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

+ Code + Text Copy to Drive Connect Gemini

```
You: What qualities does the HR candidate have?
Bot:

    The HR candidate, based on the provided information, possesses the following qualities:

    1. **People-oriented and Confidential**: The candidate is described as "people-oriented" and maintains confidentiality, which are crucial qualit

    2. **Organized and Efficient**: The candidate's experience in marketing and advertising highlights their ability to maintain organization, compl

    3. **Exceptional Multi-Tasker**: The candidate's experience also mentions exceptional multi-tasking skills, which are essential in HR roles that

    4. **Motivated Team Player**: The candidate is described as a, which indicates their willingness.

    , a quality.

    , a quality.

    1 , a quality.

    The candidate, and the candidate.

    quality.

    The candidate and quality.
```

After optimizing the prompt

+ Code + Text Copy to Drive Connect Gemini

```
    2.
You: What are the key qualities and strengths of the HR candidate in areas like employee engagement, conflict resolution,
Bot:  Based on the information provided, here are the key qualities and strengths of the HR candidate in the specified areas, along with notable ach

    **Employee Engagement:**
    - **People-oriented**: The candidate's people-oriented nature suggests a strong ability to engage with employees, understand their needs, and fo
    - **Identifying training and development opportunities**: The candidate's experience in identifying training and development opportunities indic
    - **Employee survey and secondary data analysis**: During internships, the candidate used employee surveys and secondary data to analyze growth

    **Conflict Resolution:**
    - **Provide advice and recommendations on disciplinary actions**: The candidate's experience in providing advice on disciplinary actions suggest
    - **Maintains confidentiality**: The candidate's ability to maintain confidentiality is crucial in handling sensitive conflict resolution issues
    - **Excellent communication skills**: The candidate's well-developed communication skills are essential for effective conflict resolution, as th

    **Talent Acquisition:**
    - **Prepare, develop, and implement procedures and policies on staff recruitment**: The candidate's experience in developing and implementing re
    - **Prepare notices and advertisements for vacant staff positions**: The candidate's experience in preparing job advertisements indicates a good
    - **Schedule and organize interviews**: The candidate's experience in scheduling and organizing interviews shows practical skills in the talent

    **Notable Achievements:**
    - **Nominated for Property of the Year**: In their current marketing position, the candidate led their property to be nominated for Property of
    - **3.7% increase in NOI**: The candidate's efforts contributed to a 3.7% increase in Net Operating Income (NOI) for their property, indicating

    **Certifications:**
    - While the provided information does not mention any specific HR certifications, the candidate's dual specialization in Human Resource Manageme
    - The candidate's experience in HR roles, such as HR Management Trainee, also indicates practical knowledge and application of HR principles and
```

Second example – Banking

Before optimizing the prompt

+ Code + Text | Copy to Drive | Connect ▾ | ✦ Gemini

```
[ ]

You: What are the skills of the banking candidate?
Bot:  Based on the information provided in the PDFs and the conversation history, the banking candidate possesses a diverse set of skills, including

    **Technical Skills:**

    1. **Programming and Development:**
       - Proficient in COBOL, a programming language commonly used in banking and finance industries.
       - Experience in creating CICS MAP from scratch and developing the presentation layer and service programs.
       - Familiarity with creating High-level design documents and providing architectural solutions.
       - Experience in developing new programs to handle various functions of banking activities, both online and batch.

    2. **Database and Data Management:**
       - Experience in tracking and reporting all defects in the System Integration Testing (SIT) environment.
       - Familiarity with creating Unit Test Plans (UTP) for service and online programs.
       - Ability to work on testing parts, including STUB creation and other activities for performance tuning.

    3. **Project Management:**
       - Experience in coordinating various releases and taking responsibility for daily status reporting and follow-ups.
       - Proactive resolution of functional/technical hurdles that could impact deliverables, ensuring timely and flawless delivery.
       - Ability to create project plans and manage project timelines.

    4. **Business Analysis:**
       - Experience in providing technical specifications based on business requirements.
       - Ability to perform strategic analysis and provide unique solutions to meet business needs.
```

