources. These ranked passages are then provided as input to the LLM along with the prompt. This variation from the traditional RAG framework allows for a more focused and contextually relevant extraction process by ensuring that the LLM processes only the most important passages within the document itself.

4 Experiments

4.1 Benchmark datasets

For our study, we have selected two of the most robust datasets available for dataset mention extraction, particularly the CoCon [46] and the DMDD [40] datasets. Table 1 summarises the basic statistics of the datasets. While the DMDD dataset provides pre-defined evaluation sets with 450 papers, the CoCon lacks such a designated set for the task of DME. To address this, we randomly selected 500 papers from the entire CoCon collection for our empirical study and evaluation, allowing us to report the performance of our proposed method and the baselines on this subset.

4.2 Baselines

We compare our proposed unsupervised LLM-based LLM-DME and RAG-based RAG-DME with a number of standard supervised dataset mention extraction approaches, namely SciBERT [40] and BiLSTM-CRF [45]. It is worth noting that we compare the performance of our proposed unsupervised models with the supervised ones as there are no unsupervised techniques available in the dataset mention extraction literature to the best of our knowledge [19]. The superiority of the proposed unsupervised models over the supervised baselines are depicted in Section 5.

SciBERT [40]. Following the approaches in [40], we also formulate the task of dataset mention as a token-tagging task that serves as our baselines in this paper and hence evaluate on the popular pipelines, such as SciBERT and BiLSTM with Conditional Random Fields (CRF) (see the next paragraph for more details). Specifically, we conduct experiments using the pretrained model SciBERT [4] - `'scibert-scivocab-cased'` that takes sentences as input and extracts dataset mention from input sentences as a task of token classification. SciBERT is trained on a large corpus of nearly 1.14 million scientific papers both from computer science and medical domains and is built upon the BERT architecture to create domain-specific language models. While training, all the hyper-parameters are kept same as the original SciBERT setup in [4].

Same as [40], we also leverage both sentence-level information and beyond sentences with different input lengths while training SciBERT model. We name the different variants of SciBERT in terms of input lengths as SciBERT$_{sentence}$, SciBERT$_{sample}$ and SciBERT$_{full}$ (refer to Table 2 for better understanding of the characteristics of each variant). Different from the sentence-level approach in [40], in this paper SciBERT$_{sentence}$ simply divides the whole article into multiple sentences and feed to the model with token labels and thus extracts specific tokens indicating dataset mentions. SciBERT$_{sample}$ on the other hand exploits sentence level information with limited positive (80%) and negative (20%) samples. The third variant of SciBERT, i.e. SciBERT$_{full}$ utilizes information from the whole article chunking it into 512-level-token input and not being limited to positive and negative samples like SciBERT$_{sample}$. It is worth mentioning that although both this paper and [40] leverage sentence-level information to train SciBERT model, the input pipelines are different from each other, hence, the performance reported in both the papers are not directly comparable.

BiLSTM-CRF [45]. Instead of utilizing the BERT-based pretrained models to build a domain-specific language model, BiLSTM-CRF formulates the dataset mention extraction task as a sequence labelling task with a rich representation of language using contextual embeddings. Specifically, each token in the input sequence is first mapped to a fixed sized dense vector. A BiLSTM is used to encode sequential relations between the word representations. As BiLSTM operates in both directions, the hidden state vectors leverage information both from its succeeding and preceeding terms, which subsequently present a contextual vector representation of the input text. As Conditional Random Fields (CRF) proved to be efficient to improve the performance of various sequence labelling tasks [26, 29, 45, 59], in this paper we apply CRF on BiLSTM score output sequences following its superiority in the sequence labelling task in the literatire [26, 45].

4.3 Experimental and evaluation settings

LLM parameters. As presented in Section 3, both proposed methods for DME rely on large language models. In our initial study, we have tested the methods on multiple LLMs, particularly `WizardLM-7B-uncensored-GPTQ`[6], `Llama-2-7B-32K-Instruct-GPTQ`[7], `Mistral-7B-Instruct-v0.1-GPTQ`[8], and Google's `Attributed Question Answering` (AQA) model[9]. We observed some variation in performance along with execution time for certain models, with Google's AQA model performing comparatively optimally. Hence

[6] https://huggingface.co/TheBloke/WizardLM-7B-uncensored-\GPTQ
[7] https://huggingface.co/TheBloke/Llama-2-7B-32K-Instruct-GPTQ
[8] https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.1-GPTQ
[9] https://ai.google.dev/gemini-api/docs/semantic_retrieval

in this study, we reported the performance of the proposed methods (LLM-DME and RAG-DME) applying Google's AQA model as LLM.

