LLMDap: LLM-based Data Profiling and Sharing

Shanshan Jiang*
SINTEF AS
Trondheim, Norway
Shanshan.Jiang@sintef.no

Sondre Sørbø
SINTEF AS
Trondheim, Norway
Sondre.Sorbo@sintef.no

Phil Tinn
SINTEF AS
Trondheim, Norway
Phil.Tinn@sintef.no

Shang Ferheng Karim
Oslo Metropolitan University
Oslo, Norway
kakashang96@msn.com

Dumitru Roman
SINTEF AS
Oslo, Norway
dumitru.roman@sintef.no

ABSTRACT

To boost data economy and harness the potential of the rapid expansion of available datasets, data description with rich, high quality and interoperable metadata is essential to facilitate data discovery and integration across multiple sources. Traditional keyword-based data search has limitations due to a mismatch between published data description and the terms used in data queries. In this paper, we explore the use of Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) to enable automatic metadata enrichment and improve dataset discoverability. We present LLMDap, an LLM-based pipeline for high quality data annotation and semantic discovery. The LLM pipeline automates the generation of structured and interoperable metadata from scientific publications, leveraging RAG and prior knowledge to enhance output accuracy. For data profiling, LLMDap allows data providers to efficiently generate "standardized", semantically enriched metadata for data publishing. When integrated with a data catalogue, LLMDap supports data consumers to discover and explore datasets. The method is LLM-agnostic and domain-independent, and we validated it in the biomedical domain. This work contributes to improving data discoverability, usability, and interoperability within a data sharing ecosystem.

PVLDB Workshop Format:
Shanshan Jiang, Sondre Sørbø, Phil Tinn, Shang Ferheng Karim, and Dumitru Roman. LLMDap: LLM-based Data Profiling and Sharing. VLDB 2025 Workshop: 3rd Data EConomy Workshop (DEC).

PVLDB Artifact Availability:
The source code has been made available at https://github.com/SINTEF-SE/LLMDap.

1 INTRODUCTION

In the emerging data economy, data is a critical asset for driving innovation and informed decision-making. To fully harness the potential of increasingly shared datasets, effective data discovery mechanisms are essential for identifying suitable datasets for value creation.

Data is commonly shared via data sharing platforms or data marketplaces that use catalogues to publish dataset descriptions with predefined metadata schemas. However, these schemas are often heterogeneous in type, format, and semantics, which hinders effective data sharing across distributed sources. Furthermore, such catalogues typically support only keyword-based search, leading to a mismatch between user queries and the published metadata. As a result, users may fail to locate relevant datasets simply because they do not know or use the exact metadata terms for search, even when the data is available in the catalogue [7]. This highlights a growing need for more flexible and effective data discovery methods based on semantic rather than keyword matching. Addressing this challenge requires semantically rich, context-aware, and high-quality metadata to bridge the gap between data providers and consumers.

With rapid advances in generative AI, in particular, Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG), natural language processing (NLP) has been significantly transformed. In this work, we investigate *how generative AI can be leveraged to generate rich, context-aware, high-quality dataset descriptions in a more automated and efficient manner to enhance data discoverability.* We propose a generic LLM-based pipeline (LLMDap) designed to extract accurate, interoperable metadata from scientific publications and other natural language sources. The pipeline is domain-agnostic and adaptable across disciplines. To demonstrate its applicability, we developed an LLMDap-based system for the biomedical domain with an intuitive user interface that supports (1) data providers through LLM-assisted automated generation of dataset profiles (a list of metadata describing the dataset), incorporating human-in-the-loop for quality assurance of the output, and (2) data consumers with features for effective dataset discovery and exploration. The system was evaluated with domain experts to assess its validity and effectiveness.

The main contributions of this work are:

- The LLMDap pipeline that provides a generic, domain-agnostic approach to automated metadata generation using LLMs and RAG, producing consistent, high-quality metadata aligned with domain ontologies to enhance dataset discoverability.
- The approach of using LLMDap for data sharing validated in the biomedical domain.