After optimizing the prompt

+ Code + Text | Copy to Drive | Connect ▾ | ✦ Gemini

```
    In conclusion, the banking candidate brings a strong combination of technical and soft skills to the table, making them a valuable asset to any
[ ] You: What unique skills and qualities make this candidate suited for banking? Highlight their experience in financial products, risk management, and
    Bot:  The banking candidate possesses several unique skills and qualities that make them well-suited for a career in banking, with a particular focu

    **Financial Products:**

    1. **Product Development and Management:**
       - The candidate has experience in developing new programs to handle various functions of banking activities, both online and batch. This expe
       - Their familiarity with creating High-level design documents and providing architectural solutions indicates a strong understanding of produ

    2. **Financial Operations:**
       - The candidate's experience in providing design proposals and developing new programs for various functions of banking activities suggests a

    **Risk Management:**

    1. **Risk Assessment and Mitigation:**
       - The candidate's experience in proactively resolving functional/technical hurdles that could impact deliverables demonstrates their ability
       - Their involvement in creating High-level design documents and providing architectural solutions indicates a proactive approach to risk mana

    2. **Compliance and Regulatory Knowledge:**
       - While the provided information does not explicitly mention compliance or regulatory knowledge, the candidate's experience in working with e

    **Customer Service:**

    1. **Customer Service Focus:**
       - The candidate's experience in customer service-focused roles, such as their position in marketing and public relations, indicates a strong
       - Their ability to maintain confidentiality, as mentioned in their marketing experience, is a crucial skill in banking, where customer data a
```

Third example – IT

Before optimizing the prompt

```
+ Code  + Text   Copy to Drive                                    Connect ▾   ✦ Gemini

[ ]
You: What makes the IT candidate a good fit?
Bot:  The IT candidate is a strong fit for various IT roles due to their extensive experience, diverse skill set, and proven track record of success

    **1. Proven IT Management and Leadership Experience:**

    - The IT candidate has held multiple leadership positions, including Vice President of Information Technology, System Administrator, and IT Dire
    - In their role as Vice President of Information Technology, the candidate built and improved IT organizations, turning around underperforming t
    - As an IT Director, the candidate managed a team of six, overseeing IT infrastructure and business intelligence operations for a large United W

    **2. Broad IT Skill Set:**

    - The IT candidate possesses a wide range of technical skills, including:
      - IT strategy and management
      - Project management
      - Business intelligence
      - Cloud computing
      - IT infrastructure
      - Enterprise platforms
      - Business continuity and disaster recovery
      - Budgeting and cost control
      - IT compliance (SOX, HIPAA, GLBA)
    - This diverse skill set enables the candidate to contribute effectively to various IT roles and projects, from strategic planning to day-to-day

    **3. Successful IT Project Delivery:**
```

