Q&A LLM Chat Bot Using LangChain Framework

A capstone project submitted in partial fulfilment for the award of the degree

of

Master of Science

in

Data Science & Analytics

by

Anuj Tiwari

Roll Number: 2207160003

System ID: 2022419534

UNDER THE SUPERVISION

of

Dr. Surya Kant Pal

SHARDA UNIVERSITY

Beyond Boundaries

Department of Mathematics

Sharda School of Basic Sciences & Research

Sharda University, Greater Noida, Uttar Pradesh-201310

April, 2024

Q&A LLM Chat Bot Using LangChain Framework

A capstone project submitted in partial fulfilment for the award of the degree

of

Master of Science

in

Data Science & Analytics

by

Anuj Tiwari

Roll Number: 2207160003

System ID: 2022419534

UNDER THE SUPERVISION

of

Dr. Surya Kant Pal

Department of Mathematics

Sharda School of Basic Sciences & Research

Sharda University, Greater Noida, Uttar Pradesh-201310

April, 2024

DECLARATION

I, Anuj Tiwari, declare that this **Capstone Project** entitled **"Q&A LLM Chatbot Using LangChain Framework"** is a bonafide work prepared in partial fulfillment of the requirements for the award of the degree of Master of Science in Data Science & Analytics by Sharda University, Greater Noida, under the supervision of **Dr. Surya Kant Pal** at the Department of Mathematics, Sharda University, Greater Noida.

I further certify that this work has not been submitted by me for the award of any other degree/diploma of this or any other university.

Anuj Tiwari

Date: 23 April 2024

Anuj Tiwari

Place: Greater Noida

2022419534

SHARDA UNIVERSITY
Beyond Boundaries

ACCREDITED WITH GRADE A+ NAAC

CERTIFICATE

This is to certify that this capstone project work entitled "**Q&A LLM Chatbot Using LangChain Framework**" is a bonafide work carried out by Anuj Tiwari in partial fulfillment of the requirements for the award of the degree of Master of Science in Data Science & Analytics by Sharda University, Greater Noida during the year 2023-24. It is certified that the capstone project work has been approved as it satisfies the academic requirements with respect to the capstone project work prescribed for the said degree.

Supervisor 23/04/24

Head of the Department

Head
Department of Mathematics
Sharda School of Basic Sciences & Research
Sharda University, Greater Noida, India

Dean

Dean
Sharda School of Basic Sciences & Research
Sharda University, Greater Noida, India

Dr. B K Singh

Name of the Examiners

23/04/24

Signature with Date

CERTIFICATE OF PLAGIARISM CHECK

Title of Capstone Project:	**Q&A LLM Chatbot Using LangChain Framework**
Name of Student:	**Anuj Tiwari**
System ID:	2022419534
Name of Guide:	**Dr. Surya Kant Pal**
Software Name:	**Turnitin**
Date and Time of Check:	14th April 2024 and 10:30 pm
Similarity Index:	9%
Acceptable maximum limit(%) of similarity	10%

I hereby certify that this capstone project has been evaluated using TURNITIN software. I have analysed the report produced by the system and based on it, I certify that the references in the capstone project are in accordance with good scientific practice.

Date and Sign of Student: Anuj Tiwari
23 April 2024

Date and Sign of Guide: 23/04/2024

ACKNOWLEDGEMENTS

A capstone is always a great opportunity to learn and develop. We are very grateful to have had the opportunity to work with such high-profile faculty members. It has given us a chance to explore and utilize our academic knowledge in reality efficiently. It helped us to gain more information about real-world problems and solutions than any academic theory.

First and foremost, we would like to thank **Prof. Khursheed Alam**, HOD of the Mathematics department who allowed us to undertake this project. His guidance and suggestions were always remarkable which led us to perform the project perfectly.

We would like to thank our supervisor **Dr. Surya Kant Pal** for all of his help with our project. He was extremely busy with academic work, yet he took the time to listen to and guide us. He always led us throughout the completion of the project.

The faculty members of the mathematics department made sure that everything was going smoothly. They also thanked them for their efforts. Each evaluation with a proper pattern helped us to complete our project in the proper time.

Anuj Tiwari

Name of Student

Anuj Tiwari (2022419534)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The advent of advanced natural language processing (NLP) models has revolutionized the field of conversational agents, enabling the development of sophisticated chatbots capable of understanding and generating human-like responses. In this project, we propose the creation of a Q&A LLM (Language Model) chatbot utilizing the LangChain framework. The chatbot will be powered by the GPT-3.5-turbo-instruct model, leveraging the API provided by OpenAI. Through the integration of the Streamlit library, the chatbot will be deployed, providing users with an interactive platform to seek answers on a wide range of topics. The primary objective of this project is to develop a conversational assistant that emulates human-like interaction, delivering accurate and relevant responses to user inquiries. By harnessing the capabilities of state-of-the-art NLP technology, this chatbot aims to enhance user experience and accessibility to information in diverse domains.

