Data Gatherer Revisited: Scalable Dataset Reference Extraction from Biomedical Literature

Pietro Marini
Datatecnica
pietro@datatecnica.com

Mike A. Nalls
Datatecnica
mike@datatecnica.com

Mette Peters
Datatecnica
mette@datatecnica.com

Nicole Contaxis
NYU Langone
nicole.contaxis@nyulangone.org

Aécio Santos
Centrum Wiskunde & Informatica
asrs@cwi.nl

Juliana Freire
New York University
juliana.freire@nyu.edu

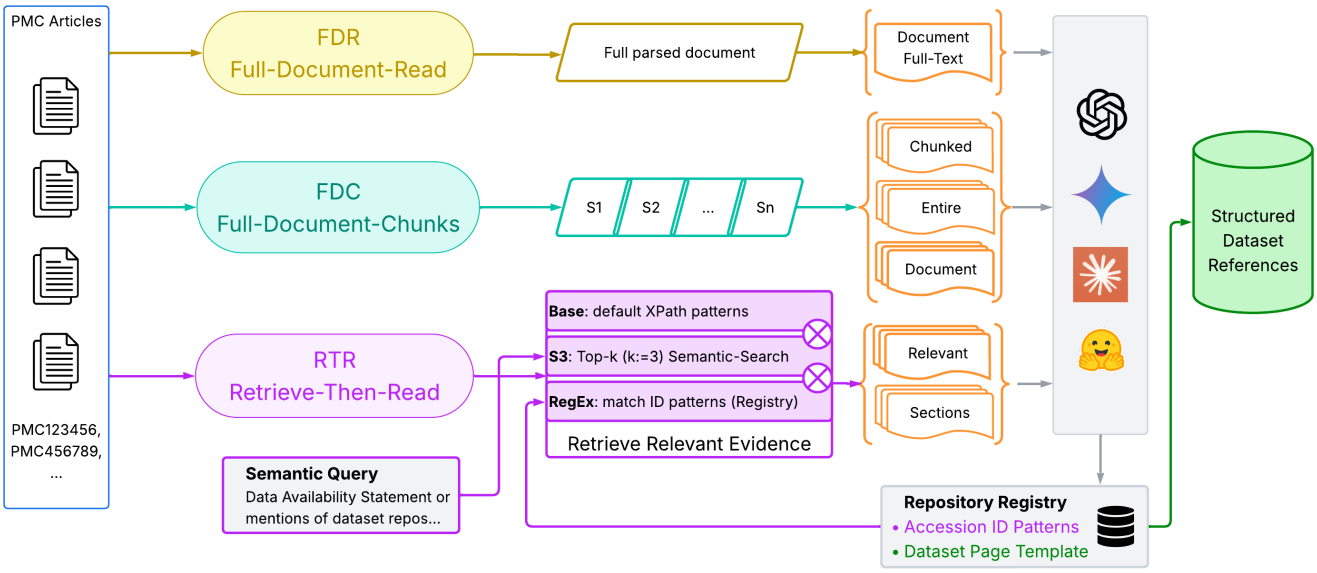

Figure 1: Data Gatherer pipeline

ABSTRACT

Scientific papers often refer to datasets through accession codes, repository links, or informal descriptions scattered across the document. Identifying these references at scale is important for reproducibility and data discovery, but sending every full article to a commercial large language model can be expensive.

Data Gatherer addresses this problem by combining evidence retrieval with language model extraction: it first selects the passages most likely to contain dataset references and then applies a language model only to those passages. Our earlier version reduced processing costs but sometimes missed relevant evidence. In this work, we improve evidence retrieval by combining semantic ranking and repository-specific identifier pattern matching. Together, these additions recover most of the recall lost by the earlier retrieve-then-read pipeline while avoiding full-document processing. We also fine-tune Flan-T5-base to perform extraction locally without relying on commercial model APIs.

On a held-out set of 249 biomedical articles, the enhanced pipeline increases recall from 0.54 to 0.98 against a reverse-engineered gold standard and from 0.31 to 0.66 against a broader automatically derived benchmark. The fine-tuned model achieves recall close to that of proprietary models while reducing estimated inference cost by 330× to 1,300×, although with lower precision.

We release the model weights, training data, repository registry, and fine-tuning and evaluation code.

VLDB Workshop Reference Format:
Pietro Marini, Mike A. Nalls, Mette Peters, Nicole Contaxis, Aécio Santos, and Juliana Freire. Data Gatherer Revisited: Scalable Dataset Reference Extraction from Biomedical Literature. VLDB 2026 Workshop: Biomedical Data Management Systems (BioDMS).

VLDB Workshop Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/VIDA-NYU/data-gatherer.

Proceedings of the VLDB Endowment. ISSN 2150-8097.

1 INTRODUCTION

Standardized dataset deposition is not a new concern: as early as 1988, Nucleic Acids Research[1] required authors to submit sequence data to EMBL and cite the resulting accession number [13]. Nearly four decades later, however, dataset references remain inconsistently reported across disciplines and publishers, making programmatic dataset mention extraction difficult at scale.

