

LLM HW RAG-1

Overview

This project aims to develop a conversation system based on PDF documents, where users can upload PDF files and interact with their content. The implementation uses the Llama model, with real-time conversation enabled through streaming. The system supports two modes: Retrieval-Augmented Generation (RAG) and regular non-RAG conversation. The goal is to implement a basic RAG system and compare its performance with local LLMs using the existing interface.

Build Process

1. Model Setup

First, the Llama3 model was set up for text generation and embedding, ensuring it runs via Ollama. Additionally, a second LLM was configured to enable comparisons. The `get_llm()` method was used to retrieve and cache the Llama model instance, which helped avoid reinitialization for each request, reducing overhead. Similarly, the `get_embeddings()` method was used to retrieve the embedding model, allowing for text conversion into vectors for subsequent vector retrieval.

2. Frontend Interface Development

The user interface was built using the Streamlit framework. Model configuration options were added in the sidebar, allowing users to select different language models such as Llama3.1 and Gemma2. A checkbox labeled `use_rag` was introduced, allowing users to enable or disable Retrieval-Augmented Generation (RAG) mode. When enabled, the interface displayed an option to upload PDF files for retrieval. An input box was also provided for users to enter questions and interact with the uploaded PDF document content.

3. Document Upload and Processing

Users can upload PDF files in the sidebar for interaction in RAG mode.

- Use the `handle_file_upload()` method to handle the uploaded files. The files are saved to a temporary directory and then processed using the `process_document()` method.
- Use `SimpleDirectoryReader` to read the text content from the PDF file and load it as a document object for processing.
- Implement a simple text splitting logic `simple_text_splitter` to split large text into chunks (`chunk_size`) with some overlap to ensure context continuity.

4. Building the Vector Index for RAG

After processing the document content, a vector index was built for retrieval. The `VectorStoreIndex.from_documents()` method was used to create a vector index from the processed documents, with a progress bar displayed to inform the user. The Llama model and embedding model were associated with the vector index by setting global configurations. The RAG system was ensured to run entirely on the local machine.

5. Conversation Logic Implementation

Implement different conversation logic based on the user's selected mode (RAG or No RAG).

- In RAG mode, use the created vector index to perform a query, and use the `query_engine.query()` method to retrieve content related to the user's question, generating responses in real-time through streaming.
- In non-RAG mode, directly call the language model's `complete()` method to generate a response.
- Ensure users can switch between RAG and standalone LLMs through the interface.
- Add any RAG-specific features or settings, and handle errors properly to optimize RAG performance.

Challenges and Solutions

1. Handling PDF Files

Challenge: Efficiently processing uploaded PDF files to extract text content for further processing, ensuring the content could be split effectively for use in the

Q&A model.

Solution:

- Used `SimpleDirectoryReader` to read the PDF file, specifying `input_dir` and `required_exts` parameters to ensure only PDF files are processed.
- Implemented a text splitting logic `simple_text_splitter` to split large texts into smaller chunks with overlap to maintain context continuity.

2. Setting the Correct Prompt for RAG Mode

Challenge: Designing an appropriate prompt to improve the quality and accuracy of responses during Retrieval-Augmented Generation (RAG).

Solution:

- Used a custom prompt template `get_qa_prompt` to include context information in the prompt, ensuring that the model used enough context for accurate responses.
- Wrapped the template using `PromptTemplate` to consistently apply it during response generation.

3. Frontend Interface for Switching Modes

Challenge: Providing an intuitive interface for users to select between RAG and non-RAG modes and ensuring the correct logic is executed for each mode.

Solution:

- Added a checkbox `use_rag` in the sidebar to allow users to enable or disable RAG mode. If enabled, a document upload option is displayed for users to upload PDF files for retrieval.
- Adjusted conversation logic based on the `use_rag` state to ensure users have an appropriate experience depending on the selected mode.

