JumpSTART LLM: Smart Application Informing and Educating Entrepreneurs and Business Professionals About Startups

Project Link: https://genai-assignment-7-capstone-project.streamlit.app/

Intro and Objective:

Entrepreneurs and various business leaders want to conduct startup research about competitors when starting a new company or business venture. By using the Gen AI Q&A project, JumpSTART LLM, they can perform market research to understand what has already been created so they can find a niche in the market and figure out how their company can be competitive.

Additionally, another important aspect is to verify the legal rules and regulations required when forming a new company. There may be many national and state specific government requirements and regulations the businesses have to follow including tax filing, incorporation law, funding and anti-trust laws. JumpSTART LLM can help break down complex legal jargon to answer specific targeted questions about legal requirements and regulations.

When starting a new business, entrepreneurs may have many questions about current startups (their competitors), the market landscape, etc. These entrepreneurs and business leaders can ask their question to the Gen AI Q&A project and receive thoughtful and impactful responses to make data-driven decisions. Some example questions may include:

- How should I split company equity with my co-founder? Should the split be based on ongoing contributions.
- When is the right time to start hiring employees?
- How do I calculate my personal finance burn rate?
- Should I take a salary?
- How do I approach investors. What non-monetary help can investors provide?
- What steps do I take to become an LLC?
- What is the process for registering my company as a non-profit?

The objective and intended use of JumpSTART LLM and web application is for CEOs, founders, CFOs, COOs, CTOs and other C-suite leaders to ask specific questions about startups especially during the early stages of company formation. These individuals will be able to receive concise and targeted responses to minimize their time spent researching information relevant to their questions. Instead, these business leaders can spend more time focused on implementation of business processes thereby speeding up their overall productivity and efficiency.

Our Gen AI model will include relevant context for startups by incorporating startup specific data fine-tuned to the use cases selected. Specifically, including information from Y-Combinator's How to Start a Startup course (*How to Start a Startup - a Course Y Combinator Taught at Stanford - YC Library | Y Combinator*, 2024). Therefore, the web application will be able to help users build startups by sharing advice and procedures recommended by Y-Combinator as well as the other data sources.

Selection of Generative AI Model:

Google's Gemini 1.5 Flash model has been selected for the Q&A Startup project. This model is a perfect text-based generative AI that can be used to create a chatbot designed to help startups and entrepreneurs perform thorough business research. This tool is relevant because, via its multimodal capabilities and extensive context window, it can efficiently process and summarize comprehensive business documents, spreadsheets, course transcripts with various speakers, and basic legal documents essential for in-depth market and competitor analyses. Gemini 1.5 Flash can quickly provide summaries and answers that are usable. It is helpful for entrepreneurs to navigate complex legal frameworks and startup strategies quickly. The affordability and easy integration of the solution through Langchain and its lightweight architecture make it feasible for on-budget startups or independent open-source projects to implement quickly and at scale.

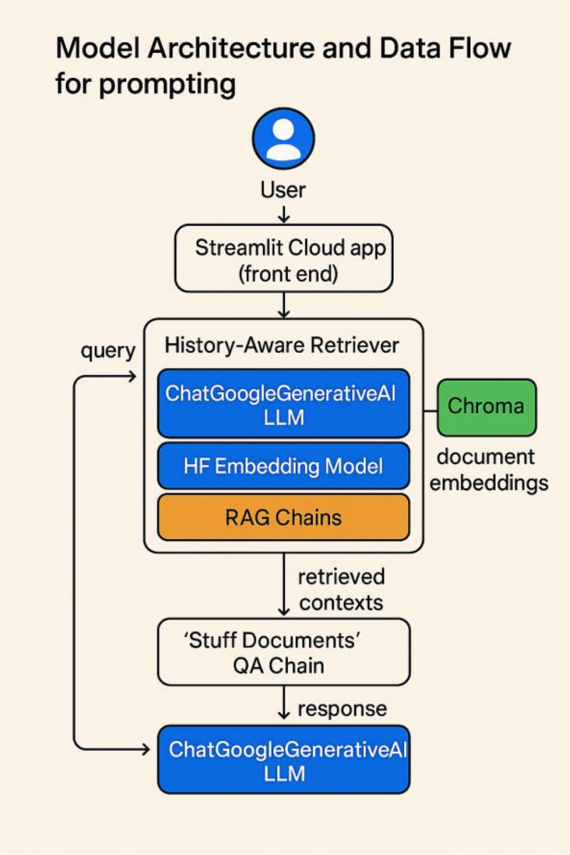
Project Definition and Use Case:

The project is an AI-powered chatbot and question-answering system designed for entrepreneurs and business leaders to help them research startups, competitor analysis, market landscape, legal regulations for new businesses, etc. Using Google's Gemini 1.5 Flash generative AI model, the chatbot will provide quick and detailed answers to user inquiries, being able to read extensive amounts of multimodal data such as detailed market research spreadsheets, legal papers, and presentations of course transcripts. Gemini 1.5 Flash can handle very long inputs, recall more of the talk or documentation, effectively utilize that context to arrive at pertinent and coherent replies, and combine data from various sources. This enables it to provide intelligent and actionable answers. For instance, it can consider information from Y-Combinator's "How to Start a Startup" course (*How to Start a Startup - a Course Y Combinator Taught at Stanford - YC Library | Y Combinator*, 2024) and provide expert responses as if it were a startup expert. This allows individuals (especially C-suite executives and founders) to access the required information quickly. This significantly lessens the time spent on research, so they can use that energy more strategically to implement business decisions and processes that contribute to overall business growth.

