

# Report on the RAG-Powered AI Assistant with LLaMA3 and GPT-2

## 1)Local LLM Setup and Interaction:

- 1.1) Run “ollama run llama3.2” to verify the installation of llama on the local machine.
- 1.2) Use GPT-2 by installing hugging face’s transformers library and setting up a python to load GPT-2 from transformers library and read the input prompt and generate response.

## 2)Implementation of the RAG system with Llama3 and GPT-2:

We are integrating a Retrieval-Augmented Generation (RAG) workflow using LLaMA and GPT-2 models to enhance response generation in a Streamlit chatbot interface. The RAG system retrieves relevant contextual information from Wikipedia via the [WikipediaAPI](#) and generates sentence embeddings using Alibaba's Sentence Transformer model to find the most relevant documents. Based on the selected model (LLaMA or GPT-2), the retrieved context is incorporated into the prompt, and either LLaMA or GPT-2 is used to generate a response. This process allows the chatbot to provide more contextually aware and accurate answers by combining Wikipedia-based document retrieval with model-based generation. Users can choose between RAG-enhanced responses or direct model responses, and response times for each interaction are tracked.

### 2.1)Sentence Embeddings Model Setup:

```
import streamlit as st
import subprocess
import re
import numpy as np
import time
from sentence_transformers import SentenceTransformer
from wikipediaapi import Wikipedia
from transformers import GPT2LMHeadModel, GPT2Tokenizer

# 1. Setup the Sentence Embeddings Model
model = SentenceTransformer("Alibaba-NLP/gte-base-en-v1.5", trust_remote_code=True)
```

This initializes the **SentenceTransformer** model, which is used to convert text into vector embeddings. In this case, it uses Alibaba's **GTE (General-purpose Text Embedding) model**. These embeddings are later used to compare and rank Wikipedia paragraphs based on their similarity to the user’s query. This is essential for the **retrieval** part of the RAG system.

### 2.2)Wikipedia API Setup:

```
# 2. Setup Wikipedia API
wiki = Wikipedia('RAGBot/Rag-1', 'en')
```

This sets up the Wikipedia API to allow searching for relevant articles in the English version of Wikipedia. It's part of the retrieval step, where Wikipedia is used as the external knowledge base to fetch contextual information related to the user's query. This provides the external data that the model may need to generate a relevant response.

### 2.3)GPT-2 Model and Tokenizer Setup:

```
# 3. Load GPT-2 Model and Tokenizer
gpt2_model = GPT2LMHeadModel.from_pretrained("gpt2")
gpt2_tokenizer = GPT2Tokenizer.from_pretrained("gpt2")
```

This loads the pre-trained **GPT-2 language model** and its tokenizer, which will be used to generate text based on the query and the retrieved Wikipedia context. GPT-2 can function in two ways: with RAG (enhanced with external context) or standalone. The tokenizer converts the input text into tokens for the model to process and later decodes the output tokens into human-readable text.

### 2.4)Post and Preprocessing the Query:

```
# 4. Function to preprocess the prompt
def preprocess_prompt(prompt: str) -> str:
    prompt = prompt.lower().strip()
    prompt = re.sub(r'^\w\s', '', prompt) # Remove punctuation
    return prompt

# 5. Function to post-process the response
def postprocess_response(response: str) -> str:
    response = re.sub(r'<\|endoftext\|>', '', response) # Clean end of text token
    return response.strip()
```

These functions preprocess and postprocess the user's input query by converting it to lowercase and removing punctuation. This step ensures that the query is clean and consistent before being used in embedding generation or as a prompt for the model. It improves the retrieval and generation process by standardizing input.

### 2.5)LLaMA Model Integration and Text generation (via Ollama):

```
# 6. Function to generate text using local LLaMA model via Ollama
def generate_llama_response(prompt):
    process = subprocess.run(
        ['ollama', 'run', 'llama3.2:latest'],
        input=prompt,
        capture_output=True,
        text=True,
        encoding='utf-8',
        errors='replace'
    )

    if process.returncode != 0:
        error_message = process.stderr.strip()
        return f"Error running Ollama: {error_message}"

    response = process.stdout.strip()
    return response
```

This function interacts with the local **LLaMA** model via the **Ollama** CLI, which runs the LLaMA model. It takes the query or the RAG-enhanced prompt, sends it to the LLaMA model, and captures the output (generated response). It handles potential errors from the process and returns either the response or an error message.

## 2.6)GPT-2 Model Integration and Text Generation:

```
# 7. Function to generate text using GPT-2
def generate_gpt2_response(prompt):
    # Encode the input prompt
    inputs = gpt2_tokenizer.encode(prompt, return_tensors="pt")

    # Set maximum input length and new tokens to generate
    max_input_length = 512 # or another suitable length
    max_new_tokens = 50 # Adjust this based on how much output you want

    # Truncate input if it exceeds the max input length
    if inputs.size(1) > max_input_length:
        inputs = inputs[:, :max_input_length]

    # Generate the response
    outputs = gpt2_model.generate(inputs, max_new_tokens=max_new_tokens, num_return_sequences=1)

    # Decode and return the generated response
    response = gpt2_tokenizer.decode(outputs[0], skip_special_tokens=True)
    return response
```

This function generates a response using **GPT-2**. It first tokenizes the prompt (either the raw query or the RAG-enhanced query) and generates text using GPT-2. The model can produce up to 50 new tokens for the response. After the response is generated, the tokenizer decodes it back into a human-readable format.

## 2.7)Wikipedia Search for Document Retrieval:

```
# 8. Function to search Wikipedia
def search_wikipedia(query):
    stop_words = {"what", "who", "how", "is", "was", "are", "which", "when", "where", "why", "did", "do", "does"}
    words = [word for word in query.split() if word.lower() not in stop_words]

    paragraphs_list = []

    for word in words:
        doc = wiki.page(word).text
        if doc:
            paragraphs = doc.split('\n\n')
            paragraphs_list.extend(paragraphs)

    if not paragraphs_list:
        return None

    docs_embed = model.encode(paragraphs_list, normalize_embeddings=True)
    query_embed = model.encode(query, normalize_embeddings=True)
    similarities = np.dot(docs_embed, query_embed.T)
    top_3_idx = np.argsort(similarities, axis=0)[-3:][::-1].tolist()
    most_similar_documents = [paragraphs_list[idx] for idx in top_3_idx]

    context = "\n\n".join(most_similar_documents)
    return context
```

This function performs the **document retrieval** part of RAG. It queries Wikipedia for relevant documents based on the user's query, filters stop words (e.g., "what", "who"), and uses the **SentenceTransformer** to convert the text of each paragraph and the query into embeddings. By computing the cosine similarity between the query and document embeddings, it retrieves the top 3 most relevant paragraphs. These paragraphs are used as contextual information in the RAG prompt. The stopwords filtering will help in reducing the response time.

