Data Gatherer: LLM-Powered Dataset Reference Extraction from Scientific Literature

Pietro Marini[1], **Aécio Santos**[1], **Nicole Contaxis**[2], **and Juliana Freire**[1,3]
[1]Tandon School of Engineering
[2]Grossman School of Medicine
[3]Center for Data Science
New York University

Abstract

Despite the advantages of data sharing and the increasing availability of open datasets, finding datasets for reuse and identifying their references in research papers can still be challenging and time-consuming. To address these issues, we propose an LLM-powered system that automates the identification and extraction of ~~structured~~ dataset records from scientific publications. Additionally, we have curated two new high-quality datasets to evaluate dataset extraction methods. Our experimental evaluation using these datasets indicates that our system can achieve high precision and recall in the dataset reference extraction task.

1 Introduction

The increasing availability of data has accelerated scientific progress. Genomic and proteomic data sharing, for example, has enabled scientists to develop approaches that rely on access to large amounts of data (JB et al., 2020). Policies and frameworks like the FAIR Principles (Wilkinson et al.) and FORCE11's Joint Declaration of Data Citation Principles (Altman et al.) and changing researchers practices have contributed to increasing the amount of data available. Yet, finding datasets for re-use and identifying datasets referenced in research papers remain a challenging and labor-intensive task (Tsueng et al.; Griffiths et al.).

In contrast to journal article and book citation practices that use standardized formats (e.g., citation styles, DOIs), dataset references are inconsistent, ambiguous, and dispersed throughout scholarly documents, making systematic discovery difficult. PubMed and PubMed Central, for example, make some dataset mentions available through LinkOut Resources which links to external resources. They also allow researchers to search for articles that include Data Availability Statements (DASs), structured sections of articles that describe datasets used. However, these indexes are not currently able to surface dataset mentions fully, especially those embedded in the article text.

Even when datasets are explicitly referenced, their mentions are often ambiguous. The same dataset may be cited under different names, abbreviations, or project titles across multiple papers. Some papers provide only partial accession codes or omit repository information, making it difficult to resolve the dataset's location. DAS's, for example, may erroneously state that all data from a study is included in the paper (Federer et al., 2018). Common issues like typos, incorrect identifiers, and broken links further hinder discovery.

To locate datasets included in papers, researchers, librarians and data curators then have to undertake the labor-intensive process of manually searching, cross-referencing, and verifying dataset mentions. Mentions may include metadata such as accession codes, repository names, URLs, or informal descriptions. They can be embedded in figure captions, tables, supplementary materials, citations, or structured article sections like a DAS rather than explicitly listed in the main text.

Recent advancements in Large Language Models (LLMs) present an opportunity for automatic discovery and extraction of dataset mentions. LLMs excel at recognizing patterns in natural language, allowing them to identify dataset references in non-standard formats and differentiate them from similar-looking text, like gene names or experiment IDs, thus enhancing extraction precision.

Contributions. In this paper, we introduce Data Gatherer[1], an LLM-powered system that automates the identification and extraction of structured dataset records from scientific publications. We aim to ease the labor-intensive process typically undertaken by researchers and librarians. The development of our tool initially focused on pro-

[1]https://github.com/VIDA-NYU/data-gatherer

teomics and genomics data, as our collaborators have subject expertise in those areas. Leveraging this expertise, we develop two benchmark datasets to evaluate the extraction quality of our tool. One dataset is a small, high-quality collection curated by an expert librarian, while the other is a larger dataset created by automatically merging existing databases that contain datasets associated with research article references. Our experimental evaluation shows that our methods can achieve a recall of up to 0.99 and precision of up to 0.87.

In summary, our main contributions are: (1) an LLM-based pipeline for identifying and extracting dataset references from scholarly documents (Section 5); (2) curation of two new datasets for evaluation of dataset extraction methods (Section 4); and (3) an experimental evaluation of our data extraction methods using different LLMs (Section 6).

2 Related Work

We review previous work related to the extraction of dataset mentions from scientific literature and categorize them into the following two groups.

Scientific Literature Related Datasets. Several datasets have been developed to facilitate research in scientific information extraction. Anzaroot and McCallum (2013) introduced a dataset for fine-grained citation field extraction, focusing on segmenting citation strings into components like title and authors. Cheung et al. (2024) presented PolyIE, a dataset for extracting entities and relations specific to polymer materials. Zhang et al. (2024) developed SciER, a dataset for entity and relation extraction focusing on datasets, methods, and tasks. While these datasets facilitate various aspects of scientific information extraction, including citation parsing, domain-specific entity extraction, and general dataset-related information extraction, the datasets produced in this paper are focused on extraction of dataset references from the scientific literature on proteomics and genomics.