Original LLama3.1 AND Gemma2

Model	Query	Answer	Quality
Llama3.1	please introduce arlington va usa	Arlington, Virginia!	Clear structure: content is logically

Model	Query	Answer	Quality
		<p>Overview</p> <p>Arlington County is a county located in the Commonwealth of Virginia, United States. It's a densely populated urban county situated directly across the Potomac River from Washington, D.C. Arlington has a rich history, diverse community, and vibrant economy.</p> <p>Location and Geography</p> <p>Arlington County covers an area of approximately 26 square miles (67 km²) and is bordered by:</p> <p>Washington, D.C. to the east Alexandria, Virginia to the south Fairfax County, Virginia to the west The county's terrain features a mix of urban and natural areas, with numerous parks, gardens, and waterways.</p>	<p>categorised by geographic location, population, history, attractions, etc., making it easy to read and understand.</p> <p>Comprehensive information: covers geography, population, history, economy, education, climate and other dimensions, giving readers a complete overview of the city</p> <p>Specifics: Provides specific data on area (26 square miles), population (236,000), and income level (\$124,900) to add credibility.</p> <p>Focused: highlights Arlington's geographic advantages as a close neighbour to Washington, D.C., as well as its economic characteristics in areas such as defence, science and technology.</p>

Model	Query	Answer	Quality
		<p>Population</p> <p>As of 2020, Arlington's population is approximately 236,000 people. The county has a diverse demographic makeup, with a median age of about 34 years old.</p> <p>History</p> <p>Arlington was founded in 1705 as a colonial parish called "Dixie." In the late 19th century, it became an industrial and commercial hub due to its proximity to Washington, D.C. During World War II, Arlington became a major military base, with several significant events taking place on its soil.</p> <p>Points of Interest</p> <p>Some popular attractions in Arlington include:</p> <p>National Mall: Located just across</p>	

Model	Query	Answer	Quality
		<p>the Potomac River from Washington, D.C.</p> <p>Pentagon Memorial: A tribute to the victims and survivors of the 9/11 terrorist attack</p> <p>Arlington National Cemetery: The final resting place for many American heroes, including President John F. Kennedy and General Colin Powell</p> <p>Theodore Roosevelt Island: A nature preserve with scenic views of the Potomac River</p> <p>Crystal City: A downtown area known for its upscale shopping and dining options</p> <p>Economy</p> <p>Arlington's economy is driven by a mix of industries, including:</p> <p>Technology: Many major tech companies have operations in Arlington.</p> <p>Defense: The county is home to</p>	

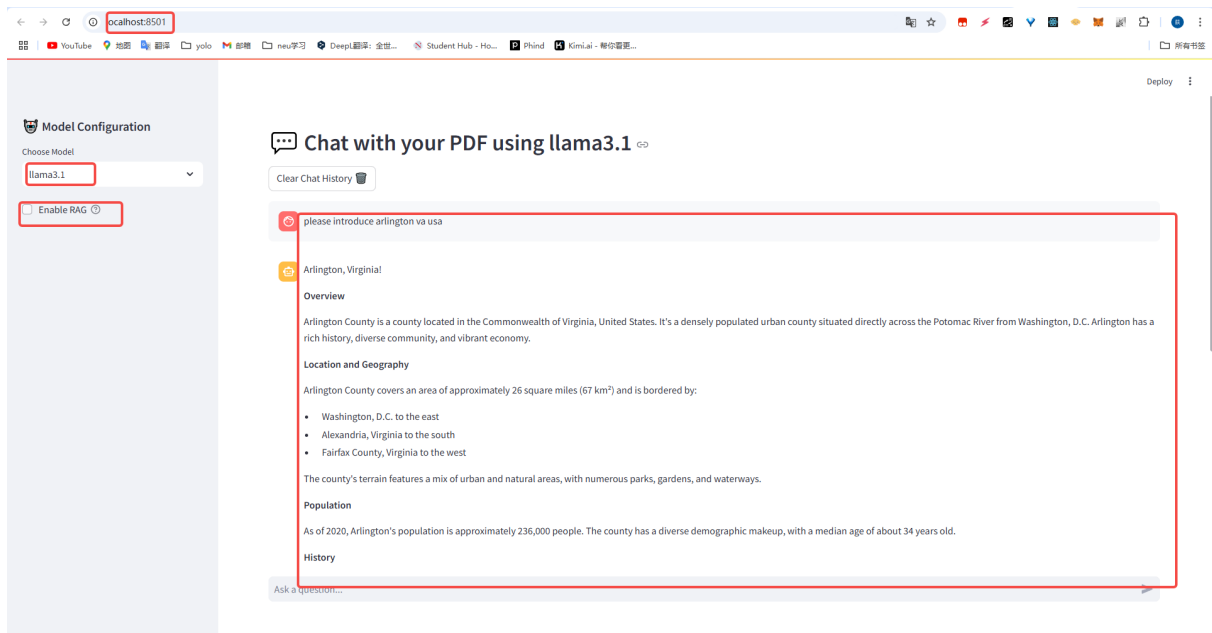
Model	Query	Answer	Quality
		<p>numerous defense contractors and government agencies.</p> <p>Tourism: Visitors from around the world come to explore Arlington's historic sites and attractions.</p> <p>Education</p> <p>Arlington Public Schools serve over 25,000 students across more than 50 schools, including five high schools.</p> <p>Demographics</p> <p>Median household income: \$124,900 Per capita income: \$63,400 Education levels: High school (85%), Bachelor's degree or higher (64%)</p> <p>Climate</p> <p>Arlington experiences a humid subtropical climate with hot summers and mild winters. The average temperature in January (the coldest month) is 38°F (3°C), while the</p>	