*Corresponding author.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents the general LLMDap framework for data profiling and sharing. Section 4 discusses the validation of the work and potential applications such as federated catalogues. Finally, Section 5 concludes the paper and suggests future work.

2 RELATED WORK

Natural language processing (NLP) techniques have been exploited to facilitate discovery of published datasets based on semantic matching instead of keyword matching. For example, [7] proposed an approach to ontology-based semantic search on open data catalogues with automatic dataset linking and indexing utilizing NLP techniques. However, imprecise or incomplete metadata of the published datasets remains a barrier to improved discoverability.

Recent advances in LLMs have significantly enhanced NLP capabilities in understanding natural language inputs and generating contextually relevant outputs. LLMs have been applied to various biomedical applications, such as question answering tasks, generation of medical terms and descriptions, and medical data analysis, demonstrating great potential for automating the generation of metadata compliant with the requirements of standards or data sharing systems. These applications exploit general models (e.g., ChatGPT, BERT [2]) or domain-specific models (e.g., PubMedBERT [3],BioBERT [10], BioMedLM [1], BioGPT [13], ClinicalBERT [5]).

Tool-augmented approaches like GeneGPT [9] integrate LLMs with biomedical APIs (e.g., NCBI[1]) to support in-context learning. An LLM-powered workflow enables natural language querying and automated analysis of cBioPortal biomedical datasets through integrated Python and LLM modules [6]. There are also efforts on Question & Answer (Q&A) benchmarks such as MedQA [8], MedMCQA [15], and MMLU [4].

However, existing work has not addressed answer generation with the precision required for metadata standardization. Retrieval-Augmented Generation (RAG) [11] offers a promising solution by grounding LLM outputs in evidence, enhancing factual accuracy and transparency—critical for scientific applications. An example RAG pipelines for biomedical data is described in [18].

Furthermore, current data catalogs or repositories lack the support for automating metadata generation or auto-filling the profiles based on natural text input. Our work addresses this gap.

3 LLMDAP FOR DATA PROFILING AND SHARING

3.1 Data Profiling with LLMDap

The idea for a generic LLM-based pipeline for data profiling is illustrated in Figure 1. The LLMDap framework consists of five main components:

(1) *Auxiliary Data Referencing:* Get additional data that may help in the query process, e.g., extraction of domain knowledge using an ontology.
(2) *Document Indexing & Embedding:* Split the full textual input document into chunks (i.e., text segments) in a string format and create vector embeddings for each chunk to capture semantic content for downstream tasks.
(3) *Context Retrieval:* Compare the user input (e.g., a schema field or a natural language query) semantically against precomputed text embeddings to retrieve the most relevant chunk from the source corpus. This retrieval process is assisted by referencing appropriate auxiliary data. Different (re)ranking methods, such as anthropic's ranking[2] or newer ranking methods can be adopted in this process.
(4) *LLM-based Query:* Provide the LLM with a prompt based on the user input and the context retrieved, and get LLM predictions based on the given context.
(5) *Pipeline Configuration:* Adapt pipeline configurations to align with specific tasks and application contexts; for instance, selecting different LLMs based on domain-specific constraints or anticipated usage patterns. Such considerations can be cost and hardware availability when choosing an LLM, or domain-specific prompt optimization.

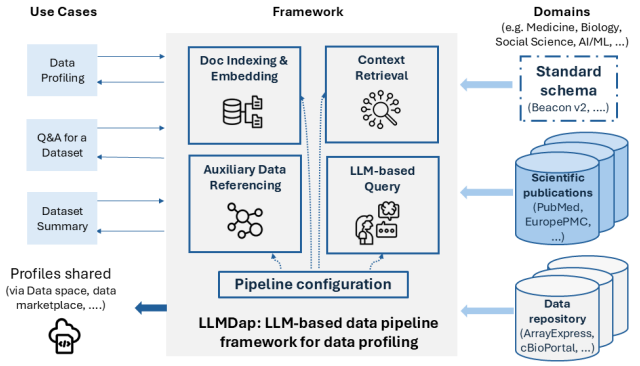

Figure 1: LLMDap Framework.