The Data Citation Corpus, created by the organization Make Data Count in collaboration with the Chan Zuckerberg Initiative, is a comprehensive list of data citations from articles and preprints meant to facilitate the creation and use of data metrics similar to bibliometrics used to measure the impact of other scholarly outputs (e.g., H-index, Impact Factor, and the RCR) (Make Data Count, 2025). The Data Citation Corpus is in part compiled using machine learning methods that leverage SciBERT-based Named Entity Recognition (Istrate, 2023). Make Data Count does not make these data citation location tools publicly available. In contrast, our tool is provided as an open-source data discovery tool that is freely accessible to researchers. It helps users locate all mentions of datasets within a collection of articles relevant to their work, facilitating the discovery of relevant datasets rather than focusing on the creation of data metrics.

Dataset Discovery and Citation Analysis. Early approaches to dataset mention extraction relied on statistical methods. Loizides and Schmidt (2016) present a semi-automatic approach using a dictionary and similarity measures to identify and link the identified references to a dataset registry. Zeng and Acuna (2020) propose using a bidirectional LSTM with a CRF inference mechanism, to detect dataset mentions. Kumar et al. (2021) propose DataQuest, a BERT-based entity recognition model with POS-aware embeddings, utilizing a two-stage pipeline for dataset sentence classification and mention extraction. These methods often rely on domain knowledge and small models trained on limited data, which restricts their adaptability to new domains. We aim to address these limitations by using the recent information extraction advancements enabled by LLMs trained on large corpora.

3 Problem Definition

We aim to automatically discover and extract dataset references from scholarly publications, focusing on citations accessible in academic documents available on the Web.

Definition 1. Given a publication identifier P (e.g., a URL or DOI that refers to a scholarly article), the goal is to build a function $\mathcal{F}$ that extracts a structured set of records $\{(d_i, r_i)\}$ from P, i.e., $\mathcal{F}(P) = \{(d_1, r_1), (d_2, r_2), ..., (d_n, r_n)\}$, where d_i is the dataset identifier, typically an accession code or another type of dataset reference, and r_i is the repository name or reference (e.g., a plain text string or a URL pointing to the repository). □

We consider a dataset reference valid if its identifier d_i exists in the repository r_i. To evaluate the ability of different approaches to identify and extract valid dataset references correctly, we built two benchmark datasets that are detailed in Section 4.

4 The `DataRef` Benchmarks

To evaluate `Data Gatherer`, we constructed two datasets using distinct methodologies: (1) `DataRef-EXP` is a manually curated by an expert librarian who identified and reviewed

publication webpages on PubMed Central, selecting articles to ensure a diverse representation of dataset citation formats; (2) DataRef-REV was built by combining metadata from two online resources: ProteomeCentral,[2] a portal that aggregates dataset information from repositories within the ProteomeXchange consortium (Deutsch et al., 2023) and the Gene Expression Omnibus (GEO) repository.[3] Below we detail the data curation approach for each of these datasets.

4.1 DataRef-EXP Dataset

The DataRef-EXP dataset was created by manually selecting and reviewing scholarly journal articles to ensure a diverse representation of dataset citation formats. Data were exclusively sourced from PubMed Central (PMC)[4]. While additional publication indexes exist, PMC was chosen as the source for two reasons. First, it provides open access to the full text of articles via an API, side-stepping possible issues with copyright and systematic downloading of journal articles. Second, it allows to filter searches to articles that contain associated data references, usually in their *Data Availability Statements* (DAS). A DAS is one type of dataset mention. It is a part of a journal article that outlines the datasets used in the paper and how to access them. Although many DAS's are incomplete or inaccurate, their availability eased the manual process of generating this DataRef-EXP dataset.

A total of 22 journal articles were chosen, resulting in 50 dataset references. Journal articles were chosen in order to maximize the variation in how included datasets were referenced, enabling a comprehensive evaluation of the Data Gatherer tool's ability to extract dataset mentions across various formats. For example, some journal articles were chosen where all dataset mentions were included in the DAS while other journal articles included dataset mentions in figures or within the full text. Additionally, some articles were chosen due to errors in dataset mentions, like inaccurate accession numbers or incomplete dataset information (e.g., an accession number but no named repository).

