Reducing Data Curation Time with an LLM-Augmented ETL System
A GEO Case Study

Ardigen: Ida Kupś, Sergiusz Wesołowski, Jakub Widawski, Michał Stachowiak

ABSTRACT

Streamlining data curation and annotation with LLM augmented system

Data annotation and curation in the biomedical field is a challenging and time-consuming process. For instance, annotating a single cell study from a GEO database can consume up to 60% of an experts time.. This burdensome task diverts scientists' valuable time away from testing novel hypotheses.

Manual annotation of GEO studies can take up to 4 days; Our pipeline reduces this to under 30 minutes, turning a multi-day bottleneck into a near real-time task.

To address this, Ardigen developed a scalable LLM-augmented ETL (Extract, Transform, Load) system that transforms complex multi-source knowledge into reusable data products. Our system combines state-of-the-art AI models like LLMs with traditional methods such as fuzzy matching, rule-based parsing, and controlled vocabularies to minimize the risk of AI hallucinations reaching 80% accuracy, as confirmed by experts.

BACKGROUND AND INTRODUCTION

NCBI's Gene Expression Omnibus (GEO) provides valuable gene expression data across various data types, but a major challenge is the lack of consistent, standardized annotation. Proper annotations, including experimental conditions and sample types, are essential for making datasets searchable, comparable, and usable across studies. These annotations are crucial for integrating data from multiple sources, facilitating accurate analysis, and ensuring reproducibility, which is key for advancing scientific discoveries. However, manual annotation is time-consuming, prone to errors, and slows down scientific progress.

To address these challenges, we developed an automated tool that extensively leverages large language models (LLMs) to streamline the annotation process. This tool detects and extracts relevant metadata, ensuring consistency and reducing human error. A minimum viable product (MVP) was developed to automatically annotate four key fields within GEO studies: Condition, Tissue, Species and CellType. The Ardigen system demonstrates the potential of AI-driven techniques to enhance accuracy and accelerate biological research.

METHODS / TECHNOLOGIES

Step 1: The pipeline first applies a conventional ETL process to capture metadata elements with known structure, predictable patterns,

Step 2: The pipeline then applies NLP and ontology based harmonization to map the input into a predefined schema.

Step 3: For unresolved records, we employ a multi-LLM strategy that selects models of varying complexity depending on task difficulty, ranging from lightweight to advanced models for complex cases.

- To optimize efficiency, the system employs context batching and concept reuse. Entries with identical hashed context signatures are grouped so that a single LLM call serves multiple records, and multiple extraction targets are combined for the same text block to further minimize LLM calls.
- This way we support context switching between reasoning and extraction stages.
- The LLM extraction modul is model-agnostic and can integrate with different LLM providers, adapting to evolving AI capabilities and pricing.
- The LLM module is based on n extraction concept definitions (an example definition is shown in Figure 6.)

RESULTS AND DISCUSSION

Key Outcomes

- Automated metadata extraction, harmonization, and enrichment on GEO single-cell datasets.
- Hybrid approach (ETL + NLP/fuzzy + LLM) achieved **high accuracy** with reduced curator bias.
- Processing time cut from **days to minutes**, lowering costs and enabling scalable deployment.

Impact

- Curators shift from manual annotation to supervisory roles.
- More **objective, reproducible annotations** across datasets.
- Modular design adapts easily to diverse biomedical domains and ontologies.
- Companion web app supports inspection, curator feedback, and continuous improvement.

CONCLUSIONS

- LLM-augmented ETL **transforms metadata curation** from a manual bottleneck into a rapid, scalable process.
- Delivers **time and cost savings**, while improving accuracy and reducing bias.
- Provides a **generalizable framework** for AI-ready biomedical data repositories, accelerating drug discovery and clinical research.