In our previous work, we introduced Data Gatherer [20], an open-source tool that leverages LLMs to identify cited datasets. The system supported two strategies. Full-Document Read (FDR) sends the entire article to a large-context language model. Retrieve-then-Read (RTR) first selects sections that are likely to contain dataset references and sends only those sections to the model. The second strategy substantially reduces token usage, but its effectiveness depends on the quality of retrieval.

In our previous evaluation, RTR matched full-document processing on curated articles but lost recall on broader benchmarks because the retrieval rules did not cover the diversity of real-world citation practices [20]. This recall-cost tradeoff is a key obstacle to scaling for large corpora.

For example, Data Gatherer is used in production to populate the Datasets page of the CARD Catalog (https://alzheimersdatahub.org), automatically surfacing dataset references across the Alzheimer's disease literature for researchers and funders. Sustaining this kind of continuous, corpus-scale ingestion is exactly what motivates this paper's cost reductions.

This paper closes most of the recall gap between Retrieve-then-Read and full-document processing by combining higher-recall retrieval with a low-cost extraction model. We evaluate these enhancements on a held-out subset of DataRef-REV [21], which contains 1,242 PubMed Central articles. We use two complementary reference signals: curated links derived from GEO and ProteomeXchange records and broader accession annotations from SciLite [31].

Our contributions are:

(1) **Higher-recall retrieval for RTR.** We add semantic section ranking and regex anchoring against a repository registry to the retrieval step, substantially improving recall over the rule-based retrieval used in [20] without resorting to full-document processing.
(2) **A fine-tuned, low-cost extraction model.** We release a training dataset and fine-tune Flan-T5-base for dataset mention extraction, and show that its recall approaches that of commercial LLMs at substantially lower inference cost, though with lower precision.
(3) **Evaluation on an expanded benchmark.** We assess performance on a held-out split of 249 DataRef-REV articles, comparing against two reference signals that differ in coverage and precision: GEO- and ProteomeXchange-derived DataRef-REV identifiers, and the wider set of accession numbers drawn from SciLite [31] text-mining annotations.
(4) **Public release.** Model weights, training data, the repository registry, and fine-tuning and evaluation code are released as open source.

[1] https://academic.oup.com/nar

2 RELATED WORK

Scholarly Document Processing. Over the last three decades, several text-mining approaches have been developed for automated information extraction and entity-relationship mapping. Most of these methods targeted the biomedical literature [25] and biocuration workflows [10]. The earliest attempts revolved around exact string matching of keywords and metadata combined with Machine Learning models for Named Entity Recognition [15], whereas more recent work leverages transformer architectures [4, 30] and in-context learning. The better generalization capabilities of these methods allowed researchers to extend scientific information extraction beyond the biomedical domain into other fields. Zhang et al. [35] introduced SciER, a collection of entities (datasets, methods, and tasks) and relations extracted from a corpus of ML papers from Papers with Code[2]. Cheung et al. [2] presented PolyIE, targeting the same task applied to publications about polymer materials. Xu et al. [33] developed an LLM-driven paper-dataset networking system for publications on arXiv. Adjacent tasks face the same underlying obstacle: heterogeneous mention styles make software mention detection [5], the SOMD shared task [29], and fine-grained citation field extraction [1] — which segments citation strings into components such as title and authors — comparably difficult to standardize. Despite the abundance of resources supporting Scholarly Document Processing research, the resource closest to our evaluation setting is DataRef [21], which we extend and release with this paper for use in our experimental evaluation (§5).

Methods for dataset mention identification. The initial framing of this problem was binary classification of words, such that each token was labeled as a dataset, concatenated to its neighboring tokens with the same tag, and finally resolved to a URL contained in the text [17]. In this context, Ghavimi et al. [6] combine tf-idf cosine similarity scoring and dictionary-based matching to identify dataset mentions and link them to existing registries, Zeng and Acuna [34] use a bidirectional LSTM with a CRF inference layer for dataset mention detection, and Kumar et al. [14] propose a BERT-based recognizer with POS-aware embeddings and a two-stage pipeline for sentence classification followed by mention extraction. Closest in spirit to our system's rule-based component is the Accession Numbers Annotation pipeline [12], based on Whatizit [26], which uses contextual cues and accession-pattern matching to annotate Europe PMC articles; to our knowledge, it remains the most directly comparable prior extraction system, despite predating modern NLP techniques.

These methods share two limitations that motivate our approach. First, the requirement of domain-specific training data limits transferability to other disciplines' scientific documents. Second, rule- and pattern-based methods such as Whatizit/Accession Number Annotation pipeline generalize only as far as their hand-curated patterns cover[3] – exactly the recall ceiling we quantify for predefined-section and regex-only retrieval in §5 and address by adding semantic retrieval on top. Our retrieval step is, structurally, a domain-specific retrieval-augmented generation (RAG) pipeline [16]: chunking, embedding, and top-k similarity ranking, combined with regex

[2] https://paperswithcode.co/
[3] https://github.com/EuropePMC/EuropePMC-Identifier-Extractor/blob/master/automata/acc181210.mwt

anchoring against a repository identifier registry to recover matches that similarity ranking alone misses.