4.2 DataRef-REV Dataset

The second dataset was constructed using a reverse-engineering methodology, leveraging structured metadata from ProteomeCentral and Gene Expression Omnibus (GEO). ProteomeCentral is a valuable source for ground truth data, offering curated metadata for over 40,000 publicly available datasets, including dataset identifiers and related paper DOIs. It aggregates datasets from various repositories and links them to citing publications, making it a great starting point for locating papers that contain dataset references. Similarly, GEO is a public functional genomics data repository managed by the National Center for Biotechnology Information (NCBI). GEO provides programmatic access through a REST API that allows us to retrieve dataset identifiers along with references to publications that mention them.

A limitation of this dataset is that it only contains references to datasets deposited in repositories that are part of the ProteomeXchange consortium or in the GEO repository – it is possible that there may be other datasets mentioned in the paper that are not deposited in these repositories. However, a significant advantage is that it is automatically generated, allowing us to obtain a much larger number of dataset references compared to the manually created DataRef-EXP dataset.

Dataset Construction Details. Each dataset entry includes a unique identifier, typically an accession code, along with the corresponding repository name, such as PRIDE, MassIVE, jPOST, PeptideAtlas, or PASSEL. Additionally, the metadata contains information about citing publications, including their DOI or PubMed Central ID (PMCID) when available, as well as the title and keywords associated with the dataset. To ensure high-quality metadata, entries lacking a DOI or publication link were discarded, guaranteeing that each dataset-reference pair has an associated paper reference.

To supplement the structured metadata, we implemented an automated data-fetching pipeline to retrieve full-text HTML versions of citing publications. Using Selenium, we systematically accessed publisher websites and extracted the HTML source of each article whenever it was available. By integrating full-text data with structured repository metadata, we ensure that our dataset reflects both formally registered dataset citations and real-world citation practices in scholarly writing.

5 The Data Gatherer Tool

Data Gatherer was designed to automatically extract dataset references from scientific publications in HTML format as discussed in Section 3. It employs LLMs to identify and reconstruct dataset information. The two main strategies for the fun-

[2] https://proteomecentral.proteomexchange.org/
[3] https://www.ncbi.nlm.nih.gov/geo/
[4] https://pmc.ncbi.nlm.nih.gov/

damental functionality are Full-Document Read (FDR) and Retrieve-Then-Read (RTR).

5.1 Retrieve-Then-Read (RTR)

Juliana: we should discuss why we use this strategy The RTR method is a two-step process that leverages the structural elements of scientific articles webpages: (1) first it locates specific target sections of the papers where dataset mentions are likely to appear, such as the data availability statements (DAS); and then (2) it collects textual content from the target sections and feeds them to an LLM using a few-shot prompt to extract dataset references (we provide prompts in Appendix A.1). Given that this method leverages the structure of PubMed Central documents, it is currently only applicable to the open-access papers available at the PMC digital repository. However, this approach can be extended to other repositories by adapting the target sections.

Rule-Based Section Retrieval. We employ a rule-based retrieval system to identify HTML document sections likely containing dataset references. This system uses a combination of CSS selectors and XPath expressions, tailored to the specific structure of various publisher websites. Our retrieval rules are in a JSON configuration for easy modification and extension. It includes general rules for all websites and domain-specific rules that depend on the publisher's webpage structure.

LLM-Based Dataset Extraction. Following section retrieval, we applied LLM-based extraction. We instructed the LLM to output dataset references in JSON format using a structured few-shot prompting approach. Multiple prompt variations were tested and refined to improve extraction precision.

5.2 Full-Document Read (FDR)

To avoid the costs associated with manually defining rules for locating target sections, we consider an alternative approach that utilizes an LLM-based extraction pipeline to process the entire document. Instead of processing only specific sections of the article, we use the entire document text. While this method is more adaptable to various publishers, it has some drawbacks. Specifically, it only works with LLMs that support relatively long context windows and requires them to handle a significantly larger input, which increases costs.

HTML Preprocessing & Filtering. Before passing documents to the LLM, we perform an HTML normalization step to remove non-informative elements, such as scripts, styles, images, iframes, buttons, and metadata tags. This preprocessing ensures that only relevant text-based content is considered, reducing noise and improving dataset extraction accuracy and costs.