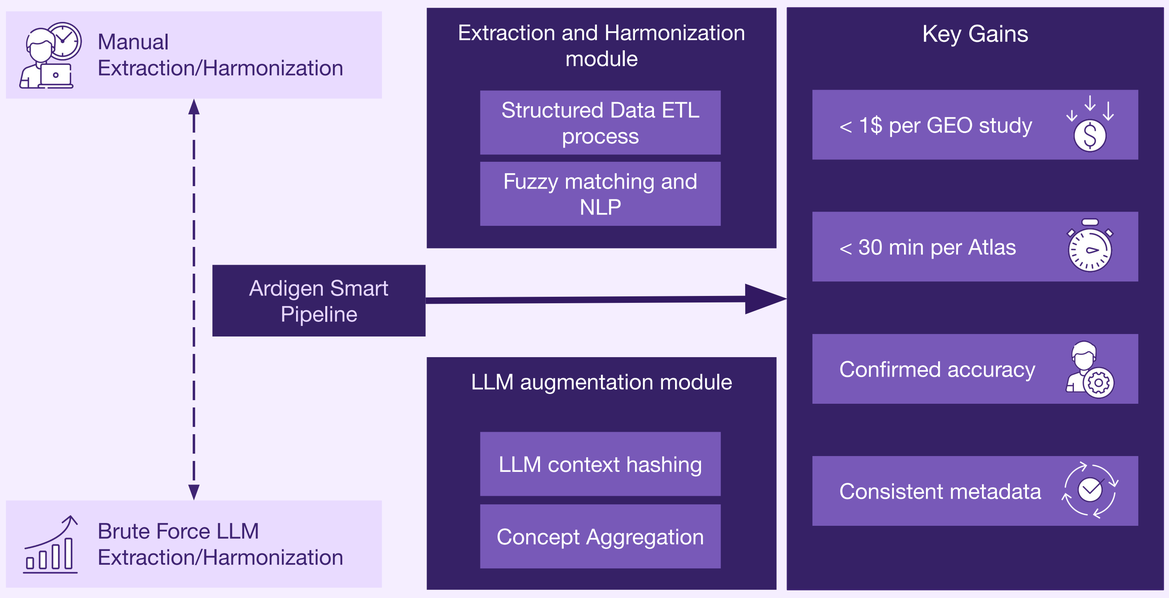

Figure 1. *Manual data curation is time-consuming, whereas applying LLMs to the entire dataset is costly. In the first iteration of the pipeline, metadata are extracted from selected GEO studies using rule-based methods, fuzzy matching, and NLP, producing a structured metadata table. In the second iteration, an LLM augmentation module fills gaps for columns that failed extraction or harmonization. This two-step process reduces time and cost while ensuring consistent ontology terms across GEO studies.*

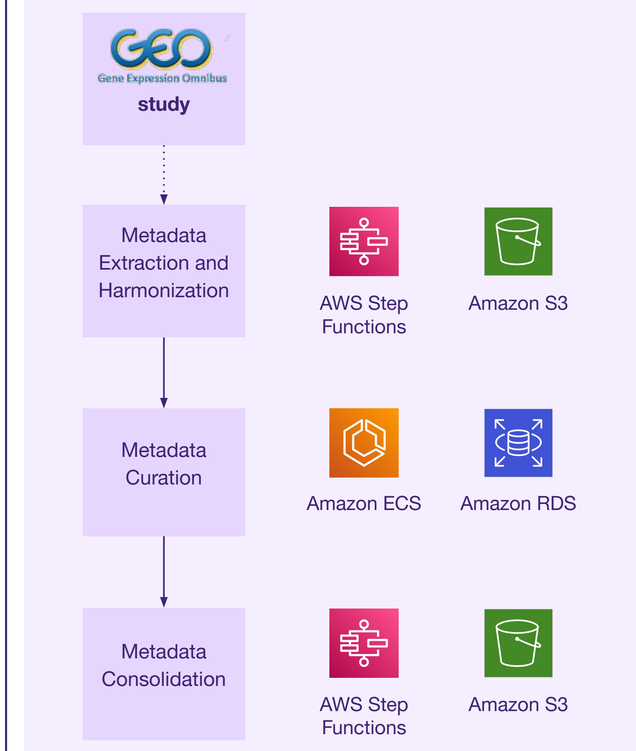

Figure 2. *Through the usage of Step Functions for orchestration, ECS web-based application and RDS SQL database with S3 bucket storage we provided a semi-automated and deployed on AWS workflow.*