Domain-specific information such as standard schemas, scientific publications and datasets can be used when setting up and configuring the LLMDap pipeline for a specific domain. For example, for the biomedical domain as illustrated in Figure 1, the schema can be based on Beacon v2 [16], the scientific publications can be from PubMed[3], Europe PMC[4] or other portals, and the datasets from portals such as ArrayExpress[5] and cBioPortal[6].

The generated dataset profiles (i.e., the metadata descriptions) can be shared on data spaces, data marketplaces or other data sharing platforms, typically via catalogues of datasets and services, to facilitate discovery. LLMDap can be used not only for data profiling, but also other use cases, such as querying and summarization of datasets.

Figure 2 illustrates how the LLMDap pipeline works. The details are presented in the following subsections.

3.1.1 Metadata Schema. The input to the LLMDap pipeline is a *metadata schema* and one or several documents describing the dataset to be profiled, such as scientific papers and lab protocols.

[1]https://www.ncbi.nlm.nih.gov/home/develop/api

[2]https://www.anthropic.com/news/contextual-retrieval
[3]https://pubmed.ncbi.nlm.nih.gov
[4]https://europepmc.org
[5]https://www.ebi.ac.uk/biostudies/arrayexpress
[6]https://www.cbioportal.org

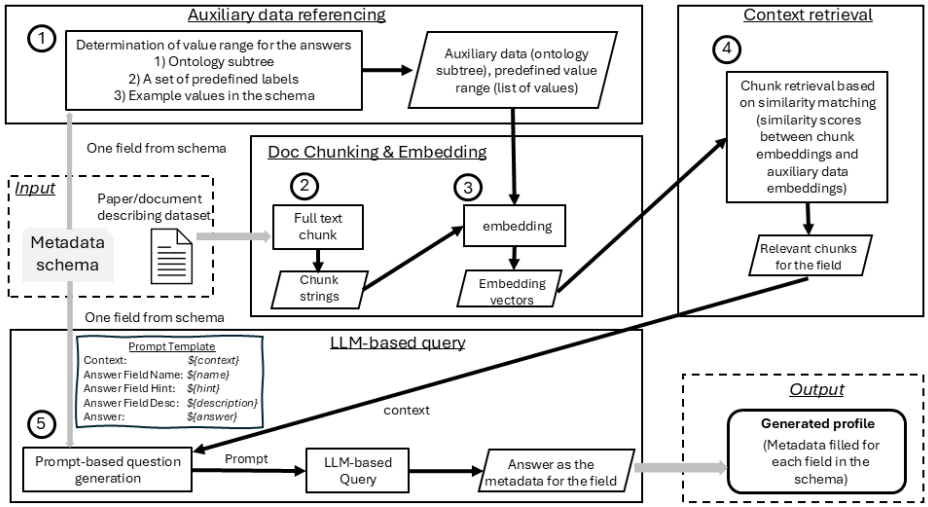

Figure 2: LLMDap Pipeline.

The schema consists of a list of *metadata fields* for LLM query and their *descriptions* as well as *example values*.

For the LLMDap execution, any schema can be used, for example, defined by standards, existing catalogues, or users. To facilitate interoperability and discoverability, the metadata fields should represent the widely used and recognized domain concepts and terms for domain-specific tasks. Typically, the fields should be linked to domain ontologies, such as Experimental Factor Ontology (EFO)[7] and Ontology for Biomedical Investigations (OBI)[8].

When profiled using standardized or interoperable schemas, data profiles generated by LLMDap can more easily be shared via data spaces, data marketplaces, or other data sharing platforms.

3.1.2 Auxiliary Data Referencing. Step 1 determines the value range for the answers to be generated by LLMs for each field defined in the metadata schema. Three types of auxiliary information can be utilized to get additional data that may assist in the context retrieval process:

- A predefined set of metadata tags associated with each schema field, either known a priori to the user or derivable from related annotated instances available on existing data catalogues;
- A domain-specific ontology with nodes similar to the expected values of the metadata fields;
- Example values provided in the input metadata schema.