Google's AQA model employs specific hyperparameters to control its behaviour during the generation process. The `temperature` parameter controls the randomness of the generated text. A lower value makes the output more focused and deterministic, while a higher value introduces more randomness. In our initial experiments, we varied the temperature parameter from 0.02 to 0.2 in steps of 0.02. The optimal performance was obtained with both LLM-DME and RAG-DME when the temperature was set to 0.06 on randomly selected 50 queries from the DMDD benchmark, indicating a lower level of randomness and a more focused output. Hence, we report the performance of the proposed models when `temperature` parameter is set to 0.06.[10] A comparative study of the impact on performance on varying this parameter is elaborated later in this section. Another parameter is `answer_style` (with options `abstractive, extractive` and `verbose`) which determines how the model generates its responses. For our task of extracting dataset names, it is set to `extractive` answer style for obvious reasons. This means that the model generates responses by selecting and extracting relevant information from the input data rather than generating entirely new text. We utilized ChromaDB vectorDB along with the `all-MiniLM-L6-v2` SentenceTransformer embedder model[11] to persist vector representations of text. This approach allowed us to efficiently store and retrieve vector embeddings, enabling effective analysis and processing of textual data

Evaluation protocol. For the evaluation of the DME methods, 5-fold cross-validation is employed for the supervised techniques to ensure robust performance assessment. For the unsupervised techniques (LLM-DME and RAG-DME), the same test sets from the cross-validation folds are utilized, and the results are averaged to provide a comprehensive and comparable evaluation.

Evaluation metrics. Evaluating the performance of systems for DME traditionally requires reporting both partial and exact match scores. Exact matches capture the full and precise dataset name, representing ideal extractions, while partial matches are important as academic papers often use abbreviations, synonyms, or slightly different phrasings when referring to datasets. However, in this study, we focus exclusively on exact matches to ensure that the extracted dataset names are fully accurate. By doing so, we aim to provide a clear and rigorous assessment of the model in identifying dataset mentions that precisely match the gold standard annotations. In this study, we report precision, recall, and F-scores based solely on exact matches.

5 Results and analysis

Table 3 presents the empirical results of all the DME methods on the two benchmark datasets. Among the baseline models, the BiLSTM-CRF [45] (discussed in Section 4.2) outperformed other baseline methods in our experiments. For the DMDD dataset, when the entire paper was considered as a query, we observed that the precision of LLM-DME in general improved as compared to most of the baseline methods, except BiLSTM-CRF, however, the difference is a mere 1% as compared to BiLSTM-CRF. In contrast, the recall improved a noticeable 26% highlighting the significant potential of large language models for this task. This translated into an overall F-score improvement of 13% over the CRF-based method by LLM-DME. Noticeably, the integration of retrieval augmented generation-based framework with a large language model yielded significantly better results compared to baselines as well as traditional LLM prompting. The RAG-based approach (RAG-DME) achieved a precision of 0.8922 on the DMDD dataset, surpassing the performance of BiLSTM-CRF as well asLLM-DME by over 14%. Additionally, recall improved by 8% over the LLM-DME and by a substantial 37% compared to BiLSTM-CRF. Overall, RAG-DME achieved an F-score of 0.8094 which is over 25% improvement as compared to the best-performing baseline.

For the CoCon dataset, the BiLSTM-CRF [45] achieved an F-score of 0.7287 and a precision of 0.8322. While BiLSTM-CRF excelled in precision, SciBERT$_{sentence}$ demonstrated the highest recall of 0.6573, closely leading the performance of BiLSTM-CRF. A similar improvement over baselines was recorded by the LLM-DME method for the CoCon datasets as well. Notably, RAG-DME surpassed all baselines as well as LLM-DME, exhibiting a 3% improvement in precision and a substantial 26% increase in recall over BiLSTM-CRF. This translates to an overall F-score improvement of over 15% for RAG-DME compared to all baseline models on the CoCon dataset.

