AI Chatbot for Life Sciences Research
Powered by Retrieval-Augmented Generation

Venkata Mani Sivasai Shanmukha Rajendra

Abstract

In this work, I address the challenge of efficiently retrieving and synthesizing relevant information from vast collections of life sciences research papers. Traditional methods of searching and retrieval in this domain are often slow, manual, and can result in inaccurate or incomplete responses. To overcome these limitations, I implement a Retrieval-Augmented Generation (RAG) model that combines the strengths of retrieval-based and generative approaches for answering complex research queries. The model is evaluated against a baseline BERT model using various metrics such as relevance, fluency, coverage, and BERT score. Results show that the RAG model outperforms the BERT model in terms of query response accuracy and efficiency, providing more relevant and contextually accurate answers. This approach offers significant improvements in the retrieval and presentation of life sciences information, making it more accessible and actionable for researchers.

Introduction

In the rapidly evolving field of life sciences, researchers are continually faced with the challenge of accessing vast amounts of scientific literature and data. The process of searching for relevant information, synthesizing findings, and extracting meaningful insights is often time-consuming and can lead to inefficient use of resources. Traditional search methods, based on keyword matching and Boolean logic, fail to capture the complex and nuanced nature of scientific queries, often returning irrelevant or incomplete results. This is particularly problematic for complex questions, where information is distributed across multiple papers and data sources.

To address these challenges, there is a growing need for more intelligent systems capable of understanding the context of research queries and retrieving accurate, relevant information quickly. Retrieval-Augmented Generation (RAG) models offer a promising solution, as they combine the strengths of both retrieval-based and generative approaches. By retrieving relevant information from a knowledge base and using a generative model to produce contextually rich responses, RAG models are well-suited to answer complex queries with high accuracy.

In this work, I propose a system that leverages a RAG model for answering life sciences research queries. The system first retrieves relevant documents using a dense retrieval mechanism and then uses a generative model to generate an accurate, coherent, and contextually appropriate response. The solution aims to streamline the research process by reducing the time spent on data retrieval and synthesis, while also improving the quality of answers provided to researchers.

Background

1. Information Retrieval in Life Sciences:

- **Concept Overview:** Life sciences research relies on retrieving specific information from large databases (e.g., PubMed). Traditional methods like TF-IDF and Boolean search match query keywords to documents but lack semantic understanding, often resulting in incomplete or irrelevant answers.

- **Previous Work:** TF-IDF and Boolean search are efficient but unable to grasp contextual meaning, limiting their ability to handle complex queries in life sciences.

2. Retrieval-Augmented Generation (RAG) Framework:

- **Concept Overview:** RAG combines retrieval-based methods with generative models for more accurate query responses. The retriever fetches relevant documents, and the generator synthesizes a response using the retrieved information.

- **Previous Work:** Introduced by Lewis et al., RAG improves question answering and fact verification by integrating Dense Passage Retrieval (DPR) and GPT-based models for contextual, multi-source answers.

3. BIOASQ Dataset for Life Sciences Query Answering:

- **Concept Overview:** The BIOASQ dataset is a large-scale collection of life sciences-related questions, documents, and ideal answers, designed for the development and evaluation of question-answering systems. Each question is associated with a set of relevant documents (e.g., PubMed articles) and a detailed ideal answer. The questions in BIOASQ span various life sciences topics, requiring systems to retrieve and synthesize information from multiple sources to generate accurate and relevant answers.
- **Previous Work:** BIOASQ has been used to evaluate systems in the life sciences domain, specifically in tasks

like biomedical question answering and document retrieval. The dataset's detailed structure, including multiple document snippets per question, enables comprehensive testing of retrieval-augmented systems. Researchers have applied methods like traditional IR, BERT, and RAG to improve the accuracy and fluency of responses based on this dataset.

