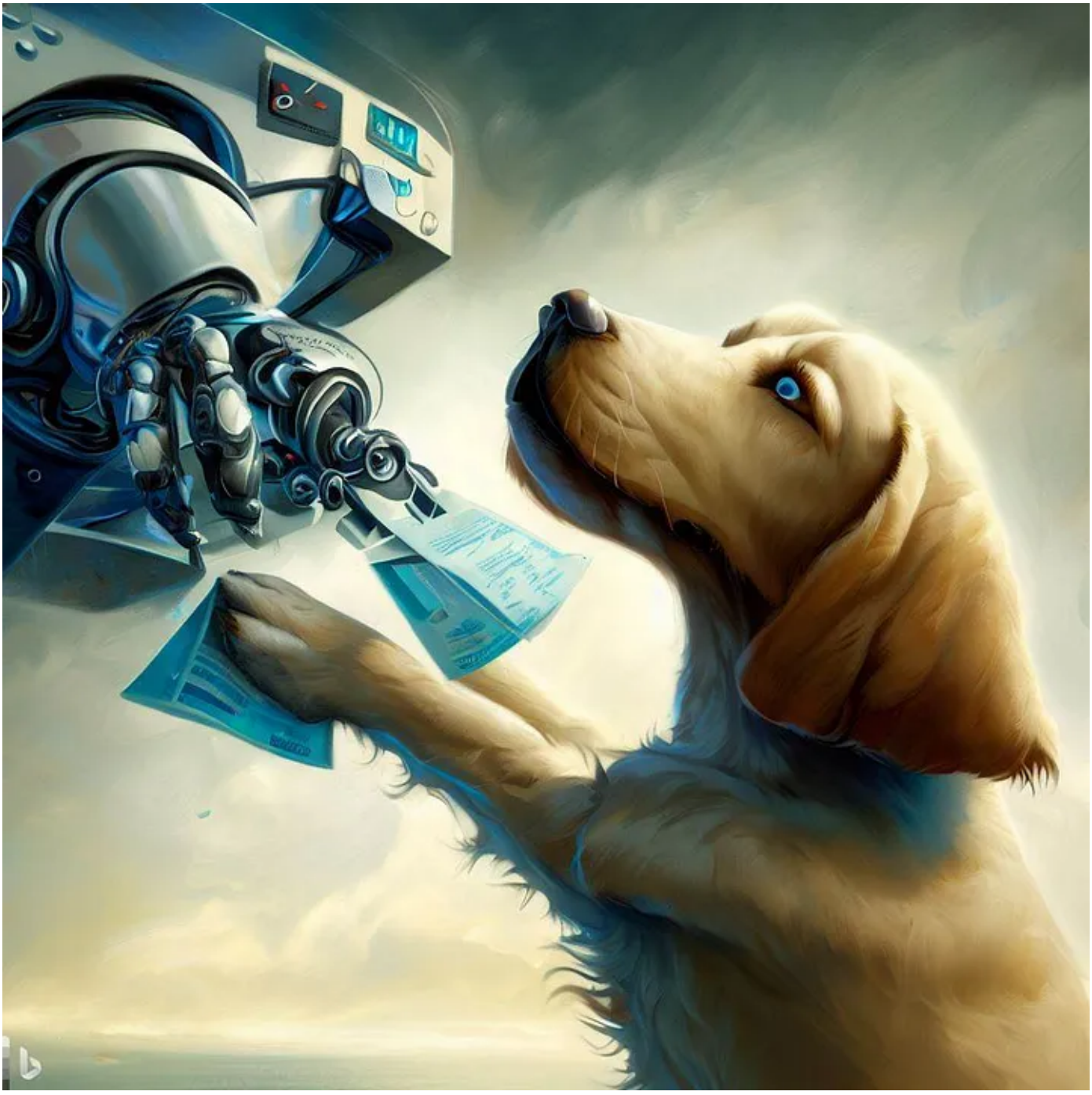

Retrieval-Augmented Generation with OpenAI Chat GPT and FAISS: A Game-Changer for Clinical Research

FHIRFLY · Follow
3 min read · Aug 25, 2023

Introduction

The advent of large language models like GPT-4 has revolutionized many fields, including healthcare and clinical research. However, these models are not perfect; one of the significant challenges they face is “hallucinating” or generating information that is not accurate or even plausible. In clinical research, where the stakes are incredibly high, such hallucinations can be problematic. So how can we harness the power of these models while also ensuring the accuracy of the information they generate? Enter Retrieval-Augmented Generation (RAG) coupled with FAISS.

The Challenge: Hallucination in Language Models

In machine learning, “hallucination” refers to the generation of statements or predictions that are not supported by the data. For example, if asked to summarize a medical paper, a hallucinating model might include details or conclusions that are not present in the paper. These hallucinations pose a significant risk, especially in fields like clinical research where accuracy is paramount.

What is Retrieval-Augmented Generation (RAG)?

Retrieval-Augmented Generation (RAG) is a paradigm that combines the best of both worlds: the powerful generative capabilities of models like GPT-4 and the retrieval capabilities of information databases. In a typical RAG setup, the model first retrieves relevant documents or data snippets from a large corpus and then uses that retrieved information to generate a response. This two-step approach significantly reduces the risk of hallucination as the model has a “reality check” against real-world data.

How does RAG Work?

1. Retrieval Phase: Given a query (e.g., a clinical question), the model searches a large database to find relevant documents or snippets.

2. Generation Phase: The model uses the retrieved information to generate a response, ensuring that the output is grounded in real-world data.

FAISS: The Secret Sauce for Efficient Retrieval

FAISS (Facebook AI Similarity Search) is an efficient similarity search and clustering library. It plays a crucial role in the retrieval phase of RAG. FAISS enables the model to search through large databases quickly, making the entire process more efficient and accurate.

Putting it All Together: A Practical Example

Let's look at a snippet from a Jupytr Notebook that demonstrates how RAG and FAISS can be used for clinical research.

In this example, the clinical question serves as the query for the RAG model. The model would then retrieve relevant medical papers or data snippets using FAISS before generating an evidence-based response. The Notebook constructs a PICO Clinical Query from any free text clincial query. and searches the NCBI database for mesh term in the individual sentence elements. It then queries PubMed for a list of Documents that meet the Meshed terms. The results are then vectorized and searched using FAISS next nearest neighbor search and the top x number of results are returned. The results are used as part of the prompt for Chat GPT to output and Evidenced Based Medicine result.

Conclusion

Retrieval-Augmented Generation (RAG) combined with FAISS offers a powerful solution for mitigating the risks of hallucination in large language models. By grounding the generative capabilities of models like GPT-4 in real-world data, we can significantly improve the accuracy and reliability of the information generated. This is particularly beneficial in high-stakes fields like clinical research, where every piece of information can have far-reaching implications.

Note: For a complete example, you can find the full notebook on GitHub.

Rag | Machine Learning | ChatGPT | OpenAI | Healthcare