## 2.8) Combining Retrieval and Generation (RAG Workflow):

```
# 9. Combining retrieval and generation (RAG workflow)
def rag_response(query, model_choice='llama'):
    start_time = time.time() # Start timing

    context = search_wikipedia(query)

    if context is None:
        return "I couldn't find relevant information from Wikipedia."

    prompt = f"""
    I have some information that might be relevant to answering a question.
    Here's the information:

    {context}

    Based on this information, please answer the following question:

    {query}

    If the information provided doesn't contain the answer, please say "I don't have enough information to answer this question."
    """

    if model_choice == 'LLaMA with RAG':
        response = generate_llama_response(prompt)
    elif model_choice == 'GPT-2 with RAG':
        response = generate_gpt2_response(prompt)
    else:
        # Direct model usage without RAG
        if model_choice == 'LLaMA':
            response = generate_llama_response(query) # Direct usage without RAG
        else: # GPT-2
            response = generate_gpt2_response(query)

    end_time = time.time() # End timing
    response_time = end_time - start_time

    return response, response_time
```

This is the **core RAG function**. First, it retrieves context from Wikipedia using `search_wikipedia()`. If context is found, it is combined with the user's query to form a prompt for the model. Depending on the selected model (**LLaMA with RAG** or **GPT-2 with RAG**), either LLaMA or GPT-2 is used to generate the final response. It also tracks the time taken for the entire process (retrieval + generation), which helps analyze performance. If no context is retrieved, a fallback message is returned.

## 2.9) Streamlit Interface for User Input and Output:

```
st.title("RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration")

# Create a session state to store chat history
if 'chat_history' not in st.session_state:
    st.session_state.chat_history = []

# Tabs for Chat and Conversation History
tab1, tab2 = st.tabs(["Chat", "Conversation History"])

# Chat Tab
with tab1:
    query = st.text_input("Your Query:", key="query_input")

    # Select Model
    model_choice = st.selectbox("Select Model:", ["LLaMA with RAG", "GPT-2 with RAG", "LLaMA", "GPT-2"])

    if st.button("Submit"):
        if query:
            with st.spinner("Fetching response..."):
                # Preprocess the user input
                preprocessed_input = preprocess_prompt(query)
                response, response_time = rag_response(preprocessed_input, model_choice=model_choice)

                # Post-process the response
                cleaned_output = postprocess_response(response)

                # Store the conversation in chat history
                st.session_state.chat_history.append({
                    "user": query,
                    "chatbot": cleaned_output,
                    "model": model_choice, # Add the selected model to chat history
                    "response_time": response_time # Store the response time
                })

                st.subheader("Response:")
                st.write(cleaned_output)

            # Display the response time
            st.write(f"Response Time: {response_time:.2f} seconds")
```

This section handles the **Streamlit UI**, allowing users to input queries, select a model (LLaMA with RAG, GPT-2 with RAG, or standalone versions of both), and view the chatbot's response. It stores chat history in `session_state` and processes user queries by calling the RAG workflow or direct generation functions, depending on the user's model choice.

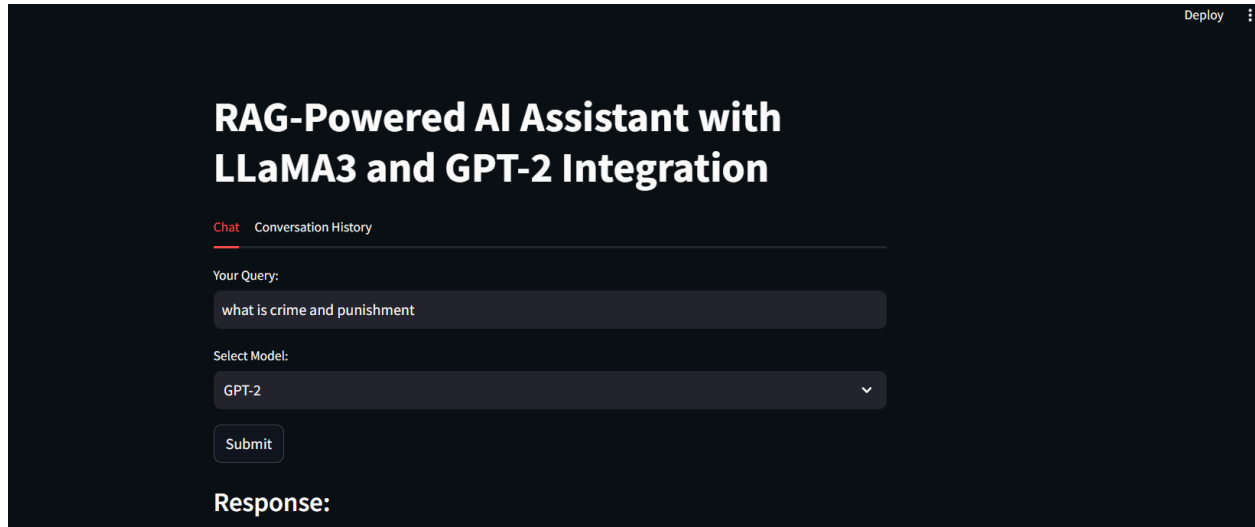
To summarize for

**Document Retrieval:** We use **Wikipedia** and a **SentenceTransformer model** to find relevant documents that match the user's query.

**Text Generation:** Once relevant documents are retrieved, the context is combined with the user's query, and text generation is performed by either **LLaMA** or **GPT-2**.

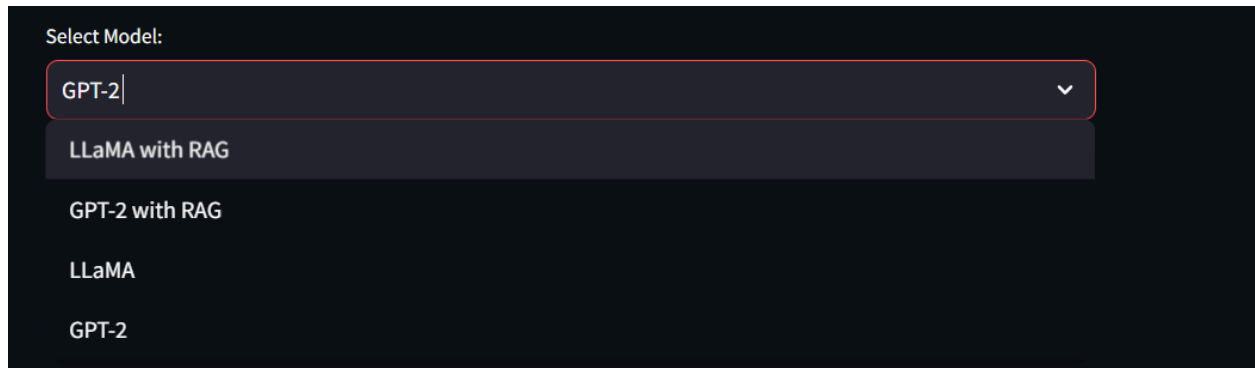
**RAG Workflow:** We implemented a **Retrieval-Augmented Generation (RAG)** system, which enhances text generation by feeding external context into the model, improving accuracy and relevance.

**UI Integration:** The process is wrapped in a **Streamlit** app, where users can interact with the system and choose between RAG-enhanced or standalone model responses. Once up and running the UI of the application looks



The screenshot shows a web application titled "RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration". It features a dark theme with a "Deploy" button in the top right corner. Below the title, there are two tabs: "Chat" (active) and "Conversation History". The "Your Query:" section contains a text input field with the text "what is crime and punishment". Below this is a "Select Model:" dropdown menu currently set to "GPT-2". A "Submit" button is located below the dropdown. At the bottom, the "Response:" section is visible but empty.

