

Conversational Analytics BOT using ChatGPT

Authors: Prakash Reddy, Pavan Alluri, Anoop Siripragada, Ankit Chhabra, Anirudh Sharma, CustomerInsights.AI

Presenters: [To be updated], CustomerInsights.AI



Introduction

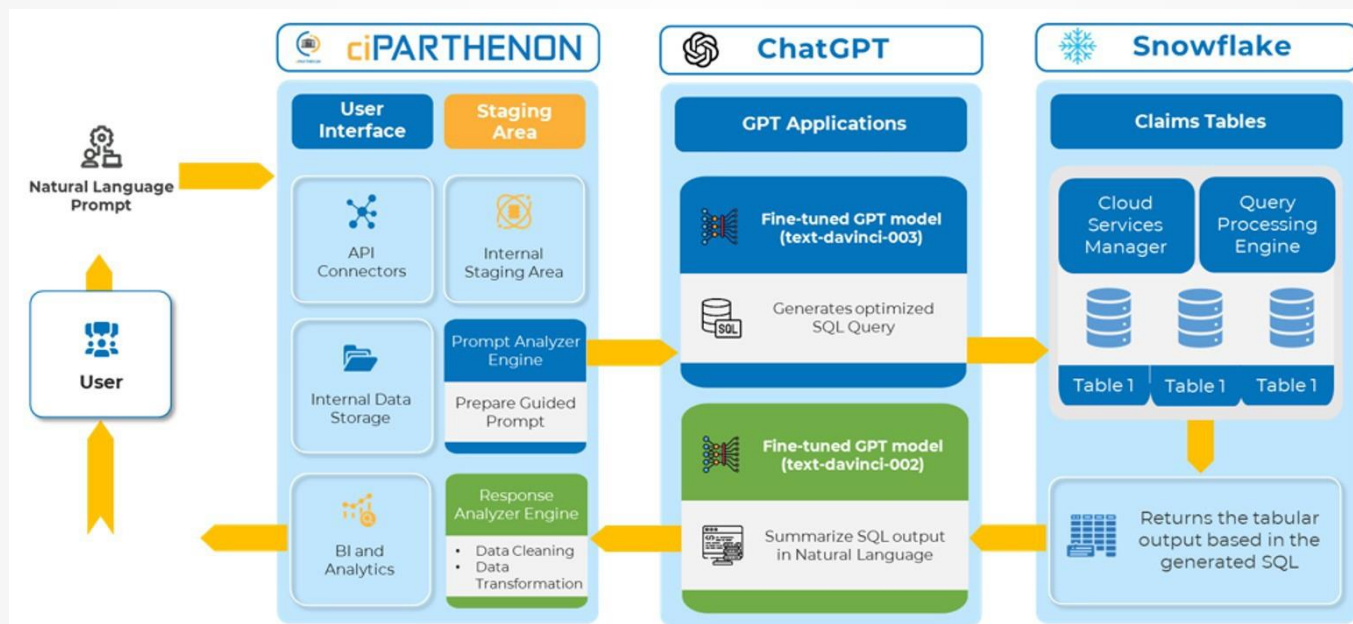
Chatbots are becoming increasingly popular in various domains, including healthcare and life sciences. However, most chatbots rely on pre-defined rules or scripts, which limit their ability to handle complex and dynamic user queries. In this paper, we propose a novel approach to develop a conversational chatbot for the life sciences domain, based on fine-tuning ChatGPT, a large-scale pre-trained language model with the goal to engage with users in a natural conversation, providing them with relevant and accurate information related to the pharmaceutical industry.

We utilized the anonymized patient claims data containing information on drugs, patients, physicians, and associated insurance plans for each of the diagnostic, prescription, and procedure codes to train the ChatGPT model on domain-specific knowledge and pertinent language for accelerating time to insights. The fine-tuned model has been made capable of consuming natural language questions as input to subsequently return the results.

Post the successful alpha testing, the product is now being used by one of our large-pharma customers to fetch accurate and reliable results across the plethora of other data assets.

Architecture

The well-designed architecture automatically scales up the underlying computational resources to generate the final insights on the queried data.



Three major components involved in developing this Conversational Analytics BOT:

1. **ciPARTHENON** – Cloud-based Life Sciences Commercial Analytics platform by CustomerInsights.AI, that allows users to input natural language query for deriving insights out of the underlying data assets; it is integrated with ChatGPT through API connectors. This platform has been added with two main components:
 - a. PromptAnalyzerEngine – to convert the natural language query to the guided prompt which is fed as actual input to the fine-tuned ChatGPT models.
 - b. Response Analyzer Engine – to help convert the response from the ChatGPT models into user- friendly format.
2. **ChatGPT Applications** – Choosing the right models for text (code) generation and summarizing the text is important because it can affect the quality of final output. Text generation has become one of the most important yet challenging tasks in natural language processing (NLP). We have utilized the following generative pretrained GPT models to generate optimized SQL query and summarize the output.

- a. Text-davinci-003 – is one of the Large Language Models (LLM) that can generate accurate and long text. It can handle more complex instructions and produces detailed responses. This model has been fine-tuned with numerous prompts and responses that are specific to our internal data to ensure that the model is well-versed with the context.
- b. Text-davinci-002 – is also a family of LLMs, is utilized to summarize the SQL outputs in the natural language format.

3. **Snowflake** – is a cloud data platform where all our internal pharma tables are hosted.

All three components are seamlessly integrated and hosted on Azure cloud to facilitate smooth communication between them.

How did we fine-tune chatGPT models?

