RAG (Retrieval-Augmented Generation)

Find the project link here:

https://colab.research.google.com/drive/1baGm3Yzd4sotGa2Y9BGaYXCM8IXTc8l1?usp=sharing

This RAG setup combines a retrieval system using a FAISS index with a generation model to produce informed responses based on the content of PDF documents.

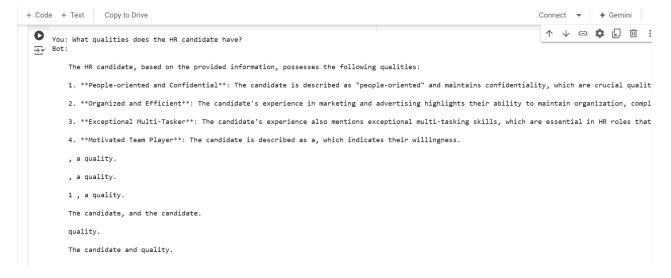
Setup Process

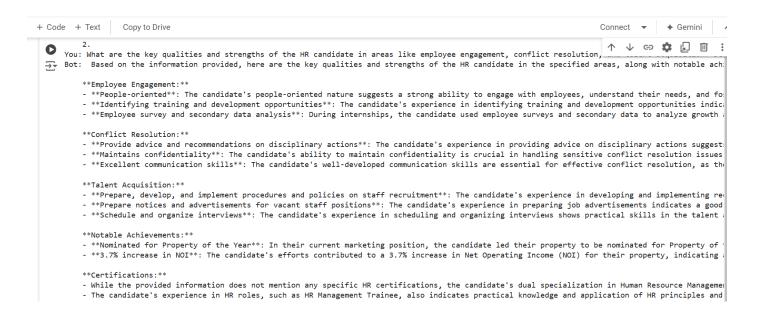
- 1. PDF Text Extraction: The `extract_text_from_pdf` and `extract_text_from_multiple_pdfs` functions read and extract text from PDF files, preparing the content for indexing and retrieval.
- 2. Text Chunking and Splitting: The `recursive_split_text` function chunks the text into manageable sizes, ensuring smooth processing in the FAISS index.
- 3. Vector Store Creation: Using the `create_faiss_index` function, FAISS indexes are created from text chunks, which are embedded using the `HuggingFaceEmbeddings` model, allowing for efficient retrieval.
- 4. Query Handling: The `handle_user_input` function integrates retrieval with generation. Relevant document content is retrieved, and conversation history is formatted before being sent to the generation model to produce a detailed response.
- 5. Response Generation: The `generate_response_from_inference_api` function uses an inference API to generate a response, incorporating both retrieved text and conversation context.