Keywords: NLP, GPT-3.5-turbo-instruct, Python, LangChain, Streamlit

DESCRIPTION OF THE PROJECT

This project proposes creating a Q&A LLM chatbot using the LangChain framework, powered by the GPT-3.5-turbo-instruct model via OpenAI API. Integrated with Streamlit, it offers an interactive platform for users to access accurate information across various topics, aiming to simulate human-like interaction and enhance accessibility to diverse domains.

1. INTRODUCTION

The rapid evolution of Artificial Intelligence (AI) has significantly impacted various aspects of modern life. Among its most promising applications, Q&A (Question Answering) chatbots have emerged as powerful tools for information access and interaction. These chatbots leverage natural language processing (NLP) capabilities to converse with users, understand their queries, and provide informative responses.

However, building effective Q&A chatbots comes with inherent challenges. Conventional approaches often struggle with providing accurate and comprehensive answers, particularly for open-ended or complex questions. Additionally, maintaining information freshness can be difficult, as these chatbots rely solely on pre-trained models and might struggle to answer queries related to recent events or newly discovered information.

To address these limitations, this capstone project delves into the development of a novel Q&A LLM chatbot utilizing the LangChain framework and the GPT-3.5-turbo-instruct model. The project aims to create a personal assistant-like chatbot capable of engaging in meaningful conversations, understanding diverse user queries, and providing informative and up-to-date responses across various domains.

LangChain offers a unique approach to building chatbots by empowering them with retrieval-augmented generation (RAG) capabilities. This framework distinguishes itself by allowing the integration of external knowledge sources with the powerful language processing abilities of LLMs (Large Language Models) like GPT-3.5-turbo-instruct. By leveraging LangChain, the chatbot can access and retrieve relevant information from external sources when necessary, enabling it to answer user queries with greater accuracy and comprehensiveness.

Furthermore, the choice of GPT-3.5-turbo-instruct as the underlying LLM plays a crucial role in the project's effectiveness. This model boasts exceptional capabilities in understanding and generating human-like language, making it well-suited for natural dialogues and comprehensive answer generation. Its ability to access and process information from the real world through the OpenAI API further enhances its potential to provide relevant and up-to-date responses.

The deployment of the chatbot using the Streamlit library ensures a user-friendly experience. Streamlit facilitates the creation of interactive web applications, making the chatbot readily accessible through a web interface. This allows users to engage with the chatbot seamlessly, asking questions and receiving responses in a real-time and user-friendly manner.

This project presents a compelling exploration of the potential for building a highly versatile and informative Q&A LLM chatbot. By utilizing the combined strengths of the LangChain framework for context-aware information retrieval and the advanced capabilities of the GPT-3.5-turbo-instruct model for language processing and generation, the project aims to create a valuable tool for users seeking information and engaging in meaningful conversations.

Furthermore, the project seeks to contribute to the ongoing advancements in the field of Q&A chatbots by demonstrating the effectiveness of the LangChain framework and the GPT-3.5-turbo-instruct model in a practical application. By evaluating the chatbot's performance on various metrics like accuracy, fluency, and user satisfaction, the project can offer valuable insights for further development and refinement of similar chatbots in the future.

Throughout this introduction, we have discussed the motivations and objectives of this project. We have delved into the specific technologies employed, including the LangChain framework, GPT-3.5-turbo-instruct model, and Streamlit library. Additionally, we have highlighted the potential contributions and expected outcomes of this project. In the subsequent sections, we can delve deeper into the particular technique employed, the consequences acquired, and the discussion regarding the project's findings and limitations.

2. OBJECTIVE

The main objectives of this project are to:

- Develop a Q&A Language Model (LLM) chatbot utilizing the LangChain framework and integrating the GPT-3.5-turbo-instruct model provided by OpenAI.
- Train the chatbot to accurately understand and respond to user queries across a diverse range of topics, demonstrating proficiency in natural language processing.
- Implement the chatbot deployment using the Streamlit library to create a user-friendly interface for seamless interaction and enhanced accessibility.
- Evaluate the performance of the chatbot in terms of response accuracy, relevance, and user satisfaction through rigorous testing and feedback analysis.
- Iterate and refine the chatbot based on user feedback and performance metrics, striving for continuous improvement in functionality and user experience.
- Explore opportunities for future enhancements and extensions of the chatbot, such as incorporating additional features, integrating with external APIs, or adapting to specific user requirements or domains.