Implementation Plan:

The application uses a technology stack enhanced with popular artificial intelligence tools and libraries. The various tools used are as follows: LangChain, to create AI chain modules matching different datasets. Hugging Face Embeddings are used to implement semantic searches while the Chroma vectors are stored inside SQLite3. Similarly, the ChatGoogleGenerativeAI interface uses Google's Gemini 1.5 Flash generative AI model as the main model in the project. The app uses Streamlit Cloud's free tier as a web framework to host the application for the various users, entrepreneurs, and business executives. The application is easy to deploy through GitHub and safely secures the API keys with the help of the secrets function in Streamlit. The first step in the development process is establishing three different data chains in LangChain corresponding to the spreadsheets on startup data, legal act papers, and course transcripts of Y-Combinator's "How to Start a Startup" class (*How to Start a Startup - a Course Y Combinator Taught at Stanford - YC Library | Y Combinator*, 2024). Their retrieval mechanisms are context-specific, which helps generate accurate responses. The testing simulates queries an average user might make if acting as a startup founder or other business professional. The testing phase involves many metrics related to the performance of a system, like faithfulness, answer relevancy, context precision, and context recall, as well as the scores of BLEU and ROUGE (more on this in the next section). Manual checks are printed out during the testing phase as a part of the model outputs to confirm that a particular query is calling the proper chain.

From the user's perspective, when a user opens the Streamlit Cloud app, the front end immediately ensures that the ChatGoogleGenerativeAI LLM and HuggingFace embedding models are loaded (or re-used) in session state alongside three RAG chains (startup data, legal data, and startup masterclass) as a part of the ReAct framework (Reasoning and Acting). As soon as the user submits a query through the Streamlit chat widget, that input is wrapped together with the existing chat history and sent to LangChain's history-aware retriever, which first uses the LLM to reformulate the question into a standalone prompt and then fetches the top-k semantically similar document embeddings from Chroma. The retrieved contexts are concatenated and passed into a “stuff documents” QA chain. This prompts the LLM with a system template that instructs it to answer concisely (or in more detail for the masterclass content) based on the provided inputs. The LLM’s response is then streamed back to Streamlit, displayed in the chat UI, and appended to the session’s chat history so that follow-up questions can continue to build on prior questions in the same session. The above process is repeated until the user finishes engaging with the app, and the application sleeps or hibernates after 24 hours of no activity. After that, to reengage the app, the user can reload the page to begin submitting further questions to the Q&A project application.


Model Architecture and Data Flow for prompting
User
Streamlit Cloud app (front end)
query
History-Aware Retriever
ChatGoogleGenerativeAI LLM
HF Embedding Model
RAG Chains
Chroma
document embeddings
retrieved contexts
‘Stuff Documents’ QA Chain
response
ChatGoogleGenerativeAI LLM

The Streamlit UI design uses Streamlit's native layout presets, with a pinch of HTML / CSS finesse to give a polished chat-like feel to the app. According to the app's home page, the left sidebar is rendered via st.sidebar calls. Next, the app has an "About" section and bullet points for every tool with a "Clear Conversation" button to restart the chat. The main header (which says "LangChain Agent Assistant") is placed inside a block. It has rounded borders, padding, and a light-blue background (that has a dark-mode alternative color palette), shown in the code, which is placed in a style tag and unsafe_allow_html=True. The chat messages are wrapped in the containers <div class=" chat-message user> for the user and the bot, respectively. Each has its background color and text color, along with an avatar image, which is dynamically pulled from Dicebear to obtain icons. The CSS modifies the margins and flexes, as well as the neat alignment of the stTextInput and the markdown. Lastly, a chat input is placed at the bottom as st.chat_input, prompting to "Ask me about startups, legal information, or get expert startup advice from course materials..." The responses are streamed back into CSS-styled containers, which can be seen in the live UI when the app is used.

Model Evaluation and Performance Metrics:

Several test criteria were used to evaluate the performance of the Gemini 1.5 Flash generative AI model embedded in the chatbot. Inference time, latency, and speed are monitored in the metrics notebook to ensure the inference is fast enough for real-time user interactions. Most importantly, the metrics notebook measures the usage of local CPU and memory, and the GCP console dashboards keep track of the cloud GPU memory consumption (which can be viewed via the online portal) to optimize the infrastructure and costs. The metrics notebook performs an in-depth analysis of model accuracy using a range of advanced text-based metrics. The metrics computed are BLEU scores (using SacreBLEU) and ROUGE scores, followed by retrieval-augmented generation-specific metrics, namely, faithfulness, answer relevancy, context precision, and context recall (from the RAGAS module, similar to those provided in the course modules). These above metrics indicate that the bot can produce accurate, relevant, and dependable responses. Finally, user experience assessment is performed via ease of interaction, usability, and the chatbot responses of the Streamlit application to investigate startup and business-related research for the day-to-day end users.