And the user can choose a model of choice from the dropdown



This image is a close-up of the "Select Model:" dropdown menu. The dropdown is open, showing a list of options: "GPT-2" (the currently selected option), "LLaMA with RAG", "GPT-2 with RAG", "LLaMA", and "GPT-2". The dropdown has a dark background with light-colored text.

And can see the conversation history from the Conversation history tab as shown below.

## RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration

Chat Conversation History

---

### Conversation History

**User:** What are the latest advancements in quantum computing

**Chatbot:** There have been many recent advancements in quantum computing, which can be broadly categorized into several areas:

- Quantum Error Correction:** Researchers have made significant progress in developing techniques to correct errors that occur during quantum computation. This is a crucial step towards large-scale quantum computing, as errors can quickly accumulate and render the computer useless.
- Superconducting Qubits:** Superconducting qubits are one of the most widely used types of quantum bits (qubits) in current quantum computers. Researchers have made significant improvements to their design, leading to higher coherence times and more efficient operations.
- Ion Traps:** Ion traps are another type of quantum computing architecture that uses ions instead of qubits. Recent advancements have led to improved control over the ions and increased scalability.

And the for each query model used and response time can also be seen in the Conversation history

Some future directions in quantum computing include:

- Quantum-Classical Hybrids:** Developing hybrid quantum-classical computers that can take advantage of both classical and quantum processing power.
- Neuromorphic Computing:** Creating quantum computers that mimic the structure and function of biological neural networks.
- Adiabatic Quantum Computers:** Developing adiabatic quantum computers that can solve specific problems more efficiently than traditional quantum algorithms.

The future of quantum computing is exciting and rapidly evolving.

**Model Used:** LLaMA

**Response Time:** 48.16 seconds

### 3)Queries and Responses

Queries across various domains and different complexities are used to test both LLaMa and GPT-2 with and without RAG implementation

For the query  
3.1)What is cricket?  
Llama with RAG:

# RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration

**Chat** Conversation History

Your Query:

what is cricket

Select Model:

LLaMA with RAG

Submit

## Response:

According to the information provided, cricket is a bat-and-ball game played between two teams of eleven players on a field, at the centre of which is a 22-yard pitch with a wicket at each end. The objective of the game is for one team to score runs by hitting the ball with a bat and then switching places with another player, while the opposing team tries to prevent them from scoring by dismissing batters (i.e., getting them out).

Response Time: 64.06 seconds



## GPT with RAG:

Your Query:

what is cricket

Select Model:

GPT-2 with RAG

Submit

### Response:

I have some information that might be relevant to answering a question. Here's the information:

Cricket is a bat-and-ball game played between two teams of eleven players on a field.

The fielding team tries to prevent runs from being scored by dismissing batters (so they are "out"). Means of dismissal include being bowled, when the ball hits the striker's wicket and dislodges the bails, and by the fielding side either catching the ball after it is hit by the bat but before it hits the ground, or hitting a wicket with the ball before a batter can cross the crease line in front of the wicket. When ten batters have been dismissed, the innings (playing phase) ends and the teams swap roles. Forms of cricket range from traditional Test matches played over five days to the newer Twenty20 format (also known as T20), in which each team bats for a single innings of 20 overs (each "over" being a set of 6 fair opportunities for the batting team to score) and the game generally lasts three to four hours. Traditionally, cricketers play in all-white kit, but in limited overs cricket, they wear club or team colours. In addition to the basic kit, some

## LlaMa:

Your Query:

what is cricket

Select Model:

LLaMA

Submit

### Response:

Cricket is a popular team sport played with a bat, ball, and wickets (three vertical stumps and two horizontal stumps) on a rectangular field. It is one of the most widely played sports in the world, especially in countries like India, Australia, England, Pakistan, and South Africa.

Here's a brief overview:

**Objective:** The objective of cricket is to score runs by hitting the ball with a bat and running between the wickets, while the opposing team tries to stop them by getting the batsmen out.

**Basic rules:**

1. The game is played by two teams, each with 11 players.
2. One team bats, while the other team fields and bowls (throws) the ball.
3. The batting team sends two batsmen onto the field, who take turns to hit the ball.

GPT-2:

# RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration

Chat Conversation History

**Your Query:**

## what is cricket

Select Model:

GPT-2

Submit

**Response:**

what is cricket?"

[illegible]

Response Time: 64.58 seconds

3.2)Who invented the telephone?

LlaMa with RAG:

# RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration

**Chat** Conversation History

Your Query:

Who invented the telephone?

Select Model:

LLaMA with RAG

Submit

## Response:

While the text mentions several individuals as pioneers or contributors to the development of the telephone, it does not explicitly state that any one person invented the telephone.

However, according to the text, Alexander Graham Bell was the first to be awarded a patent for the electric telephone by the United States Patent and Trademark Office (USPTO) in March 1876.

Response Time: 110.65 seconds

## GPT2 with RAG:

Your Query:

Who invented the telephone?

Select Model:

GPT-2 with RAG



Submit

### Response:

I have some information that might be relevant to answering a question. Here's the information:

#### Early history

Before the development of the electric telephone, the term telephone was applied to other inventions, and not all early researchers of the electrical device used the term. Perhaps the earliest use of the word for a communications system was the telephon created by Gottfried Huth in 1796. Huth proposed an alternative to the optical telegraph of Claude Chappe in which the operators in the signaling towers would shout to each other by means of what he called "speaking tubes", but would now be called giant megaphones. A communication device for sailing vessels, called telephone, was invented by Captain John Taylor in 1844. This instrument used four air horns to communicate with vessels in foggy weather. Johann Philipp Reis used the term in reference to his invention, commonly known as the Reis telephone, in c. 1860. His device appears to be the first device based on the conversion of sound into electrical impulses.

LlaMa:

# RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration

Chat

Conversation History

Your Query:

Who invented the telephone?

Select Model:

LLaMA

Submit

## Response: ↔

The invention of the telephone is credited to Alexander Graham Bell, a Scottish-born inventor and scientist. He patented the first practical telephone on March 7, 1876.

However, there was some controversy over who actually invented the telephone first. Elisha Gray, an American inventor, had filed a caveat for a telephone invention at the United States Patent Office on February 14, 1876, just hours after Bell. Additionally, Antonio Meucci, an Italian inventor, is also credited by some with inventing the first telephone-like device in the 1840s.

GPT-2:

# RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration

**Chat** Conversation History

Your Query:

Who invented the telephone?

Select Model:

GPT-2

Submit

## Response:

who invented the telephone) and the fact that the telephone was invented by a man who was not a scientist.

The fact that the telephone was invented by a man who was not a scientist. The fact that the telephone was invented by a man who was not a

Response Time: 91.45 seconds



3.3)explain gravity?

LlaMa with RAG:

Your Query:

explain gravity

Select Model:

LLaMA with RAG

Submit

## Response:

Based on the provided information, gravity can be explained as follows:

Gravity is a fundamental interaction primarily observed as mutual attraction between all things that have mass. It is the most significant interaction between objects at the macroscopic scale and determines the motion of planets, stars, galaxies, and even light.

According to Einstein's general theory of relativity, gravity is not a force, but rather the curvature of spacetime caused by the uneven distribution of mass. This curvature causes masses to move along geodesic lines, which are the shortest paths possible in curved spacetime. The most extreme example of this curvature is a black hole, from which nothing—not even light—can escape once past the event horizon.