3. LITERATURE REVIEW

Koh Matsuda et al. (2024) focusesed on leveraging Large Language Model (LLM) applications in education, particularly through LangChain's integration capabilities. It introduces a system integrating Retrieval Augmented Generation (RAG) with external data from Pinecone database, aiming to enhance educational interactions. The paper discusses methods for LangChain integration in education, highlighting benefits and potential challenges. Future improvements target refinement based on user feedback, utilizing LangSmith.

Mahyar Abbasian et al. (2024) addressesed the limitations of current Conversational Health Agents (CHAs), particularly their lack of multi-step problem-solving and personalized interactions. It introduces openCHA, an open-source framework powered by Large Language Models (LLMs), designed to enhance CHAs' capabilities. openCHA enables integration with external sources for data analysis and knowledge acquisition, fostering personalized, multi-modal conversations in healthcare settings. The framework's proficiency in handling complex healthcare tasks is demonstrated through three illustrative examples.

Lianmin Zheng (2023) examined the challenge of evaluating large language model (LLM) based chat assistants, proposing the use of strong LLMs as judges. It investigates biases and limitations in LLM judgment, suggesting methods to address them. Two benchmarks, MT-bench and Chatbot Arena, are introduced to verify LLM judgment against human preferences. Results indicate strong agreement between LLM judges and humans, highlighting the scalability and cost-effectiveness of LLM-based evaluation methods.

Qingyun Wu et al. (2023) introduced AutoGen2, an open-source framework enabling the creation of Large Language Model (LLM) applications with customizable, conversable agents capable of collaborative task completion. AutoGen2 offers flexibility in agent interaction behaviors, supporting both natural language and code-based programming for diverse applications. Empirical studies validate its effectiveness across domains such as mathematics, coding, question answering, operations research, and entertainment, showcasing its versatility and utility in various contexts.

Oguzhan Topsakal et al. (2023) explored the utilization of Large Language Models (LLMs) for rapid application development, with a focus on LangChain, an open-source library. It discusses LLMs' diverse capabilities and LangChain's modular design, facilitating swift development of bespoke AI applications. Practical examples illustrate LangChain's potential in creating LLM-based applications efficiently. The study underscores LangChain's significance in advancing LLM application development methodologies.

Arjun Pesaru (2023) integrated LangChain and the Large Language Model (LLM) to develop a PDF chatbot. LangChain simplifies chatbot creation and scalable AI/LLM applications, while LLM offers diverse text generation capabilities. The chatbot, trained on a PDF dataset, leverages LLM for text responses and Google Search for broader information access. Utilizing Pinecone for storage and React JS for frontend, the project demonstrates LangChain and LLM's potential in creating informative, versatile chatbots.

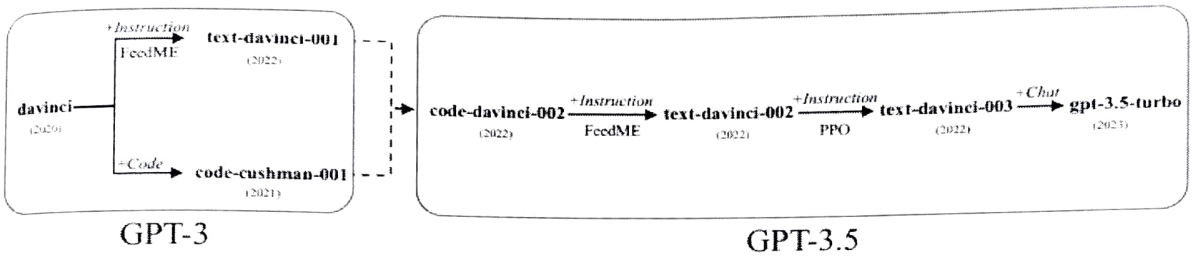
4. METHODOLOGY

The methodology for this project entails a systematic approach to developing and evaluating the Q&A LLM chatbot using the LangChain framework. Firstly, the LangChain framework is carefully selected based on its compatibility with the chosen LLM model, GPT-3.5-turbo-instruct, and its support for deployment through the Streamlit library. Following framework selection, the integration of the LLM model into LangChain is executed, ensuring seamless functionality for generating responses across diverse topics. A curated dataset comprising questions and corresponding answers is then prepared to train the chatbot, leveraging transfer learning techniques to expedite training. The deployment environment is configured using Streamlit to create an intuitive user interface for interacting with the chatbot, ensuring optimal performance across various devices. Rigorous testing is conducted to evaluate the chatbot's performance in understanding user queries and providing relevant answers, with metrics such as accuracy, relevance, and user satisfaction assessed through simulated interactions and feedback. Iterative improvements are implemented based on testing results and user input, with enhancements made to address any identified shortcomings and optimize performance continually. Finally, comprehensive documentation and reporting are compiled to document the development process, methodologies employed, and findings obtained, providing insights into the project's execution and outcomes.

