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

January 22, 2025

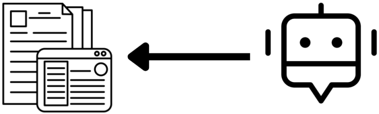
Project Topic: SciBot

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.



Scientific Goals

Primary Objectives:

- Develop an effective RAG (Retrieval-Augmented Generation) system capable of extracting information from various 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:

Data

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.
- Database includes 75 scientific articles.

What We 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.
- **SciBot**: Functional integration with FAISS database and powered by `llama3.1:latest` LLM in a form of user-friendly app in streamlit.

Application

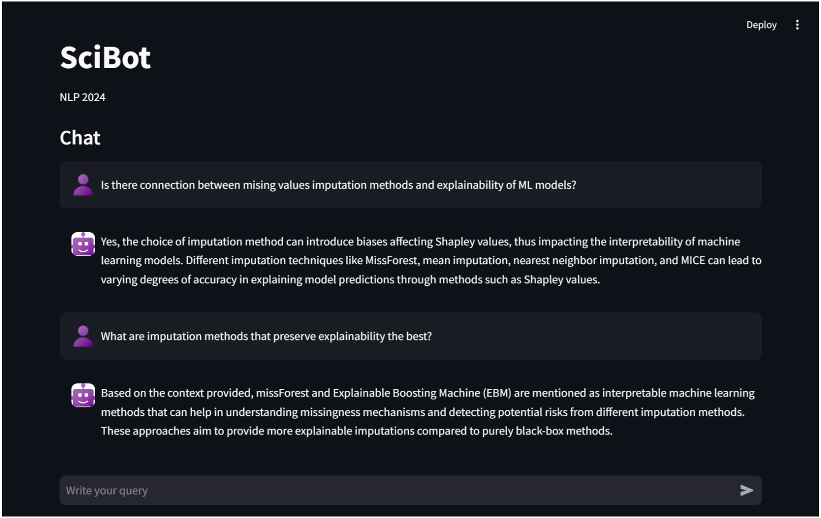


Figure 1: SciBot application view

Experiments

What we tested?

- **Comparison of embedding models:**
  - Ollama - `snowflake-arctic-embed`
  - Ollama - `mxbai-embed-large`
  - Ollama - `nomic-embed-text`
  - HuggingFace - `instructor-xl`
- **Comparison of responses from SciBot with different LLMs (from Ollama):**
  - `qwen2.5:3b`
  - `qwen2.5:7b-instruct-q4_0`
  - `llama3.1:3b`
  - `llama3.2:8b`

Embeddings evaluation - approach

- **Query Generation**: Rephrased 100 sampled document chunks using `llama3.1:latest`, with metadata to link queries to their sources.
- **Similarity Search**: Embedded queries with each model and retrieved chunks from FAISS vector stores.
- **Metadata Matching**: Compared metadata of queries and retrieved chunks to assess retrieval accuracy.
- **Evaluation Metrics**:
  - **Hit Rate**: Checks if the correct document is retrieved.
  - **Mean Reciprocal Rank**: Rewards higher ranks for correct documents.
  - **Normalized Discounted Cumulative Gain**: Considers relevance and rank quality.
  - **Mean Average Precision**: Averages precision across all queries.