In simpler terms, gravity can be understood as a force that attracts objects with mass towards each other, with the strength of the gravitational field numerically equal to the acceleration of objects under its

GPT2 with RAG:

explain gravity

Select Model:

GPT-2 with RAG

Submit

## Response:

I have some information that might be relevant to answering a question. Here's the information:

In physics, gravity (from Latin *gravitas* 'weight') is a fundamental interaction p

On Earth, gravity gives weight to physical objects, and the Moon's gravity is responsible for sublunar tides in the oceans. The corresponding antipodal tide is caused by the inertia of the Earth and Moon orbiting one another. Gravity also has many important biological functions, helping to guide the growth of plants through the process of gravitropism and influencing the circulation of fluids in multicellular organisms. The gravitational attraction between the original gaseous matter in the universe caused it to coalesce and form stars which eventually condensed into galaxies, so gravity is responsible for many of the large-scale structures in the universe. Gravity has an infinite range, although its effects become weaker as objects get farther away. Gravity is most accurately described by the general theory of relativity, proposed by Albert Einstein in 1915, which describes gravity not as a force, but as the curvature of spacetime, caused by the uneven distribution of mass, and causing masses to move along geodesic lines. The most extreme



LLaMa:

Your Query:

Select Model:

LLaMA

Submit

## Response:

Gravity is a fundamental force of nature that attracts two objects with mass towards each other. It is a universal force that affects everything with mass or energy, from the smallest subatomic particles to the largest galaxies.

**What causes gravity?**

According to Albert Einstein's theory of general relativity, gravity is not a force that acts between objects, but rather a curvature of spacetime caused by the presence of mass and energy. The more massive an object is, the greater its gravitational pull on other objects.

Imagine spacetime as a fabric that stretches and curves in response to massive objects. The curvature of this fabric creates a "gravitational field" around the object, which affects the motion of nearby objects.

GPT2:

# RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration

Chat

Conversation History

Your Query:

Select Model:

GPT-2

Submit

## Response:

explain gravity and the effect of gravity on the body.

The first step is to determine the mass of the body. The mass of the body is determined by the mass of the body's internal organs. The body's internal organs are the organs that control

Response Time: 33.56 seconds

3.4)What is the boiling point of water?

LlaMa with RAG:

# RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration

**Chat** Conversation History

Your Query:

What is the boiling point of water

Select Model:

LLaMA with RAG

Submit

## Response:

According to the information provided, the boiling point of water is typically considered to be 100 °C (212 °F; 373 K) at sea level. However, it's worth noting that pressure and a change in the composition of the liquid may alter the boiling point of the liquid, and high elevation cooking generally takes longer since boiling point is a function of atmospheric pressure.

Response Time: 92.98 seconds

GPT2 with RAG:

What is the boiling point of water

Select Model:

GPT-2 with RAG

Submit

## Response:

I have some information that might be relevant to answering a question. Here's the information:

In cooking

Boiling is the method of cooking food in boiling water or other water-based liquids such as stock or milk. Simmering is gentle boiling, while in poaching the cooking liquid moves but scarcely bubbles. The boiling point of water is typically considered to be 100 °C (212 °F; 373 K), especially at sea level. Pressure and a change in the composition of the liquid may alter the boiling point of the liquid. High elevation cooking generally takes longer since boiling point is a function of atmospheric pressure. At an elevation of about one mile (1,600 m), water boils at approximately 95 °C (203 °F; 368 K). Depending on the type of food and the elevation, the boiling water may not be hot enough to cook the food properly. Similarly, increasing the pressure as in a pressure cooker raises the temperature of the contents above the open air boiling point.

Boiling or ebullition is the rapid phase transition from liquid to gas or vapour; the reverse of boiling is condensation. Boiling occurs when a liquid is heated to its boiling point, so that the vapour pressure of

LLaMa:

What is the boiling point of water

Select Model:

LLaMA

Submit

## Response:

The boiling point of water is 212 degrees Fahrenheit (°F) or 100 degrees Celsius (°C) at standard atmospheric pressure. However, this value can vary depending on factors such as altitude and pressure.

At higher altitudes, the air pressure is lower, which means that the boiling point of water will be lower. For example:

- At sea level: 212°F (100°C)
- At 1,000 meters (3,300 feet): 205°F (96°C)
- At 2,000 meters (6,600 feet): 201°F (94°C)

On the other hand, at higher altitudes or under increased pressure, the boiling point of water will be higher.

Response Time: 84.83 seconds

GPT2:

# LLaMA3 and GPT-2 Integration

**Chat** Conversation History

Your Query:

What is the boiling point of water

Select Model:

GPT-2

Submit

## Response:

what is the boiling point of water?

The boiling point of water is the point at which the water boils. The boiling point of water is the point at which the water boils.

The boiling point of water is the point at which the water boils.

The

Response Time: 77.77 seconds

### 3.5) What are the benefits of regular exercise?

LLaMa with RAG:

Select Model:

LLaMA with RAG

Submit

**Response:**

Based on the provided information, the benefits of regular exercise include:

- Reducing the risk of health problems, such as cardiovascular disease, stroke, and cancer
- Maintaining physical fitness and overall health
- Contributing to maintaining a healthy weight, regulating the digestive system, building and maintaining healthy bone density, muscle strength, and joint mobility
- Promoting physiological well-being and reducing surgical risks
- Strengthening the immune system
- Increasing life expectancy and the overall quality of life
- Preventing aging by reducing inflammatory potential
- Reducing the risk of premature mortality worldwide (approximately 9%)
- Decreasing the burden of disease from coronary heart disease, type 2 diabetes, breast cancer, and colon cancer (approximately 6%, 7%, 10%, and 10% respectively)

GPT2 with RAG:

Select Model:

GPT-2 with RAG

Submit

**Response:**

I have some information that might be relevant to answering a question. Here's the information:

```
Exercise is physical activity that enhances or maintains fitness and overall health
```

In terms of health benefits, usually, 2.5 hours of moderate-intensity exercise per week is recommended for reducing the risk of health problems. At the same time, even doing a small amount of exercise is healthier than doing none. Only doing an hour and a quarter (11 minutes/day) of exercise could reduce the risk of early death, cardiovascular disease, stroke, and cancer.

Health effects Physical exercise is important for maintaining physical fitness and can contribute to maintaining a healthy weight, regulating the digestive system, building and maintaining healthy bone density, muscle strength, and joint mobility, promoting physiological well-being, reducing surgical risks, and strengthening the immune system. Some studies indicate that exercise may increase life expectancy and the overall quality of life. People who participate in moderate to high levels of physical exercise have a lower mortality rate compared to individuals who by comparison are not physically active. Moderate



LLaMa:

Select Model:

LLaMA

Submit

## Response:

Regular exercise provides numerous physical and mental health benefits, which can improve overall quality of life. Some of the most significant benefits of regular exercise include:

**Physical Benefits:**

- Weight Management:** Regular exercise helps burn calories, maintain weight loss, and reduce body fat.
- Improved Cardiovascular Health:** Exercise strengthens the heart and lungs, reducing blood pressure, improving circulation, and lowering the risk of heart disease.
- Increased Strength and Flexibility:** Regular physical activity builds muscle mass, bone density, and flexibility, making daily activities easier and less painful.
- Better Sleep:** Exercise helps regulate sleep patterns, leading to improved sleep quality, duration, and depth.
- Reduced Risk of Chronic Diseases:** Regular exercise has been shown to reduce the risk of developing conditions such as type 2 diabetes, certain types of cancer, and osteoporosis.