4.1 Study About the model

The trajectory leading to the inception of GPT-3.5-Turbo-Instruct traces back to the seminal development of the GPT-3 model. Introduced by OpenAI in June 2020, GPT-3 represented a monumental leap in the realm of natural language processing. With an unprecedented scale of 175 billion parameters, GPT-3 showcased unparalleled linguistic prowess, adeptly comprehending and generating human-like text across a diverse array of topics and tasks. Its release heralded a new era in AI-driven language models, demonstrating the transformative potential of large-scale deep learning architectures in understanding and generating human-like text.

Despite its remarkable capabilities, GPT-3 exhibited certain limitations, particularly in scenarios necessitating precise instruction following and task completion. This recognition catalyzed the development of GPT-3.5-Turbo-Instruct, a specialized variant tailored explicitly for task-oriented interactions. Building upon the foundation laid by its predecessor, GPT-3.5-Turbo-Instruct inherits GPT-3's impressive linguistic abilities while introducing refinements optimized for instruction following and task completion.



Source: *https://arxiv.org/ftp/arxiv/papers/2303/2303.10420.pdf*

Figure 1: The evolutionary relationship of the GPT series models

The evolution from GPT-3 to GPT-3.5-Turbo-Instruct signifies a deliberate progression towards more specialized and purpose-driven language models. While GPT-3 demonstrated the feasibility and potential of large-scale language models in natural language understanding and generation, subsequent iterations sought to refine and optimize these capabilities for targeted applications. GPT-3.5-Turbo-Instruct represents a strategic response to the demand for precision and efficiency in specific use cases, where clear directives and accurate responses are paramount.

In the landscape of AI-driven language models, GPT-3.5-Turbo-Instruct occupies a unique niche, distinct from its conversational counterpart, GPT-3.5-Turbo. While GPT-3.5-Turbo excels in engaging in natural language conversations, the Instruct variant prioritizes task-oriented functionality. This dichotomy underscores the versatility of OpenAI's language models, catering to a diverse spectrum of user needs and preferences.

One of the defining features of GPT-3.5-Turbo-Instruct is its emphasis on instruction following and task completion, minimizing the need for additional prompts or clarification. This precision and efficiency render it an invaluable tool for scenarios necessitating clear directives and accurate

responses. Leveraging its multi-turn capability, users can provide context and refine instructions over successive interactions, augmenting the model's understanding and efficacy.

The development of GPT-3.5-Turbo-Instruct underscores OpenAI's commitment to innovation and user-centric design. By addressing specific challenges and demands in task-oriented interactions, the Instruct model exemplifies OpenAI's dedication to advancing the capabilities of its language models. Despite the subsequent release of the GPT-4 model, the continued relevance and efficacy of GPT-3.5-Turbo-Instruct underscore its utility and value in specialized applications.

Furthermore, the integration of training records as much as September 2021 guarantees that GPT-3.5-Turbo-Instruct is informed with the aid of the today's available data, enabling it to offer up-to-date and correct responses. This commitment to leveraging the most comprehensive dataset available underscores OpenAI's dedication to advancing the capabilities of its language models and meeting the evolving needs of its users.

Table 1: Performances of LLM Models

		Model			
Task	**Dataset**	Gemini Pro	GPT 3.5 Turbo	GPT 4 Turbo	Mixtral
Knowledge-based QA	MMLU (5-shot)	64.12	67.75	**80.48**	-
	MMLU (CoT)	60.63	70.07	**78.95**	-
Reasoning	BIG-Bench-Hard	65.58	71.02	**83.90**	41.76
Mathematics	GSM8K	69.67	74.60	**92.95**	58.45
	SVAMP	79.90	82.30	**92.50**	73.20
	ASDIV	81.53	86.69	**91.66**	74.95
	MAWPS	95.33	**99.17**	98.50	89.83
Code Generation	HumanEval	52.44	65.85	**73.17**	-
	ODEX	38.27	42.60	**46.01**	-
Machine Translation	FLORES (0-shot)	29.59	37.50	**46.57**	-
	FLORES (5-shot)	29.00	38.08	**48.60**	-
Web Agents	WebArena	7.09	8.75	**15.16**	1.37

Source: *https://arxiv.org/pdf/2312.11444.pdf*

OpenAI's GPT-3.5-turbo-instruct is a language model designed to excel in understanding and executing particular commands correctly. Unlike GPT-3.5-turbo, which is in general geared in the direction of undertaking conversations, GPT-3.Five-turbo-instruct shines in completing numerous obligations and answering questions directly. This new practise language version is designed for effectively following precise instructions, just like the chat-targeted GPT-3.5-turbo. It gives the identical value and performance as different GPT models, all inside a 4K context window, and it uses education facts up to September 2021.