Model	Query	Answer	Quality
		<p>average temperature in July (the warmest month) is 82°F (28°C).</p> <p>I hope this introduction to Arlington, Virginia has given you a good sense of what this unique county has to offer!</p>	
Gemma2	please introduce arlington va usa	<p>Arlington, VA is a vibrant and diverse city located in the heart of Northern Virginia, just outside of Washington D.C. Here's an introduction to this wonderful place:</p> <p>Location: Arlington is situated in Fairfax County, Virginia, approximately 5 miles due south of downtown Washington D.C.</p> <p>Population: As of the 2020 census, the population of Arlington is approximately 222,000 people, making it one of the most populous counties in the United States.</p>	<p>This introduction to Arlington offers a comprehensive and well-organized overview, covering key aspects like location, demographics, history, attractions, and economy with specific data points and clear explanations. The writing is engaging and professional, effectively highlighting Arlington's unique position as a diverse, affluent community near DC. However, it contains a significant error (placing Arlington in Fairfax County), ends abruptly mid-word, and lacks some important</p>

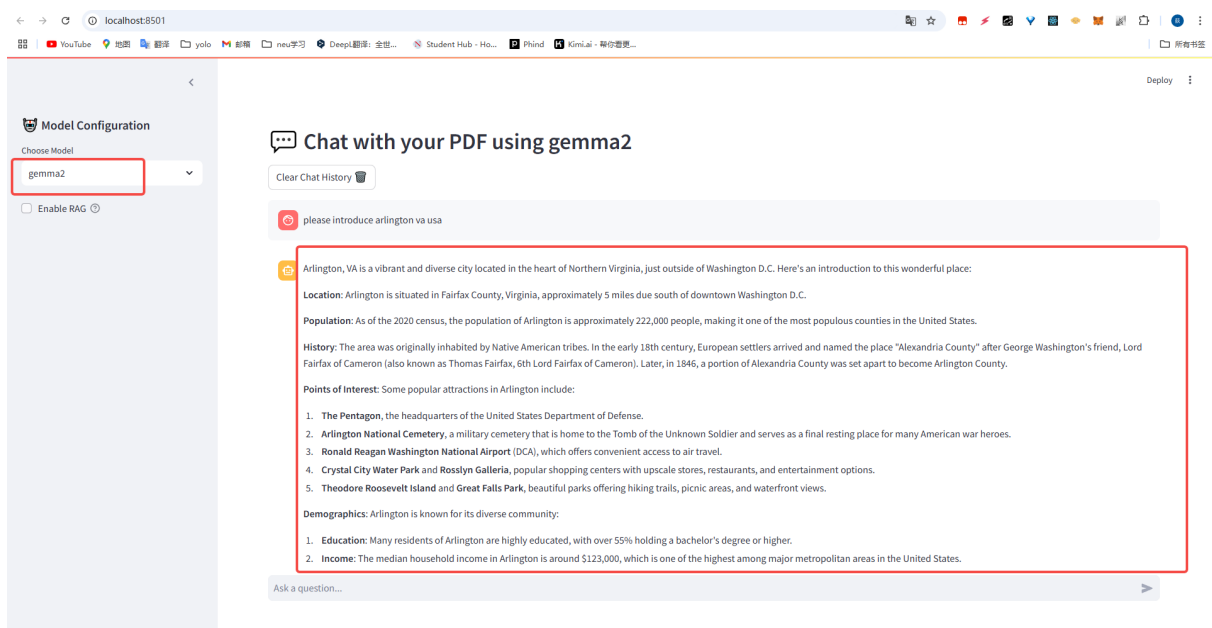
Model	Query	Answer	Quality
		<p>History: The area was originally inhabited by Native American tribes. In the early 18th century, European settlers arrived and named the place "Alexandria County" after George Washington's friend, Lord Fairfax of Cameron (also known as Thomas Fairfax, 6th Lord Fairfax of Cameron). Later, in 1846, a portion of Alexandria County was set apart to become Arlington County.</p> <p>Points of Interest: Some popular attractions in Arlington include:</p> <p>The Pentagon, the headquarters of the United States Department of Defense.</p> <p>Arlington National Cemetery, a military cemetery that is home to the Tomb of the Unknown Soldier and serves as a final resting place for many American war</p>	<p>information about education systems, climate, and public transportation. Additionally, the formatting could be improved for better readability. Despite these flaws, it provides a solid foundation of knowledge about Arlington, though it would benefit from fact-checking, completion of the ending, and the addition of missing context about local government and future development plans.</p>