While performance improvements were observed on both datasets, the gains achieved on the DMDD dataset were significantly more substantial than those on the CoCon dataset. It is also worth noting that the overall performance of all the DME methods are better on the CoCon dataset than DMDD. This is because DMDD contains larger, more complex papers with a wider variety of dataset mentions; also the size of the queries is significantly bigger for DMDD (whole paper) than CoCon (abstract) resulting in making the task more challenging. This also resulted in providing more opportunities for the improvement of the proposed models. Overall, both proposed models substantially outperformed baselines in comprehensively identifying dataset names in both benchmarks with varying sizes, leading to substantial recall as well as precision improvements.

Comparing model turnaround time. Table 4 presents a comparative analysis of average execution time for the proposed and baseline DME methods. We report the total execution time for the 5-fold cross-validation in the fourth column, while the last column contains the average per-query execution time. Execution times for both the proposed LLM-DME and RAG-DME methods are reported for the complete query set on each dataset. The same comparison is graphically presented in Figure 2 where the average per-query execution time for both benchmarks is presented using a line plot. The table and figure show that the proposed RAG-DME method outperforms all other methods discussed in terms of speed. Notably, both the proposed methods exhibit lower turnaround time than all the baseline methods on both benchmarks. Further, the LLM-DME method exhibits lower turnaround than the RAG-supported proposed method as well. While the LLM in RAG-DME processes less text per query compared to the traditional prompt-based LLM

[10] A similar trend was observed in CoCon dataset: a random sample of 100 papers was used during the tuning of `temperature`. We ensured that the papers selected for parameter tuning were distinct from the test query set for unbiased training.
[11] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Table 3: Performance metrics (Precision, Recall, F-score) of the proposed DME methods and baseline models on two benchmark datasets. The best performance for each dataset is highlighted in bold, and the top-performing baseline metrics are underlined. The percentage of improvement over the best-performing baseline is presented in brackets for the two proposed methods.

		DMDD Dataset			CoCon Dataset		
	Methods	Precision	Recall	F-score	Precision	Recall	F-score
Baselines	SciBERT$_{sentence}$ [40]	0.6583	0.4987	0.5675	0.8046	0.6573	0.7235
	SciBERT$_{sample}$ [40]	0.7537	0.5389	0.6285	0.8176	0.6513	0.7250
	SciBERT$_{full}$	0.7011	0.5099	0.5904	0.7944	0.6431	0.7108
	BiLSTM-CRF [45]	0.7801	0.5425	0.6400	0.8322	0.6481	0.7287
Ours	LLM-DME	0.7701 (-1%)	0.6828 (26%)	0.7238 (13%)	0.8539 (2%)	0.8250 (26%)	0.8392 (15%)
	RAG-DME	**0.8922** (14%)	**0.7408** (37%)	**0.8094** (26%)	**0.8562** (2%)	**0.8315** (27%)	**0.8436** (16%)

Table 4: Comparison of performance of baselines and our proposed unsupervised approaches (LLM-DME and RAG-DME) in terms of the total time taken to respond to the full query set and the average inference time per query for two benchmark datasets, DMDD and CoCon. The least execution time for each dataset is highlighted in bold.

Dataset	Methods	# queries	Total (min.)	Average (sec.)
DMDD	SciBERT$_{sentence}$	450	296.17	39.49
	SciBERT$_{sample}$		321.10	42.81
	SciBERT$_{full}$		389.70	51.96
	BiLSTM-CRF		411.08	54.81
	LLM-DME		**99.36**	**13.52**
	RAG-DME		151.05	20.14
CoCon	SciBERT$_{sentence}$	500	135.82	16.30
	SciBERT$_{sample}$		158.31	18.99
	SciBERT$_{full}$		182.08	21.85
	BiLSTM-CRF		201.30	24.16
	LLM-DME		**17.57**	**2.11**
	RAG-DME		29.19	3.50