Corpus-scale dataset citation tracking. A separate line of work tracks dataset citations at the scale of bibliometric graphs rather than per-article extraction. Microsoft Academic Graph conformed dataset citations to "dataset paper" citations [32], a pattern inherited by Web of Science, Scopus, and OpenAlex [24], whose cross-system document-type inconsistencies [9, 18] further limit their use as a reverse-engineered gold standard. DataCite-based efforts [7, 28] and the resulting Data Citation Corpus [19] partially close this gap; the corpus is compiled in part using SciBERT-based named entity recognition [11], but the underlying extraction tooling is not made publicly available. In contrast, our system is open source.

Finally, we fine-tune a small model (§4) to reduce commercial-API inference cost. A more principled alternative is task-specific knowledge distillation [22], which trains directly against the teacher's soft output distribution; we leave this for future work.

3 DATA GATHERER PIPELINE

Figure 1 illustrates the pipeline's three processing paths — Retrieve-then-Read (RTR), Full-Document-Read (FDR), and Full-Document-Chunks (FDC) — and its four components: a Fetcher and Parser (§3.1), the Retrievers (§3.2), the LLM Client (§3.3), and the Repository Registry (§3.4). All paths share the Fetcher and Parser and diverge at retrieval: FDR passes the full parsed document to a large-context commercial API; FDC, used for limited-context models like the fine-tuned T5, passes every chunk without a retrieval filter; RTR selects a candidate pool of snippets as context for the LLM. All three paths output a normalized dataset-mention record per match, detailed in §3.3.

3.1 Fetcher and Parser

The Fetcher retrieves raw content for PMC article URLs via up to three strategies, tried in priority order: a bulk NCBI Entrez API call (structured XML), a lightweight HTTP GET, and a headless Selenium WebDriver. Entrez XML is tried first; if it fails quality checks (minimum section count, required section types), the fallbacks run in order. HTTP GET is new to this paper — our prior implementation was based on Selenium alone. Since Selenium web-scraping has grown unreliable as NCBI blocks automated browser traffic, we store a fixed Parquet snapshot of the articles used in our evaluation (§5) for stability.[4] The *Parser* converts raw XML or HTML strings into an element tree: XML documents are parsed with `lxml` under either the JATS or TEI schema and HTML documents are processed with `BeautifulSoup`.

3.2 Retrievers

Three retrieval methods contribute independently to the candidate pool. Our previous version [20] used only predefined patterns: a list of publisher- and layout-specific tags (e.g., selectors targeting "Data Availability" and related blocks), configured in an external patterns file[5], adding all matched sections verbatim. Here, we introduce two new retrieval strategies.

Semantic retrieval. Each section is split into chunks that fit within the token window of a sentence-transformer model [27] and gets embedded (default: `all-MiniLM-L6-v2`, 256-token context window). Chunk embeddings are cached on disk to avoid recomputation across runs. They are ranked based on their l2-distance with respect to a fixed query (Figure 3) and the top-k snippets (default $k = 3$) are added to the candidate pool. We set $k = 3$ because an average PMC article spans roughly 60K tokens (in HTML; fewer in XML) against a 256-token context window, and a single top-1 chunk frequently missed portions of Data Availability statements split across multiple chunks; $k = 3$ balances recall against injecting excess unrelated context into the LLM prompt.

Regex anchoring. The full text of the document is scanned with regular expressions compiled from the `id_pattern` fields of the repository registry (§3.4). Every section containing at least one match is added to the candidate pool, ensuring that accession IDs appearing in unexpected section types are not missed.

The union of the candidate sets produced by the three strategies is passed as context to the LLM extraction step.

3.3 LLM Client

The LLM Client is a model-agnostic interface that sends the extraction prompts assembled by the pipeline and parses the responses returned by the models. We support two types of backend. The first is commercial APIs (Anthropic, OpenAI, Gemini, and Portkey), for which provider-specific few-shot prompts[6] request structured JSON schema responses[7]. The second, introduced in this paper, is a small open model that we fine-tune specifically for this extraction task (Flan-T5-base; § 4), for which a plain-text prompt is rendered and passed to the model directly, rather than the few-shot, schema-constrained prompts used for the commercial backends. Regardless of backend, normalized responses are tuples (`dataset_identifier`, `data_repository`, `dataset_webpage`) and are deduplicated before output. To ensure the reproducibility of this stochastic inference step, we set the model temperature parameter to 0.0, use response schemas, and prompts with role definitions. Users may modify prompts and response schemas to extract additional or different information about datasets.

3.4 Repository Registry

The Repository Registry[8] is a curated JSON dictionary of biomedical data repositories, keyed by either an internet domain (e.g., massive.ucsd.edu, www.ebi.ac.uk) or a short alias (e.g., `geo`, `dbgap`, `pdb`). Each entry provides a human-readable `repo_name` and, when available, a `dataset_webpage_url_ptr` template with an `__ID__` placeholder, used post-hoc to construct and validate the dataset page link once an extracted repository reference resolves to a known entry — not every repository defines one. The `id_pattern` attribute, defined for most entries, is used by the RegEx retriever (§3.2) and validates extracted identifiers against the expected accession format. Other optional fields — `download_root`, `access_mode`,

[4] https://zenodo.org/records/21389083

[5] https://github.com/VIDA-NYU/data-gatherer/blob/main/data_gatherer/config/retrieval_patterns.json

[6] https://github.com/VIDA-NYU/data-gatherer/tree/main/data_gatherer/prompts/prompt_templates/dataset_prompts

[7] https://github.com/VIDA-NYU/data-gatherer/blob/main/data_gatherer/llm/response_schema.py

[8] https://github.com/VIDA-NYU/data-gatherer/blob/main/data_gatherer/config/open_bio_data_repos.json

and `javascript_load_required` — support optional downstream processing such as dataset-page content load. Cross-entry mappings (`repo_root`, `repo_mapping`) are used for repository resolution. The Repository Registry is a controlled vocabulary [8] rather than a formal ontology [23]: it supports pattern matching and URL templating, not class hierarchies or logical inference. It ships with the system and is updated manually.