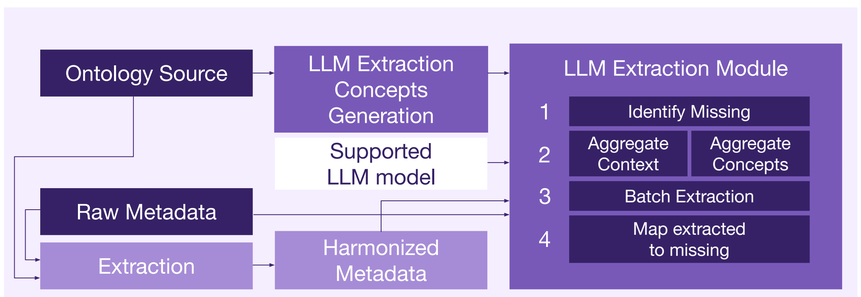

Figure 3. *LLM module in the high level context of the ETL pipeline.*

Sample ID	Original Value	Harmonized Value
sample1	Homo sapiens	Human
sample3	9606	Human
sample4	F Fly (D. melan)	⚠

→

Sample ID	Original Value	Harmonized Value
sample1	Homo sapiens	Human
sample3	9606	Human
sample4	F Fly (D. melan)	Fruit fly

Figure 4. *The web application displays the extracted (original) values alongside their harmonized versions. The data curator reviews the results and corrects metadata for rows that the pipeline failed to harmonize.*

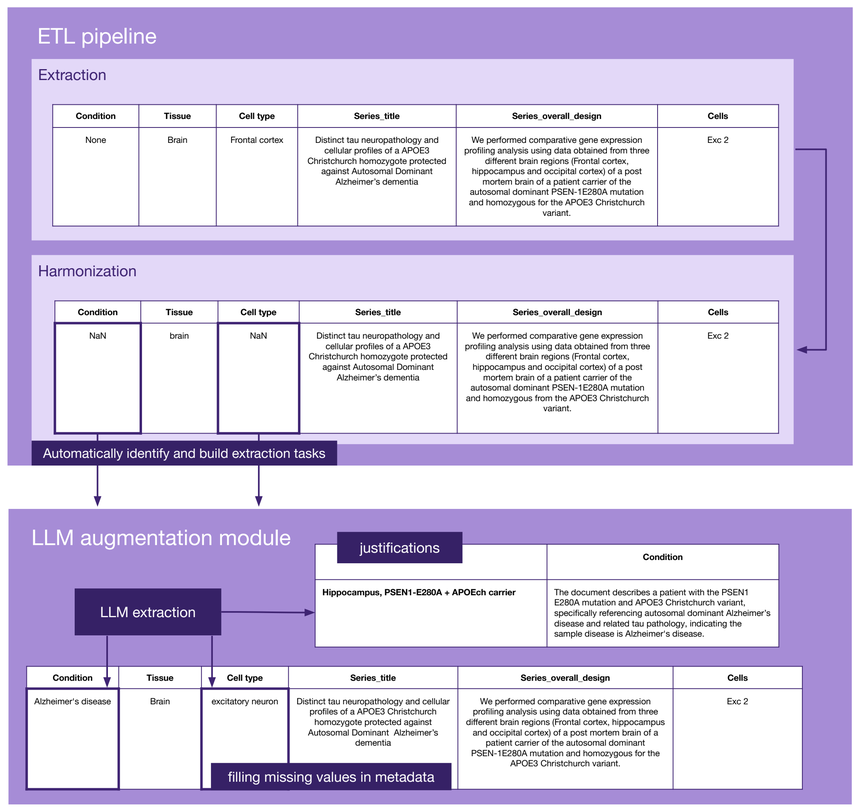

Figure 5. *The diagram shows how the data progresses through the LLM module within the ETL pipeline from raw metadata through harmonization and LLM re-harmonization. In each stage we see gradual improvement in the data structure and completeness.*

