

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 AI

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

Ask a question about your documents:

what are the jobs available ?

what are the jobs available ?

Based on the given context, here are some of the available job options:

1. Physical Therapist
2. UX/UI Designer
3. Environmental Scientist
4. Financial Planner
5. Aerospace Engineer
6. Copywriter
7. Nurse Practitioner
8. Occupational Therapist
9. IT Security Analyst
10. Content Strategist
11. Medical Laboratory Technologist
12. IT Systems Administrator
13. Event Planner
14. IT Project Manager
15. Landscape Architect

Ask a question about your documents:

what is the meaning of life?

what is the meaning of life?

I'm an AI language model and don't have the ability to experience life or hold philosophical beliefs. However, I can tell you that the meaning of life is a question that has puzzled philosophers, scientists, and theologians for centuries. There are many different theories, ranging from the philosophical (e.g., existence for its own sake, seeking happiness, finding purpose) to the scientific (e.g., evolution and survival) to the religious (e.g., fulfilling a divine plan). Ultimately, the answer may depend on one's personal beliefs and values.

2. Using HuggingfaceHub

Embedding Model - “hkunlp/instructor-xl”

LLM - “google/flan-t5-xxl”

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

Ask a question about your documents:

what are the jobs available?

what are the jobs available?

Copywriter Creates written content for advertising, marketing, or promotional purposes. Develops compelling copy, adheres to brand guidelines, and engages target audiences. Nurse Practitioner Provides advanced nursing care and healthcare services. Diagnoses and treats medical conditions, prescribes medications, and collaborates with healthcare teams.

what is the meaning of life?

I don't know

Ask a question about your documents:

who is the president of sri lanka?

who is the president of sri lanka?

Ranil Wickramasinghe

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	Response Quality
Less than 100KB	3.3secs	4 secs	Response is accurate to the document provided. Answered the questions not related to the PDF. Answers are unnecessary lengthy.
About 25MB	75 secs	5.3 secs	

Responses for Mistral 7B,

Ask a question about your documents:

what are the available jobs?

what are the available jobs?

Based on the context provided, here are the different jobs mentioned:

1. Physical Therapist
2. UX/UI Designer
3. Environmental Scientist
4. Financial Planner
5. Aerospace Engineer
6. Content Strategist
7. Medical Laboratory Technologist
8. IT Systems Administrator
9. Event Planner
10. Copywriter
11. Nurse Practitioner
12. Occupational Therapist
13. IT Security Analyst
14. IT Project Manager
15. Landscape Architect
16. Forensic Scientist

Ask a question about your documents:

what is the meaning of life?

what is the meaning of life?

I'm an artificial intelligence and don't have the ability to experience or contemplate the meaning of life. The meaning of life is a philosophical question that has been debated throughout history, and there are many different perspectives on this topic. Some people believe that life has no inherent meaning, while others believe that it is up to each individual to give their life meaning through their experiences and actions. Ultimately, the answer to this question is subjective and may depend on one's personal beliefs and values. The professions listed in the context are all focused on various aspects of improving the human experience or making a positive impact on the world.

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	6 secs	Response is accurate to the document provided. Answered the questions not related to the PDF. Answers are unnecessary lengthy.
About 25MB	129 secs	8.4 secs	

Responses for Llama 13B Model,

Ask a question about your documents:

what are the jobs available?

what are the jobs available?

Based on the given context, the following are the job available:

1. IT Project Manager
2. Landscape Architect
3. Forensic Scientist
4. Physical Education Teacher
5. Content Strategist
6. Aerospace Engineer
7. Copywriter
8. Nurse Practitioner
9. Occupational Therapist
10. IT Security Analyst
11. Physical Therapist
12. UX/UI Designer
13. Environmental Scientist
14. Financial Planner

Ask a question about your documents:

what is the meaning of life?

what is the meaning of life?

The question of the meaning of life is a complex and abstract topic that has been debated by philosophers, theologians, scientists, and many other thinkers throughout history. There is no one definitive answer to this question, as it can be interpreted in many different ways depending on one's beliefs, values, and experiences. However, some possible approaches to understanding the meaning of life might include:

1. Religious or spiritual perspectives: Many people believe that the meaning of life is connected to a higher power or divine being, and that our purpose in life is to fulfill a divine plan or follow a set of moral principles.
2. Personal fulfillment: Others might argue that the meaning of life is found in personal fulfillment and the pursuit of happiness, whether through material success, relationships, or personal achievements.
3. Social or cultural contexts: Our understanding of the meaning of life can also be influenced by our social and cultural contexts, such as our family, community, or society as a whole.

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 AI

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 accurately with the information provided in the PDF. Has followed the prompt instructions to not give unnecessary long answers and avoided answering unrelated questions
About 25MB	6 seconds	About 6 seconds	

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

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?

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 accurately with the information provided in the PDF. Has followed the prompt instructions to not give unnecessary long answers and avoided answering unrelated questions
About 25MB	75 secs	50 secs	

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.

48.9828996677246 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!

59.436715841293335 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 accurately with the information provided in the PDF. Has followed the prompt instructions to not give unnecessary long answers and avoided answering unrelated questions
About 25MB	129 secs	8.4 secs	

Responses for Llama 13B Model,

Ask a question about your documents:

what are the available jobs?

what are the available jobs?

Here are some available jobs based on the provided document:

1. IT Project Manager
2. Landscape Architect
3. Forensic Scientist
4. Physical Education Teacher
5. Content Strategist
6. Aerospace Engineer
7. Copywriter
8. Nurse Practitioner
9. Occupational Therapist
10. IT Security Analyst
11. Physical Therapist
12. UX/UI Designer
13. Environmental Scientist
14. Financial Planner
15. Medical Laboratory Technologist
16. IT Systems Administrator

Ask a question about your documents:

who is the president in china?

who is the president in china?

Sure, I'd be happy to help! The answer to the question "Who is the president in China?" is not related to the document provided, which lists various careers and their descriptions. Therefore, I cannot provide an answer to this question based on the information given. If you have any other questions or need further assistance, please feel free to ask!

7.510854244232178 seconds to generate response