After optimizing the prompt

```
+ Code  + Text   Copy to Drive                                    Connect ▾   ✦ Gemini

You: What qualities and technical skills make this candidate a strong fit for IT? Focus on their programming expertise, problem-solving abilities, a
Bot:  The IT candidate's strong fit for IT roles is evident in their extensive programming expertise, exceptional problem-solving abilities, and exp

    **1. Programming Expertise:**

    - The IT candidate has a solid foundation in programming, with experience in various languages and technologies. Their experience includes:
      - COBOL: Proficient in COBOL, a programming language commonly used in banking and finance industries, with experience in creating CICS MAP fro
      - JCL: Familiarity with Job Control Language (JCL) for submitting and controlling jobs in a mainframe environment.
      - DB2: Experience with DB2, a relational database management system, including tools such as QMF, SPUFI, FILE-AID, DB2 Admin, and DB2 command
      - UNIX: Experience working in a UNIX environment, with hands-on experience using debugging tools such as XPEDITOR and IBM DEBUGER.
      - Configuration management tools: Expertise in using configuration management tools like Endevor and SCLM.
    - The candidate's programming expertise enables them to contribute effectively to software development, maintenance, and improvement projects in

    **2. Problem-Solving Abilities:**

    - The IT candidate has demonstrated exceptional problem-solving abilities throughout their career, as evidenced by:
      - Proactively resolving functional/technical hurdles that could impact deliverables, ensuring timely and flawless delivery of banking projects
      - Identifying problems and providing effective solutions, as demonstrated by their ability to turn around underperforming IT teams and make IT
      - Their experience in creating High-level design documents and providing architectural solutions indicates a strong problem-solving approach,
      - The candidate's ability to work effectively in both online and batch environments showcases their adaptability and problem-solving skills in
    - The IT candidate's strong problem-solving abilities enable them to tackle complex IT challenges, identify root causes, and implement effective

    **3. Experience with Recent Technology Projects:**

    - The IT candidate has worked on various recent technology projects, demonstrating their ability to stay current with the latest trends and tool
      - **Cloud Computing**: The candidate has experience in cloud computing, with a focus on leveraging best-in-class infrastructures to create IT
```

Fourth example – Finance

Before optimizing the prompt

+ Code + Text Copy to Drive Connect ✦ Gemini

```
[ ] while True:
        user_input = input("You: ")

        if user_input.lower() == "exit":
            break

        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
```

After optimizing the prompt

+ Code + Text Copy to Drive Connect ✦ Gemini

```
You: What are the finance candidate's strengths in areas like financial analysis, regulatory compliance, and cost management? Include any noteworthy
Bot:  The finance candidate's strengths in financial analysis, regulatory compliance, and cost management are evident in their extensive experience

    **1. Financial Analysis:**

    - **Expertise in Financial Analysis Tools**: The candidate is proficient in various financial analysis tools and software, including:
      - Microsoft Excel: The candidate has a strong command of Excel, using it to create complex financial models, perform data analysis, and genera
      - Hyperion Planning: The candidate has experience with Hyperion Planning, a budgeting and forecasting tool, enabling them to create and manage
      - Cognos TM1: The candidate is familiar with Cognos TM1, a multidimensional database and planning tool, allowing them to build and maintain co
      - Anaplan: The candidate has experience with Anaplan, a cloud-based planning and performance management platform, demonstrating their ability
    - **Financial Analysis Skills**: The candidate possesses strong financial analysis skills, as demonstrated by:
      - Their ability to analyze financial statements, identify trends, and make data-driven decisions.
      - Their experience in creating and maintaining financial models, budgets, and forecasts.
      - Their ability to perform variance analysis, comparing actual results to budgeted or forecasted figures, and identify the reasons for any dif
      - Their experience in conducting break-even analysis, cost-volume-profit analysis, and other financial analysis techniques to support strategi
    - **Noteworthy Achievements**: The candidate's notable achievements in financial analysis include:
      - Developing and implementing a new budgeting and forecasting process that reduced the time required for budget preparation by 50% and improve
      - Creating a comprehensive financial model for a major acquisition, enabling the candidate's organization to make an informed decision about t

    **2. Regulatory Compliance:**

    - **Knowledge of Regulatory Requirements**: The candidate has a solid understanding of regulatory requirements in the finance industry, includin
      - Sarbanes-Oxley Act (SOX): The candidate is familiar with the requirements of SOX, such as internal controls and financial reporting, and has
      - Generally Accepted Accounting Principles (GAAP): The candidate has a strong understanding of GAAP and experience in preparing financial stat
      - Other regulatory requirements: The candidate is familiar with other regulatory requirements, such as those related to anti-money laundering
    - **Regulatory Compliance Experience**: The candidate's experience in regulatory compliance includes:
```