Handling Long HTML Documents. We use only LLMs that support long-context windows, with GPT-4 (128K tokens) being the model with the smallest context limit. In cases where documents exceed this limit, the content is truncated until it fits the model context size constraints.

Dataset	Model	Method	Precision	Recall
DataRef-EXP	gpt-4o-mini	FDR	0.835	0.840
		RTR	0.911	0.905
	gemini-2.0-flash	FDR	0.770	0.852
		RTR	0.905*	0.641*
DataRef-REV	gpt-4o-mini	FDR	0.864	0.985
		RTR	0.898	0.622
	gemini-2.0-flash	FDR	0.776	0.995
		RTR	0.913*	0.427*

Table 1: Comparison of different LLMs, and methods (FDR, RTR) on DataRef-EXP vs DataRef-REV.

6 Experimental Evaluation

We evaluated Data Gatherer's performance on two datasets, DataRef-EXP and DataRef-REV (described in Section 4). To evaluate extraction quality, we use *precision* and *recall* metrics, which assess the accuracy of extracted references and the ability to identify all references in an article, respectively.

We report the results in Table 1, which include a comparison of different LLMs and extraction approaches in the two datasets. Comparing the LLMs on the DataRef-EXP dataset, gpt-4o-mini achieves higher precision than gemini-2.0-flash in both methods. Conversely, gemini-2.0-flash achieves higher recall, a trend that also happens for DataRef-REV. We also note that recall on DataRef-EXP is generally lower, which is expected since the dataset was designed to include a high variety of difficult cases. Finally, the RTR method seems to help improve gpt-4o-mini's precision. This suggests that reducing the input size helps models that don't scale well for long inputs, at the cost of decreasing the recall in some cases.

7 Conclusion

Researchers, librarians, and data curators currently spend significant amounts of time locating dataset mentions in scholarly papers. They perform this work both to locate datasets for secondary analysis projects and also to ensure a paper's conclusions are well-supported by the data. To ease this

time-intensive and difficult task, we designed Data Gatherer to automatically find and parse dataset mentions in articles. As new methodologies in the sciences increasingly rely on access to large amounts of open data this tool can have a notable impact on the way that researchers, data curators, and librarians find, review, and aggregate data to meet the promise of these new methods.

Limitations

As in any LLM-powered system, our work has several limitations. In particular, the retrieve-then-read (RTR) approach depends on the PubMed Central (PMC) document structure, so it requires additional effort to be extended to other repositories. The full-document read (FDR) approach aims to resolve this limitation by processing the full document, however, this limits the number of LLMs that can be used and may increase processing costs. Regardless of the strategy, the system can miss dataset references or output incorrect references. It also relies on LLM capabilities, which can be limited in ambiguous contexts. Additionally, our evaluation datasets, `DataRef-EXP` and `DataRef-REV`, may not fully represent all dataset citation practices since their size is limited and mainly cover papers related to proteomics and genomics research fields.

References

Micah Altman, Christine Borgman, Mercè Crosas, and Maryann Matone. An introduction to the joint principles for data citation. 41(3):43–45.

Sam Anzaroot and Andrew McCallum. 2013. A new dataset for fine-grained citation field extraction.

Jerry Cheung, Yuchen Zhuang, Yinghao Li, Pranav Shetty, Wantian Zhao, Sanjeev Grampurohit, Rampi Ramprasad, and Chao Zhang. 2024. Polyie: A dataset of information extraction from polymer material scientific literature. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2370–2385.

Eric W Deutsch, Nuno Bandeira, Yasset Perez-Riverol, Vagisha Sharma, Jeremy J Carver, Luis Mendoza, Deepti J Kundu, Shengbo Wang, Chakradhar Bandla, Selvakumar Kamatchinathan, and 1 others. 2023. The proteomexchange consortium at 10 years: 2023 update. *Nucleic Acids Research*. PMID: 36370099, DOI: `https://doi.org/10.1093/nar/gkac1040`.

Lisa M. Federer, Christopher W. Belter, Douglas J. Joubert, Alicia Livinski, Ya-Ling Lu, Lissa N. Snyders, and Holly Thompson. 2018. Data sharing in PLOS ONE: An analysis of data availability statements. 13(5):e0194768.

Emily Griffiths, Rebecca M Joseph, George Tilston, Sarah Thew, Zoher Kapacee, William Dixon, and Niels Peek. Findability of UK health datasets available for research: a mixed methods study. 29(1):e100325.