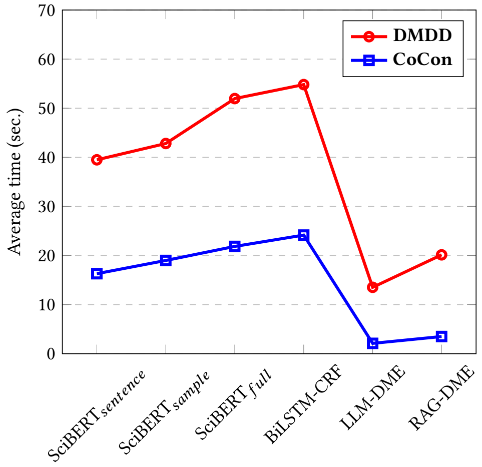


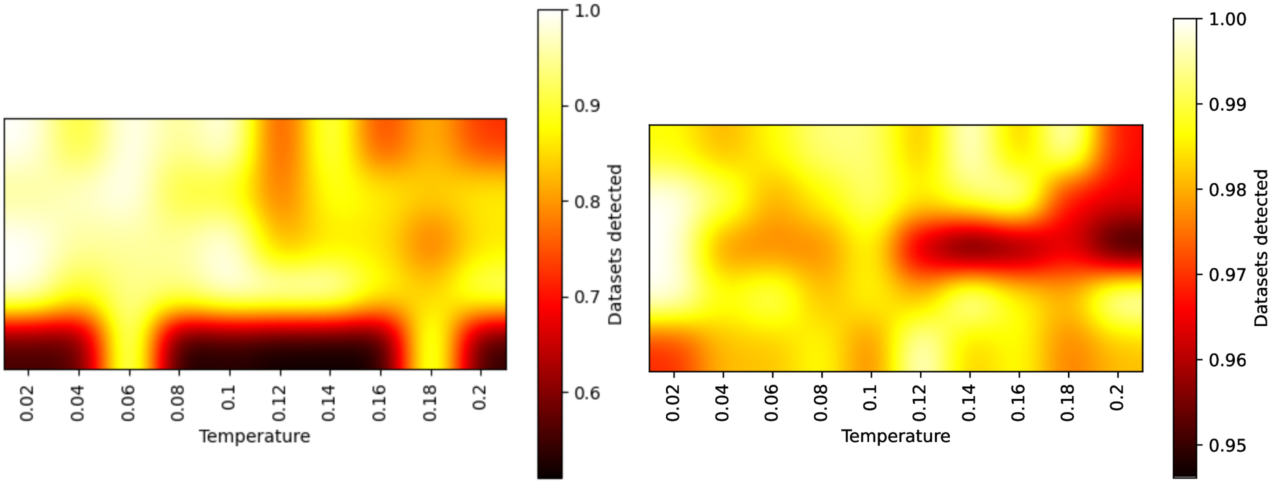
Figure 2: Average execution time for each DME method when processing an entire DMDD paper or a CoCon abstract as a single query.

approach LLM-DME, there is an increase in execution time reported per-query in both benchmarks. The increase in turnaround is approximately 7 seconds per query from DMDD and less than 2 seconds for CoCon.

Effect of varying `temperature` on performance. We demonstrate the variation in DME performance when `temperature` parameter of LLM is varied. A higher `temperature` value introduces more randomness and makes LLMs less deterministic in their text generation. This can result in diverse and creative outputs, which are beneficial for tasks requiring novelty and exploration, such as creative writing. However, for the task of DME, the primary objective is to accurately identify specific keywords, requiring precision and consistency over creativity.

Given the deterministic nature of keyword selection in DME, there is minimal need for the variability introduced by higher temperature settings. Hence, we have kept the value of `temperature` set near the lower end of the spectrum. Particularly, a grid search was conducted within the range of 0.02 to 0.2, with increments of 0.02 to optimize the `temperature` parameter. For each temperature value, the RAG-DME model was applied to 500 randomly selected papers from the DMDD and CoCon datasets (excluding evaluation sets). We generated two heatmaps visualizing the percentage of correctly extracted datasets for each temperature setting in Figure 3. In these heatmaps, colour intensity correlates with the number of extracted datasets, with lighter shades representing a higher percentage. The optimal temperature was determined to be 0.06 for the DMDD dataset, as indicated by the overall lighter shades in Figure 3a. While the CoCon dataset also exhibited less sensitivity to temperature variations, with acceptable performance observed between 0.04 and 0.1 (see Figure 3b), the value of 0.06 was adopted for consistency across both datasets.

A similar trend in performance was noted for LLM-DME when the temperature was varied in a grid structure. Particularly, we noticed indistinguishable variation in performance when `temperature` was set to values between 0.02 and 0.08 with a major performance deterioration after that. Hence, the same `temperature` value was set (0.06) for the LLM-DME method as well.



(a) DMDD

(b) CoCon

Figure 3: Heatmap visualizing the dataset extraction results obtained by varying the Temperature parameter from 0.02 to 0.2, with a step of 0.02. Colour intensity represents the number of datasets extracted for each Temperature value. The highest percentage of datasets extracted at `temperature` value near 0.06, indicating the optimal value for the DME tasks.