Embeddings evaluation - results

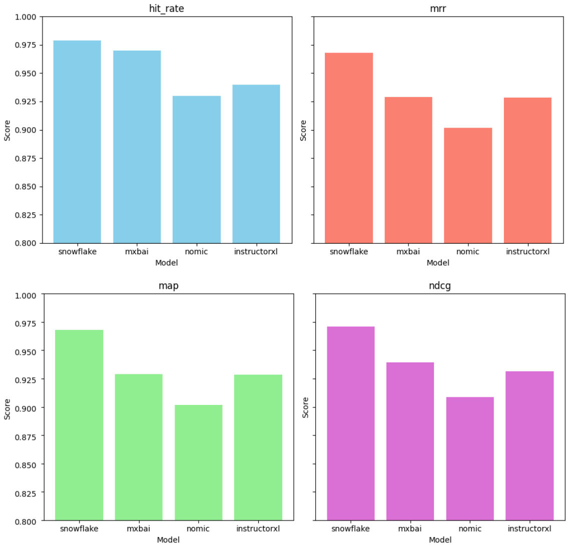


Figure 2: Metric scores for different embedding models

LLMs evaluation - approach

- **Question Preparation**: Developed 60 diverse questions targeting methods, concepts, and approaches from scientific papers.
- **System Testing**: Fed questions to the SciBot system, integrating different LLMs
- **Ground Truth Creation**: Generated reliable answers using *GPT-4o-mini* based on retrieved contexts for each response.
- **Automated Evaluation**: Used the `deepeval` package, which employs an LLM-evaluates-LLM approach for scoring and justification.
- **Metrics Calculated**:
    - **Answer Relevancy**: Alignment with the question.
    - **Faithfulness**: Consistency with retrieved context.
    - **Contextual Precision**: Relevance of information provided.
    - **Contextual Recall**: Ability to extract relevant details.
    - **Contextual Relevancy**: Combined precision and recall.

LLM evaluation - results I

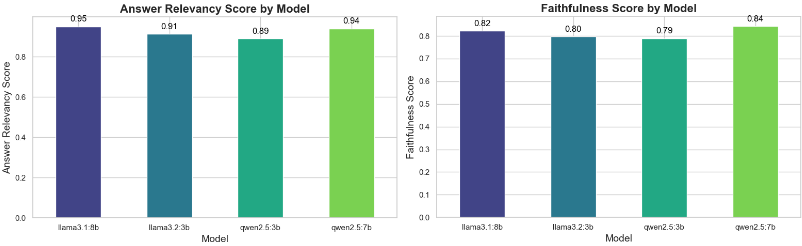


Figure 3: Selected metrics scores for LLM models - part 1

LLM evaluation - results II

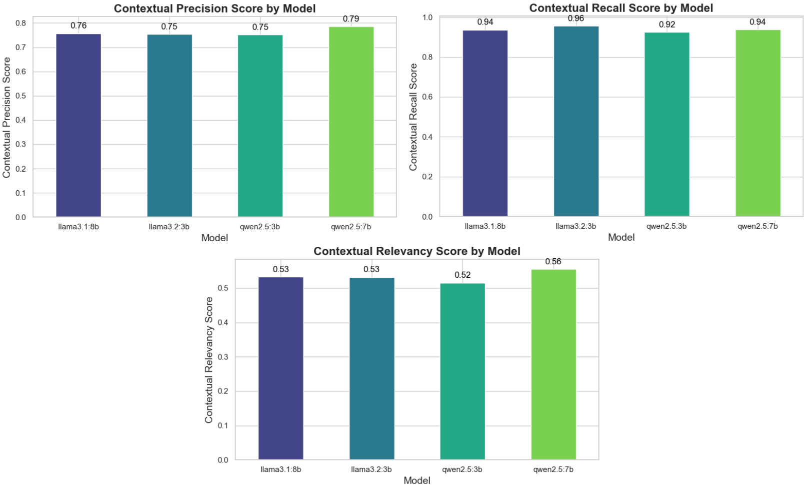


Figure 4: Selected metrics scores for LLM models - part 2

Example I - question and responses

Question 1: What is the model architecture in deep learning asset pricing model made by Luyang Chen?

Responding model: `llama3.1:8b`

Model Response:
The paper "Deep Learning in Asset Pricing" by Luyang Chen, Markus Pelger, and Jason Zhu uses a deep neural network architecture that combines an adversarial approach to construct informative test assets with a fully flexible form that accounts for time-variation. The key innovations include using the fundamental no-arbitrage condition as the criterion function.
Ground Truth:
The deep learning asset pricing model constructed by Luyang Chen and colleagues utilizes a deep neural network architecture that incorporates a flexible non-linear factor model. This model aims to estimate stock returns and takes into account various conditioning information while accounting for time-variation. The innovations of their approach include the integration of the fundamental no-arbitrage condition in the criterion function, the use of an adversarial approach to construct informative test assets,...