Ana-Maria Istrate. 2023. Building the open global data citation corpus – chan zuckerberg initiative. Publisher: Zenodo Version Number: 1.0.

Byrd JB, Greene AC, Prasad DV, Jiang X, and Greene CS. 2020. Responsible, practical, genomic data sharing that accelerates research. *Nature Reviews Genetics*.

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: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings 23*, pages 43–53. Springer.

F Loizides and B Schmidt. 2016. Identifying and improving dataset references in social sciences full texts. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, page 105.

Make Data Count. 2025. Open data metrics require open infrastructure: Data citation corpus.

Ginger Tsueng, Marco A. Alvarado Cano, José Bento, Candice Czech, Mengjia Kang, Lars Pache, Luke V. Rasmussen, Tor C. Savidge, Justin Starren, Qinglong Wu, Jiwen Xin, Michael R. Yeaman, Xinghua Zhou, Andrew I. Su, Chunlei Wu, Liliana Brown, Reed S. Shabman, Laura D. Hughes, the NIAID Systems Biology Data Dissemination Working Group, and Serdar Turkarslan. Developing a standardized but extendable framework to increase the findability of infectious disease datasets. 10(1):99.

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, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, and 34 others. The FAIR guiding principles for scientific data management and stewardship. 3(1):160018.

Tong Zeng and Daniel Acuna. 2020. Finding datasets in publications: the syracuse university approach. In *Rich Search and Discovery for Research Datasets*, pages 158–165. SAGE.

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Latecki, and Eduard Dragut. 2024. Scier: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13083–13100.

A LLM Prompts

A.1 Retrieve-Then-Read (RTR) Prompts

```
[
    {
        "role": "system",
        "content": "You are a specialized assistant that extracts dataset references from the content of scientific papers. You must output a JSON array of objects, where each object has the following keys: 'dataset_identifier', 'data_repository', and 'dataset_webpage'. Follow the structure of the provided examples exactly."
    },
    {
        "role": "user",
        "content": "Extract dataset references based on the examples below:\n\nExample 1:\nContent: 'The study used dataset EGAS00001000925, which is available at the European Genome Archive.'\nResponse:\n[\n    {\n        \"dataset_identifier\": \"EGAS00893672193\",\n        \"data_repository\": \"European Genome Archive\",\n        \"dataset_webpage\": \"https://ega-archive.org/studies/EGAS00001000925\"\n    }\n]\n\nExample 2:\nContent: 'Proteomics data was obtained from PRIDE, accession PXD029821.'\nResponse:\n[\n    {\n        \"dataset_identifier\": \"PXD029821\",\n        \"data_repository\": \"PRIDE\",\n        \"dataset_webpage\": \"https://www.ebi.ac.uk/pride/archive/projects/PXD029821\"\n    }\n]\n\nExample 3:\nContent: 'The repository dbGaP hosts the dataset phs001366.v1.p1 at this location.'\nResponse:\n[\n    {\n        \"dataset_identifier\": \"phs001366.v1.p1\",\n        \"data_repository\": \"dbGaP\",\n        \"dataset_webpage\": \"https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001366.v1.p1\"\n    }\n]\n\nNow process the following content:\n\nContent: {content}"
    }
]
```

Listing 1: Prompt to extract dataset references from small HTML elements.

