Universidad
Carlos III de Madrid

Machine Learning for Health

Natural Language Processing

Academic Research Assistance System

Authors:

Pablo Peiro Corbacho

Ana González Aranda

Celia de la Fuente Montero

Paula Martín Palomeque

May 2025

1. GitHub Repository

The full source code and relevant resources for the project are publicly available on GitHub at the following URL: https://github.com/PpeiroUC3M/RAG_NLP_MasterMLH

2. System Design Overview

This project uses the arXiv dataset from Kaggle, which includes over 1.7 million abstracts and metadata records from arXiv.org. The dataset was selected for its academic relevance, large scale, and well-structured content, making it well-suited for retrieval tasks in scientific domains.

The backend is implemented in Python and follows a Retrieval-Augmented Generation (RAG) approach. It uses langdetect to identify the input language and translates non-English queries to English via the Helsinki-NLP translation models. The translated query is embedded using nomic-embed-text and compared against stored embeddings in ChromaDB. Top matching documents (the number is chosen by the user) are retrieved and filtered by category and a similarity threshold before being passed as context to the prompt and then to the Large Language Model (LLM) "Llama 3.1:8B", which generates an informative grounded response to the researcher. The final output includes the retrieved documents (with metadata) and the LLM-generated response, which is translated back to the input language (excluding titles).

3. Configuration and Optimization

To prepare the arXiv dataset, we implemented a preprocessing pipeline. Initially, we parsed the metadata file to extract essential fields (id, title, abstract, category, authors, and doi). Entries lacking a title or abstract were discarded, and missing values in other fields were filled with a placeholder to maintain consistency. We then removed LaTeX commands, comments, and non-textual environments, replacing mathematical formulas with a [FORMULA] token and normalizing whitespace. Finally, we concatenated each paper's title and abstract into a single content field.

Data is stored in a persistent ChromaDB instance located at ./vector_database, structured as a single collection that holds abstract texts and associated metadata of arXiv papers. Document and query embeddings are generated using the nomic-embed-text model, which supports a context window of 8192 tokens and produces embeddings with a dimensionality of 768.

In the system, the prompt is crafted with explicit instructions that constrain the LLM only to use the information contained in the retrieved documents, summarize them, and compare with the research direction. Furthermore, a similarity threshold of 0.7 cosine similarity, is applied during document retrieval to filter out irrelevant or misleading context, a common cause of hallucination in RAG setups.

The llama3.1:8b model was used for text generation with the parameters temperature = 0.2, top_p = 0.85, and repeat_penalty = 0.8. This configuration was chosen to ensure accurate and coherent responses by limiting randomness and reducing repetition, making it well-suited for a research assistant context.

As part of its multilingual functionality, the system incorporates an automatic translator based on Helsinki-NLP (opus-mt) models, which support input in more than 25 languages, including Spanish, French, German, Arabic or Russian. This broadens the system's accessibility to a wide range of international users, allowing them to formulate queries in their native language and receive appropriately translated responses.

5. System Evaluation

To evaluate the performance and robustness of the system we assessed its behavior focusing on how effectively the system retrieves relevant documents, generates accurate summaries, maintains language fidelity and prevents hallucinated content.

In terms of relevance, each user query is embedded and compared against a vector database using cosine similarity. The predefined similarity threshold ensures that only documents with a strong semantic match to the query are considered, minimizing the inclusion of unsupported or speculative content. To evaluate this step we implement the recall@K evaluation metric on a small test set of 10 queries verifying if the model could retrieve relevant documents for those queries and we obtain: recall@1 = 0.8, recall@3=0.9, recall@5=1.0 and recall@10=1.0.

Language consistency is a critical feature. Query language is automatically and accurately detected using langdetect when there is sufficient context (more than 30 words preferred). Then queries and responses are correctly translated using Helsinki-NLP translation models, only having problems when they encounter technical concepts.

In terms of time performance, document retrieval via RAG is highly efficient. The LLM inference speed is largely dependent on system resources, but it remains generally responsive. Translation, however, tends to be slower than expected.

6. Conclusions and Limitations

In conclusion, our research assistance system successfully integrates semantic search based on embeddings with local LLM reasoning, providing an effective environment for exploring scientific literature. The pre-indexing of documents and the structured prompt design ensure that the model's responses are relevant and aligned with the retrieved context.

In summary, concrete measures have been taken to minimize hallucinations, one of the most common challenges in generative models. The combination of a similarity threshold to filter relevant documents and a strict policy that prevents the model from being invoked in the absence of adequate information reinforces the system’s reliability. These decisions, along with explicit prompt instructions to restrict responses to the available content, help ensure accurate answers.

However, the system does present several limitations. First, while the Helsinki models cover a wide range of languages, their accuracy may degrade slightly in scientific domains or when handling very short phrases. Related to this, language detection based on langdetect may be unreliable when dealing with short queries (less than 30 words), which can result in incorrect classifications and, consequently, unnecessary or inaccurate translations.

In addition, although nomic-embed-text yields good results in general semantic similarity tasks, it is not specifically optimized for the scientific or academic domain. This may hinder precise retrieval in specialized areas, especially when dealing with technical terminology or idiomatic expressions common in research contexts. Also as the vector database consists of abstracts, the retrieved documents have limited information and the LLM has less knowledge for the response.

In future works, it would be advisable to replace langdetect with more robust language classifiers to expand translation support, and also to integrate models trained specifically on scientific corpora.