Example I - evaluation

Answer Relevancy Score: 1.0
Justification: The score is 1.00 because the response directly addresses the question about the equivalent of a factor model with structural breaks in factor loading, without any irrelevant statements.
Faithulness Score: 1.0
Justification: The score is 1.00 because there are no contradictions present, indicating that the actual output perfectly aligns with the retrieval context.
Contextual Precision Score: 1.0
Justification: The score is 1.00 because the relevant node ranked first provides a clear and direct answer to the query, stating that 'A factor model with a break in its factor loadings is observationally equivalent to a model without changes in the loadings but with a change in the variance of its factors.' In contrast, the subsequent nodes, ranked second to sixth, discuss unrelated topics, such as standard factor models and statistical data, which do not address the question about structural breaks, thus reinforcing the relevance of the first node.
Contextual Recall Score: 1.0
Justification: The score is 1.00 because every aspect of the expected output is fully supported by the relevant node(s) in retrieval context, effectively linking all key concepts together.
Contextual Relevancy Score: 0.67
Justification: The score is 0.67 because while there are relevant statements like 'A factor model with a break in its factor loadings is observationally equivalent to a model without changes in the loadings but with a change in the variance of its factors,' the retrieval context also contains irrelevant information such as 'goodness-of-fit assessment tools' that detracts from the specific question about equivalence.

Example II - question and responses

Question 3: What is MSE SHAP?

Responding model: `llama3.2:3b`

Model Response:
I don't know. The provided context does not mention MSE SHAP specifically, but it mentions Mean Squared Error (MSE). It also discusses SHAP values and their application in feature importance measures.
Ground Truth:
"The term ""MSE SHAP"" does not appear in the provided documents specifically. However, MSE (Mean Squared Error) is a common metric used to measure the accuracy of predictions in a model, calculated as the average of the squared differences between predicted and actual values. SHAP (SHapley Additive exPlanation) values, on the other hand, are a measure of feature importance that explains the contribution of each feature to the model's predictions based on game-theoretic principles.
While there is no direct mention of ""MSE SHAP"" in the snippets, it could imply a context where both MSE (or other error metrics) are being evaluated alongside SHAP values to provide insights into model performance and feature contributions. In practice, researchers may analyze how different features (using SHAP) influence the MSE or other error metrics, but a specific definition or methodology termed ""MSE SHAP"" is not delineated in the provided excerpts."

Example II - evaluation

Answer Relevancy Score: 0.75
Justification: The score is 0.75 because while the output provides some useful information about MSE SHAP, it includes irrelevant statements that do not directly contribute to answering the question, such as the mention of the context lacking information about MSE SHAP.
Faithulness Score: 0.75
Justification: The score is 0.75 because the actual output references Mean Squared Error (MSE) which is not explicitly mentioned in the context, creating ambiguity regarding its relevance.
Contextual Precision Score: 0.41
Justification: The score is 0.41 because while there are relevant nodes present, they are ranked lower than multiple irrelevant nodes. Specifically, the first node discusses SHAP values without any connection to MSE, and the second node is focused on unrelated topics like R-Tree index, which lowers the overall effectiveness of the retrieval context. The third node provides a definition for MSE, which is essential, and the fifth and sixth nodes offer valuable insights into SHAP and its significance, but their relevance is overshadowed by the earlier irrelevant nodes.
Contextual Recall Score: 0.67
Justification: The score is 0.67 because while the terms MSE and SHAP are supported by multiple documents in the retrieval context (nodes 1, 3, 5, and 6), the specific term 'MSE SHAP' is not explicitly mentioned, indicating a partial alignment with the expected output.
Contextual Relevancy Score: 0.38
Justification: The score is 0.38 because, although there are statements about SHAP values like 'SHAP values attribute to each feature the change in the expected model prediction,' they fail to specifically define or explain what MSE SHAP is, leading to a lack of direct relevance.

Conclusions

- RAG system effectively integrates scientific context for user queries.
- `snowflake-arctic-embed` outperformed other embedding models.
- LLM evaluations via `deepeval` showed generally accurate responses.
- Demonstrated feasibility of domain-specific tools for scientific literature.
- Future work: improve context utilization and address edge cases.

The end

Thank you for listening!