4 LOW-COST, OPEN-SOURCE LLM INFERENCE

Commercial LLMs incur per-token costs that scale with corpus size, making continuous, large-scale extraction runs expensive (Table 2). As a low-cost alternative, we fine-tune Flan-T5-base [3], a small open model (250M parameters); batched inference on a single GPU (NVIDIA H100 NVL) avoids per-token API charges when run locally, making large-scale processing over PubMed Central practical. Each training example consists of an input snippet and a target JSON array of (`dataset_identifier`, `repository_reference`, `dataset_webpage`) records.

Training data. We release a snippet-level training dataset built from the same PMC article IDs as the 80% training split of DataRef-REV [21]. The target dataset references were annotated using GPT-4o-mini and subsequently reviewed by the authors before fine-tuning[9]. It is publicly available on Hugging Face[10].

Training procedure. Fine-tuning runs for 5 epochs on a single GPU in bf16 precision, using standard seq2seq decoding with post-hoc normalization and deduplication.[11] Inputs are truncated to 512 tokens with the prefix "Extract dataset information: "; outputs are capped at 256 tokens. Checkpoints are selected by exact-match accuracy on the held-out split rather than loss alone. We release the fine-tuned model on Hugging Face[12].

Grounding Check. During fine-tuning, we observed that some model outputs did not regex-match any text in the source snippet. To determine whether these outputs were unsupported by the source text, we manually reviewed all the flagged outputs and report a representative sample in Table 1. The examples show that non-matches arise from (i) identifiers that are implicitly specified via ranges or lists, (ii) formatting variants in the source text (e.g., whitespace/hyphenation), and (iii) genuinely unsupported outputs such as placeholders or non-dataset identifiers.

5 EXPERIMENTAL EVALUATION

5.1 Setup

Dataset. We evaluate on the held-out 20% split of DataRef-REV [21], comprising 249 PMC articles (`random_state=42`). The fine-tuning data use the corresponding 80% training-set PMC IDs, but with separately generated and reviewed snippet-level annotations.

Retrieval configurations. base—no semantic retrieval, no regex anchoring (replicates the prior system and results); **S3**—semantic retrieval (k=3), no regex anchoring; **RegEx**—regex flagging of section snippets; **RS3**—regex anchoring and semantic retrieval.

[9]https://huggingface.co/datasets/vida-nyu/pmc-articles-dataset-mentions-snippets/blob/main/Build_dedup_training_dataset.ipynb
[10]https://huggingface.co/datasets/vida-nyu/pmc-articles-dataset-mentions-snippets
[11]https://github.com/VIDA-NYU/data-gatherer/tree/main/scripts/Local_model_finetuning
[12]https://huggingface.co/vida-nyu/flan-t5-base-dataref-info-extract

Table 1: Representative flagged examples (predicted identifiers not matching any accession regex in the source snippet).

Predicted ID	Snippet evidence (abridged)
`GSM2816665`	Data Availability ... "... series `GSE105030` ... samples `GSM2816664 to GSM2816669`."
`GSE41306`	Data availability ... "... (`GSE 41306`) ... (`GSE 217222`) ... deposited ..."
`E-MTAB-8556`	Data availability ... "... accession number `E-MTAB- 8556`."
`GSExxxxxx`	Data and code availability ... "... deposited at `ncbi.nlm.nih.gov/geo/` ..."
`Yin et al. dataset`	Materials and methods ... "... data sets were downloaded ... the Yin et al ..."
`No. 2021YFA1301601`	Funding ... "... (`2021YFA1302604` and `2021YFA1301601`) ..."
`MsigDB v7.0`	Methods ... "... Molecular Signatures Database (`MsigDB`) `v7.0` ..."

Ground truths. SciLite [31]: automatically derived Europe PMC text-mining annotations [26], filtered to remove non-data subtypes (GO terms, RefSNP, etc.). **Gold GT**: 388 paper–dataset links, reverse-engineered from ProteomeXchange and GEO database entries cited across the 249 test articles.

Metrics. Macro-averaged precision (P), recall (R), and F1 per article [20]. For analysis of model reliability, we additionally use the anchoring audit (§4) to characterize non-grounded outputs.

5.2 Results

Table 4 reports precision, recall, and F1 for all model–configurations on the 249-article held-out test set. Figure 2 summarizes the recall on the Gold benchmark across retrieval configurations and models paired with cost, as shown in Table 2.