GPT2:

## RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration

Chat

Conversation History

Your Query:

What are the benefits of regular exercise

Select Model:

GPT-2

Submit

## Response:

what are the benefits of regular exercise?

The benefits of regular exercise are not as clear cut as they might seem. The main benefit of regular exercise is that it helps you to get more out of your day. It also helps you to get more out of your life.

Response Time: 28.89 seconds

### 3.6) Why is recycling important?

Llama with RAG:

Your Query:

Why is recycling important

Select Model:

LLaMA with RAG

Submit

## Response:

Based on the information provided, recycling is important for several reasons:

1. **Environmental sustainability:** Recycling helps remove raw material input and redirect waste output in the economic system, promoting environmental sustainability.
2. **Conservation of natural resources:** Recycling can conserve materials by reducing the need for new raw materials, which helps to lower greenhouse gas emissions, energy use, air pollution, and water pollution.
3. **Prevention of waste:** Recycling prevents the waste of potentially useful materials and reduces the consumption of fresh raw materials.
4. **Economic benefits:** Recycling can generate income from the sale of quality recyclates, supporting local governments, households, and businesses.
5. **Job creation and investment:** Pursuing high-quality recycling can promote consumer and business confidence in the waste and resource management sector, encouraging investment in this area.

## GPT2 with RAG:

Why is recycling important

Select Model:

GPT-2 with RAG

Submit

### Response:

I have some information that might be relevant to answering a question. Here's the information:

Recycling is the process of converting waste materials into new materials and objects.

Recycling is a key component of modern waste reduction and is the third component of the "Reduce, Reuse, and Recycle" waste hierarchy. It promotes environmental sustainability by removing raw material input and redirecting waste output in the economic system. There are some ISO standards related to recycling, such as ISO 15270:2008 for plastics waste and ISO 14001:2015 for environmental management control of recycling practice. Recyclable materials include many kinds of glass, paper, cardboard, metal, plastic, tires, textiles, batteries, and electronics. The composting and other reuse of biodegradable waste—such as food and garden waste—is also a form of recycling. Materials for recycling are either delivered to a household recycling center or picked up from curbside bins, then sorted, cleaned, and reprocessed into new materials for manufacturing new products. In ideal implementations, recycling a material produces a fresh supply of the same material—for example, used office paper would be converted into new office

## LlaMa:

Why is recycling important

Select Model:

LLaMA

Submit

### Response:

Recycling is important for several reasons:

- Conservation of Natural Resources:** Recycling helps to conserve natural resources by reusing materials that would otherwise be extracted, processed, and consumed. This reduces the need for mining, drilling, and logging, which can help preserve ecosystems and habitats.
- Reduction of Waste:** Recycling helps to reduce the amount of waste sent to landfills, where it can take hundreds of years to decompose. By recycling, we can minimize the amount of waste that ends up in landfills and oceans.
- Energy Savings:** Recycling requires significantly less energy than producing new materials from raw resources. For example, recycling aluminum cans uses 95% less energy than producing new aluminum from raw materials.
- Greenhouse Gas Reductions:** The production of new materials from raw resources requires large amounts of energy and generates significant greenhouse gas emissions. By recycling, we can reduce the amount of energy needed to produce new materials, which in turn reduces greenhouse gas emissions.



GPT2:

## RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration

**Chat** Conversation History

Your Query:

Why is recycling important

Select Model:

GPT-2

Submit

**Response:**

why is recycling important?"

"It's important to recycle, but it's not the only thing that's important," he said. "It's important to recycle, but it's not the only thing that's important."

The city's recycling program is

Response Time: 55.12 seconds

3.7)Who directed RRR?

LLaMa with RAG:

## RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration

**Chat** Conversation History

Your Query:

who directed RRR

Select Model:

LLaMA with RAG

Submit

**Response:**

According to the text, RRR was directed by S. S. Rajamouli.

Response Time: 86.62 seconds

GPT2 with RAG:

Select Model:

GPT-2 with RAG

Submit

**Response:** ↗

I have some information that might be relevant to answering a question. Here's the information:

RRR (subtitled onscreen as Roudram Ranam Rudhiram) is a 2022 Indian Telugu-language

Rajamouli conceived the story by reimagining Rama Raju and Bheem meeting in 1920 Delhi as two unstoppable forces in an epic fight with colonial tyranny. The film was announced in March 2018. Principal photography of the film began in November 2018 in Hyderabad and continued until August 2021, owing to delays caused by the COVID-19 pandemic. It was filmed extensively across India, with a few sequences filmed in Ukraine and Bulgaria. The film's soundtrack and background score were composed by M. M. Keeravani, with cinematography by K. K. Senthil Kumar and editing by A. Sreekar Prasad. Sabu Cyril is the film's production designer while V. Srinivas Mohan supervised the visual effects. Made on a budget of ₹550 crore, RRR was the most expensive Indian film at the time of its release. The film was released theatrically on 25 March 2022. With ₹223 crore worldwide on its first day, RRR recorded the highest opening-day earned by an Indian film. It emerged as the highest-grossing film in its home market of Andhra Pradesh

LlaMa:

**RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration** ↗

Chat Conversation History

Your Query:

who directed RRR

Select Model:

LLaMA

Submit

**Response:**

I couldn't find any information on a movie titled "RRR". However, I found that there is a Bollywood film titled "RRR" released in 2022, which was directed by S.S. Rajamouli.

Response Time: 83.35 seconds

GPT2:

**Chat** Conversation History

Your Query:  
who directed RRR

Select Model:  
GPT-2

Submit

**Response:**  
who directed rrrr)  
(3) The person who directed rrrr)  
(4) The person who directed rrrr)  
(5) The person who directed rrrr)  
(6) The person who directed  
Response Time: 89.09 seconds

3.8)What are the main festivals celebrated in India?

LLaMa with RAG:

**RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration**

**Chat** Conversation History

Your Query:  
What are the main festivals celebrated in India

Select Model:  
LLaMA with RAG

Submit

**Response:**  
I don't have enough information to answer this question. The text provides information on various religious festivals from different cultures and religions around the world, but it does not specifically mention the main festivals celebrated in India.  
Response Time: 71.45 seconds

GPT2 with RAG:

Select Model:

GPT-2 with RAG

▼

Submit

## Response:

I have some information that might be relevant to answering a question. Here's the information:

Religious festivals

Among many religions, a feast is a set of celebrations in honour of God or gods. A feast and a festival are historically interchangeable. Most religions have festivals that recur annually and some, such as Passover, Easter, and Eid al-Adha are moveable feasts – that is, those that are determined either by lunar or agricultural cycles or the calendar in use at the time. The Sed festival, for example, celebrated the thirtieth year of an Egyptian pharaoh's rule and then every three (or four in one case) years after that. Among the Ashantis, most of their traditional festivals are linked to gazette sites which are believed to be sacred with several rich biological resources in their pristine forms. Thus, the annual commemoration of the festivals helps in maintaining the buoyancy of the conserved natural site, assisting in biodiversity conservation. In the Christian liturgical calendar, there are two principal feasts, properly known as the Feast of the Nativity of our Lord (Christmas) and the Feast of the Resurrection (Easter), but minor festivals in honour of local