Despite the availability of the GPT-4 model, This Model remains a powerful and value-effective alternative. It powers the broadly popular ChatGPT and gives users the ability to create their very own chatbot with comparable abilties. One of the key benefits of that version is its multi-turn capability, allowing it to just accept a sequence of messages as input. This characteristic is an improvement over the GPT-3 model, which only supported single-turn textual content activates. With this option, customers can make use of preset situations and previous responses as context to beautify the satisfactory of the generated reaction.

The number one intention of this model is to effectively follow commands. It isn't always intended to be a conversational version but rather a project-orientated one. This difference units it apart from chat fashions and makes it fantastically green at imparting precise responses. Whether you

need particular responsibilities accomplished or have questions that require specific solutions, this model has were given you blanketed.

Perhaps the most exceptional difference between GPT-3.5-turbo and the model used in project is their method to interactions. While GPT-3.5-turbo is conversational and chatty, the Instruct variation is more undertaking-orientated. It excels in following commands with out requiring additional prompts, making it an invaluable device for particular responsibilities and queries.

Why OpenAI Developed GPT-3.5 Turbo Instruct

Here's why this model came into existence:

Clearer and On-Point Answers: OpenAI recognized that older models sometimes generated responses that were either unclear or strayed off-topic. GPT-3.5 Turbo Instruct was trained to address these issues, ensuring that the answers it provides are not only accurate but also straightforward.

Versatility for Everyone: Whether you're a tech guru or someone with limited technical knowledge, GPT-3.5 Turbo Instruct is designed to cater to your needs. It's about making AI accessible and useful for a wide range of users.

Heightened Accuracy: One of the primary objectives of GPT-3.5 Turbo Instruct is to provide coherent and contextually relevant responses. This means you can expect even more accurate answers to your queries.

Reduced Toxicity: GPT-3 models, while revolutionary, had a tendency to generate outputs that could be untruthful or harmful. The refinement process of GPT-3.5 Turbo Instruct includes reducing this toxicity, ensuring that the information it generates is both reliable and safe.

Reinforcement Learning from Human Feedback (RLHF): OpenAI employed a cutting-edge technique known as reinforcement learning from human feedback. This approach involves real-world demonstrations and evaluations by human labelers, enabling the model to learn and improve from human interactions.