Fifth example – Consultant

Before optimizing the prompt

+ Code + Text | Copy to Drive   Connect ▾ | ✦ Gemini

```
[ ] You: What are the qualities of the consultant candidate?
    Bot:  The consultant candidate possesses a unique blend of skills, experiences, and qualities that make them an excellent fit for consulting roles.

        **1. Strong Business Acumen:**

        - **Industry Experience**: The candidate has extensive experience in various industries, including property management, public speaking, adverti
        - **Strategic Thinking**: The candidate's experience in strategic brand management and PR campaigns demonstrates their ability to think strategi
        - **Financial Awareness**: The candidate's experience in budget creation and management, as well as their ability to optimize rents based on mar

        **2. Excellent Communication Skills:**

        - **Superb Writing and Speaking Skills**: The candidate's exceptional writing and speaking skills enable them to communicate effectively with ta
        - **Charismatic and Persuasive**: The candidate's charismatic and persuasive nature allows them to build strong relationships, influence stakeho
        - **Excellent Presentation Skills**: The candidate's experience in public speaking and media relations indicates their ability to deliver compel

        **3. Proven Leadership and Team Management Skills:**

        - **Motivated Team Player**: The candidate's experience in leading teams and driving results, as demonstrated by their nomination for Property o
        - **Mentoring and Leadership**: The candidate's experience in mentoring team members and providing advice on disciplinary actions indicates thei
        - **Project Management**: The candidate's experience in overseeing marketing efforts, managing advertising budgets, and organizing financial rep

        **4. Adaptability and Continuous Learning:**

        - **Quick Learner**: The candidate's ability to adapt to new situations and learn quickly, as evidenced by their experience in POC (Proof of Con
        - **Versatile Experience**: The candidate's diverse experience in marketing, advertising, public relations, and property management indicates th
```

After optimizing the prompt

+ Code + Text | Copy to Drive   Connect ▾ | ✦ Gemini

```
        In conclusion, the consultant candidate's unique blend of strong business acumen, excellent communication skills, proven leadership and team man
    You: What are the top qualities of this consulting candidate, particularly in client relationship management, strategic planning, and project leader
    Bot:  Based on the information provided, here are the top qualities of the consulting candidate in client relationship management, strategic plannin

        **1. Client Relationship Management:**

        - **Exceptional Communication Skills**: The candidate's superb writing and speaking skills, along with their charismatic and persuasive nature,
        - **Active Listening**: The candidate's ability to listen actively to clients and stakeholders, as demonstrated by their experience in customer
        - **Trust and Confidentiality**: The candidate's experience in maintaining confidentiality in marketing and public relations roles indicates the
        - **Proactive Approach**: The candidate's proactive approach to identifying and addressing client needs, as evidenced by their experience in stu

        **Achievements in Client Outcomes:**

        - **Increased Occupied Rent**: The candidate's efforts in increasing their property's occupied rent by over $100 and continuing to rise demonstr
        - **Property of the Year Nomination**: The candidate's nomination for Property of the Year indicates their ability to exceed client expectations

        **2. Strategic Planning:**

        - **Strategic Thinking**: The candidate's experience in strategic brand management and PR campaigns, as well as their ability to optimize rents
        - **Data-Driven Decision Making**: The candidate's experience in analyzing financial data, identifying trends, and making data-driven decisions
        - **Cross-Functional Collaboration**: The candidate's experience in working with cross-functional teams, as demonstrated by their role in overse

        **Achievements in Strategic Planning:**

        - **Successful Acquisition Integration**: The candidate's experience in creating a comprehensive financial model for a major acquisition, enabli

        **3. Project Leadership:**
```

