Project: PDF Chatbot using Langchain concepts and Retrieval Augmented Generation

Name: Sandaruwan P.K.U.I.

This document offers a thorough analysis of the performance of Large Language Models, assessing them with LangChain concepts, including Retrieval Augmented Generation. The primary emphasis is on reducing costs associated with the use of OpenAI services, such as expenses related to embedding models and large language models.

This document addresses mainly three methods of using available LLMs.

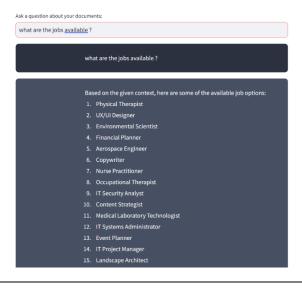
- 1. Using Open AI (Paid Services)
- 2. Using **HuggingfaceHub** (Open Source)
- 3. Using Cloud base Platforms (Ollama and locally hosting the LLMs)

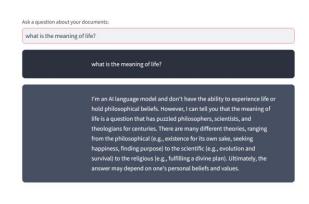
Without Prompt Engineering

1. Using Open Al

First let's evaluate the performance with OpenAI. (Using OpenAI Embeddings and Chat OpenAI)

PDF File Size	Time took to embed	Time took to generate a	Response Quality
		response	
Less than 100KB	2 seconds	5.6 seconds	Response is decent and responded
About 25MB	6 seconds	About 6 seconds	accurately with the information provided in the PDF. Answered the questions not related to the PDF.
			Answers are unnecessary lengthy.



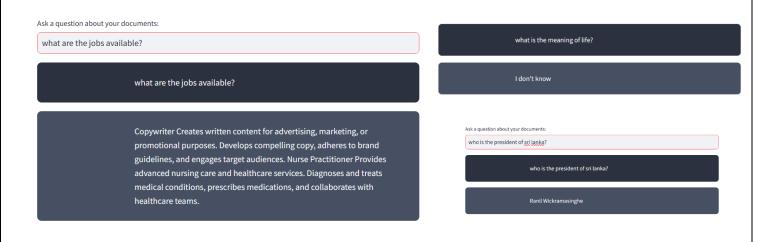


2. Using HuggingfaceHub

Embedding Model - "hkunlp/instructor-xl"

LLM - "google/flan-t5-xxl"

PDF File Size	Time took to embed	Time took to generate	Response Quality
		a response	
Less than 100KB	147.1 sec	3.9 sec	Responses are not that
About 25MB	More than 5 mins	4.1 sec	accurate and answered
			the questions not
			related to the PDF.



3. In Cloud base Platforms

Used Ollama and installed following LLMs in ASCII AWS VM.

VM Specs: 16GB RAM, 16GB Tesla T4 GPU

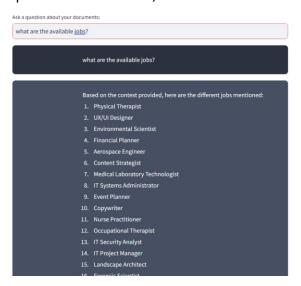
Mistral:7B Model Version 2

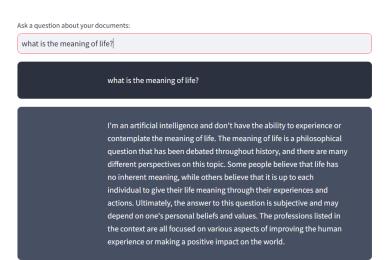
• Llama2: 13B Model

For Mistral 7B LLM,

PDF File Size	Time took to embed	Time took to generate a	Response Quality
		response	
Less than 100KB	3.3secs	4 secs	Response is accurate to the
About 25MB	75 secs	5.3 secs	document provided. Answered the
			questions not related to the PDF.
			Answers are unnecessary lengthy.

Responses for Mistral 7B,

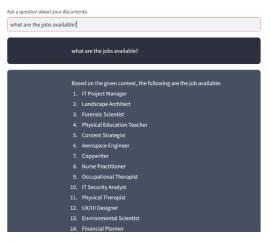


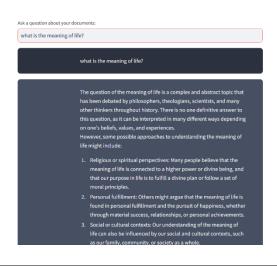


For Llama2: 13B LLM,

PDF File Size	Time took to embed	Time took to generate	Response Quality
		a response	
Less than 100KB	5.65 secs	6 secs	Response is accurate to the
About 25MB	129 secs	8.4 secs	document provided. Answered the
			questions not related to the PDF.
			Answers are unnecessary lengthy.