Table 2: Comparison of different gpt model

Comparison of different OpenAI Models

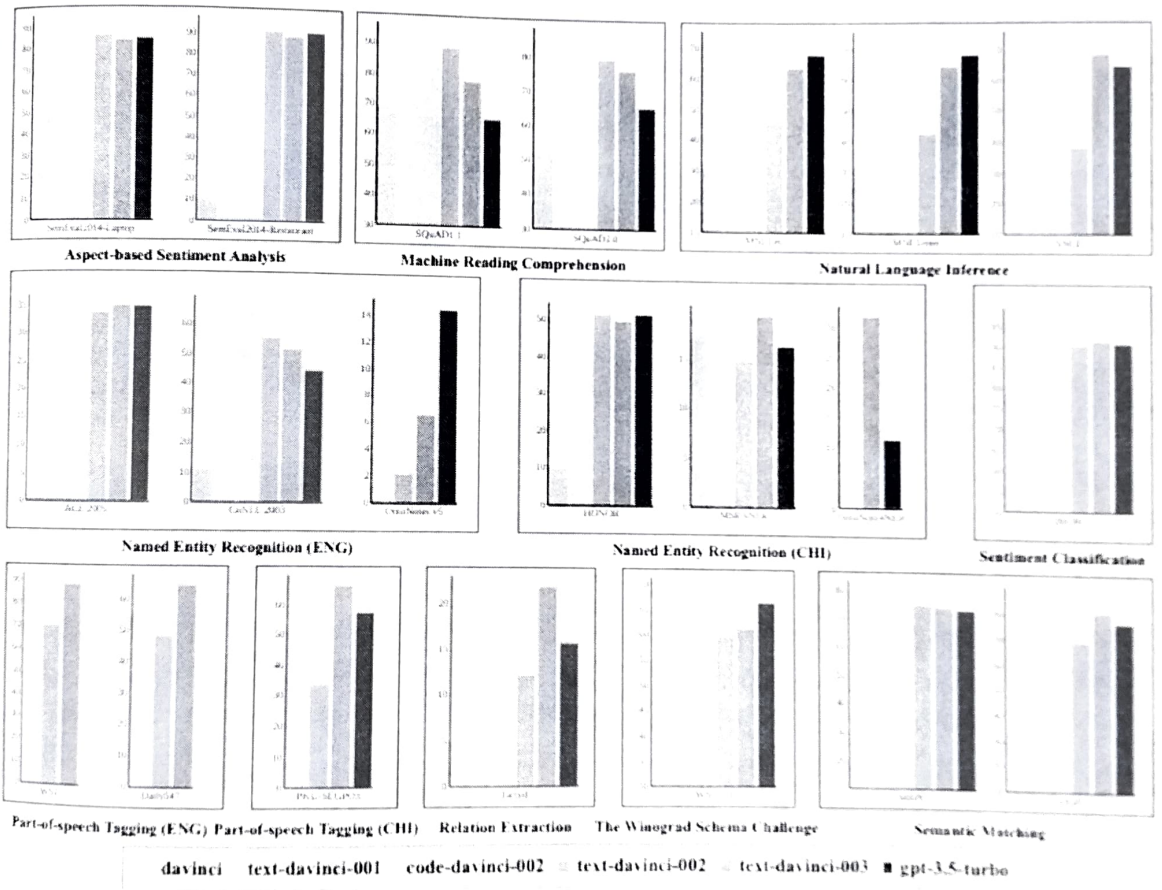
MODEL	DESCRIPTION	CONTEXT WINDOW	TRAINING DATA
gpt-3.5-turbo	Currently points to gpt-3.5-turbo-0613. The gpt-3.5-turbo model alias will be automatically upgraded from gpt-3.5-turbo-0613 to gpt-3.5-turbo-0125 on February 16th.	4,096 tokens	Up to Sep 2021
gpt-3.5-turbo-1106	GPT-3.5 Turbo model with improved instruction following, JSON mode, reproducible outputs, parallel function calling, and more. Returns a maximum of 4,096 output tokens. Learn more.	16,385 tokens	Up to Sep 2021
gpt-3.5-turbo-instruct	Similar capabilities as GPT-3 era models. Compatible with legacy Completions endpoint and not Chat Completions.	4,096 tokens	Up to Sep 2021
gpt-3.5-turbo-16k	Legacy Currently points to gpt-3.5-turbo-16k-0613.	16,385 tokens	Up to Sep 2021
gpt-3.5-turbo-0613	Legacy Snapshot of gpt-3.5-turbo from June 13th 2023. Will be deprecated on June 13, 2024.	4,096 tokens	Up to Sep 2021
gpt-3.5-turbo-16k-0613	Legacy Snapshot of gpt-3.5-16k-turbo from June 13th 2023. Will be	16,385 tokens	Up to Sep 2021

How does gpt-3.5-turbo-instruct work?

GPT-3.5 Turbo Instruct, an advancement in the field of large language models, is designed to excel at following specific instructions provided by humans. This capability stems from its foundation on the GPT (Generative Pre-trained Transformer) architecture, but with an additional layer of training focused on adhering to user guidance. This fine-tuning process utilizes human feedback to instill in the model a deeper understanding of how to interpret and execute instructions effectively.

When a user interacts with GPT-3.5 Turbo Instruct, they typically provide a clear and concise set of instructions alongside any relevant information or context. This information is then processed by the model, enabling it to grasp the user's intent and desired outcome. Subsequently, GPT-3.5 Turbo Instruct leverages its vast knowledge base and language processing capabilities to generate text or complete tasks in alignment with the provided instructions.

For instance, a researcher might instruct the model to "Write a concise summary of the scientific research on the effects of climate change on coral reefs, focusing on the past decade and highlighting potential solutions." GPT-3.5 Turbo Instruct would then access and analyze relevant scientific articles, synthesize the information within the specified timeframe, and present a focused summary that incorporates potential solutions for coral reef preservation.



Source: *https://arxiv.org/ftp/arxiv/papers/2303/2303.10420.pdf*

Figure 2: The performance of different models in zero-shot scenario

Overall, the core strength of GPT-3.5 Turbo Instruct lies in its ability to interpret and act upon human instructions. This empowers users to leverage the model's capabilities for various applications, including generating creative text formats like poems or code, translating languages, summarizing complex information, and even assisting with research endeavors. It's important to note that while GPT-3.5 Turbo Instruct demonstrates significant advancements in instruction following, it's still under development and continuous improvement, with ongoing efforts to refine its accuracy and effectiveness.