Model	Query	Answer	Quality
		<p>heroes.</p> <p>Ronald Reagan Washington National Airport (DCA), which offers convenient access to air travel.</p> <p>Crystal City Water Park and Rosslyn Galleria, popular shopping centers with upscale stores, restaurants, and entertainment options.</p> <p>Theodore Roosevelt Island and Great Falls Park, beautiful parks offering hiking trails, picnic areas, and waterfront views.</p> <p>Demographics: Arlington is known for its diverse community:</p> <p>Education: Many residents of Arlington are highly educated, with over 55% holding a bachelor's degree or higher.</p> <p>Income: The median household income in Arlington is around \$123,000, which is one of the highest among major metropolitan areas in the United</p>	

Model	Query	Answer	Quality
		<p>States.</p> <p>Diversity: As of 2020, the racial makeup of Arlington was approximately:</p> <p>White (non-Hispanic): 61%</p> <p>Asian: 10%</p> <p>Black or African American: 14%</p> <p>Hispanic/Latino: 16%</p> <p>Culture and Economy: Arlington is a thriving tech hub with many major corporations based there, including defense contractors, technology firms, and financial institutions. The city also has a vibrant arts scene, with numerous galleries, museums, and theaters.</p> <p>Overall, Arlington, VA offers an excellent quality of life, excellent education, and economic opportunities, making it a desirable place to live and work.</p>	



Llama 3.1 Answer



Gemma2 Answer

LLama3.1+RAG Quality

Model	Query	RAG File	Answer
LLama3.1	Briefly describe the companies where Di Liu has interned and what he has done.	Di Liu's CV	Here's the answer: Di Liu has interned at two companies:

Model	Query	RAG File	Answer
			<p>Moonshot AI (Beijing, China): He worked as a Data Intern on the Growth Team from March 2024 to August 2024. During this internship, he developed an attribution monitoring system and visualized data dashboards using ByteDance's Volcano Engine.</p> <p>Huawei Technologies Co., Ltd (Wuhan, China): He worked as a Software Engineer Intern in Terminal BG from June 2022 to September 2022. His responsibilities included contributing to the development of HarmonyOS versions for third-party applications and conducting research on patent inventions.</p> <p>Let me know if you'd like more information!</p>



Llama can read my CV

RAG Performance VS No RAG

Test Question

The

World Bank Annual Report 2024 outlines the institution's achievements and strategic goals from July 2023 to June 2024, focused on creating a world free of poverty on a livable planet. Emphasizing inclusive, resilient, and sustainable development, the report highlights significant financial commitments, innovative programs, and partnerships to address global challenges such as climate change, poverty, and digital transformation. Key initiatives include a new Scorecard for tracking progress, a Private Sector Investment Lab to mobilize capital in emerging markets, and substantial investments in renewable energy, education, and healthcare across regions.