6 `exData`: A Dataset Extractor and Explainer

We present `exData`, a web-based application designed to streamline the identification and annotation of dataset mentions within research papers. It leverages RAG-DME, the RAG-supported LLM-based dataset mention extraction method described earlier in this paper (see Section 3.2). The tool can be accessed from this link: https://jcdl2024.streamlit.app/. [12] Followings are the notable features of `exData`.

- **User-friendly interface:** `exData` offers a simple yet intuitive web-based interface where users can easily upload research papers in PDF format. The tool also supports direct PDF crawling and processing from a given URL.[13]
- **Dataset name extraction:** After receiving an input PDF, `exData` utilizes the RAG-DME method (discussed in Section 3.2 to extract dataset names from the text. The underlying algorithm processes the document, converting it into a textual format using GROBID [31], preprocessing the text, splitting it into sentences and passages, and then applying the LLM with the specified prompt after performing dense retrieval based on the list of queries, to identify dataset mentions.
- **PDF annotation:** After extracting the dataset names, `exData` visually enhances the input PDF by highlighting identified dataset mentions. This feature provides the users with the context in which the dataset is mentioned within the paper enhancing the *explainability* and accessibility of the information. Users can also download the annotated PDF with highlighted dataset mentions for easy incorporation into their research workflow or sharing with collaborators.

To accurately assess the computational efficiency of the RAG-based DME process as utilized in `exData`, it is important to note that the result caching is explicitly turned off in the current implementation. However, caching is an essential feature that `exData` will incorporate in future releases. A simple walk-through of `exData` can be found in the following link: https://tinyurl.com/jcdl24-exData.

7 Conclusion and Future Work

In this paper, we have introduced two novel approaches for extracting dataset names from academic papers. The first method utilizes the capabilities of large language models in understanding textual patterns based on prompts. The second method builds upon this by incorporating a retrieval-based filtering framework similar to the retrieval augmented generation (RAG) approach.We have evaluated both methods against state-of-the-art baselines on two benchmark datasets, demonstrating significant improvements in precision, recall, and F-score. The retrieval-supported RAG-DME model achieved the best overall performance, with a substantial recall enhancement of up to 37% compared to the best-performing baseline. Both the proposed methods demonstrated efficient processing times, with minimal computational overhead. In addition, we introduce `exData`, a tool that automatically annotates the input PDFs with the extracted dataset names, highlighting them within the context of the text. This feature allows users to have a deeper understanding of how datasets are mentioned and utilized within academic papers.

The rationale behind choosing 'dataset mention' over the other is motivated by the fact that this domain of study is less explored in the literature. As part of future work, we plan to apply the proposed methods for similar entity extraction tasks such as baseline, tasks, software ([49]) etc. extractions. Additionally, `exData` will incorporate result caching to avoid the repetition of computation for the same paper, together with entity ranking.

[12] `exData` is deployed on a free platform https://streamlit.io/. Due to potential server load, `exData` may experience temporary unavailability. In such cases, users are encouraged to revisit at a later time.

[13] This has been tested with papers from https://arxiv.org/ and https://aclanthology.org/; users may face some technical issues if the source is restricted.

A Appendix

A.1 Prompt used in LLM

In this section, we present the exact prompt used by the LLM for Dataset Mention Extraction (DME) in this paper. In our initial experiments, we varied the prompt and manually evaluated the performance. After careful consideration and comparison, we finalized the following prompt, which has been carefully tailored to identify and extract dataset mentions from the input text.

Prompt used in LLM

```
Extract all named datasets used or mentioned in the provided passages from a research paper as it is. Do not change or modify the extracted dataset. Ensure that the output is in csv format and that only datasets with explicit names are included from the passages. For clarity, a dataset refers to a collection of organized data points or records that serve a specific purpose. Datasets are commonly utilized in various fields such as science, research, machine learning, statistics, economics, and more. They can be structured or unstructured and are often referenced in research papers to support findings, validate hypotheses, or provide evidence for arguments. Datasets may be explicitly mentioned within the passages, such as "We utilize the <Dataset> collected from <Source> for our analysis." or "The <Dataset> provided by <Provider> contains valuable information for our research." Additionally, datasets can be constructed from other datasets through aggregation, transformation, or combination processes. For instance, "We constructed our dataset by merging data from multiple sources, including <Dataset1> and <Dataset2>." In some cases, the word "dataset" may be implicit, and datasets may be referred to by other terms such as "data collection", "data source", or "data repository". Datasets are NOT methods. Methods are something which is applied. Datasets are used on methods. So, extract datasets and ignore methods. Ensure that the extraction process focuses on identifying datasets with specific names and excludes general descriptions of data sources or collections. Datasets are alphanumeric words that may not have any meaning.
```