4. NLP and ML in Life Sciences Research:

- **Concept Overview:** NLP and ML techniques are transforming life sciences research by automating data extraction, analysis, and synthesis from vast scientific literature, enabling faster and more accurate insights.
- **NLP for Information Extraction**: Identifies and classifies entities (e.g., genes, diseases) and their relationships, making unstructured data more accessible.
- **ML for Data Mining**: Uncovers patterns in large biomedical datasets, revealing complex biological relationships.
- **Deep Learning for Text Generation**: Models like GPT and BERT generate context-aware answers, improving query responses.
- **Integration in Research Tools**: Enhances features like paper summarization, drug discovery, and disease prediction, speeding up research.
- **Previous Work:** NLP and ML have been applied in drug repurposing, genomic analysis, and literature mining, showing their potential to automate and refine research tasks in life sciences.

5. Evaluation Metrics:

- RAG model performance is assessed using relevance (quality of the response), fluency (grammatical correctness), coverage (completeness of the answer), and BERT score (measuring similarity to reference answers using BERT embeddings).
- These metrics are standard in NLP and QA, with BERT Score providing an advanced method to compare generated text to reference answers.

Related Work

In the life sciences domain, several advancements have been made in information retrieval, question answering, and text generation. This section reviews key academic papers, case studies, and projects that form the foundation for the work presented in this project.

1. **Information Retrieval in Life Sciences:**

- **PubMed and Biomedical Search:** Traditional search engines, like PubMed, rely heavily on keyword-based search methods such as TF-IDF and Boolean search. These methods are still widely used but struggle with complex, domain-specific queries due to their lack of semantic understanding and contextual relevance.
- **Case Study**: A study by Cohan et al. (2019) in "A Survey on Information Retrieval for Biomedical Text Mining" shows the limitations of traditional retrieval models and discusses the need for more advanced models that understand medical contexts.

2. **Retrieval-Augmented Generation (RAG):**
   - **Lewis et al. (2020):** The introduction of RAG models demonstrated the efficacy of combining retrieval-based techniques with generative models for answering complex queries. This hybrid approach has since been widely adopted for tasks like question answering and fact verification. Their approach integrates Dense Passage Retrieval (DPR) to fetch relevant documents and then passes them through a generative model (such as BART or GPT-3) to synthesize coherent and context-aware responses.
   - **Relevance to our Work:** This paper is foundational in supporting the use of RAG for life sciences research. I apply a similar RAG-based architecture to answer domain-specific queries, leveraging both retrieval and generation.

3. **BERT and Transformers for Question Answering:**
   - **BERT (Devlin et al., 2019):** BERT, a transformer-based model, revolutionized NLP by enabling bidirectional understanding of text. Fine-tuned for question answering tasks, BERT efficiently retrieves and predicts answer spans. However, it struggles with multi-document or complex queries requiring external knowledge.
   - **Case Study:** The SQuAD (Stanford Question Answering Dataset) benchmark demonstrated BERT's effectiveness in QA tasks but highlighted its limitations when data spans multiple documents. This is a key reason to opt for RAG in the project, as it can pull out from various sources, offering better context coverage.

4. **Other Methods and Why They Were Not Used:**
   - **TF-IDF and BM25:** These traditional retrieval methods were not chosen due to their inability to understand the semantic meaning of queries. They rely solely on surface-level keyword matching, which often leads to irrelevant results in complex life sciences queries.
   - **Knowledge Graphs:** Knowledge graphs have shown promise in structuring biomedical knowledge. However, their reliance on predefined entities and relationships limits flexibility when addressing dynamic, emerging research questions. My RAG model offers more adaptability by integrating both retrieval and generative capabilities.
   - **Rule-Based Systems:** While rule-based systems could offer precise control over query responses, they are time-intensive to design and difficult to scale in the life sciences domain. These systems also lack the flexibility to handle the complexity of natural language, which RAG better accommodates.

Project Description

Overview

The aim of this project is to develop a chatbot tailored for biomedical research powered by Retrieval Augmented Generation (RAG), designed to assist researchers and clinicians by delivering accurate, contextually rich answers derived from a wide range of biomedical research papers. This tool aims to accelerate scientific discoveries and enhance decision-making processes in life sciences and biomedical fields.