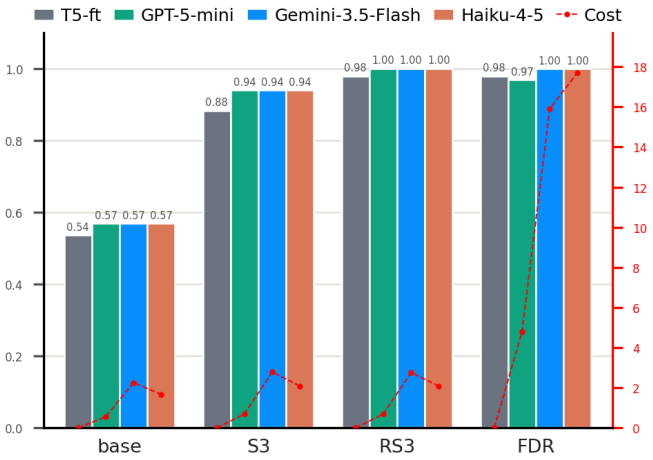

Figure 2: Cost (in US$) and Recall performance on the Gold benchmark, by model and pipeline configuration.

5.2.1 Can RTR be high-recall without full-document LLMs? Across all four models, the recall increases monotonically from base to S3 to RS3 on both benchmarks. For Flan-T5, Gold-GT recall rises from 0.537 (base) to 0.883 (S3) to 0.979 (RS3); SciLite recall follows

the same shape (0.313 → 0.577 → 0.658). Commercial models show the same pattern at a higher absolute level – for Gemini-3.5-flash, Gold-GT recall rises from 0.569 (base) to 0.939 (S3) to 1.000 (RS3) – and GPT-5-mini and Claude-Haiku-4.5 follow the same base < S3 < RS3 ordering on both benchmarks.

To isolate the contribution of regex anchoring from semantic retrieval, we also ran each model with regex flagging alone (RegEx), without semantic retrieval. RS3's recall matches or exceeds RegEx's for every model on both ground truths (Gold: 0.000–0.016 higher; SciLite: 0.013–0.058 higher; Table 5) – confirming that semantic retrieval was not recall-negative in any evaluated model–benchmark combination, and recovers additional true positives on SciLite's broader accession taxonomy. One possible explanation is that semantic retrieval captures mentions with formatting variants or without exact matches to any repository registry `id_pattern`, as illustrated by some cases in §4.

RS3's recall also compares favorably to the full-document configurations (FDR for the commercial models, FDC for Flan-T5), which read the entire article rather than a small set of retrieved snippets. On Gold GT, RS3 matches or exceeds FDR/FDC for every model. On SciLite GT, RS3 exceeds FDR for Claude-Haiku-4.5 (0.693 vs. 0.661) and Gemini-3.5-flash (0.680 vs. 0.641), and trails moderately for Flan-T5 (0.658 vs. 0.758) and GPT-5-mini (0.653 vs. 0.679).

These results confirm that the recall losses previously observed for RTR are primarily driven by missed evidence in retrieval, and that combining regex anchoring with semantic retrieval closes most of this gap – recovering recall comparable to reading the full document, at a fraction of the context cost, for most model/ground-truth combinations.

5.2.2 Can a fine-tuned small model compete with commercial APIs? At RS3, Flan-T5's recall is comparable to the three commercial models on both ground truths: Gold-GT recall of 0.979, versus 1.000 for GPT-5-mini, Claude-Haiku-4.5, and Gemini-3.5-flash; SciLite recall of 0.658, within the 0.653–0.693 band spanned by the three commercial models. Flan-T5 reaches this recall at a fraction of the cost: $0.0021 per 249-article run at RS3, versus $0.69 (GPT-5-mini), $2.09 (Claude-Haiku-4.5), and $2.76 (Gemini-3.5-flash) (Table 2) – a 330× to 1,300× cost reduction for comparable recall.

Precision is lower under RS3 than under the regex-only ablation (RegEx) for every model, which is expected: semantic retrieval is designed to surface snippets the regex anchor does not match – typically the harder, less formally stated citations – and some of that additional retrieved context can yield additional false positives. This effect is most visible on SciLite, whose broader accession taxonomy contains more such informally stated cases, and comparatively muted on Gold GT, whose 388 curated links are drawn from GEO and ProteomeXchange entries with well-defined accession formats and few hard cases. We therefore do not treat the RS3-vs-RegEx precision gap as a deficiency of semantic retrieval, but as the expected cost of the additional recall it is designed to recover.

Flan-T5 makes recall at this scale affordable, matching commercial-model recall within a few points on both ground truths.

5.3 Estimated Costs

Table 2 shows the costs of each model. The commercial API costs are based on the number of tokens (as presented in Table 3) multiplied by the corresponding provider rates. For the T5 model running locally, we used the electricity-only cost – estimated as the running time from the logs times the mean GPU energy consumption in Wh times $0.17/kWh – but the actual cost of renting it from a cloud provider for that time would lead to higher costs.

Table 2: Estimated inference cost (USD) per model and configuration, evaluated on the 249-article REV test set.

Model	Base	S3	RS3	FDR/FDC
Flan-T5 (fine-tuned)	$0.0009	$0.0020	$0.0021	$0.0088
GPT-5-mini	$0.56	$0.70	$0.69	$4.80
Claude-Haiku-4.5	$1.67	$2.09	$2.09	$17.71
Gemini-3.5-flash	$2.26	$2.79	$2.76	$15.90

Table 3: Commercial APIs token usage and Flan-T5 GPU energy draw per configuration (249 articles).