LlaMa:

Select Model:

LLaMA

▼

Submit

## Response:

India is a diverse country with a rich cultural heritage, and it celebrates numerous festivals throughout the year. Here are some of the main festivals celebrated in India:

- Diwali** (Festival of Lights): Celebrated over 5 days, Diwali is a significant festival in Hinduism, marking the triumph of light over darkness and good over evil.
- Holi** (Festival of Colors): A vibrant celebration of spring, Holi marks the arrival of new life and renewal. It's celebrated with colors, music, and dancing.
- Navratri** (Nine Nights): A 9-day festival honoring the divine feminine energy, Navratri is a time for music, dance, and worship.
- Ganesh Chaturthi** (Birth of Lord Ganesha): Celebrated to commemorate the birth of Lord Ganesha, the remover of obstacles, this festival is marked by offering prayers and sweets.
- Baisakhi** (Spring Festival): A significant festival in Sikhism, Baisakhi marks the beginning of the new year and the harvest season.
- Dussehra** (Victory of Good Over Evil): Celebrated over 10 days, Dussehra commemorates the victory of Lord Rama over the demon king Ravana.

GPT2:

## RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration

**Chat** Conversation History

Your Query:

What are the main festivals celebrated in India

Select Model:

GPT-2

Submit

### Response:

what are the main festivals celebrated in india?

I think it's a good question. I think it's a good question. I think it's a good question. I think it's a good question. I think it's a good question. I think it's a good question. I think it's a good question

Response Time: 69.56 seconds



### 3.9)What is inflation?

LlaMa with RAG:

# RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration

**Chat** Conversation History

Your Query:

what is inflation

Select Model:

LLaMA with RAG

Submit

## Response:

Inflation is a general increase in the prices of goods and services in an economy. It is usually measured using a consumer price index (CPI). When the general price level rises, each unit of currency buys fewer goods and services; consequently, inflation corresponds to a reduction in the purchasing power of money.

Response Time: 58.88 seconds

GPT2 with RAG:

what is inflation

Select Model:

GPT-2 with RAG

Submit

## Response:

I have some information that might be relevant to answering a question. Here's the information:

In economics, inflation is a general increase in the prices of goods and services

Changes in inflation are widely attributed to fluctuations in real demand for goods and services (also known as demand shocks, including changes in fiscal or monetary policy), changes in available supplies such as during energy crises (also known as supply shocks), or changes in inflation expectations, which may be self-fulfilling. Moderate inflation affects economies in both positive and negative ways. The negative effects would include an increase in the opportunity cost of holding money, uncertainty over future inflation, which may discourage investment and savings, and, if inflation were rapid enough, shortages of goods as consumers begin hoarding out of concern that prices will increase in the future. Positive effects include reducing unemployment due to nominal wage rigidity, allowing the central bank greater freedom in carrying out monetary policy, encouraging loans and investment instead of money hoarding, and avoiding the inefficiencies associated with deflation. Today, some economists favour a low

LlaMa:

Select Model:

LLaMA

Submit

**Response:**

Inflation is a complex economic concept, but I'll break it down simply:

**What is inflation?**

Inflation is a sustained increase in the general price level of goods and services in an economy over time. It's a rise in the average cost of living, where prices for everyday items like food, housing, clothing, and other necessities tend to go up.

**Example:**

Imagine you had  
100 last year and could buy 10 burgers with it. But this year, due to inflation, the price of burgers has increased by 20. Now 100 can only buy 8 burgers. That's inflation!

GPT2:

**RAG-Powered AI Assistant with LLaMA3 and GPT-2 Integration**

Chat

Conversation History

Your Query:

what is inflation

Select Model:

GPT-2

Submit

**Response:**

what is inflation?"

"It's not inflation," he said. "It's inflationary."

"I'm not going to say that it's inflationary," he said. "I'm going to say that it's inflationary."

Response Time: 48.15 seconds

### 3.10) What are the latest Advancements in quantum computing?

LlaMa with RAG:

**User:** What are the latest advancements in quantum computing

**Chatbot:** Based on the provided information, one of the latest advancements mentioned is that by 2011, researchers had entangled 14 qubits. This indicates a significant progress in the field of quantum computing.

**Model Used:** LLaMA with RAG

**Response Time:** 24.54 seconds

GPT2 with RAG:

Deploy

What are the latest advancements in quantum computing

Select Model:  
GPT-2 with RAG

Submit

**Response:**  
I have some information that might be relevant to answering a question. Here's the information:  

Research and emerging technologies

DNA-based computing and quantum computing are areas of active research for both computing hardware and software, such as the development of quantum algorithms. Potential infrastructure for future technologies includes DNA origami on photolithography and quantum antennae for transferring information between ion traps. By 2011, researchers had entangled 14 qubits. Fast digital circuits, including those based on Josephson junctions and rapid single flux quantum technology, are becoming more nearly realizable with the discovery of nanoscale superconductors. Fiber-optic and photonic (optical) devices, which already have been used to transport data over long distances, are starting to be used by data centers, along with CPU and semiconductor memory components. This allows the separation of RAM from CPU by optical interconnects. IBM has created an integrated circuit with both electronic and optical information processing in one chip. This is denoted CMOS-integrated nanophotonics (CINP). One benefit of optical interconnects is that motherboards, which formerly required

LlaMa:

**User:** What are the latest advancements in quantum computing

**Chatbot:** There have been many recent advancements in quantum computing, which can be broadly categorized into several areas:

1. **Quantum Error Correction:** Researchers have made significant progress in developing techniques to correct errors that occur during quantum computation. This is a crucial step towards large-scale quantum computing, as errors can quickly accumulate and render the computer useless.
2. **Superconducting Qubits:** Superconducting qubits are one of the most widely used types of quantum bits (qubits) in current quantum computers. Researchers have made significant improvements to their design, leading to higher coherence times and more efficient operations.
3. **Ion Traps:** Ion traps are another type of quantum computing architecture that uses ions instead of qubits. Recent advancements have led to improved control over the ions and increased scalability.
4. **Topological Quantum Computing:** Topological quantum computing is a new paradigm for quantum computing that uses exotic materials called topological insulators. This approach has shown promise for creating more robust and scalable quantum computers.



GPT2:

# LLaMA3 and GPT-2 Integration

Chat Conversation History

Your Query:

What are the latest advancements in quantum computing

Select Model:

GPT-2

Submit

## Response:

what are the latest advancements in quantum computing?

The first step is to understand the quantum state of the system. The second step is to understand the quantum state of the system. The third step is to understand the quantum state of the system.

The first step is to understand

Response Time: 21.57 seconds

#### 4) Query-Wise Comparative Analysis

1. For the query, **“What are the benefits of regular exercise”**, LLaMA with RAG provided a highly relevant and accurate answer, offering deep insights into physical and mental health benefits, with strong context awareness and a response time of 39.6 seconds. GPT-2 with RAG responded faster (28.68s) but was moderately relevant, with approximate accuracy, and provided good context awareness. LLaMA without RAG also performed well in relevance and accuracy, though with a slightly slower response time (41.15s), while GPT-2 without RAG offered low relevance and accuracy, lacking sufficient context.
2. When asked about **“the latest advancements in quantum computing”**, LLaMA with RAG again excelled with high relevance, accuracy, and context awareness, responding in 24.45 seconds. GPT-2 with RAG showed moderate relevance and approximate accuracy, answering in 21.78 seconds. LLaMA, without RAG, was slightly slower (52.85s) but produced moderately relevant and accurate information. GPT-2 without RAG struggled to offer relevant and accurate details, responding with low relevance and limited context.