5 optimized prompts

First Example - HR

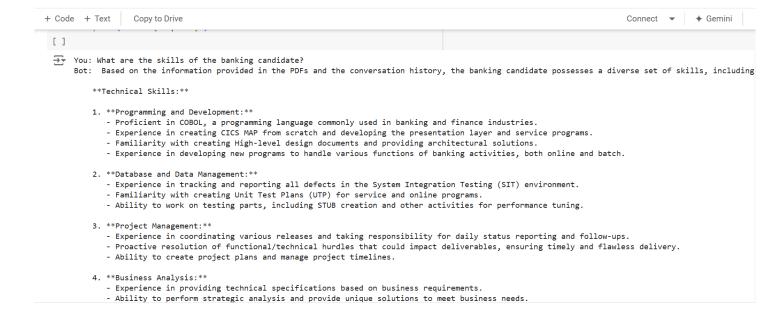
Before optimizing the prompt

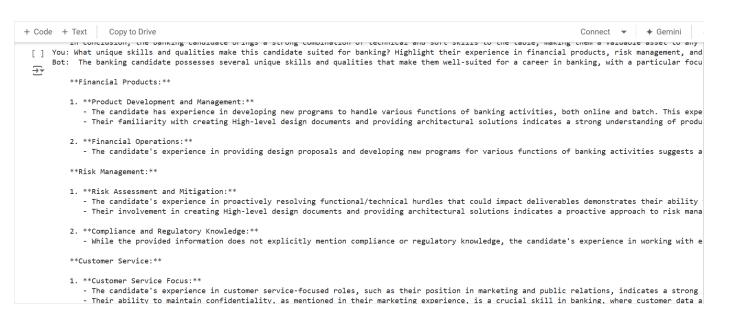




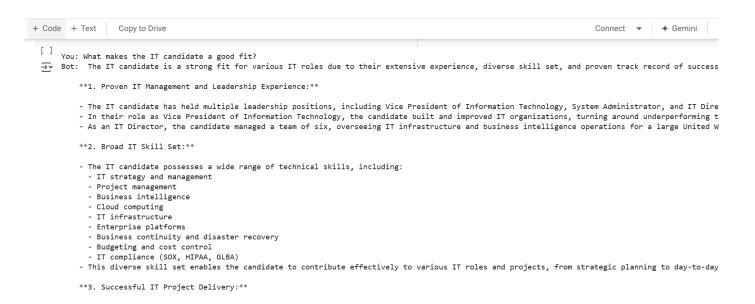
Second example – Banking

Before optimizing the prompt

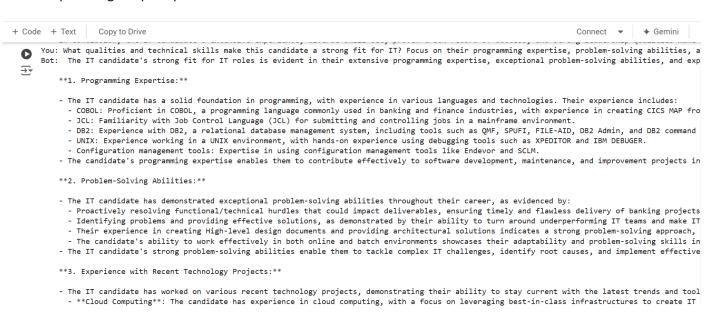




Before optimizing the prompt



After optimizing the prompt



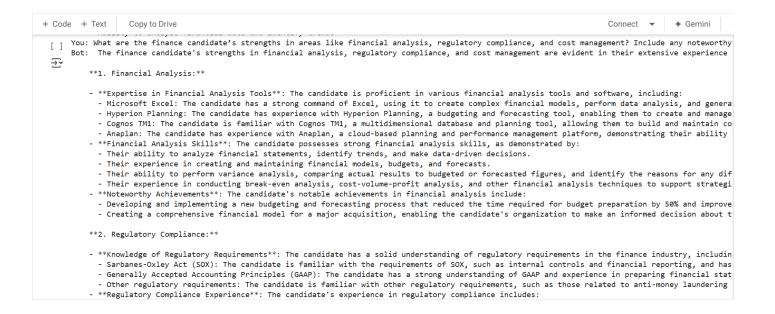
Fourth example - Finance

Before optimizing the prompt

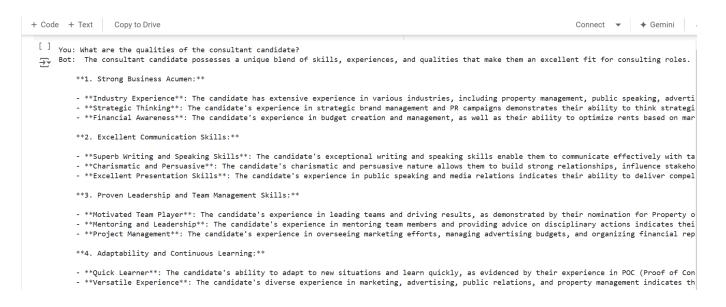
```
+ Code + Text Copy to Drive
                                                                                                                              Connect ▼

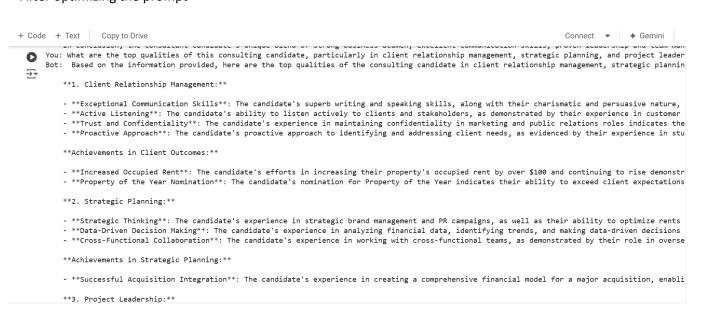
→ Gemini

          user_input = input("You: ")
          if user input.lower() == "exit":
          response = handle_user_input(user_input, vectorstore, conversation_history)
          print(f"Bot: {response}")
 You: What are the strengths of the finance candidate?
      Bot: The finance candidate's strengths are the following:
          - Strong financial skills and knowledge
          - Experience in financial analysis and reporting
          - Ability to manage and analyze financial data
          - Strong financial analysis skills
          - Ability to analyze financial data and identify trends
          - Strong financial analysis skills
          - Ability to analyze financial data and identify trends
          - Ability to analyze financial data and identify trends
          - Ability to analyze financial data and identify trends
```