4.2 Apply gpt-3.5 turbo instruct model

To integrate GPT-3.5-turbo-instruct into our chatbot, we followed a meticulous approach. Firstly, we established a secure connection to the OpenAI API using the provided authentication credentials. Subsequently, we designed prompts and instructions tailored for the chatbot's functionalities. These prompts provided GPT-3.5-turbo-instruct with the necessary context and guidance to effectively understand user queries and generate informative responses.

Next, we utilized the OpenAI API's functionalities to send user queries as prompts to the GPT-3.5-turbo-instruct model. The model then processed the prompts, accessed and retrieved relevant information through the LangChain framework if necessary, and finally generated responses in natural language that addressed the user's intent. This seamless interaction between the OpenAI API, GPT-3.5-turbo-instruct model, and the LangChain framework proved instrumental in building a robust and informative Q&A LLM chatbot.

4.3 Code on Visual Studio

To build the Q&A LLM chatbot, the development process involves several steps using the specified Python libraries within Visual Studio Code. Here's a breakdown of how the coding and building of the chatbot are executed:

Setting up the Environment: Begin by creating a virtual environment within Visual Studio Code to isolate the project dependencies. Install the necessary Python libraries using pip, including langchain, openai, huggingface_hub, python-dotenv, and streamlit.

Downloading and Configuring Models: Download the LLM model, GPT-3.5-turbo-instruct, from OpenAI or Hugging Face and configure it for use within the project. This involves setting up API keys, environment variables, and any other necessary configurations.

Integrating LangChain: Implement LangChain to bridge the gap between user prompts and model understanding. Utilize LangChain's capabilities to convey user queries effectively to the LLM model, ensuring accurate and contextually relevant responses.

Building the Streamlit Application: Utilize the streamlit library to develop the user-friendly web interface for the chatbot. Create input fields for user queries and options for selecting desired functionalities, such as asking questions or receiving answers.

Deploying the Chatbot: Once the chatbot application is developed and tested locally, deploy it using Streamlit's built-in deployment capabilities.

4.4 Deploying the App on Streamlit

Once the development phase concludes, the next step involves running our application. The following image illustrates the application's user interface.

Figure 3: Interface of the application built with Streamlit

5. INTERPRETATION

The successful deployment of our application on the Streamlit interface, accessible through localhost, marks a significant achievement in bringing our project to life. Streamlit's intuitive user interface empowers users to interact with the chatbot effortlessly. Designated input fields allow users to submit queries and personalize their experience by selecting from various chatbot roles.

The inclusion of personalized chatbot options, such as Data Scientist, Engineer, Researcher, English Tutor, and Fitness Coach, highlights the versatility and adaptability of our chatbot application. By tailoring responses based on the chosen persona, the chatbot enhances user engagement and satisfaction, catering to individual needs and preferences across different domains.

Our project's success lies in the effective integration of advanced natural language processing capabilities, facilitated by the LangChain framework and the GPT-3.5-turbo-instruct model. Through careful coding and configuration within Visual Studio Code, we harnessed the power of these technologies to create a functional and user-centric chatbot application.

The project's impact extends beyond its technical aspects. By demonstrating the feasibility and effectiveness of deploying sophisticated chatbot solutions on accessible platforms like Streamlit, we pave the way for broader adoption of AI-powered conversational agents in various domains and applications.

Figure 4: Running application on local host

Therefore our application runs perfectly and provides the desired output.

6. CONCLUSION

We succesfully built and deployed a Q&A chatbot utilizing LangChain and GPT-3.5-turbo-instruct, accessed via OpenAI's API and deployed in a user-friendly interface. It explored the potential of this combination to create a personal assistant-like chatbot for answering diverse questions across various domains. The project successfully demonstrated the effectiveness of LangChain in retrieving relevant information for the chatbot and the GPT-3.5-turbo-instruct model's ability to understand user queries and generate informative responses. This contributes to the advancement of Q&A chatbot development by showcasing the potential of these technologies in a practical application. Future work could involve data analysis tools like LangSmith to refine the chatbot based on user interaction patterns and explore incorporating domain-specific models for even more specialized applications. Overall, this project successfully explored the potential of LangChain and GPT-3.5-turbo-instruct for building a versatile Q&A chatbot, paving the way for further advancements in the field.