3. For **“the main festivals celebrated in India”**, LLaMA with RAG offered a highly relevant and accurate response (65.86s), covering cultural context well. GPT-2 with RAG was moderately relevant and accurate, with good context but a longer response time (71.2s). LLaMA without RAG matched relevance and accuracy but was slower (85.65s). GPT-2 without RAG, however, lacked cultural awareness and relevance, providing limited context.
4. Regarding **“the boiling point of water”**, LLaMA with RAG provided a highly relevant and accurate response, noting temperature variations under different conditions, and completed in 78.39 seconds. GPT-2 with RAG was moderately relevant, with approximate accuracy, and responded slightly slower (80.16s). LLaMA maintained high relevance without RAG (76.3s), but GPT-2 without RAG delivered a low relevance and accuracy response with limited context.
5. For **“Who invented the telephone?”**, LLaMA with RAG offered a detailed response with high relevance, accuracy, and context awareness, though with a lengthy response time of 167.79 seconds. GPT-2 with RAG was faster (85.87s) but had less contextual depth. LLaMA without RAG was accurate but provided limited context (74.01s), while GPT-2 without RAG had low relevance and accuracy.
6. On the question, **“Why is recycling important?”**, LLaMA with RAG gave a well-rounded response (78.07s), explaining environmental, economic, and social impacts accurately and with high context. GPT-2 with RAG, while accurate, offered only moderate relevance and limited context in a faster 62.64 seconds. LLaMA without RAG had similar relevance and accuracy, while GPT-2 without RAG missed the deeper benefits, providing limited details.
7. For **“Explain gravity”**, LLaMA with RAG delivered a relevant, accurate answer with good scientific context (43.66s). GPT-2 with RAG was faster at 30.43 seconds but had moderate relevance with approximate accuracy. LLaMA matched the RAG-enabled relevance and accuracy but was slower at 50.3 seconds, while GPT-2 without RAG was less informative.
8. In response to **“What is cricket?”**, LLaMA with RAG answered with high relevance and accuracy, covering rules and history well (45.35s). GPT-2 with RAG was moderately relevant and accurate with good context (44.3s). LLaMA without RAG also maintained high relevance and accuracy, though slightly slower (59.86s), whereas GPT-2 without RAG lacked sufficient detail and context.
9. For **“What is Crime and Punishment?”**, LLaMA with RAG gave a high-relevance but approximate answer in 70.01 seconds. GPT-2 with RAG showed moderate relevance but similar approximation in less time (65.72s). LLaMA, without RAG, provided accurate context and high relevance faster (56.85s). GPT-2 without RAG’s response, however, was limited in accuracy and context.
10. On **“What is inflation?”**, LLaMA with RAG offered a timely, relevant, and accurate response (61.96s), with good economic context. GPT-2 with RAG was moderately relevant, with approximate accuracy, responding in 53.92 seconds. LLaMA without RAG was equally relevant but slightly slower, while GPT-2 without RAG lacked context depth.

11. For “**Who directed RRR?**”, LLaMA with RAG provided a relevant and accurate answer with cultural context in 85.08 seconds. GPT-2 with RAG, moderately relevant and accurate, took slightly less time (80.39s). LLaMA without RAG answered accurately but with slightly less relevance, while GPT-2 without RAG’s response was brief and lacked accuracy.

### Comparative Analysis Summary:

| Query   | Model          | Response Time (s) | Relevance | Accuracy    | Context Awareness |
|---|----------------|-------------------|-----------|-------------|-------------------|
| What are the benefits of regular exercise             | LLaMA with RAG | 39.6              | High      | Accurate    | High              |
| What are the benefits of regular exercise             | GPT-2 with RAG | 28.68             | Moderate  | Approximate | Good              |
| What are the benefits of regular exercise             | LLaMA          | 41.15             | High      | Accurate    | Good              |
| What are the benefits of regular exercise             | GPT-2          | 28.56             | Low       | Inaccurate  | Limited           |
| What are the latest advancements in quantum computing | GPT-2          | 21.27             | Low       | Inaccurate  | Limited           |
| What are the latest advancements in quantum computing | GPT-2 with RAG | 21.78             | Moderate  | Approximate | Good              |
| What are the latest advancements in quantum computing | LLaMA          | 52.85             | Moderate  | Accurate    | Good              |
| What are the latest advancements in quantum computing | LLaMA with RAG | 24.45             | High      | Accurate    | High              |
| What are the main festivals celebrated in India       | LLaMA with RAG | 65.86             | High      | Accurate    | High              |
| What are the main festivals celebrated in India       | GPT-2 with RAG | 71.2              | Moderate  | Accurate    | Good              |
| What are the main festivals celebrated in India       | LLaMA          | 85.65             | High      | Accurate    | Good              |
| What are the main festivals celebrated in India       | GPT-2          | 67.57             | Low       | Inaccurate  | Limited           |
| What is the boiling point of water                    | LLaMA with RAG | 78.39             | High      | Accurate    | Good              |
| What is the boiling point of water                    | GPT-2 with RAG | 80.16             | Moderate  | Approximate | Limited           |
| What is the boiling point of water                    | LLaMA          | 76.3              | High      | Accurate    | Good              |
| What is the boiling point of water                    | GPT-2          | 77.22             | Low       | Inaccurate  | Limited           |
| Who invented the telephone?                           | LLaMA with RAG | 167.79            | High      | Accurate    | High              |
| Who invented the telephone?                           | GPT-2 with RAG | 85.87             | High      | Accurate    | Limited           |