KEY FEATURES

- **ETL Pipeline with Fuzzy Matching**
  Our pipeline accelerates metadata extraction by first applying synonym and fuzzy matching to handle obvious cases. This means LLMs are only invoked when needed, focusing their power on non-obvious or complex scenarios while reducing the costs of maintenance of such a system.
- **Ontology Mapping**
  Extracting metadata is only half the challenge; making it actionable and interoperable is a critical process. We ensure that all extracted metadata is mapped to controlled vocabularies, accounting for synonyms, acronyms, and multiple meanings. This guarantees consistency across datasets and platforms.
- **Context Aggregation and Cost Control**
  Brute-force LLM extraction of GEO metadata would be prohibitively expensive—datasets often contain hundreds of thousands of records. Our pipeline uses smart context and concept aggregation so that each unique context is processed only once. In practice, this reduces the number of LLM queries by up to **16,000×**, while built-in cost monitoring (genai-prices) keeps token usage under control.
- **Model-Agnostic Design**
  The system is not tied to any single LLM provider. Its modular, model-agnostic design ensures compatibility with all major architectures, making it future-proof in a rapidly evolving AI landscape.
- **Justifications and References**
  Beyond producing metadata, the pipeline generates justifications and references for each extracted element. These outputs give curators transparency and guidance, allowing them to supervise the LLM process with confidence.
- **Web Application**
  The application serves as an interface for exploring and understanding metadata produced by the pipeline. Its features allow data curators to navigate, review and edit data with ease. Data curators can seamlessly replace metadata values across hundreds of thousands of records by selecting the correct alternatives from controlled vocabularies and validate data integrity.
- **Consolidation Step**
  After the metadata has been reviewed by the data curator, the final stage of the pipeline is metadata integration with the GEO study count data. The resulting datasets are saved in standard formats such as .h5ad, .h5mu, or .rds. At this stage, data has undergone quality review and is prepared for the downstream analysis.

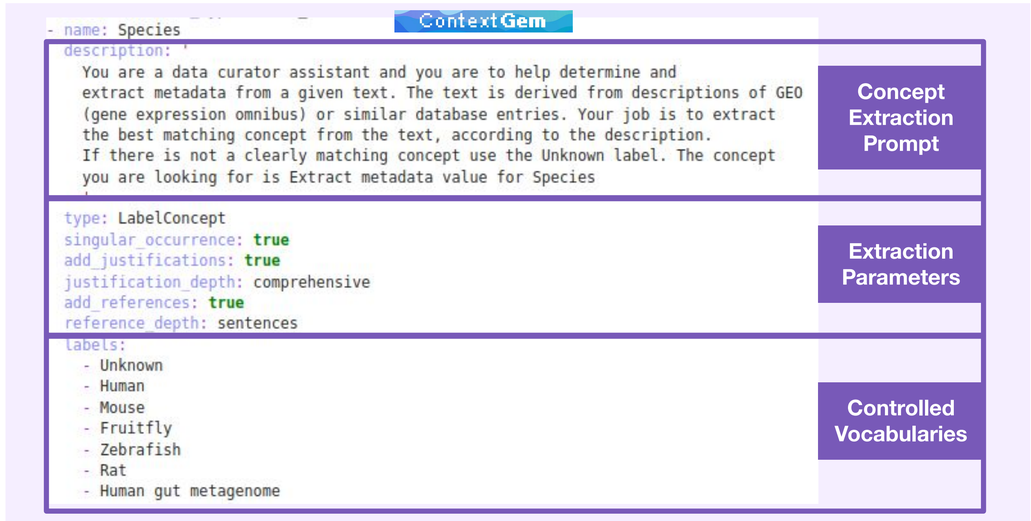

Figure 6. *The LLM augmentation module uses the open-source tool ContextGem to extract metadata from the files. Each category contains a prompt, parameters, and the controlled vocabularies that should be used to perform harmonization."*

IMPACT

Ardigen's approach drastically reduces the annotation time of the GEO Study to just 30 minutes, empowering scientists to focus on higher-level research. The annotated data is easily accessible through an intuitive user interface. By delivering significant operational value through reduced curation overhead, our system accelerates the creation of AI-ready data repositories for drug discovery and clinical research.

REFERENCES

[1] Home - GEO - NCBI
[2] PubMed