The output of this process is auxiliary data and a predefined value range for each field, e.g., as an ontology subtree or a list of values.

3.1.3 Document Chunking & Embedding. In step 2, the full text of the input document is split into chunks based on document structures, i.e., section titles such as "METHODS", "RESULTS". This produces a set of TextNodes [12] objects representing the chunk strings and metadata. Afterwards, chunks are embedded into vectors (step 3) for use in the next steps, where document chunks and auxiliary data are embedded into one shared embedding space.

3.1.4 Context Retrieval. Step 4 identifies the most relevant chunks for each field defined in the schema. Semantic matching is performed based on cosine similarity scores between chunk embeddings and auxiliary data embeddings. The output is a concatenated string of high-relevance chunks per field, providing the next step of LLM query with candidate context for prediction.

3.1.5 LLM-based Query. Step 5 provides the LLM with a prompt based on a template that includes the contexts retrieved from the prior step and user input (in the case of Q&A and summarization) or field descriptions from the target data schema (in the case of data profiling). The prompt allows the LLM to generate a prediction of the answer based on the given contexts. This process is independent of the LLM models used. The output is a generated profile with metadata filled for each field in the schema.

3.2 Data Sharing with LLMDap

To facilitate data sharing, additional components are needed to work with LLMDap, as shown in Figure 3. In particular, a data catalogue is used to store the profiles of the shared datasets so that users can search and access them. The architecture of LLMDap-based data sharing consists of:

- *User Interface (UI)* intended for the users to generate metadata using LLMDap (as data provider), browse and find datasets (as data consumer), and explore datasets with chat-style question answering interface (as data consumer).
- *LLM-based pipeline,* including components providing functionality of LLMDap as described in Section 3.1, and API for interaction with the UI. In *LLM query*, LLM is used for both dataset metadata generation and Q&A.
- *Data catalogue* with the generated metadata for datasets and associated documents. It can be realised by a centralised

[7] https://www.ebi.ac.uk/efo/index.html
[8] https://obi-ontology.org

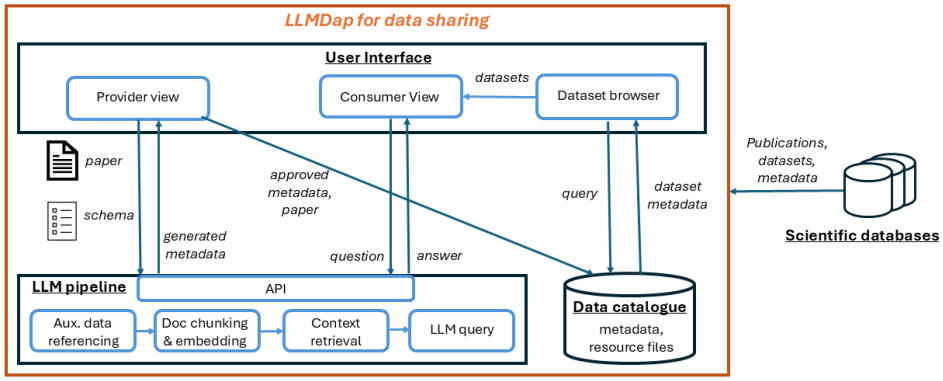

Figure 3: LLMDap for data sharing.

data storage, or as a federated solution from distributed data storage.

- *Scientific databases* with APIs for retrieving metadata from relevant databases.

Figure 3 illustrates the interaction between system components as well. For data profiling, the LLM pipeline API accepts a metadata schema and a paper as input for metadata generation; users can verify the generated metadata and make changes if needed before the metadata with associated files are stored in the data catalogue. For data discovery, the UI interacts with the data catalogue, e.g., using database queries. For dataset exploration, the LLM pipeline API accepts queries in natural language and returns LLM generated answers. APIs for external scientific databases are used to retrieve more metadata and resources file to enrich the metadata and associated resources that consumers can explore.