A.2 Full-Document Read (FDR) Prompts

```
[
  {
    "role": "model",
    "parts": [
      {
        "text": "I am a large language model trained to be informative and comprehensive. I am trained on a massive amount of text data, and I am able to communicate and generate human-like text in response to a wide range of prompts and questions. For this task, I will act as a specialized assistant that can identify datasets mentioned in a publication and create a summary suitable for non-specialists.\n\nThe output should be a JSON array of objects, where each object has the following keys:\n- \"dataset_identifier\": This is any alphanumeric string (maybe including punctuation marks) that uniquely identifies or provides access to a dataset.\n- \"repository_reference\": This is the URL or reference to the data repository where the dataset can be found.\n\nHere are some examples for reference:\n\n['dataset_identifier'=> 'EGAS00001000925', 'repository_reference'=> 'https=>//ega-archive.org/datasets/EGAS00001000925',\n 'dataset_identifier'=> 'GSE69091', 'repository_reference'=> 'https=>//www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69091',\n 'dataset_identifier'=> 'PRJNA306801', 'repository_reference'=> 'https=>//www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA306801',\n 'dataset_identifier'=> 'phs003416.v1.p1', 'repository_reference'=> 'dbGaP',\n 'dataset_identifier'=> 'PXD049309', 'repository_reference'=> 'https=>//www.ebi.ac.uk/pride/archive/projects/PXD049309',\n 'dataset_identifier'=> 'IPX0004230000', 'repository_reference'=> 'http=>//www.iprox.org',\n 'dataset_identifier'=> 'MSV000092944', 'repository_reference'=> 'https=>//massive.ucsd.edu/',\n 'dataset_identifier'=> 'n/a', 'repository_reference'=> 'https://data.broadinstitute.org/ccle_legacy_data/mRNA_expression/'\n]"
      }
    ]
  },
  {
    "role": "user",
    "parts": [
      {
        "text": "Given the information that I am going to share:\n 1) the webpage in HTML format that you have to extract datasets information from.\n 2) a sample of already known data repositories.\n\nPlease return a JSON array of objects where each object has the following structure:\n- `dataset_identifier`: The dataset identifier (a code). If not found, set it to \"n/a\".\n- `repository_reference`: The URL or reference to the data repository. If not found, set it to \"n/a\".\n\
```

```
    nPlease follow these strict
    instructions:\n- The output must be
    a valid JSON array of objects.\n-
    Each object must contain the keys `
    dataset_identifier` and `
    repository_reference`.\n- Any other
    output format will be considered
    invalid.\b\bBelow is the input data
    that you will use to generate the
    output:\n1) html => {content}\n2)
    repos => {repos}."
      }
    ]
  }
]
```

Listing 2: Prompt for Gemini to extract dataset references from full documents normalized.

```
[
    {
        "role": "system",
        "content": "I am a large
    language model trained to be
    informative and comprehensive. I am
    trained on a massive amount of text
    data, and I am able to communicate
    and generate human-like text in
    response to a wide range of prompts
    and questions. For this task, I will
     act as a specialized assistant that
     can identify datasets mentioned in
    a publication and create a summary
    suitable for non-specialists.\n\nThe
     output should be a JSON array of
    objects, where each object has the
    following keys:\n- \"
    dataset_identifier\": This is any
    alphanumeric string (maybe including
     punctuation marks) that uniquely
    identifies or provides access to a
    dataset.\n- \"repository_reference
    \": This is the URL or reference to
    the data repository where the
    dataset can be found.\n\nHere are
    some examples for reference:\n\n['
    dataset_identifier'=> '
    EGAS00001000925', '
    repository_reference'=> 'https://
    ega-archive.org/datasets/
    EGAS00001000925',\n '
    dataset_identifier'=> 'GSE69091', '
    repository_reference'=> 'Gene
    Expression Omnibus (GEO)',\n '
    dataset_identifier'=> 'PRJNA306801',
     'repository_reference'=> 'https://
    www.ncbi.nlm.nih.gov/bioproject/?
    term=PRJNA306801',\n '
    dataset_identifier'=> 'phs003416.v1.
    p1', 'repository_reference'=> 'dbGaP
    ',\n 'dataset_identifier'=> '
    PXD049309', 'repository_reference'=>
     'https://www.ebi.ac.uk/pride/
    archive/projects/PXD049309',\n '
    dataset_identifier'=> 'IPX0004230000
    ', 'repository_reference'=> 'http
    ://www.iprox.org',\n '
    dataset_identifier'=> 'MSV000092944
    ', 'repository_reference'=> 'https
    ://massive.ucsd.edu/',\n '
    dataset_identifier'=> 'n/a', '
    repository_reference'=> 'https://
    data.broadinstitute.org/
    ccle_legacy_data/mRNA_expression/'\n
    ]"
    },
    {
        "role": "user",
        "content": "I have a webpage in
    HTML format ({content}) and a list
    of known data repositories ({repos})
    . Please return a JSON array of
    objects, where each object has the
    structure:\n- `dataset_id`: The
    dataset identifier (a code). If not
    found, set it to 'n/a'.\n- `
    repository_reference`: The URL or
    reference to the data repository. If
     not found, set it to 'n/a'.\nEnsure
     the output is a plain JSON array,
    not nested inside another structure,
     and not an Unterminated string."
    }
]
```

Listing 3: Prompt for gpt-4o-mini to extract dataset references from full documents normalized.