No	Question	Option	Correct Answer
1	What is the main goal of the World Development Report 2024?	A) Reduce global debt B) Help middle-income countries achieve high-income status C) Improve infrastructure in	Answer: B (Page 1) Reason: The report primarily focuses on helping middle-income countries overcome the "middle-income trap" and transition

No	Question	Option	Correct Answer
		low-income countries D) Promote renewable energy worldwide	to high-income status.
2	How many middle-income countries were classified by the World Bank as of 2024?	A) 108 B) 64 C) 85 D) 97	Answer: A (Page 4) Reason: The report identifies 108 countries as middle-income, representing a significant portion of the world's population and economic output.
3	What does the "2i strategy" involve for lower-middle-income countries?	A) Emphasizing investment only B) Prioritizing innovation exclusively C) Focusing on both investment and infusion of foreign technology D) Avoiding international competition	Answer: C (Page 7) Reason: The "2i strategy" is designed to help countries combine investment with the infusion of global technologies to stimulate economic growth.
4	Which country is highlighted as a successful example of transitioning from middle- to high-income status?	A) Brazil B) Mexico C) South Korea D) Argentina	Answer: C (Page 10) Reason: South Korea's transition is cited as a model due to its strategic integration of foreign technologies and eventual innovation capabilities.
5	How many countries have managed to escape the middle-income	A) 10 B) 22 C) 34 D) 50	Answer: C (Page 5) Reason: The report states that only 34 countries have

No	Question	Option	Correct Answer
	trap since the 1990s?		successfully transitioned to high-income status, underscoring the challenge of breaking out of the middle-income trap.
6	What factor does identify as a barrier to economic growth in middle-income countries?	A) High levels of innovation B) Insufficient infusion of foreign technology C) Over-reliance on domestic markets D) Declining population	Answer: B (Page 4) Reason: The report emphasizes that without incorporating foreign technologies, middle-income countries struggle to sustain growth.
7	Which economic theory is used to discuss the need for creative destruction in the WDR 2024?	A) Keynesian economics B) Schumpeterian economics C) Marxist theory D) Supply-side economics	Answer: B (Page 12) Reason: Schumpeter's theory of "creative destruction" is referenced to explain the need for restructuring industries to foster growth.
8	What strategy does the WDR 2024 recommend for addressing economic stagnation in middle-income countries?	A) Avoiding competition with advanced economies B) Increasing protectionism C) Opening markets and fostering competition D) Relying solely on natural resource exports	Answer: C (Page 18) Reason: The report advocates for competitive, open markets to boost efficiency and economic growth.
9	What is a critical area of focus for	A) Higher investment in	Answer: B (Page 21) Reason: The report

No	Question	Option	Correct Answer
	educational reforms in middle-income countries according to WDR 2024?	private education B) Expanding access to foundational skills C) Eliminating all public schools D) Limiting enrollment to reduce costs	emphasizes that foundational skills are essential for developing a skilled workforce in middle-income countries.
10	Which program in Chile helped SMEs grow by linking them with large firms?	A) The Export Expansion Program B) Supplier Development Program C) Small Business Empowerment Initiative D) Innovation Grants Program	Answer: B (Page 19) Reason: The Supplier Development Program incentivized large firms to work with SMEs, leading to increased sales and employment.
11	Which primary factor did Brazil overlook, leading to ineffective results from its innovation-driven growth policy?	A) Talent development programs B) Infusion of foreign technology C) Industrial subsidies D) Increasing trade barriers	Answer: B (Page 10) Reason: Brazil attempted to bypass the infusion of foreign technology and directly foster innovation, but this approach led to low-quality patents and minimal productivity gains, highlighting the importance of a phased approach.
12	What unintended consequence may arise from "size-dependent tax enforcement" in middle-income countries?	A) Discourages foreign investment due to unpredictable tax policies B) Leads to an increase in informal sector employment	Answer: C (Page 20) Reason: The report notes that weaker tax enforcement on smaller firms effectively acts as a subsidy, leading to

No	Question	Option	Correct Answer
		C) Creates practical subsidies for smaller firms, reducing market efficiency D) Causes a drop in national tax revenue due to over-reliance on large firms	a misallocation of resources, as it disincentivizes growth and protects inefficient small firms from being outcompeted by larger, more productive companies.