4 VALIDATION IN THE BIOMEDICAL DOMAIN

To validate the LLMDap approach, we implemented a system for the biomedical domain based on the design outlined in Section 3.2 using Python, Streamlit[9] and SQLite. User uploaded files (e.g., scientific papers or other documentations) are stored as local files. The SQLite database is used to implement the data catalogue, and it stores the generated metadata and the links to the local files. Details about the user interfaces and the realization of the data profiling and sharing are described in the following subsections. In addition, to access external documents, APIs of biomedical databases are used, such as NCBI E-utilities (PubMed[10]) and EBI services (BioStudies[11], ArrayExpress, EuropePMC[12]). Two LLM pipelines are implemented, one for data profiling and another one for data exploration using Q&A (details are described below).

4.1 User Interface

A modular, intuitive UI[13] is implemented using Streamlit with a multi-page design. The system offers two primary user interfaces:

- *Provider View:* Enables the submission of biomedical documents and the automatic generation of dataset metadata through the LLMDap backend pipeline.
- *Consumer View:* Facilitates querying, searching, and retrieving dataset-related information via a ChatGPT-style interface using the Q&A LLM pipeline. This includes dataset overview, browsing, and semantic search, as well as extraction of additional insights from associated scientific papers to support research activities.

The key components and operational steps of the Streamlit-based UI as supported by the respective UI pages (illustrated in Figure 4) are described in the following subsections.

Key Human-Machine Interaction principles [17][14] have guided the design and implementation of the user interface for effective data visualization.

4.1.1 Provider View. The Provider View facilitates the creation and submission of metadata derived from scientific literature. This functionality is accessible via the Provider page (Figure 4(a)), and supports a stepwise process from document input to metadata storage. This component prepares valid input for profiling through PDF or XML parsing, metadata extraction (e.g., PMID parsing), and schema validation using Pydantic[14].

A typical provider interaction with the system is:

(1) *Input of Scientific Paper:* Users can supply a scientific publication for metadata profiling by either uploading a local file in PDF or XML format, or by entering a corresponding URL or PubMed ID.
(2) *Schema Selection:* A metadata schema is required for guiding the LLM-based profiling process. Users may utilize a predefined default schema or upload a custom schema formatted in JSON. Pydantic models from the uploaded JSON schemas are dynamically generated to provide schema flexibility.
(3) *Initiation of Metadata Generation:* Upon clicking the *Process Input* button, the system transmits both the selected document and metadata schema to the backend LLM pipeline. The LLM processes the input to generate structured metadata relevant to the dataset.

[9] https://streamlit.io
[10] https://www.ncbi.nlm.nih.gov/home/develop/api
[11] https://www.ebi.ac.uk/biostudies
[12] https://europepmc.org/RestfulWebService
[13] https://github.com/SINTEF-SE/LLMDap/tree/main/llm_ui
[14] https://docs.pydantic.dev/latest