7. FUTURE SCOPE AND RECOMMENDATIONS

The successful development and deployment of the Q&A chatbot utilizing LangChain and GPT-3.5-turbo-instruct mark a significant step forward in conversational AI. Here are some key areas for future exploration and enhancement:

Leveraging User Interaction Data: Integrating data analysis tools like LangSmith can offer valuable insights into user interaction patterns and preferences. By analyzing user feedback and engagement metrics, the chatbot's performance can be refined and personalized to better meet user needs.

Tailoring the Chatbot to Specific Domains: Future iterations of the chatbot could explore integrating domain-specific models. By fine-tuning the language model on domain-specific data, the chatbot can provide more accurate and tailored responses in various fields, potentially enhancing its applicability in sectors like healthcare, finance, or legal services.

Expanding Multimodal Communication: Expanding the chatbot's capabilities beyond text-based interactions to include support for multimedia inputs (such as images, audio, or video) can enrich the user experience and enable more diverse applications. Integration with multimodal language models can facilitate comprehension and generation of responses across different modalities, leading to richer interactions.

Continuous Improvement through User Feedback: Implementing mechanisms for collecting and analyzing user feedback on an ongoing basis is crucial for the chatbot's continuous improvement. By continuously soliciting user input and adapting to evolving user preferences, the chatbot can stay relevant and effective in addressing user queries and needs.

Addressing Ethical and Privacy Concerns: Addressing Ethical and Privacy Concerns: As conversational AI technologies retain to boost, proactively addressing ethical and privateness concerns is important. Future research have to focus on ensuring transparency, equity, and user privacy in chatbot interactions. This includes incorporating mechanisms for knowledgeable

consent, records protection, and algorithmic duty, fostering trust and responsible development in AI-powered interactions.

REFERENCES

[1] Matsuda, K. and Frank, I., 2024. LangChain Unleashed: Advancing Education Beyond ChatGPT's Limits. arXiv preprint arXiv:2410.02474.

[2] Abbasian, M., Azimi, I., Rahmani, A.M. and Jain, R., 2024. Conversational health agents: A personalized llm-powered agent framework. arXiv preprint arXiv:2410.02474.

[3] Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. and Zhang, H., 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36. arXiv preprint arXiv:2306.05685v4.

[4] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X. and Wang, C., 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. arXiv preprint arXiv:2308.08155.

[5] Topsakal, O. and Akinci, T.C., 2023. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In International Conference on Applied Engineering and Natural Sciences. arXiv preprint arXiv:2308.04472.

[6] Pesaru, A., Gill, T.S. and Tangella, A.R., 2023. AI assistant for document management Using Lang Chain and Pinecone. International Research Journal of Modernization in Engineering Technology and Science. arXiv preprint arXiv:2310.05476.

Q&A LLM Chat Bot Using LangChain Framework

ORIGINALITY REPORT

9%	7%	4%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.clarifai.com Internet Source	5%
2	Pegah Safari, Mehrnoush Shamsfard. "Data augmentation and preparation process of PerInfEx: a Persian chatbot with the ability of information extraction", IEEE Access, 2024 Publication	1%
3	clarifai.com Internet Source	1%
4	journal.achsm.org.au Internet Source	<1%
5	Antonio Di Maria, Lorenzo Bellomo, Fabrizio Billeci, Alfio Cardillo et al. "A web-based platform for extracting and modeling knowledge from biomedical literature as a labeled graph", Bioinformatics, 2024 Publication	<1%
6	pdfs.semanticscholar.org Internet Source	<1%
7	Submitted to Curtin University of Technology	

Anuj Tewari
23 April 2024

—	Student Paper	<1 %
8	machinelearningmodels.org Internet Source	<1 %
9	Momen Yaseen M. Amin. "AI and Chat GPT in Language Teaching: Enhancing EFL Classroom Support and Transforming Assessment Techniques", International Journal of Higher Education Pedagogies, 2023 Publication	<1 %
10	nips.cc Internet Source	<1 %
11	www.ncbi.nlm.nih.gov Internet Source	<1 %
12	Bruce Hopkins. "ChatGPT for Java", Springer Science and Business Media LLC, 2024 Publication	<1 %
13	Michael Balas, Edsel B. Ing. "Conversational AI Models for ophthalmic diagnosis: Comparison of ChatGPT and the Isabel Pro Differential Diagnosis Generator", JFO Open Ophthalmology, 2023 Publication	<1 %
14	Christopher J. Lynch, Erik J. Jensen, Virginia Zamponi, Kevin O'Brien, Erika Frydenlund, Ross Gore. "A Structured Narrative Prompt	<1 %

for Prompting Narratives from Large Language Models: Sentiment Assessment of ChatGPT-Generated Narratives and Real Tweets", Future Internet, 2023

Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		