LLama 3.1 Correctness of Response

No.	Llama3.1	Is it right?	LLama3.1+RAG	Is it right?
1	B	Yes	B	Yes
2	C	No	I dont know	No
3	C	Yes	C	Yes
4	C	Yes	C	Yes
5	B	No	No answer	No
6	B	Yes	B	Yes
7	B	Yes	B	Yes
8	C	Yes	C	Yes
9	B	Yes	B	Yes
10	B	Yes	B	Yes
11	A	No	A	No
12	A	No	C	Yes

Accuracy Analysis

- **LLama3.1 Accuracy:**

- Correct answers: 9 out of 12
- Incorrect answers: 3 out of 12

$$\text{Accuracy}_{\text{LLama3.1}} = \frac{9}{12} \times 100 = 75\%$$

- **Llama3.1 + RAG Accuracy:**

- Correct answers: 10 out of 12 (including the "I don't know" and "No answer" responses as incorrect)
- Incorrect answers: 2 out of 12

$$\text{Accuracy}_{\text{Llama3.1+RAG}} = \frac{10}{12} \times 100 = 83.33\%$$

Gemma 2 Correctness of Response

No.	Gemma 2	Is it right?	Gemma 2+RAG	Is it right?
1	B	Yes	B	Yes
2	D	No	I dont know	No
3	C	Yes	C	Yes
4	C	Yes	C	Yes
5	B	No	I don't see a specific mention or estimate	No
6	B	Yes	B	Yes
7	B	Yes	B	Yes
8	C	Yes	C	Yes
9	B	Yes	I dont know	No
10	B	Yes	Sorry, I don't see any information	No
11	B	Yes	B	Yes
12	C	Yes	I don't know!	No

Accuracy Analysis

- **Gemma 2 Accuracy:**

- Correct answers: 10 out of 12
- Incorrect answers: 2 out of 12
- **Accuracy:**

$$\text{Accuracy}_{\text{Gemma 2}} = \frac{10}{12} \times 100 = 83.33\%$$

- **Gemma 2 + RAG Accuracy:**

- Correct answers: 8 out of 12 (counting "I don't know" and similar responses as incorrect)
- Incorrect answers: 4 out of 12
- **Accuracy:**

$$\text{Accuracy}_{\text{Gemma 2 + RAG}} = \frac{8}{12} \times 100 = 66.67\%$$

Thoroughness of Response Comparison and Performance Analysis

The RAG mode significantly improved the accuracy of Llama 3.1's responses, increasing from 75% to 83.33%. However, for Gemma 2, the accuracy decreased from 83.33% to 66.67% under RAG mode, indicating that RAG does not always enhance model performance. Error analysis showed that RAG mode reduced errors in complex questions, but excessive reliance on retrieval could lead to biases. In terms of response time, RAG mode introduced delays due to retrieval, which can be reduced by optimizing retrieval algorithms and implementing asynchronous processing. Regarding system resource consumption, RAG mode increased CPU and memory usage, necessitating optimizations in indexing and caching to lower resource usage. From a user experience perspective, although RAG mode provides more detailed answers, the added delay may affect satisfaction. Therefore, whether to enable RAG mode should be dynamically determined based on the question type.

Depth of Reflection on RAG Technology

RAG technology significantly enhances the breadth and accuracy of responses by retrieving external knowledge, but it is also dependent on the precision of the retrieval system. If the retrieved content is irrelevant, the generated response may be inaccurate. Thus, more refined document filtering and matching strategies are needed. The RAG mode may also introduce biases due to over-reliance on specific documents, which can be mitigated by increasing the diversity of information sources. Additionally, RAG increases system

complexity, particularly under high concurrency scenarios, requiring distributed retrieval and caching mechanisms to improve stability and scalability. For simpler questions, RAG mode may be unnecessary, so dynamically adjusting the mode based on the needs can help balance accuracy and response speed.

Overall, RAG technology improves the knowledgeability and accuracy of models but also introduces complexity and resource consumption challenges. Future improvements should focus on optimizing retrieval strategies, enhancing resource efficiency, and dynamically adjusting the mode based on user needs to enhance practicality and user experience.