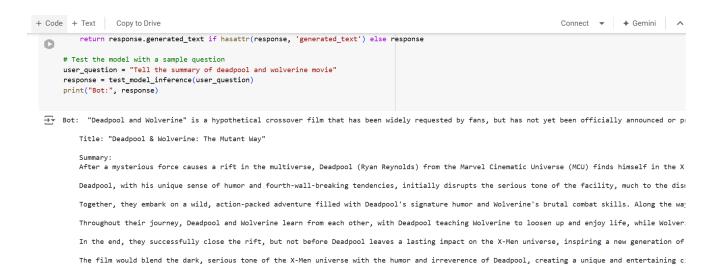


Before optimizing the prompt





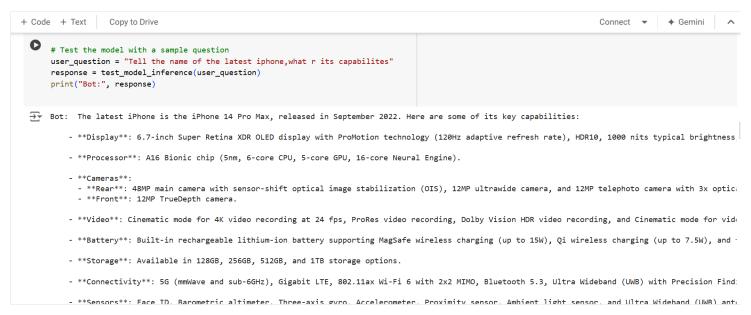
Results of model before implementing RAG



Without the RAG implementation, the language model's knowledge was only current up to the end of 2023. This led to outdated responses, such as referencing older movies like *Deadpool and Wolverine: The Mutant Way* instead of the latest releases.



The model mistakenly references the 2021 *Ghostbusters* movie as "upcoming" due to its last knowledge update only covering information up to 2023.



The model generates information on the iPhone 14 Pro, which indicates that without Retrieval-Augmented Generation (RAG) implementation, the LLM's knowledge cutoff is 2022, resulting in outdated responses for more recent queries.

Results of model after implementing RAG

The model is now capable of providing the latest information through Retrieval-Augmented Generation (RAG). Here's the latest on the Wolverine and Deadpool movies, the iPhone 16, and the Ghostbusters movie for 2024:



- I faced considerable difficulties in acquiring datasets that effectively demonstrate the usefulness of Retrieval-Augmented Generation (RAG). The limited availability of relevant, high-quality data underscored the challenges of showcasing the model's capabilities
- I also explored various prompt optimization techniques to enhance the model's responses. This process involved fine-tuning prompts for clarity and specificity to maximize the relevance and insightfulness of the generated answers

Improvements Observed

- Enhanced Answer Relevance: By providing the model with only the most relevant retrieved information, responses became more accurate and contextually appropriate.
- Efficiency Gains in Query Processing: Utilizing the FAISS index for retrieval greatly improved the processing speed for relevant information retrieval, especially when compared to a purely generative model approach.
- Scalability for Larger Document Sets: The setup is efficient and capable of handling multiple documents, with scalability made feasible by FAISS indexing and document chunking, allowing for faster and more accurate responses even as the document set grows.

This RAG setup has streamlined and enhanced responses based on document content, improving the assistant's capacity for generating informed answers on a range of queries.

Setup Instructions

You can run this project in the following way: using Google Colab Option 1: Run on Google Colab

- 1. Open Google Colab.
- 2. Upload the project files, including the data files and requirements.txt, directly into Colab.
- 3. Install the required dependencies by running:

!pip install -r requirements.txt

4. Run the application code by executing:

!python app.py

Reasoning Behind the Optimized Prompts

The optimized prompts were designed to enhance both the specificity and clarity of model responses, maximizing relevance and reducing ambiguity. Key principles applied:

1. Clarity and Precision:

- The prompts were carefully structured to be clear and direct, ensuring the model understood exactly what information was being requested.
- Redundant words or vague phrases were removed to avoid confusing the model or leading it to provide extraneous details.

2. Contextual Relevance:

 Relevant background context was embedded in the prompts, which helped the model focus on the topic at hand. This included specific product or version names, tasks, or areas (e.g., "latest iPhone model capabilities") to reduce chances of outdated or tangential responses.

3. **Descriptive Language**:

- Descriptive language was added to guide the model's response format, such as asking for lists, bullet points, or examples when needed, making the outputs more structured and easier to parse for the user.
- Phrasing was also chosen to encourage comprehensive responses without excessive detail, improving conciseness and readability.

4. Prioritization of Key Details:

 The prompts explicitly asked for the most important or unique capabilities and features, especially for high-level summaries, rather than generalized information. This directed the model to focus on noteworthy details, avoiding information overload.

5. **Iterative Testing**:

 Prompts were iteratively tested and refined to achieve consistent accuracy. For example, prompts that led to overly broad or verbose responses were adjusted to be more specific in the type of response required