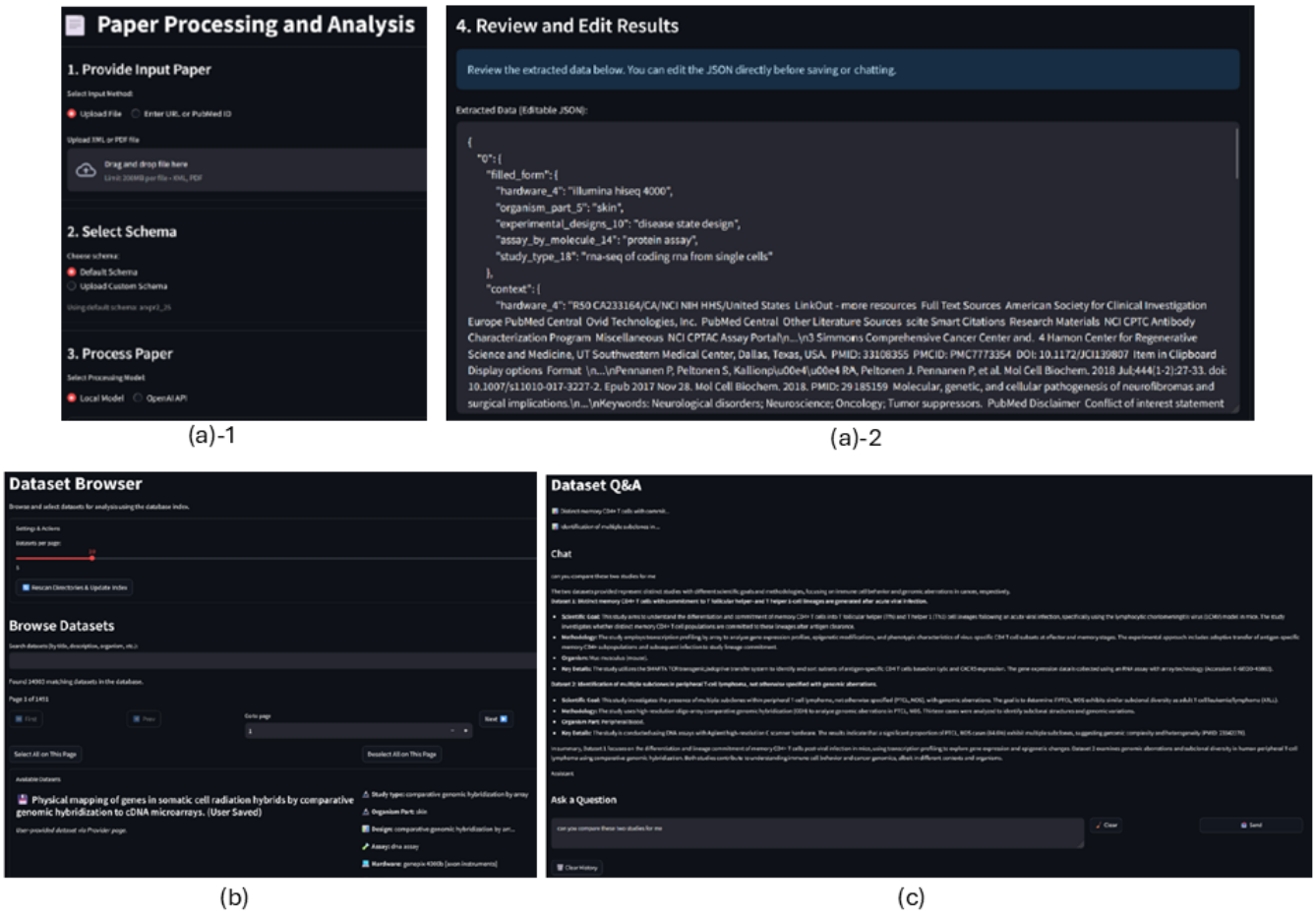

Figure 4: Screenshots of the main user interface pages: (a) Provider view, (b) Dataset browser, (c) Consumer Q&A.

(4) *Metadata Review and Editing:* The resulting metadata is displayed in an editable text field. This allows users to review the output and perform manual refinements where necessary.
(5) *Finalization and Storage:* Once satisfied, users can click the *Save to Database* button to complete the process. Supplementary metadata (e.g., PubMed-derived bibliographic data) may be appended. The finalized metadata are then stored in the catalogue database and associated with any relevant local files.

4.1.2 Dataset Browser. The Dataset Browser page (Figure 4(b)) provides an interface for viewing, searching, and selecting datasets that have been processed and indexed in the system's database. This component serves as the entry point for the consumer-oriented Q&A workflow (Section 4.1.3). This page provides the following functionalities:

(1) *Browsing Datasets:* The page presents a paginated list of available datasets along with key metadata fields, enabling rapid visual inspection.
(2) *Searching Datasets:* A search input field allows users to filter datasets by specified keywords.
(3) *Dataset Selection for Q&A:* Users may select one or more datasets via checkboxes adjacent to each entry. Clicking the *Ask Questions About Selected Datasets* button initiates the Q&A workflow based on the selected datasets.
(4) *Index Update:* A *Rescan Directories & Update Index* button triggers a background operation that scans predefined directories and updates the catalogue index accordingly.