A.2 Keyword list for passage filtering

In this section, we present the carefully curated list of keywords used to filter the passages containing potential dataset mentions as mentioned in Section 3. Note that, this list is not an exhaustive list and should be modified/expanded according to the specific research domains being evaluated.

List of keywords (in no particular order)

```
data(set|base)
anal(ytics|ysis)
resear(ch|ch paper)
stud(y|ies?)
exper(iment|iments?)
method(ology|ologies?)
collect(ion|ions?)
sampl(e|ing)
variabl(e|es?)
observ(ation|ations?)
surve(y|ys?)
popul(ation|ations?)
repositor(y|ies?)
databas(e|es?)
sourc(e|es?)
raw data
secondar(y|ies?)
primar(y|ies?)
min(e|ing)
proces(s|sing)
clean(ing|)
manipul(ation|ations?)
integrat(e|ion)
aggregat(e|ion)
visualiz(e|ation)
interpret(ation|ations?)
data(-|\s)?set
task
challenge
(knowledge|data)\s*base
benchmark
corpus
(train|test) (set)
evaluat(ed|ion)

(trained|experimented) on
(evaluated|tested|compared) on
(experiment|train|performance)[\sa-zA-Z0-9]+on
```

A.3 Query list for semantic search

As presented in Section 3.2, the RAG-DME model performs a dense retrieval in its pipeline. This step ensures filtered, more focused chunks of text are passed on to the LLM for the dataset name selection. The dense retrieval is performed using a set of predefined queries based on our observation of how the datasets are typically mentioned in academic papers. The query list is constructed using a methodology analogous to the keyword list generation process outlined in Appendix A.2 which is used for performing a filtering of the passages.

List of queries

```
"Data used in the study",
"Dataset utilized for research",
"Data collection methods",
"Datasets examined in the paper",
"Datasets referenced in the research",
"Data sources investigated",
"Dataset mentioned in the study",
"Data utilized for analysis",
"Data collection procedures",
"Dataset discussed in the paper",
"Data sources utilized",
"Data sources referenced in the paper",
"Datasets employed for investigation",
"Datasets used as benchmarks",
"Results of challenge",
"Other datasets are",
"Another dataset is"
```

References

[1] Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering Large Language Models for Machine Translation with Finetuning and In-Context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 11127–11148.

[2] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In *Proc. of 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, 546–555.

[3] Hannah Bast and Claudius Korzen. 2017. A Benchmark and Evaluation for Text Extraction from PDF. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 1–10.

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proc. of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3615–3620.

[5] Oyvind Raddum Berg. 2011. PDFExtract. https://github.com/oyvindberg/PDFExtract/. Last accessed on July 25th, 2024.

[6] Lutz Bornmann and Robin Haunschild. 2022. Empirical analysis of recent temporal dynamics of research fields: Annual publications in chemistry and related areas as an example. *Journal of Informetrics* 16, 2 (2022), 101253.

[7] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications* 8, 1 (Oct. 2021).

[8] Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain Adaptation of Rule-Based Annotators for Named-Entity Recognition Tasks. In *Proc. of 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1002–1012.

[9] Muntabir Hasan Choudhury, Jian Wu, William A. Ingram, and Edward A. Fox. 2020. A Heuristic Baseline Method for Metadata Extraction from Scanned Electronic Theses and Dissertations. In *Proc. of ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20)*. Association for Computing Machinery, 515–516.

[10] Julien Delaunay, Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Georgeta Bordea, Nicolas Sidere, and Antoine Doucet. 2023. A Comprehensive Survey of Document-level Relation Extraction (2016-2023). arXiv:2309.16396

[11] Jennifer D'Souza, Anett Hoppe, Arthur Brack, Mohmad Yaser Jaradeh, Sören Auer, and Ralph Ewerth. 2020. The STEM-ECR Dataset: Grounding Scientific Entity References in STEM Scholarly Content to Authoritative Encyclopedic and Lexicographic Sources. In *Proc. of Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, 2192–2203.