System Components

1. Data Collection and Preparation:

a) **Source Data:** The dataset consists of over 5000 records from BioASQ, which covers a broad range of biomedical topics, including genetic disorders, disease mechanisms, treatment protocols, and clinical research. Each record includes a question related to a biomedical topic, multiple relevant documents (e.g., PubMed articles), and ideal answers derived from those documents. However, due to resource constraints, the model has been trained using a subset of 200 records to begin with.

b) **Information Retrieval System:** The system utilizes embeddings generated by Dense Passage Retriever (DPR) for efficient information retrieval. DPR is used to create embeddings for both the user query and the abstracts extracted from the research papers. This allows the chatbot to retrieve contextually relevant information from the biomedical corpus and enhance the quality of responses.

2. Prompt and Query Initiation:

a) Query Processing for Relevant Information: The system begins by processing user queries related to biomedical topics. These queries are first embedded using the Dense Passage Retriever (DPR) model. The query embeddings are generated, capturing the semantic meaning of the user's question in a high-dimensional vector space. These embeddings are then used to match the query against the indexed biomedical research data stored in the system. The embeddings enable precise retrieval of relevant documents, ensuring that the system focuses on the most contextually appropriate information

b) Information Retrieval for Context Enhancement: After generating the query embeddings, the system leverages DPR to perform information retrieval by comparing the query embedding with the embeddings of abstracts stored in the system. The similarity scores (calculated through cosine similarity) between the user query embedding and the abstract embeddings are used to retrieve the most relevant documents or sections. These selected abstracts provide the necessary context, ensuring that the chatbot's responses are grounded in credible sources. This context enhancement is crucial for generating accurate, reliable answers based on the latest biomedical research.

3. Generative Model Integration:

a) **Language Models (GPT-3.5-Turbo):** The core of the chatbot's ability to generate contextually accurate responses is powered by GPT-3.5-turbo, accessed via the OpenAI API. GPT-3.5-turbo is designed to generate coherent and context-aware text, ensuring that the answers it provides are informed by the relevant biomedical context retrieved earlier.

b) **Embedding Generation:** The embeddings used for both the user query and the abstracts extracted from the biomedical research papers are generated using Dense Passage Retriever (DPR). DPR is a neural network-based approach for information retrieval that creates dense vector representations (embeddings) of both the documents and the user queries. These embeddings allow the system to compute similarity scores, which are key in identifying the most relevant documents for any given query. In addition to **DPR**, **BERT** is used to generate embeddings as a **baseline** comparison for evaluating the effectiveness of **DPR** in the retrieval process. BERT helps to provide a different perspective on embedding generation, which can be compared with **DPR** to assess performance improvements.

c) **Similarity-Based Document Retrieval:** Once the user submits a query, the system computes the embedding for the query using DPR. The next step is to calculate similarity scores between the query embedding and the embeddings of the abstracts stored in the system. These similarity scores measure how closely each abstract matches the user's query based on semantic meaning. The similarity score is computed using methods like cosine similarity, which measures the angle between two vectors in a high-dimensional space. A higher cosine similarity indicates that the abstract is more relevant to the query.

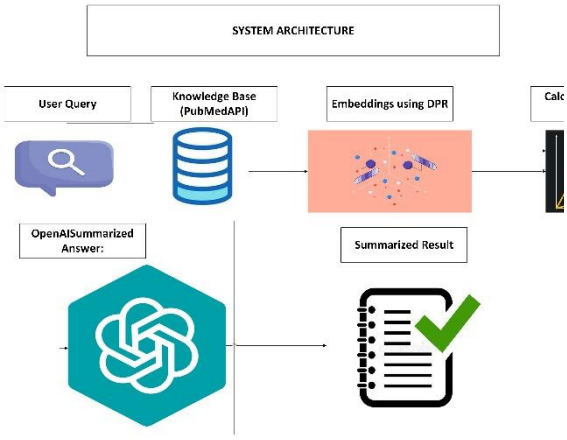
d) **Document Retrieval:** Based on these similarity scores, the system retrieves the most relevant documents or sections from the dataset. The abstracts with the highest similarity scores are selected for further processing. These documents provide the necessary context that GPT-3.5-turbo uses to generate informed, accurate answers to the user's query.