4.1.3 Consumer View. The Consumer View supports natural language interaction with the curated dataset metadata. The interface resembles a ChatGPT-style chat system and allows end-users to pose queries about selected datasets.

A typical consumer interaction with the system is:

(1) *Context Display:* Upon activation, the page displays a synthesized summary of the selected datasets, which serves as contextual information for subsequent user queries.
(2) *Question Submission:* Users enter a question in a designated text input area and submit it using the *Send* button. The query, along with contextual metadata, is forwarded to the Q&A LLM pipeline for processing.
(3) *Answer Generation and Display:* The system retrieves relevant information using the LLM and presents the answer in

the chat interface. A persistent chat history is maintained for each question.

In addition, a *configuration page* is used for specifying the system settings, including selection of the LLM model for profiling, adjustment of parameters such as temperature and maximum tokens for LLM, and update of Q&A prompt template.

4.2 LLM Pipelines Leveraging RAG

As introduced at the beginning of this section, there are two pipelines instantiated for the system. The first one is the *LLMDap pipeline* for automatic extraction of dataset metadata with four components that implement the workflow steps described in Section 3. This pipeline works in the backend. The second one is a *Q&A pipeline* that utilizes an external LLM service for consumer query.

Our approach is LLM-agnostic (independent of the underlying LLM), and different LLMs can be used for these two pipelines as demonstrated in our implementation. For example, different LLMs (GTP-4o mini, Meta-Llama-3.1, Mistral-7B) were tested in the LLM query step of the LLMDap Pipeline to assess and compare their performance, while the OpenAI service is used for Consumer Q&A. The Streamlit UI integrates with the *LLMDap pipeline* through the call_reference[15] function in the *Provider View*, while OpenAI's API is used to implement the *Consumer Q&A View*[16].

Both LLM pipelines use RAG to enhance output quality through fact-based retrieval for improved trustworthiness. During the profiling process, relevant input document segments are retrieved and provided as contextual input for field-level metadata extraction by the LLM used in the LLMDap. In the Consumer Q&A pipeline, contextual information is dynamically assembled from multiple sources, including the structured data catalogue (SQLite), previously extracted metadata (JSON), and live external resources (e.g., PubMed, EBI). This compiled context is structured into a prompt and submitted to the LLM, enabling the generation of responses grounded in curated, dataset-specific information.

4.3 Validation with Domain Experts

To validate the approach proposed in this paper, an early version of the implemented system was demonstrated to four biomedical domain experts to gather feedback from potential users. The experts considered that such a system would be useful for their research and they liked the design and usability of the system. They also suggested some improvements to the user interface design regarding new functionality and usability, which have been implemented in the current version as described in this paper. For example, they wished to have a page to manage datasets, a *select all* button from the dataset page, more search terms and possibilities, which led to the current *Browse Datasets* page with the new button and extended search options. Another example is adding more feedback to the user, e.g., this led to the additional confirmation/error messages shown in the *Provider View* page to indicate what happened and with what information after successfully uploading a dataset.

The expert feedback is considered representative of the typical requirements of potential users and serves as an indicator of the system's usability and practical utility.

[15]https://github.com/SINTEF-SE/LLMDap/blob/main/profiler/run_inference.py
[16]https://github.com/SINTEF-SE/LLMDap/blob/main/llm_ui/app/llm.py

4.4 Dataset Schema and Federated Catalogue

In our implementation, a central database is used for data catalogue. Alternatively, federated data catalogues can be implemented to facilitate discovery in a distributed data ecosystem. As current standards and catalogues use heterogenous metadata schema, schema mapping is needed to establish a harmonized domain schema with common and widely used metadata fields, so that the LLM generated profiles can be interoperable across catalogues.

Although the schema used in the implemented system for validation is specific to datasets curated from ArrayExpress for controlled experimentation, the adoption of emerging standard schemas, such as Beacon v2 defined by GA4GH [16], should be considered to enhance interoperability and alignment with community practices.