[12] Geraint Duck, Goran Nenadic, Andy Brass, David L Robertson, and Robert Stevens. 2013. bioNerDS: exploring bioinformatics' database and software use through literature mining. *BMC Bioinformatics* 14, 1 (jun 2013).

[13] Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment Any Text: A Universal Approach for Robust, Efficient and Adaptable Sentence Segmentation. *arXiv preprint arXiv:2406.16678* (2024).

[14] Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers. In *Proc. of 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 679–688.

[15] Georgi Georgiev, Preslav Nakov, Kuzman Ganchev, Petya Osenova, and Kiril Simov. 2009. Feature-Rich Named Entity Recognition for Bulgarian Using Conditional Random Fields. In *Proc. of International Conference RANLP-2009*. Association for Computational Linguistics, 113–117.

[16] Behnam Ghavimi, Philipp Mayr, Christoph Lange, Sahar Vahdati, and Sören Auer. 2017. A semi-automatic approach for detecting dataset references in social science texts. *Inf. Serv. Use* 36, 3-4 (Feb. 2017), 171–187.

[17] Dan Gillick. 2009. Sentence Boundary Detection and the Problem with the U.S.. In *Proc. of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Association for Computational Linguistics, 241–244.

[18] Carlos Gómez-Rodríguez and Paul Williams. 2023. A Confederacy of Models: a Comprehensive Evaluation of LLMs on Creative Writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 14504–14528.

[19] Jenny Heddes, Pim Meerdink, Miguel Pieters, and Maarten Marx. 2021. The Automatic Detection of Dataset Names in Scientific Articles. *Data* 6, 8 (2021).

[20] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. (2020). https://spacy.io/

[21] Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2021. TDMSci: A Specialized Corpus for Scientific Literature Entity Tagging of Tasks Datasets and Metrics. In *Proc. of 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 707–714.

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

[23] Tibor Kiss and Jan Strunk. 2006. Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics* 32, 4 (12 2006), 485–525.

[24] Sandeep Kumar, Tirthankar Ghosal, and Asif Ekbal. 2021. DataQuest: An Approach to Automatically Extract Dataset Mentions from Scientific Papers. In *Towards Open and Trustworthy Digital Societies*. Springer International Publishing, 43–53.

[25] Philippe Laban, Alexander R. Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a Haystack: A Challenge to Long-Context LLMs and RAG Systems. arXiv:2407.01370 [cs.CL] https://arxiv.org/abs/2407.01370

[26] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of Eighteenth International Conference on Machine Learning (ICML '01)*. Morgan Kaufmann Publishers Inc., 282–289.

[27] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proc. of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 260–270.

[28] Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. 2023. Few-shot In-context Learning on Knowledge Base Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 6966–6980.

[29] Wei Li and Andrew McCallum. 2003. Rapid development of Hindi named entity recognition using conditional random fields and feature induction. *ACM Transactions on Asian Language Information Processing* 2, 3 (sep 2003), 290–294.

[30] Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, Effective, and Self-Supervised: Transforming Masked Language Models into Universal Lexical and Sentence Encoders. In *Proc. of 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1442–1459.

[31] P. Lopez. 2008–2024. Grobid. https://github.com/kermitt2/grobid. Last accessed on July 25th, 2024.

[32] Meiyu Lu, Srinivas Bangalore, Graham Cormode, Marios Hadjieleftheriou, and Divesh Srivastava. 2012. A Dataset Search Engine for the Research Document Corpus. In *2012 IEEE 28th International Conference on Data Engineering*. 1237–1240.

[33] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proc. of 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 3219–3232.

[34] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proc. of 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 3219–3232.

[35] Debanjan Mahata, John Kuriakose, Rajiv Ratn Shah, and Roger Zimmermann. 2018. Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings. In *Proc. of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 634–639.

[36] Roberto Martínez-Cruz, Debanjan Mahata, Alvaro J. López-López, and José Portela. 2023. Enhancing Keyphrase Extraction from Long Scientific Documents using Graph Embeddings. *CoRR* abs/2305.09316 (2023).

[37] Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. Where's the Point? Self-Supervised Multilingual Punctuation-Agnostic Sentence Segmentation. In *Proc. of 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 7215–7235.

[38] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticæ Investigationes* 30, 1 (2007), 3–26.