Usage Metric	Base	S3	RS3	FDR/FDC
LLM input tokens	710,153	925,728	949,831	15,208,537
LLM output tokens	191,431	232,797	227,489	499,457
Flan-T5 electricity (Wh)	5.14	11.93	12.28	52.04
Flan-T5 running time (s)	111.2	219.3	228.1	728.7

6 DISCUSSION

These results confirm that high recall at low cost is achievable without full-document processing; the remaining limitations point to natural next steps.

6.1 Limitations

Regex patterns require maintenance as repository formats evolve. Coverage remains bounded by the registry; SciLite can be used as a scalable signal to identify missing repository patterns and prioritize registry enrichment. Regarding post-hoc verification of extracted identifiers: the registry's `id_pattern` field already serves this role – every prediction is checked against its repository's expected accession format, and non-matches are flagged for review rather than silently accepted.

6.2 Future Work

We plan to expand the registry automatically via an iterative enrichment loop: each iteration re-runs the extraction pipeline over the corpus in parallel across GPUs, groups previously unmatched identifier mentions, and prompts an LLM agent to propose new registry entries (repository name and regex pattern), which are accepted only if they pass an empirical coverage/false-match check against the extracted data.[13]

[13]Implemented but not yet run to convergence: https://github.com/VIDA-NYU/data-gatherer/blob/main/scripts/enrich_ontology.py

AUTHORS

Pietro Marini (Datatecnica, *data management*) is a data engineer and researcher working on LLM-based information extraction from the scientific literature and data integration for neurodegenerative diseases. He developed the Data Gatherer system and maintains the CARD Catalog (https://alzheimersdatahub.org/).

Mike A. Nalls (Datatecnica, *biomedical data science*) founded Datatecnica in 2017 after over a decade of experience in large dataset analytics and methods research in healthcare and other scientific fields. Having authored 550+ peer-reviewed publications (H-index > 150) in the field of applied statistics in large datasets, brain diseases, and genomics, he is a strong advocate for open science.

Mette Peters (Datatecnica, *biomedical*) is a Senior Advisor to the Division of Neuroscience at the NIA/NIH, bringing over 25 years of experience in data science from federal, industry, and nonprofit settings with a strong knowledge of data management architectures, data security and sharing policies, and biomedical data FAIRness.

Nicole Contaxis (NYU Langone, *data librarian*) is a medical data librarian and Head of Data Sharing and Metadata Management at the NYU Health Sciences Library. She specializes in clinical research data infrastructure, open science repository design, and navigating ethical data-sharing policies across health systems.

Juliana Freire (New York University, *data management*) is an Institute Professor of Computer Science and Data Science at New York University. An ACM and AAAS Fellow, she specializes in large-scale data visualization, data provenance, and computational reproducibility, developing AI-driven systems to extract trustworthy insights from complex data.

Aécio Santos (Centrum Wiskunde & Informatica, *data management*) is a Postdoctoral Researcher in the Database Architectures group at Centrum Wiskunde & Informatica (CWI) in Amsterdam. Previously, he was a Research Engineer at New York University (NYU), working on multiple research and development programs.

ACKNOWLEDGMENTS

This work was supported in part by the DARPA ASKEM (HR0011262087), ARPA-H BDF, and NSF (OAC-2411221). This research was also supported in part by the Intramural Research Program of the NIH, National Institute on Aging (NIA), National Institutes of Health, Department of Health and Human Services; project number Z01 AG000534, as well as the National Institute of Neurological Disorders and Stroke. This work utilized the computational resources of the NIH STRIDES Initiative (https://cloud.nih.gov) through the Other Transaction agreement – Azure: OT2OD032100, Google Cloud Platform: OT2OD027060, Amazon Web Services: OT2OD027852. This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Claude Code was used for initial code review, GitHub management, as well as README file and code annotation refactoring. Gemini in Google Docs corrected grammar and spelling errors.

Some authors contributed to this work as consultants to the NIH's CARD.

APPENDIX

Figure 3: Fixed query used for semantic snippet retrieval.

```
Data Availability Statement or mentions of dataset
repositories/portals, identifiers, or accession codes, including
PRIDE, ProteomeXchange, MassIVE, iProX, JPOST, Proteomic Data
Commons (PDC), Genomic Data Commons (GDC), Cancer Imaging Archive
(TCIA), Imaging Data Commons (IDC), Gene Expression Omnibus (GEO),
ArrayExpress, dbGaP, Sequence Read Archive (SRA), Protein Data Bank
(PDB), Mendeley Data, Synapse, European Genome-Phenome Archive
(EGA), BIGD, and ProteomeCentral.

Also include dataset identifiers or links such as PXD, MSV, GSE,
GSM, GPL, GDS, phs, syn, PDC, PRJNA, DOI, or accession code.

Look for phrases like deposited in, available at, submitted to,
uploaded to, archived in, hosted by, retrieved from, accessible
via, or publicly available.

Capture statements indicating datasets, repositories, or data
access locations.
```

Table 4: Precision (P), recall (R), and F1 against the two benchmarks, by model and run-configs, on the 249-PMCID test set.