Results of model before implementing RAG

```
+ Code  + Text  Copy to Drive                                    Connect  ✦ Gemini
    return response.generated_text if hasattr(response, 'generated_text') else response

# Test the model with a sample question
user_question = "Tell the summary of deadpool and wolverine movie"
response = test_model_inference(user_question)
print("Bot:", response)
```

```
Bot:  "Deadpool and Wolverine" is a hypothetical crossover film that has been widely requested by fans, but has not yet been officially announced or pı

    Title: "Deadpool & Wolverine: The Mutant Way"

    Summary:
    After a mysterious force causes a rift in the multiverse, Deadpool (Ryan Reynolds) from the Marvel Cinematic Universe (MCU) finds himself in the X

    Deadpool, with his unique sense of humor and fourth-wall-breaking tendencies, initially disrupts the serious tone of the facility, much to the disı

    Together, they embark on a wild, action-packed adventure filled with Deadpool's signature humor and Wolverine's brutal combat skills. Along the wa;

    Throughout their journey, Deadpool and Wolverine learn from each other, with Deadpool teaching Wolverine to loosen up and enjoy life, while Wolver:

    In the end, they successfully close the rift, but not before Deadpool leaves a lasting impact on the X-Men universe, inspiring a new generation of

    The film would blend the dark, serious tone of the X-Men universe with the humor and irreverence of Deadpool, creating a unique and entertaining c:
```

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.

```
+ Code  + Text  Copy to Drive                                    Connect  ✦ Gemini
    return response.generated_text if hasattr(response, 'generated_text') else response

# Test the model with a sample question
user_question = "Tell the summary of ghostbusters:frozen empire movie"
response = test_model_inference(user_question)
print("Bot:", response)
```

```
Bot:  "Ghostbusters: Frozen Empire" is a 2021 American supernatural comedy film directed by Jason Reitman and Gil Kenan, and written by Dan Aykroyd, Jä

    The film is set in the fictional town of Summerville, Oklahoma, where a mysterious cold snap has caused the town to freeze over. The Ghostbusters,

    As they delve deeper into the mystery, they discover that the cold snap is the work of a powerful and ancient entity known as the "Frozen Empire,"

    Along the way, they encounter a variety of ghosts and other supernatural creatures, as well as a group of rival ghost hunters who are determined tı

    In the end, the Ghostbusters are able to defeat the Frozen Empire and save the world from its icy grip. The film ends with a message about the impı

    "Ghostbusters: Frozen Empire" received mixed reviews from critics, who praised the film's humor and visual effects but criticized its plot and chaı
```

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

+ Code + Text Copy to Drive Connect ✦ Gemini

```
# Test the model with a sample question
user_question = "Tell the name of the latest iphone,what r its capabilites"
response = test_model_inference(user_question)
print("Bot:", response)
```

```
Bot:  The latest iPhone is the iPhone 14 Pro Max, released in September 2022. Here are some of its key capabilities:

    - **Display**: 6.7-inch Super Retina XDR OLED display with ProMotion technology (120Hz adaptive refresh rate), HDR10, 1000 nits typical brightness

    - **Processor**: A16 Bionic chip (5nm, 6-core CPU, 5-core GPU, 16-core Neural Engine).

    - **Cameras**:
      - **Rear**: 48MP main camera with sensor-shift optical image stabilization (OIS), 12MP ultrawide camera, and 12MP telephoto camera with 3x optic
      - **Front**: 12MP TrueDepth camera.

    - **Video**: Cinematic mode for 4K video recording at 24 fps, ProRes video recording, Dolby Vision HDR video recording, and Cinematic mode for vid

    - **Battery**: Built-in rechargeable lithium-ion battery supporting MagSafe wireless charging (up to 15W), Qi wireless charging (up to 7.5W), and

    - **Storage**: Available in 128GB, 256GB, 512GB, and 1TB storage options.

    - **Connectivity**: 5G (mmWave and sub-6GHz), Gigabit LTE, 802.11ax Wi-Fi 6 with 2x2 MIMO, Bluetooth 5.3, Ultra Wideband (UWB) with Precision Find

    - **Sensors**: Face ID, Barometric altimeter, Three-axis gyro, Accelerometer, Proximity sensor, Ambient light sensor, and Ultra Wideband (UWB) ant
```

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:

+ Code + Text Copy to Drive Connect Gemini

```
while True:
    user_input = input("You: ")

    if user_input.lower() == "exit":
        break

    response = handle_user_input(user_input, vectorstore, conversation_history)
    print(f"Bot: {response}")
```

```
You: When did the deadpool and wolverine movie released
Bot:  Deadpool & Wolverine was released on July 26, 2024, as part of Phase Five of the Marvel Cinematic Universe (MCU). The film premiered on July 22,
You: exit
```

+ Code + Text Copy to Drive Connect Gemini

```
    response = handle_user_input(user_input, vectorstore, conversation_history)
    print(f"Bot: {response}")
```

```
You: Tell the summary of ghostbusters:frozen empire movie
Bot:  Ghostbusters: Frozen Empire is a 2024 American supernatural comedy film directed by Jason Reitman and written by Gil Kenan and Jason Reitman. It

    The cast includes Paul Rudd as Gary Grooberson, Carrie Coon as Callie Spengler, Finn Wolfhard as Trevor Spengler, Mckenna Grace as Phoebe Spengler

    Principal photography began on March 20, 2023, in London, under the working title Firehouse, with Eric Steelberg serving as the cinematographer. F

    The film had its world premiere at the AMC 13 Theater at Lincoln Square in New York City on March 14, 2024, and was released in the United States

    In the film, after the events in Summerville, Oklahoma, the Spengler family moves to New York City. When a new ghost threat begins to terrorize th
You: exit
```

+ Code + Text Copy to Drive Connect Gemini

Insert code cell below (Ctrl+M B)

```
You: Tell the name of the latest iphone,what r its capabilites
Bot:  The latest iPhone models announced by Apple are the iPhone 16 and iPhone 16 Plus, as well as the iPhone 16 Pro and iPhone 16 Pro Max. Here are s

    **iPhone 16 and iPhone 16 Plus:**
    - Display: 6.1-inch and 6.7-inch Super Retina XDR OLED, 60Hz, HDR10, Dolby Vision, 1200 nits (typ), 1700 nits (HBM)
    - Sound: Stereo speakers, Dolby Atmos
    - Connectivity: Wi-Fi 6, Bluetooth 5.3, Ultra-wideband (UWB), NFC (reader mode, Express Cards), LEO satellite (Globalstar, limited)
    - Camera: Dual 12MP camera system (Wide and Ultra Wide), 4K video recording at 24 fps, 30 fps, or 60 fps
    - Battery: 3279 mAh (iPhone 16) and 4352 mAh (iPhone 16 Plus)
    - Storage: 128GB, 256GB, or 512GB
    - RAM: 6GB
    - Processor: A18 Bionic chip
    - Water resistance: IP68 (maximum depth of 6 meters up to 30 minutes)

    **iPhone 16 Pro and iPhone 16 Pro Max:**
    - Display: 6.3-inch and 6.9-inch LTPO Super Retina XDR OLED, 120Hz, HDR10, Dolby Vision, 1000 nits (typ), 2000 nits (HBM)
    - Sound: Dolby Atmos-tuned Spatial Audio
    - Connectivity: Wi-Fi 6, Bluetooth 5.3, Ultra-wideband (UWB), Thread, NFC (reader mode, Express Cards), LEO satellite (Globalstar, limited), USB-C
    - Camera: Triple 12MP camera system (Wide, Ultra Wide, and Telephoto), 4K video recording at 24 fps, 30 fps, 60 fps, or 120 fps
    - Battery: 3200 mAh (iPhone 16 Pro) and 4500 mAh (iPhone 16 Pro Max)
    - Storage: 128GB, 256GB, 512GB, or 1TB
    - RAM: 8GB (iPhone 16 Pro) and 12GB (iPhone 16 Pro Max)
    - Processor: A18 Pro Bionic chip
    - Water resistance: IP68 (maximum depth of 6 meters up to 30 minutes)
```

Challenges Faced

- 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