5 CONCLUSION AND FUTURE WORK

In this paper, we introduced the generic LLMDap pipeline for data profiling and sharing, and demonstrated how LLMs and RAG can be applied to enrich metadata and support question answering in the biomedical domain. The work proposes a solution leveraging the power of LLMs and the factual, truth-based trustworthiness enhanced by RAG to improve data discoverability in a data sharing ecosystem. The feedback provided by domain experts reflects common user requirements and highlights the system's usability and potential applicability.

For further work, we are conducting extensive experiments to evaluate and compare the performance of different configurations of LLMDap using various LLM models in terms of accuracy and cost, and will report the comparative results. Moreover, some enhancements may improve usability, facilitate insight extraction, and increase the transparency and interpretability of LLM outputs. For example, a graph-based visualization can be used to represent relationships among extracted metadata elements. Highlighting document segments and data sources used by LLM and confidence scores can improve transparency in metadata extraction and Q&A. Moreover, side-by-side charts or heatmaps can be used to enable visual comparison of key metadata fields across multiple datasets. Finally, while the LLMDap pipeline is validated using biomedical datasets, the solution can be adapted to other domains by adjusting metadata extraction to domain-specific terminology, ontologies and data formats, and integrating with domain databases.

ACKNOWLEDGMENTS

The work is funded through the projects UPCAST (HE 101093216), enRichMyData (HE 101070284), and DataPACT (HE 101189771). We also thank the colleagues from the National Hellenic Research Foundation for valuable discussion and feedback during the work.

REFERENCES

[1] Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421* (2024).

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*.

[3] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare* 3, 1, Article 2 (Oct. 2021), 23 pages.

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300 [cs.CY] https://arxiv.org/abs/2009.03300

[5] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv:1904.05342 [cs.CL] https://arxiv.org/abs/1904.05342

[6] Rong Ji, Kai Gong, Lihong Huang, Wenxian Yang, and Rongshan Yu. 2024. Leveraging LLMs for Automated Analysis of Biomedical Data. In *2024 9th International Conference on Communication, Image and Signal Processing (CCISP)*. 67–71. https://doi.org/10.1109/CCISP63826.2024.10765518

[7] Shanshan Jiang, Thomas F. Hagelien, Marit Natvig, and Jingyue Li. 2019. Ontology-Based Semantic Search for Open Government Data. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*. 7–15. https://doi.org/10.1109/ICOSC.2019.8665522

[8] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences* 11, 14 (2021). https://doi.org/10.3390/app11146421

[9] Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2024. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics* 40, 2 (2024), btae075.

[10] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (09 2019), 1234–1240.

[11] 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. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL] https://arxiv.org/abs/2005.11401

[12] Jerry Liu. 2022. *LlamaIndex*. https://doi.org/10.5281/zenodo.1234

[13] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics* 23, 6 (09 2022), bbac409.

[14] Jakob Nielsen. 1994. 10 Usability Heuristics for User Interface Design. https://www.nngroup.com/articles/ten-usability-heuristics/ Accessed: 2025-06-15.

[15] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. arXiv:2203.14371 [cs.CL] https://arxiv.org/abs/2203.14371

[16] Manuel Rueda, Roberto Ariosa, Mauricio Moldes, and Jordi Rambla. 2022. Beacon v2 Reference Implementation: a toolkit to enable federated sharing of genomic and phenotypic data. *Bioinformatics* 38, 19 (08 2022), 4656–4657. https://doi.org/10.1093/bioinformatics/btac568

[17] Euphemia Wong. 2025. Shneiderman's Eight Golden Rules Will Help You Design Better Interfaces. https://www.interaction-design.org/literature/article/shneiderman-s-eight-golden-rules-will-help-you-design-better-interfaces?srsltid=AfmBOophhbNz3afFiwW8pAgTLIElM4Fb1NKTP9KY31AqLg3FZF999n4J Accessed: 2025-06-15.

[18] Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S. Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. 2025. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems* 2, 1 (2025), 1–5.