		SciLite GT			Gold GT		
Model	**Config**	P	R	F1	P	R	F1
Flan-T5-ft	base	.186	.313	.233	.349	.537	.423
	S3	.317	.577	.409	.431	.883	.580
	RS3	.336	.658	.445	.444	**.979**	.611
	FDC	.233	**.758**	.357	.225	**.979**	.366
	RegEx	**.349**	.635	**.450**	**.511**	**.979**	**.671**
Claude-Haiku-4.5	base	.266	.319	.290	.488	.569	.525
	S3	.200	.634	.305	.224	.939	.362
	RS3	.211	**.693**	.324	.230	**1.000**	.374
	FDR	.642	.661	**.651**	.837	.999	.911
	RegEx	**.643**	.635	.639	**.856**	.996	**.921**
GPT-5-mini	base	.276	.319	.295	.483	.569	.522
	S3	.516	.600	.555	.703	.939	.804
	RS3	.540	.653	.591	.717	**1.000**	.835
	FDR	.483	**.679**	.564	.609	.968	.748
	RegEx	**.588**	.640	**.613**	**.798**	**1.000**	**.888**
Gemini-3.5-flash	base	.283	.319	.300	.491	.569	.527
	S3	.511	.622	.561	.643	.939	.764
	RS3	.530	**.680**	.596	.652	**1.000**	.789
	FDR	**.646**	.641	**.643**	**.863**	.999	**.926**
	RegEx	.629	.628	.628	.843	.984	.908

Table 5: Recall delta between RS3 and the regex-only (RegEx). Positive values indicate RS3 is recall-positive w.r.t RegEx.

	Gold GT – Recall			SciLite GT – Recall		
Model	RS3	RegEx	Δ	RS3	RegEx	Δ
Flan-T5-ft	.979	.979	0.000	.658	.635	+0.023
Claude-Haiku-4.5	1.000	.996	+0.004	.693	.635	**+0.058**
GPT-5-mini	1.000	1.000	0.000	.653	.640	+0.013
Gemini-3.5-flash	1.000	.984	+0.016	.680	.628	+0.052

REFERENCES

[1] Sam Anzaroot and Andrew McCallum. 2013. A new dataset for fine-grained citation field extraction. In *ICML Workshop on Peer Reviewing and Publishing Models (PEER)*.

[2] 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)*. 2370–2385.

[3] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. https://doi.org/10.48550/ARXIV.2210.11416

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[5] Caifan Du, Johanna Cohoon, Patrice Lopez, and James Howison. 2021. Softcite dataset: A dataset of software mentions in biomedical and economic research publications. *Journal of the Association for Information Science and Technology* 72, 7 (2021), 870–884. https://doi.org/10.1002/asi.24454 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24454

[6] Behnam Ghavimi, Philipp Mayr, Sahar Vahdati, and Christoph Lange. 2016. Identifying and improving dataset references in social sciences full texts. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press, 105–114.

[7] Michael Groenendyk and Laura Ivan. 2025. Data as scholarly output: addressing challenges in data citation tracking through natural language processing automation. *Scientometrics* (2025). https://api.semanticscholar.org/CorpusID:284033348

[8] Mark Hahnel, John Chodacki, Stan Neumann, Lars Holm Nielsen, Julian Gautier, Audrey Hamelers, Gretchen Gueguen, Mark Call, and David Scherer. 2025. GREI Metadata Recommendations from DataCite schema version 4.6. https://doi.org/10.5281/zenodo.16953589

[9] Nick Haupka, Jack H. Culbert, Alexander Schniedermann, Najko Jahn, and Philipp Mayr. 2025. Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, PubMed and Semantic Scholar. arXiv:2406.15154 [cs.DL] https://arxiv.org/abs/2406.15154

[10] Lynette Hirschman, Gully A. P. C Burns, Martin Krallinger, Cecilia Arighi, K. Bretonnel Cohen, Alfonso Valencia, Cathy H. Wu, Andrew Chatr-Aryamontri, Karen G. Dowell, Eva Huala, Anália Lourenço, Robert Nash, Anne-Lise Veuthey, Thomas Wiegers, and Andrew G. Winter. 2012. Text mining for the biocuration workflow. *Database* 2012 (01 2012), bas020. https://doi.org/10.1093/database/bas020 arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/bas020/16728680/bas020.pdf

[11] Ana-Maria Istrate. 2023. Building the Open Global Data Citation Corpus – Chan Zuckerberg Initiative. https://doi.org/10.5281/zenodo.7634893 Publisher: Zenodo Version Number: 1.0.

[12] Şenay Kafkas, Jee-Hyub Kim, and Johanna R. McEntyre. 2013. Database Citation in Full Text Biomedical Articles. *PLOS ONE* 8, 5 (05 2013), 1–9. https://doi.org/10.1371/journal.pone.0063184

[13] P Kahn and D Hazeldine. 1988. NAR's new requirement for data submission to the EMBL data library: information for authors. *Nucleic Acids Res* (1988). https://pmc.ncbi.nlm.nih.gov/articles/PMC336623/

[14] 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*. Springer, 43–53.

[15] Robert Leaman and Graciela Gonzalez. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. In *Biocomputing 2008*. World Scientific, 652–663.

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.