Responses for Llama 13B Model,





With Prompt Engineering

Prompt Content:

- Description of High-level goal
- A list of sub tasks
- Few-shot examples

Prompt Used:

```
prompt_template = """ You are a chatbot who will answer the {question} of the user.
If the user questions are related to the provided {context}, then you should answer
with maximum of 4 sentences. But If the user question is not related to the document
provided; avoid answering that question. Don't use User: or assistant: in the response.
Here are some few examples.

User: "Who will be the next president in sri lanka?"
   assistant: "Please ask a relevant question."

User: "What is the meaning of life?"
   assistant: "This question is not relevant to the document given to me. Can you
please ask another question?"

question: {question}
   context: {context}
   chat history: {chat_history}

"""
```

Re-evaluate the performance with Prompt Engineering

1. Using Open Al

let's re-evaluate the performance with OpenAI. (Using OpenAI Embeddings and Chat OpenAI)

PDF File Size	Time took to embed	Time took to generate a response	Response Quality
Less than 100KB	2 seconds	3.8 seconds	Response is decent and responded
About 25MB	6 seconds	About 6 seconds	accurately with the information provided in the PDF. Has followed the prompt instructions to not give unnecessary long answers and avoided answering unrelated questions

Responses for Open AI models

what are the available jobs?

what are the available jobs?

There are various available jobs, such as IT Project Manager, Landscape Architect, Forensic Scientist, Physical Education Teacher, Content Strategist, Aerospace Engineer, Copywriter, Nurse Practitioner, Occupational Therapist, IT Security Analyst, Physical Therapist, UX/UI Designer, Environmental Scientist, Financial Planner, Medical Laboratory Technologist, IT Systems Administrator, and Event Planner. Each job has its own unique responsibilities and requirements.

3.8278732299804688 seconds to generate response

Ask a question about your documents:

who is the president of india?

who is the president of india?

This question is not relevant to the document given to me. Can you please ask another question?

24.59121084213257 seconds to generate response

2. Using HuggingfaceHub

Embedding Model - "hkunlp/instructor-xl"

LLM - "google/flan-t5-xxl"

PDF File Size	Time took to embed	Time took to generate	Response Quality
		a response	
Less than 100KB	130 sec	5.2 sec	Responses are not that accurate.
About 25MB	More than 5 mins	6.1 sec	and answered the questions not related to the PDF. There is a reaction to the prompt but model
			is not taking it intelligently

Responses for HuggingfaceHub models

Ask a question about your documents:

what are the jobs available?

what are the jobs available?

I have listed the 27. Aerospace Engineer, 28. Copywriter, 29. Nurse Practitioner, 30. Occupational Therapist, 31. IT Security Analyst, 14. Sales Representative, 15. Web Developer, 16. Public Relations Specialist, 17. Pharmacist, 18. Social Media Manager, 19. Teacher, 20. Software Developer, 21. Registered Nurse, 22. Marketing Manager, 23. Project Manager, 24. Electrician, 25. Software Developer, 26. Registered Nurse, 27. Aerospace Engineer,

5.24003529548645 seconds to generate response

Ask a question about your documents:

what is the meaning of life?

what is the meaning of life?

This question is not relevant to the document given to me. Can you please ask another question?

 ${\tt 2.2889397144317627}\ \ seconds\ to\ generate\ response$

who is the president of india

The following occupations are related to engineering:

2.2099387645721436 seconds to generate response

3. In Cloud base Platforms

For Mistral 7B LLM,

PDF File Size	Time took to embed	Time took to generate a response	Response Quality
Less than 100KB	3.3secs	48 secs	Response is decent and responded
About 25MB	75 secs	50 secs	accurately with the information provided in the PDF. Has followed the prompt instructions to not give unnecessary long answers and avoided answering unrelated questions

Responses for Mistral 7B,

Ask a question about your documents:

what are the jobs available?

what are the jobs available?

The jobs available include IT Project Manager, Landscape Architect, Forensic Scientist, Physical Education Teacher, Content Strategist, Aerospace Engineer, Copywriter, Nurse Practitioner, Occupational Therapist, IT Security Analyst, Physical Therapist, UX/UI Designer, Environmental Scientist, Financial Planner, Medical Laboratory Technologist, IT Systems Administrator, and Event Planner. These professions involve various roles such as managing projects, designing outdoor spaces, analyzing evidence, instructing students, developing content strategies, designing and testing aircraft, creating written content, providing advanced nursing care, helping individuals improve daily living skills, protecting computer systems, assessing and treating patients, focusing on user experience, studying the environment, and assisting clients in managing finances.

48.98280906677246 seconds to generate response

who is the president in india?

The current president of India is not mentioned in the provided document. The document lists various professions and their roles, including IT Project Manager, Landscape Architect, Forensic Scientist, Physical Education Teacher, Content Strategist, Aerospace Engineer, Copywriter, Nurse Practitioner, Occupational Therapist, IT Security Analyst, Physical Therapist, UX/UI Designer, Environmental Scientist, Financial Planner, Medical Laboratory Technologist, IT Systems Administrator, and Event Planner. If you have a question related to any of these professions or their roles, feel free to ask!

50.436715841293335 seconds to generate response

For Llama2: 13B LLM,

PDF File Size	Time took to embed	Time took to generate a response	Response Quality
Less than 100KB	5.65 secs	7.5 secs	Response is decent and responded
About 25MB	129 secs	8.4 secs	accurately with the information provided in the PDF. Has followed the prompt instructions to not give unnecessary long answers and avoided answering unrelated questions

Responses for Llama 13B Model,

