

SciBot - Chatbot for analysis of scientific papers

Natural Language Processing 2024

Mid-term presentation

Michał Gromadzki, 313356

Jakub Piwko, 313451

Kacper Skonieczka, 313505

Grzegorz Zakrzewski, 313555

December 11, 2024

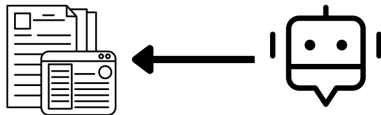
Project Topic: SciBot

SciBot: Your Research Assistant

A chatbot designed to enhance the knowledge of Large Language Models (LLMs) through a Retrieval-Augmented Generation (RAG) system, leveraging a customizable database of scientific articles.

Key Features:

- Utilizes scientific papers reviewed during our master thesis.
- Answers highly specific questions about state-of-the-art algorithms and methods.
- Provides sources for claims directly from the article database.
- Aims to assist students and researchers in their daily academic and research work.



Primary Objectives:

- Develop an effective RAG (Retrieval-Augmented Generation) system capable of extracting information from a wide range of scientific articles.
- Design a user-friendly chatbot interface for efficient, fast, and seamless interaction with the database.
- Compare various open large language models and embedding models.
- Evaluate embeddings using metrics like:
 - Hit Rate
 - Maximal Marginal Relevance (MMR)
- Assess chatbot responses using the LLM-evaluates-LLM approach with metrics:

No open dataset was used. Why?

- Articles are highly relevant to data science, aligning with our thesis topics.
- Familiarity with the articles' content facilitates easier initial evaluation.
- Continuous growth of the article database mirrors real-world knowledge expansion, ideal for testing RAG systems.

Current Progress:

- Database includes nearly 40 scientific articles.
- Articles can be accessed here: **Articles Database**.

What We Already Achieved?

Using the LangChain framework, we have implemented:

- **PDF Loader:** Reads and processes articles.
- **Text Splitter:** Divides articles into manageable chunks.
- **Embeddings Generation:** Created using `hkunlp/instructor-xl` model from Hugging Face.
- **FAISS Indexing:** Efficient storage and retrieval of document embeddings.
- **Chatbot Prototype:** Functional integration with FAISS database and powered by `llama3.1:latest` LLM.

Evaluation Strategy:

- For each question, we retrieved 6 article fragments as context using our RAG system.
- Chatbot responses were compared to **ground truth** generated by GPT-4o-mini, provided with the same context.
- Qualitative comparison of responses against the retrieved context was conducted to assess accuracy and relevance.



Example I

Question: Please give some examples of metaheuristics.

Model Response:

Some examples of metaheuristics include Simulated Annealing (SA), Tabu Search (TS), Evolutionary Algorithms (EAs), Ant Colony Optimization algorithm (ACO), and Particle Swarm Optimization (PSO).

Ground Truth:

Some examples of metaheuristics include Simulated Annealing (SA), Tabu Search (TS), Evolutionary Algorithms (EAs), Ant Colony Optimization (ACO), and Particle Swarm Optimization (PSO).

Example II

Question: "What is the idea behind the AAC approach called "GGA"?"

Model Response:

A very specific question! After some research, I found that GGA stands for **Generalized Gradient Approximation**. It's an **approximation method used in density functional theory (DFT)** to calculate the exchange-correlation energy of a system. In DFT, the total energy of a system is divided into three parts: the kinetic energy, the electron-electron interaction energy, and the exchange-correlation energy. The GGA approach is a ...

Ground Truth:

The GGA (**Genetic Algorithm with Genetic Engineering**) approach, as described in the context, integrates a surrogate model to evaluate the quality of new configurations within its operation. It combines a competitive population, which competes in races, with a non-competitive population, allowing the winning configurations from the competitive races to be recombined with individuals from the non-competitive group. The crossover operator generates new individuals based on ...

Example III

Question: What is the idea behind the PAP approach called "GLOBAL"?

Model Response:

I don't have information on a concept or methodology called 'PAP' that includes an approach called 'GLOBAL'. It's possible that it might be specific to particular fields, industries, or research. If you could provide more context or details about what PAP and GLOBAL refer to in your query, I'll do my best to offer a helpful response.

Ground Truth:

The PAP approach called "GLOBAL" focuses on optimizing performance by exploiting the advantages of different population-based algorithms through a parallel framework. Its main goal is to improve the overall performance across a diverse range of multi-objective problems by employing multiple algorithms that operate independently and in parallel. This strategy allows the algorithm portfolio to explore the search space globally, providing a higher chance of finding better solutions than any single algorithm could achieve on its own.

Key Focus Areas:

- Expand article database and generate a wider variety of test questions.
- Evaluate and optimize:
 - Embedding models for improved retrieval.
 - LLMs for enhanced response accuracy.
- Build a user-friendly Streamlit application with:
 - Intuitive interface for interaction with the chatbot.
 - Clear separation of backend and frontend functionalities.

The end

Thank you for listening!