From Text to Genes: Can LLMs Enhance Expression Annotations?

Ardigen: Zofia Długosz, Tomasz Gargas, Bohdan Bodnar, Michał Czekański, Agata Mierzaniec, Marzena Jankowska, Tomasz Jetka, Dawid Rymarczyk
Faculty of Mathematics and Computer Science, Jagiellonian University: Dawid Rymarczyk

BACKGROUND AND INTRODUCTION

NCBI's Gene Expression Omnibus (GEO) provides valuable gene expression and functional genomics data, 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 based on large language models (LLMs) that streamlines 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 in GEO studies: Condition, Tissue, Drug and Intervention, demonstrating the potential of AI-driven techniques to enhance accuracy and accelerate biological research.

DATA

Currently, 3,000 GEO experiments have been annotated. In total, there are 35,000 Homo sapiens gene expression studies available in the GEO database.

Ontologies:

Ontologies are utilized to map cells, tissues, and diseases, ensuring consistent terminology across the dataset. DrugBank is employed to provide standardized and validated names for drugs.

Quality control.

Data was annotated and verified by human specialists in biology field.

METHODS

Computational environment and data storing

Selected data sources were processed using the data pipeline, and the resulting data was securely stored in a data lake, awaiting processing to extract insights. Cloud setup is based on Google Cloud Platform.

Model

To extract necessary information advanced NLP techniques were employed, mainly focused on Large Language Models (LLM), such as Gemini, with a Retrieval-Augmented Generation (RAG) framework and underlying Vector Database.

Postprocessing

After predicting selected values the results were postprocessed using

- Ontologie mapping
- Drug bank mapping
- Diverse algorithms for cleaning and standardizing the output.

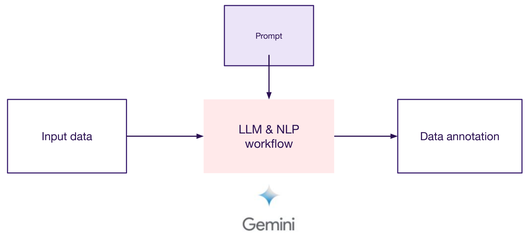
Evaluation

To evaluate the model we used strict accuracy, which checked if LLM output is exactly the same as expected target. This could cause the results to be underestimated as the LLM often returns alternative correct answers.

RESULTS AND DISCUSSION

- Design, implementation and deployment of a system for automatic dataset annotation has been proposed and performed in a short period of time
- Usage of a pretrained Large Language Model such as Gemini is enough to achieve satisfiable results but requires careful prompt engineering
- LLM annotator is able to annotate datasets automatically with good accuracy and in a short time (minutes per multiple datasets)
- Significant part of LLM's mistakes (defined as LLM's annotations that are different than annotations produced by humans) are in fact alternative correct annotations. Sometimes, these "incorrect" annotations can be even more specific.
- Future system improvement can be based on a user feedback and incorporated through active learning

LLM ANNOTATION SCHEMA



LLM PROMPT EXAMPLE

In order to extract correct information from articles it was necessary to develop prompts which were passed to LLM. Here is an example of such query for column “condition”:

You are an annotator for a biotech company, working with a dataset of biological samples.

Your task is to fill in the empty '{column_name}' column of a csv file with the relevant disease for each sample.

Use "negative" for samples that are healthy or where the health status isn't clearly mentioned. Avoid adjectives like "severe" and provide the general disease name (e.g., use "asthma" instead of "severe asthma"). If there are multiple conditions, select the primary one being studied.

Always use the full, non-abbreviated disease names.

BIO_DESCRIPTION_FRAGMENTS: {bio_description_fragments}

INPUT CSV: {csv}

TARGET/EXPECTED	PREDICTED BY MODEL
Blood leukocytes	Blood
Respiratory epithelium	Airway
Lung	Lung tumour
Non-small cell lung carcinoma	Non small cell lung cancer
Hepatocellular carcinoma	Liver cancer
Esophageal cancer	Esophageal cancer
Prostate carcinoma	Prostate cancer
Cervical cancer	Cervical squamous epithelial cancer
Breast cancer	Breast invasive ductal carcinoma
Colorectal cancer	Colorectal adenocarcinoma
Colorectal cancer	Gastroenteritis
Common cold	Hypertension
Coronary artery disease	Hyperlipidemia

Table 1: Example annotations for fields “tissue” and “condition”. One can observe that very often annotations returned from LLM are alternative, correct answers.