|  |                |       |          |             |         |
|--|----------------|-------|----------|-------------|---------|
| <b>Who invented the telephone?</b>             | LLaMA          | 74.01 | High     | Accurate    | Limited |
| <b>Who invented the telephone?</b>             | GPT-2          | 75.29 | Low      | Inaccurate  | Limited |
| <b>Why is recycling important</b>              | LLaMA with RAG | 78.07 | High     | Accurate    | High    |
| <b>Why is recycling important</b>              | GPT-2 with RAG | 62.64 | Moderate | Accurate    | Limited |
| <b>Why is recycling important</b>              | LLaMA          | 61.7  | High     | Accurate    | Good    |
| <b>Why is recycling important</b>              | GPT-2          | 54.88 | Low      | Inaccurate  | Limited |
| <b>explain gravity</b>                         | LLaMA with RAG | 43.66 | High     | Accurate    | Good    |
| <b>explain gravity</b>                         | GPT-2 with RAG | 30.43 | Moderate | Approximate | Good    |
| <b>explain gravity</b>                         | LLaMA          | 50.3  | High     | Accurate    | High    |
| <b>explain gravity</b>                         | GPT-2          | 29    | Low      | Inaccurate  | Limited |
| <b>what is cricket</b>                         | LLaMA with RAG | 45.35 | High     | Accurate    | High    |
| <b>what is cricket</b>                         | GPT-2 with RAG | 44.3  | Moderate | Accurate    | Good    |
| <b>what is cricket</b>                         | LLaMA          | 59.86 | High     | Accurate    | High    |
| <b>what is cricket</b>                         | GPT-2          | 44.98 | Low      | Inaccurate  | Limited |
| <b>what is crime and punishment</b>            | LLaMA with RAG | 70.01 | High     | Approximate | Limited |
| <b>what is crime and punishment book about</b> | LLaMA with RAG | 65.72 | Moderate | Approximate | Limited |
| <b>what is crime and punishment book about</b> | GPT-2 with RAG | 56.85 | Moderate | Approximate | High    |
| <b>what is crime and punishment book about</b> | LLaMA          | 65.52 | High     | Accurate    | Good    |
| <b>what is crime and punishment book about</b> | GPT-2          | 54.63 | Low      | Inaccurate  | Limited |
| <b>what is inflation</b>                       | LLaMA with RAG | 61.96 | High     | Accurate    | Good    |
| <b>what is inflation</b>                       | GPT-2 with RAG | 53.92 | Moderate | Approximate | Good    |
| <b>what is inflation</b>                       | LLaMA          | 62.89 | High     | Accurate    | Good    |
| <b>what is inflation</b>                       | GPT-2          | 52.44 | Low      | Inaccurate  | Limited |
| <b>who directed RRR</b>                        | LLaMA with RAG | 85.08 | High     | Accurate    | Good    |
| <b>who directed RRR</b>                        | GPT-2 with RAG | 80.39 | Moderate | Approximate | High    |
| <b>who directed RRR</b>                        | LLaMA          | 76.64 | Moderate | Approximate | High    |
| <b>who directed RRR</b>                        | GPT-2          | 78.95 | Low      | Inaccurate  | Limited |

## Observations:

### Relevance, Accuracy, and Context-Awareness

- **RAG System:** The RAG system generally provided more context-aware and relevant responses compared to standalone LLaMA and GPT-2. For example, in answering "What are the latest advancements in quantum computing?", RAG retrieved detailed information about quantum error correction, superconducting qubits, and quantum machine learning. However, the standalone models either repeated information or were less focused. The responses using RAG were also richer in factual content.
- **Standalone LLaMA:** Provided answers that were almost as good as RAG in many cases, showing decent coherence and relevance, though it may have faced some challenges in handling certain general knowledge questions. LLaMA performed well when backed by RAG.
- **Standalone GPT-2:** GPT-2 performed worst among the four variations in terms of generating coherent responses, for domain-specific queries like "Explain gravity" or "What are the latest advancements in quantum computing?", it often repeated sentences or lacked sufficient depth. GPT-2 with RAG, however, showed a notable improvement, giving more detailed and contextually accurate answers.

### Response Times and Resource Usage

- **RAG System:** Response times ranged from 30 to 80 seconds on average. While slower than standalone models due to the retrieval component, it was expected due to the additional steps involved in fetching relevant documents from the knowledge base.
- **Standalone LLaMA:** LLaMA was slightly faster than RAG but produced significantly less Focused results sometimes. Response times ranged from 52 to 85 seconds.
- **Standalone GPT-2:** GPT-2 was consistently faster, often providing responses in 28-54 seconds, but the quality of the output was mixed when not backed by retrieval. It often repeated sentences or lacked sufficient depth.

## 5) Reflection and documentation:

The integration of a Retrieval-Augmented Generation (RAG) system significantly enhances chatbot performance compared to using standalone local LLMs, addressing key limitations in relevance, accuracy, and context awareness. By adding a retrieval mechanism to the generation process, RAG systems bring a dynamic, knowledge-informed layer to responses, particularly benefiting complex and domain-specific inquiries that require current or specialized information.

From the responses it's clear that the RAG system significantly enhances GPT-2 and LLaMA models by improving relevance, accuracy, and context-awareness. Comparing RAG-enabled versions of GPT-2 and LLaMA to their standalone counterparts reveals clear advantages, especially in handling specialized or context-sensitive queries.

For instance, with **GPT-2 using RAG**, responses in areas like “the latest advancements in quantum computing” and “the benefits of regular exercise” showed marked improvement in both relevance and detail. Standalone GPT-2 tended to give broad, surface-level responses, often lacking specific details necessary to address complex queries fully. However, RAG-equipped GPT-2 could retrieve detailed information on quantum technologies and health benefits, boosting relevance and depth. This retrieval layer effectively transformed GPT-2 from a generalist into a more informed responder, especially in specialized topics.

Similarly, **LLaMA with RAG** outperformed its non-RAG version in complex inquiries such as “festivals celebrated in India” and “who invented the telephone?” While standalone LLaMA provided factual but simplified responses, RAG-equipped LLaMA accessed multiple sources, providing richer cultural or historical details. For example, it discussed various festivals in India and explained multiple contributors to the invention of the telephone, making the response more contextually aware and nuanced. This enhancement in LLaMA’s relevance and context sensitivity turned it into a more adaptable and responsive tool.

**Accuracy** also saw clear improvement, especially evident in fact-heavy questions. For example, in “who directed RRR?” and “what is inflation?”, RAG-enabled versions of both GPT-2 and LLaMA offered precise answers. Standalone versions were more likely to generalize or give incomplete answers, particularly on detailed topics. With RAG, both models accessed exact details, including the director’s name and inflation’s economic implications, proving preferable for users seeking specific information.

**Context-awareness** was another major benefit of RAG integration. In questions like “what is gravity?” and “why is recycling important?”, RAG allowed both models to retrieve scientific or environmental contexts, resulting in responses that were accurate and relevant to the query’s intent. For example, standalone GPT-2 and LLaMA discussed gravity and recycling in broad terms. However, with RAG, these models retrieved situationally appropriate information, leading to responses that captured scientific theories or environmental and economic aspects of recycling.

The primary trade-off observed was a **slightly longer response time** for RAG-enabled models due to the retrieval step. LLaMA with RAG, for instance, took significantly longer on some responses, particularly in complex queries like “who invented the telephone?” where historical context was retrieved. However, the enhanced response quality justified the minor delay, and the time remained within a functional range for most practical uses.

Overall, the data indicates that adding RAG transforms both GPT-2 and LLaMA from basic response generators into more accurate, contextually aware assistants. This enhancement allows them to handle a broader range of topics with greater depth, accuracy, and reliability, which would be difficult to achieve with standalone LLMs alone. RAG integration thus makes both models more practical for real-world applications where detail and precision are essential, proving that retrieval is a valuable strategy for enhancing LLM capabilities.

## Challenges Faced and Solutions Devised

### 1.Package Installation Issues

- **Challenge:** During the setup phase, I encountered several issues while installing various required packages. Conflicts between package versions and dependencies made it difficult to achieve a stable environment.
- **Solution:** To mitigate these issues, I set up a dedicated Python environment using `virtualenv`. This isolated environment ensured that package dependencies did not conflict with other projects and allowed me to install the required packages without issues. Using a `requirements.txt` file also helped streamline the installation process.

### 2.Implementing GPT-2 with RAG

- **Challenge:** Integrating GPT-2 with RAG posed a significant challenge due to its input token length limitations. The model could only process a limited number of tokens at a time, which made it difficult to handle larger queries or responses effectively.
- **Solution:** To address this limitation, I had to implement manual handling of input token lengths in the code. I implemented a strategy to concatenate responses when necessary, ensuring that the chatbot could still provide comprehensive answers despite the token restrictions.