e) Augmentation via RAG: The retrieved documents are then passed to GPT-3.5-turbo as context. GPT-3.5-turbo synthesizes this information, generating the final response to the user query, ensuring that the answer is

not only relevant but also informed by authoritative biomedical sources.

4. System Architecture

a) The system integrates a user interface for researchers and clinicians to input queries.



b) The architecture includes an information retrieval layer, which fetches relevant data from the biomedical corpus using DPR-based embeddings, and a generative model layer that processes and synthesizes this data with GPT-3.5-turbo to generate responses.

5. Evaluation and Metrics:

a) **Cosine Similarity:** Used to measure the relevance of the generated responses (from both RAG and BERT models) compared to the ideal answers present in the source dataset.

[1] **Implementation:** The embeddings of the generated responses and the ideal answers are computed using the Dense Passage Retriever (DPR) model, and the cosine similarity score is calculated.

[2] **Purpose**: Higher similarity indicates that the generated response is more aligned with the ideal answer.

b) BERT Score: A metric used to evaluate the semantic similarity between the generated response and the ideal answer.

[1] **Implementation:** Bert Score is computed using the pre-trained BERT model for each generated response and its corresponding ideal answer.

[2] **Purpose:** A higher Bert Score signifies better quality in terms of semantic similarity.

c) Fluency (Grammar Checking): Evaluates the grammatical correctness of the generated response using a grammar checker tool (language_tool_python).

[1] **Implementation:** The number of grammar issues detected in the response is calculated, and a normalized fluency score is computed.

[2] **Purpose:** A higher fluency score indicates fewer grammatical issues and better overall readability.

d) ROUGE Score: Used to assess the coverage and similarity in terms of content between the generated response and the ideal answer, with a focus on recall-based evaluation.

[1] **Implementation:** The ROUGE-L score is computed using the rouge_scorer library, which evaluates the longest common subsequences between the generated response and the ideal answer.

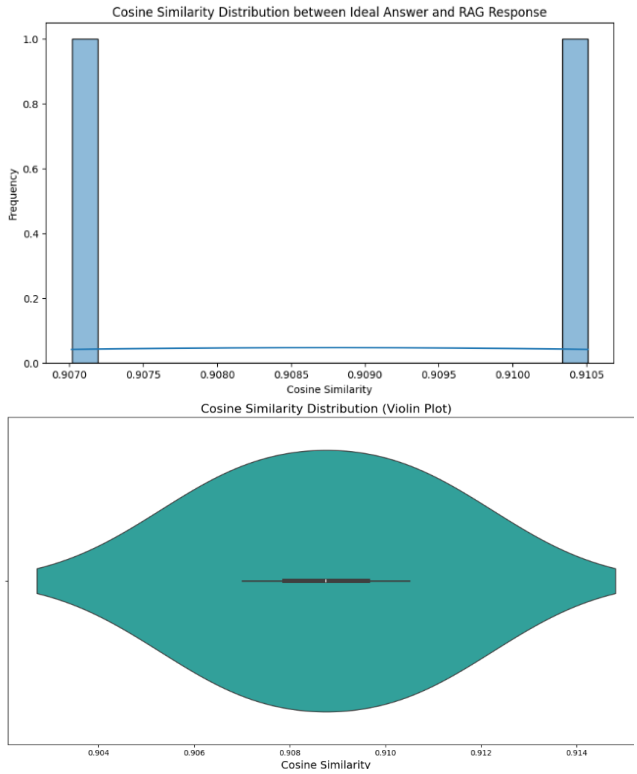
[2] **Purpose:** Higher ROUGE scores indicate better content coverage and alignment with the ideal answer.