Deployment Strategy:

The app will be publicly available in GitHub and deployed via Streamlit Cloud to ensure developers integrate and update the app and provide optimal user access. Streamlit Cloud's free tier allows easy hosting and fast deployment directly from GitHub repos, along with secure API Key management, which is perfect for users wanting to deploy any web-based application at scale rapidly. The AI assistant will have a simple and intuitive interface for users that will clearly display its data sources: Startup Data, Legal Data and How to Start a Startup masterclass expertise (*How to Start a Startup - a Course Y Combinator Taught at Stanford - YC Library | Y Combinator*, 2024). When the chatbot is initialized, it shows that it is working successfully. Users can ask questions regarding startups, legal requirements, or strategic entrepreneurial advice via the prompt bar marked. The agent implements a ReAct framework, adeptly reasoning through more complex queries with the retrieval of accurate, pertinent information, providing a responsive, effective, informative user experience for entrepreneurs and business leaders.

Expected Outcomes and Challenges:

The AI-driven chatbot will help conduct research for startups and entrepreneurs efficiently through fast, accurate, and actionable insights that will help business leaders make better decisions and improve productivity. The chatbot can access relevant information from market data, legal documents, and startup guidance courses. Entrepreneurs are expected to benefit from the decreased research work, more tailored insights suited to their startup, and quick strategy execution. During the startup phase, a dataset limitation could occur. This challenge could impact the accuracy of responses. New laws and regulations about AI are constantly being passed (in the application, the data was from 2023, and the legal data was from a few selected legal acts). Additionally, as the number of users increases, performance issues like high volume of usage can be other issues that may arise. Moreover, there may be deployment problems. These include integration problems, deployment infrastructure, and data storage problems. This is inevitable as the application is upgraded to accommodate more use cases for other types of businesses. Also, when more features are added, like mobile application design and usability. To overcome these issues, the recommended actions include; suggest constantly expanding and updating the datasets by adding the most relevant, updated-data from both the startup and legal data chains for improved answer breadth and accuracy; use efficient load-balancing and resource management strategies to mitigate performance bottlenecks; edit the application design via tweaks to the layouts, navigation and caching for the most optimal and functional mobile-user experience; utilize the Streamlit Cloud Pro tier on Snowflake for robust deployment for easy integration, scalability and stable performance.

Resources Required:

Utilizing LangChain for modular AI chain development, Hugging Face Embeddings for semantic data encoding, and Google's Gemini 1.5 Flash generative AI model for rapid and contextually appropriate response generation as crucial resources can aid in implementing the AI Chatbot. The application features important data sets that have been utilized to help ensure the accuracy of the chatbot. These include exhaustive CSV files containing data related to startups, transcripts from the "How to Start a Startup" course by Y-Combinator (*How to Start a Startup - a Course Y Combinator Taught at Stanford - YC Library | Y Combinator*, 2024), and massive .txt documents related to legal acts and regulations in the U.S. Semantic embeddings are stored and retrieved using a Chroma vector database stored in SQLite3. Good cloud infrastructure is required to deploy the Gemini model and run the vector databases efficiently. The hosting will be done with the help of the free tier of Streamlit Cloud. The storage will be small, can be scaled (comes with a Linux container), and will be responsive.

Conclusion:

To conclude, JumpSTART LLM is leveraging a Generative AI model (Google's Gemini 1.5 Flash) and Langchain as a web application to answer detailed business questions to entrepreneurs, startup founders, and other business leaders. A chatbot can provide faster, accurate, and relevant insights with the help of a holistic framework and various data sets, including startup data sets, legal documents, transcripts from expert advice, and many other sources. As a result, it allows business people to pay less attention to costs and time associated with research and more attention to key business processes, including funding, product development, and user engagement. To improve the project, regularly update datasets featuring recent startups' trends, legal compliance requirements, and feedback loop from users for better accuracy. Reinforced Learning from Human Feedback (RLHF) and other mechanisms may also help. The infrastructure will likely be optimized for scalability in the future, and improvements will be made to the mobile user interface for an interactive and intuitive user experience. Overall, JumpSTART LLM considerably reduces the time it takes startups to find accurate and valuable information, positively impacting their time to market, and will help startups and entrepreneurs make work and lifestyle decisions inside and outside the office.

References

How to Start a Startup - A course Y Combinator taught at Stanford - YC Library | Y Combinator. (2024). Y Combinator. https://www.ycombinator.com/library/carousel/How%20to%20Start%20a%20Startup%20-%20A%20course%20Y%20Combinator%20taught%20at%20Stanford