[17] Meiyu Lu, Srinivas Bangalore, Graham Cormode, Marios Hadjieleftheriou, and Divesh Srivastava. 2012. A Dataset Search Engine for the Research Document Corpus. *Proceedings - International Conference on Data Engineering*, 1237–1240. https://doi.org/10.1109/ICDE.2012.80

[18] D. A. Maisano, L. Mastrogiacomo, L. Ferrara, and F. Franceschini. 2025. A large-scale semi-automated approach for assessing document-type classification errors in bibliometric databases. *Scientometrics* 130, 3 (March 2025), 1901–1938. https://doi.org/10.1007/s11192-025-05244-y

[19] Make Data Count. 2025. Open data metrics require open infrastructure: Data Citation Corpus. https://makedatacount.org/find-a-tool/ https://makedatacount.org/find-a-tool/.

[20] Pietro Marini, Aécio Santos, Nicole Contaxis, and Juliana Freire. 2025. Data Gatherer: LLM-Powered Dataset Reference Extraction from Scientific Literature. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, Tirthankar Ghosal, Philipp Mayr, Amanpreet Singh, Aakanksha Naik, Georg Rehm, Dayne Freitag, Dan Li, Sonja Schimmler, and Anita De Waard (Eds.). Association for Computational Linguistics, Vienna, Austria, 114–123. https://doi.org/10.18653/v1/2025.sdp-1.10

[21] Pietro Marini, Aécio Santos, Nicole Contaxis, and Juliana Freire. 2025. *DataRef: A Dataset of Data Citations*. https://doi.org/10.5281/zenodo.15549086

[22] Amir Moslemi, Anna Briskina, Zubeka Dang, and Jason Li. 2024. A survey on knowledge distillation: Recent advancements. *Machine Learning with Applications* 18 (2024), 100605. https://doi.org/10.1016/j.mlwa.2024.100605

[23] N. Noy and Deborah Mcguinness. 2001. Ontology Development 101: A Guide to Creating Your First Ontology. *Knowledge Systems Laboratory* 32 (01 2001).

[24] Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833 [cs.DL] https://arxiv.org/abs/2205.01833

[25] Dietrich Rebholz-Schuhman, Anika Oellrich, and Robert Hoehndorf. 2012. Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews Genetics* 13 (11 2012), 829–839. https://doi.org/10.1038/nrg3337

[26] Dietrich Rebholz-Schuhmann, Miguel Arregui, Sylvain Gaudan, Harald Kirsch, and Antonio Jimeno. 2008. Text processing through Web services: calling Whatizit. *Bioinformatics* 24, 2 (01 2008), 296–298. https://doi.org/10.1093/bioinformatics/btm557 arXiv:https://academic.oup.com/bioinformatics/article-pdf/24/2/296/49044576/bioinformatics$_2$4$_2$2$_2$96.*pdf*

[27] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[28] Nicolas Robinson-Garcia, Philippe Mongeon, Wei Jeng, and Rodrigo Costas. 2017. DataCite as a novel bibliometric source: Coverage, strengths and limitations. *Journal of Informetrics* 11, 3 (2017), 841–854. https://doi.org/10.1016/j.joi.2017.07.003

[29] Sharmila Upadhyaya, Wolfgang Otto, Frank Krüger, and Stefan Dietze. 2025. SOMD2025: A Challenging Shared Tasks for Software Related Information Extraction. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, Tirthankar Ghosal, Philipp Mayr, Amanpreet Singh, Aakanksha Naik, Georg Rehm, Dayne Freitag, Dan Li, Sonja Schimmler, and Anita De Waard (Eds.). Association for Computational Linguistics, Vienna, Austria, 137–145. https://doi.org/10.18653/v1/2025.sdp-1.13

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL] https://arxiv.org/abs/1706.03762

[31] Venkatesan, Aravind, Kim Jee-Hyub, Talo Francesco, Michele Ide-Smith, Julien Gobeill, Jacob Carter, Riza Batista-Navarro, Sophia Ananiadou, Patrick Ruch, and Johanna McEntyre. 2017. SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data [version 2; referees: 2 approved, 1 approved with reservations]. *Wellcome Open Research* (Jul 10 2017). https://www.proquest.com/scholarly-journals/scilite-platform-displaying-text-mined/docview/2130425168/se-2 Copyright - © 2017. This work is published under http://creativecommons.org/licenses/by/4.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2024-06-10; SubjectsTermNotLitGenreText - Switzerland; United Kingdom–UK.

[32] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies* 1, 1 (02 2020), 396–413. https://doi.org/10.1162/qss_a_00021 arXiv:https://direct.mit.edu/qss/article-pdf/1/1/396/1760880/qss$_{a0}$0021.*pdf*

[33] Anjie Xu, Ruiqing Ding, and Leye Wang. 2025. ChatPD: An LLM-driven Paper-Dataset Networking System. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 5106–5116. https://doi.org/10.1145/3711896.3737202

[34] Tong Zeng and Daniel Acuna. 2020. Finding datasets in publications: the Syracuse University approach, In Rich Search and Discovery for Research Datasets. *arXiv preprint arXiv:2405.13135*, 158–165. https://doi.org/10.5281/zenodo.4402303

[35] Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan 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*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). 13083–13100. https://doi.org/10.18653/v1/2024.emnlp-main.726