Empirical Results

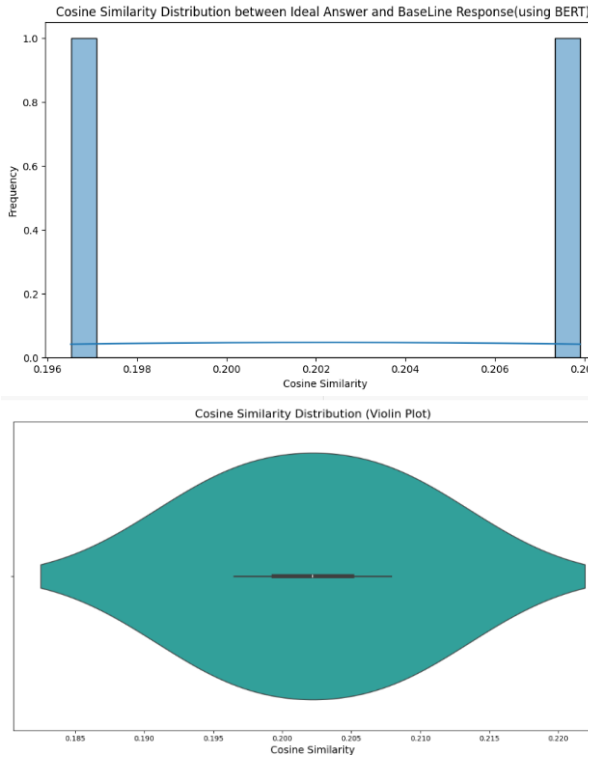
To evaluate the performance of the RAG-based model and BERT-based baseline, we calculated various metrics including Cosine Similarity, Bert Score, ROUGE Score, and Fluency, using a set of biomedical queries. The goal was to compare how well the RAG-based model and BERT model responded to these queries based on relevance, fluency, and content coverage. We also visualized the distribution of cosine similarity scores and compared the results between both models.

1. **Cosine Similarity Distribution:** The first metric used for evaluation was Cosine Similarity, which measures how close the generated responses (from both RAG and BERT) are to the ideal answers (ground truth). For this, we visualized the cosine similarity distributions for both the RAG and BERT models using histograms and violin plots.

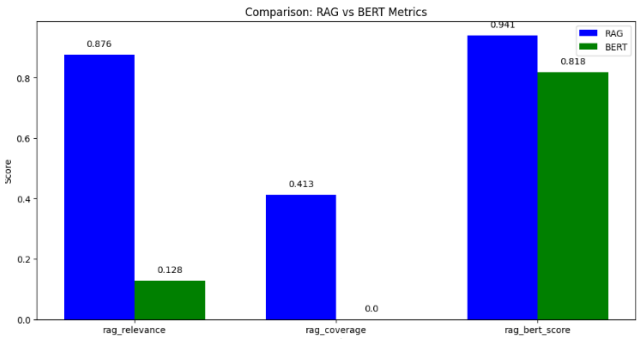
a) **RAG Response Cosine Similarity**: Computed the cosine similarity between the ideal answers and the responses generated by the RAG-based model. The resulting distribution was visualized using histogram and a violin plot. This analysis helps us understand how closely the RAG model's responses align with the ideal answers.



a) BERT Response Cosine Similarity: Similarly, calculated the cosine similarity between the ideal answers and the baseline responses generated by the BERT model. This distribution was also visualized using histogram and a violin plot.



2. **Comparison Between RAG and BERT Metrics:** Compared the performance of the RAG-based model and the BERT baseline using the following evaluation metrics:
   a) Relevance: How relevant the generated responses are to the query.
   b) Coverage: How well the responses generated cover the key information from the ideal answers.
   c) BERT Score: A metric that evaluates the semantic similarity between the generated response and the ideal answer.