REFERENCES

[1] https://www.ncbi.nlm.nih.gov/geo/
[2] https://pubmed.ncbi.nlm.nih.gov/
[3] https://www.ncbi.nlm.nih.gov/pmc/

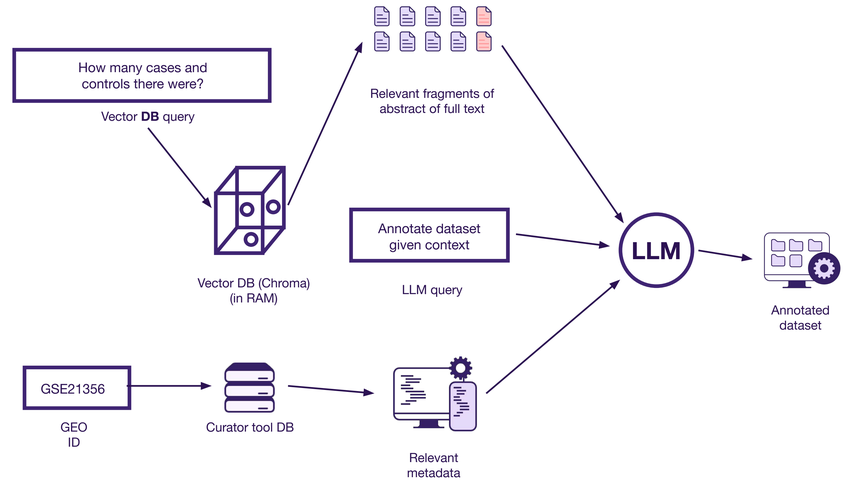


Figure 1: Technical overview of the pipeline

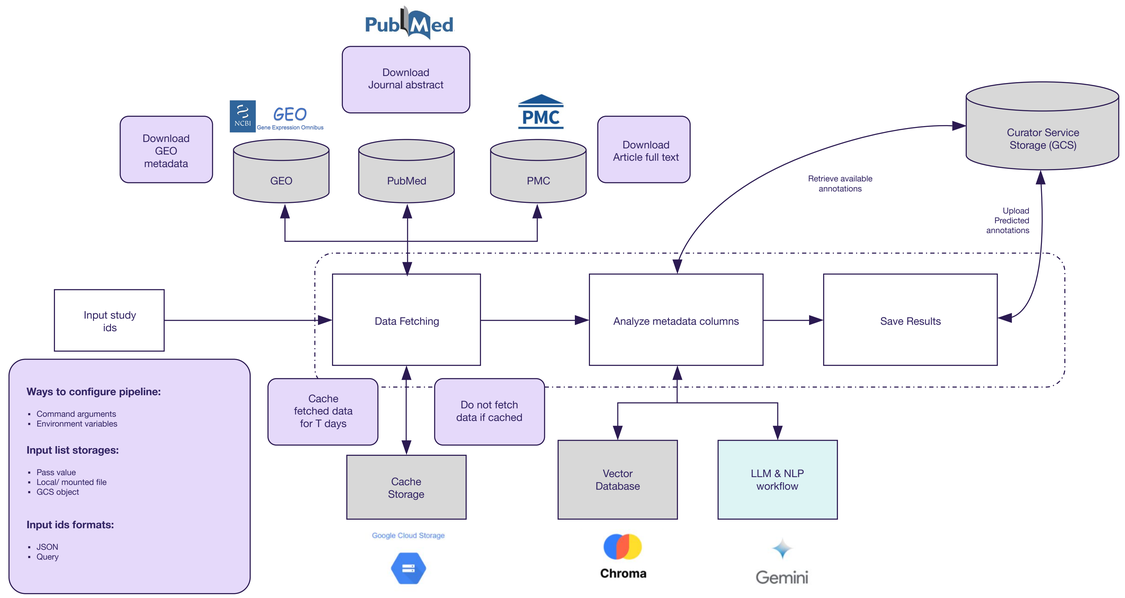


Figure 2: Pipeline overview with model and data sources [2], [3].

COLUMN	Condition	Tissue	Drug	Intervention	Average
Accuracy	0.67	0.67	0.80	0.93	> 0.80

Table 2: Accuracy of the annotations for different ontologies. One can observe that even with a strict validation regime where we require annotations to be exact, the solution yields high accuracy, and on average the accuracy metric is higher than 80%. Observed accuracy is lowered because of the validation protocol.