[39] Gabriela F. Nane, Nicolás Robinson-García, François van Schalkwyk, and Daniel Torres-Salinas. 2023. COVID-19 and the scientific publishing system: growth, open access and scientific fields. *Scientometrics* 128, 1 (2023), 345–362.

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

[41] Parth A. Patel and Mohammad Javed Ali. 2022. Characterizing Innovation in Science through the Disruption Index. *Seminars in Ophthalmology* 37, 6 (2022), 790–791.

[42] Animesh Prasad, Chenglei Si, and Min-Yen Kan. 2019. Dataset Mention Extraction and Classification. In *Proc. of Workshop on Extracting Structured Knowledge from Scientific Publications*. Association for Computational Linguistics, 31–36.

[43] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 3982–3992.

[44] Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic Sentence Boundary Disambiguation. In *Proc. of Second Workshop for NLP Open Source Software (NLP-OSS)*. Association for Computational Linguistics, 110–114.

[45] Dhruva Sahrawat, Debanjan Mahata, Mayank Kulkarni, Haimin Zhang, Rakesh Gosangi, Amanda Stent, Agniv Sharma, Yaman Kumar Singla, Rajiv Ratn Shah, and Roger Zimmermann. 2019. Keyphrase Extraction from Scholarly Articles as Sequence Labeling using Contextualized Embeddings. *ArXiv* abs/1910.08840 (2019).

[46] T. Saier, Y. Dong, and M. Farber. 2023. CoCon: A Data Set on Combined Contextualized Research Artifact Use. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE Computer Society, 47–50.

[47] Ivan Sekulic, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. 2024. Reliable LLM-based User Simulator for Task-Oriented Dialogue Systems. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*. Association for Computational Linguistics, St. Julians, Malta, 19–35.

[48] Ayush Singhal and Jaideep Srivastava. 2013. Data Extract: Mining Context from the Web for Dataset Extraction. *International Journal of Machine Learning and Computing* (2013), 219–223.

[49] Petros Stavropoulos, Ioannis Lyris, Natalia Manola, Ioanna Grypari, and Haris Papageorgiou. 2023. Empowering Knowledge Discovery from Scientific Literature: A novel approach to Research Artifact Analysis. In *Proc. of 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*. Association for Computational Linguistics, 37–53.

[50] Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Nelleke Oostdijk, Tommaso Caselli, Tadashi Nomoto, Onur Uca, Farhana Ferdousi Liza, and See-Kiong Ng. 2023. RECESS: Resource for Extracting Cause, Effect, and Signal Spans. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Nusa Dua, Bali, 66–82.

[51] Rachel Wicks and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In *Proc. of 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 3995–4007.

[52] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large Language Models for Generative Information Extraction: A Survey. arXiv:2312.17617

[53] Yang Yang, Zhilei Wu, Yuexiang Yang, Shuangshuang Lian, Fengjie Guo, and Zhiwei Wang. 2022. A Survey of Information Extraction Based on Deep Learning. *Applied Sciences* 12, 19 (2022).

[54] Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023. Complementary Explanations for Effective In-Context Learning. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 4469–4484.

[55] Hokeun Yoon and JinYeong Bak. 2023. Diversity Enhanced Narrative Question Generation for Storybooks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 465–482.

[56] Yousef Younes and Ansgar Scherp. 2023. Question Answering Versus Named Entity Recognition for Extracting Unknown Datasets. *IEEE Access* 11 (2023), 92775–92787.

[57] Tong Zeng and Daniel E. Acuna. 2020. *Finding datasets in publications: the Syracuse University approach.* Zenodo.

[58] Chengzhi Zhang, Lei Zhao, Mengyuan Zhao, and Yingyi Zhang. 2022. Enhancing keyphrase extraction from academic articles with their reference information. *Scientometrics* 127, 2 (2022), 703–731.

[59] Shunli Zhang, Yancui Li, Shiyong Li, and Fang Yan. 2022. Bi-LSTM-CRF Network for Clinical Event Extraction With Medical Knowledge Features. *IEEE Access* 10 (2022), 110100–110109.

[60] Shaowen Zhou, Bowen Yu, Aixin Sun, Cheng Long, Jingyang Li, and Jian Sun. 2022. A Survey on Neural Open Information Extraction: Current Status and Future Directions. In *Proc. of Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. International Joint Conferences on Artificial Intelligence Organization, 5694–5701.