Observations

1. **Cosine Similarity:** Observed that the RAG model had a higher cosine similarity compared to the BERT baseline, indicating that the RAG model's responses were closer to the ideal answers in terms of semantic similarity.
2. **Fluency and Quality:** Although RAG scored better on most metrics, further evaluation with metrics like fluency and ROUGE could give a deeper insight into the quality and completeness of the responses.
3. **Metrics Comparison:** When comparing the RAG and BERT models, the RAG model generally outperformed BERT in terms of relevance and coverage.

Experiments demonstrated that the RAG-based model excels in generating relevant and high-coverage responses in comparison to the baseline BERT model, especially in the biomedical research domain where context and precision are crucial.

Broader implications

1. **Accelerating Biomedical Research:** The chatbot accelerates the discovery of treatments and understanding of diseases by providing quick, context-rich answers from scientific literature, helping researchers stay ahead in complex fields like genetic disorders and gene editing.

2. **Improving Clinical Decision-Making:** Clinicians can make better-informed decisions by accessing real-time, accurate research and clinical insights, especially in rare diseases, enhancing patient outcomes.

3. **Enhancing Knowledge Accessibility:** The system democratizes access to high-quality research, supporting not only professionals but also students and non-experts in understanding complex biomedical concepts.

4. **Ethical and Privacy Concerns:** Ensuring the system handles data ethically, accurately, and without bias is crucial, especially when working with sensitive biomedical information.

5. **Reducing Human Error:** By reducing information overload, the chatbot helps minimize human error in both research and clinical practices, leading to better decision-making.

6. **Combating Misinformation:** The chatbot, rooted in credible sources, can counter misinformation in healthcare, ensuring that users receive accurate, reliable information.

7. **Long-Term Societal Impact:** Over time, the system can improve global healthcare by enabling better research collaboration, disease prevention, and public health outcomes.

8. **Expansion Potential:** Beyond biomedical research, this technology could extend to other sectors, benefiting industries like legal and educational fields with AI-driven insights.

Conclusion

1. **Summary of Results**: The project successfully implemented a chatbot leveraging advanced models like Dense Passage Retriever (DPR) and GPT-3.5 for generating meaningful responses in the biomedical domain. Through detailed performance metrics like cosine similarity, BERT score, and fluency checks, the chatbot demonstrated promising results in providing contextually relevant and accurate answers to user queries.

2. **Key Learnings**:

    a. The integration of RAG with GPT-3.5 has proven effective in answering biomedical queries, showcasing the importance of combining retrieval and generation models.

    b. Fine-tuning the model on domain-specific data (e.g., biomedical research) significantly improves accuracy and relevance.

    c. A deep understanding of NLP techniques such as embeddings, transformers, and performance evaluation metrics is crucial to building reliable systems.

Future Enhancements

1. **Scalability:** With more computational resources and time, scaling the model to handle larger datasets (beyond 200 records) could improve the chatbot's performance and accuracy.

2. **Advanced Retrieval Mechanisms:** Exploring more sophisticated retrieval techniques like hybrid search methods or incorporating multimodal data (e.g., images, graphs) could enhance the system's response quality.

3. **Real-World Application:** Extending the system to real-time clinical environments for decision support or integrating it with electronic health records (EHRs) to assist clinicians in daily practices could be an exciting avenue for future exploration

Advice for Future DS 5983 Students

1. **Focus on Data Quality**: Ensure that the data you use is clean, representative, and domain-specific for optimal performance.

2. **Experiment with Different Models:** Don't limit yourself to one approach, explore different retrieval and generation models to see which combination works best for your task.

3. **Evaluation is Key:** Always assess your model using relevant metrics and ensure to analyze your results critically to identify areas of improvement.

4. **Plan and Iterate:** Time management is crucial—define a clear roadmap but be flexible enough to adjust as new insights emerge during the project.

Collaborate and Seek Feedback: **Discuss your approach with peers or instructors for feedback to improve your project.**