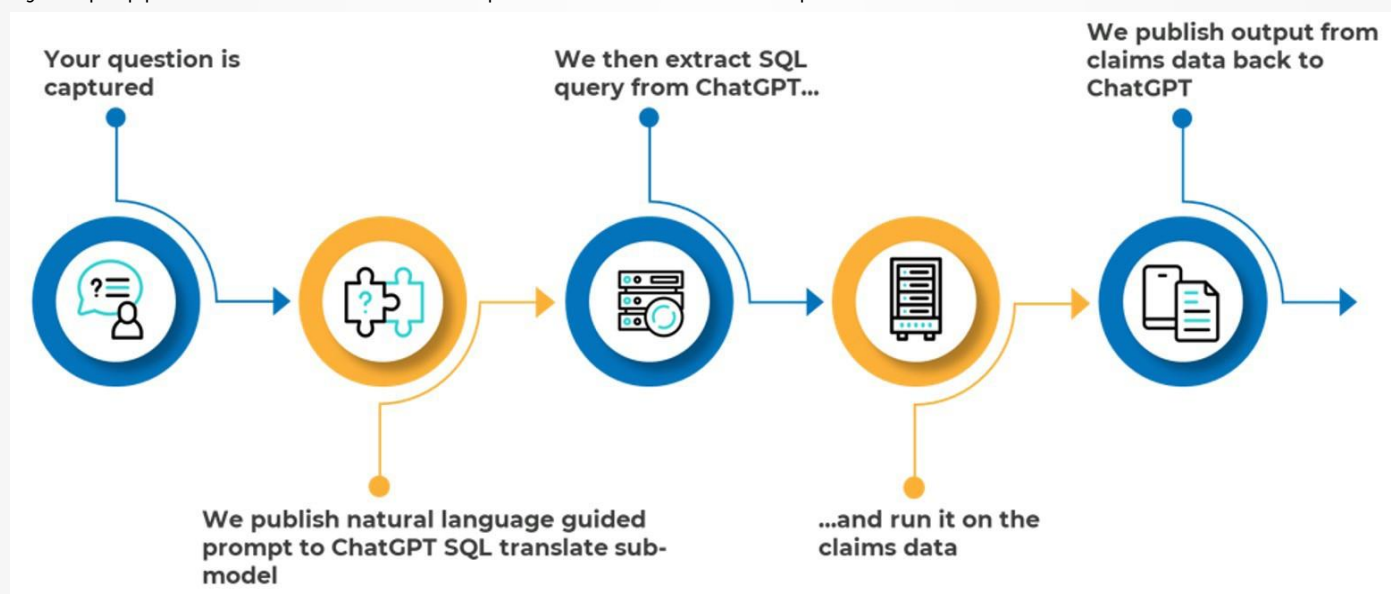
ChatGPT models have been trained on a vast corpus of open text available up until 2021 and do not contain information post that referred to process called 'Few-Shot Learning'. However, they have the flexibility to be trained on more recent data and deployed in a production environment.

Fine-tuning improves on few-shot learning by training on large number of examples that can fit in the prompt, allowing the user to achieve better results on wider range of tasks. At a high level, fine-tuning process involves the following steps:

- a. Prepare user / data specific training data – in the form of prompt and response.
- b. Train the models – existing models are trained using latest data (prompts and responses)
- c. Deploy the models – fine-tuned models are deployed on Azure cloud infrastructure and are accessed through OpenAI APIs.

How did we set up the whole process?

The process of fine-tuning ChatGPT on internal pharmaceutical data involved feeding the language models with large amounts of prompts and responses which includes the extensive information of the internal pharma tables schema and guidelines to generate optimized Snowflake SQL queries. The model has been trained to recognize patterns and associations within data, allowing it to generate meaningful responses to user queries. Below is the step-by-step approach we utilized to develop a conversational bot for pharmaceutical data.



The trained models are incorporated into our internal PLaaS (Platform Led Analytics as Service) platform – ciPARTHENON, which serves as the front-end interface for user interaction. The chatbot is specifically engineered to comprehend natural language queries and provide relevant information in a conversational tone. Our chatbot possesses the capability to offer users a personalized and informative experience that assists them in navigating the intricate landscape of the pharmaceutical industry.

Some of the key analyses using the Conversational BOT:

Here are a few examples where the BOT was able to retrieve data and accurately summarize the results.

Market Sizing:

- a) How many patients that got diagnosed with immunodeficiency are likely to use a specific drug?
- b) What is the unique count of patients that have been prescribed a specific drug in 2022?
- c) What is the national-level payer mix for a specific drug in 2022?

Propensity of physicians and accounts based on drug utilization:

- a) Which geographies have the greatest number of physicians diagnosing patients with a specific condition?
- b) What is the share of voice of a specific drug's competitor nationally versus split by states?
- c) Provide a list of accounts that enable 75% of a specific drug's sales nationally.

Diagnostic-level trends for disease areas of interest

- a) Give me the claim count diagnosed with a given condition?
- b) Give me the claim count by top payer type for RA in 2022?

Evolution of patient trends over months or weeks:

- a) Give me the monthly trend of unique patient counts who have used a drug in 2022?
- b) Give me the monthly trend of unique patient counts who have diagnosed with a given condition in 2022?

To accurately extract and summarize the data, for example questions like above, the user must possess coding proficiency, knowledge of the table schema, and a comprehensive understanding of the data, which is both skill-intensive and time-consuming. Our Conversational Analytics BOT helps eliminate such constraints by assisting business leaders or executives to quickly ask queries and receive responses in natural language.

Conclusion //

We conclude that ChatGPT is a powerful and flexible framework for developing conversational chatbots for the pharma domain and other potential applications. Overall, this work presents an innovative approach to developing conversational chatbots that leverage the power of advanced language models like ChatGPT. The model was trained on internal pharmaceutical data to create a chatbot that is both accurate and informative, providing users with the information they need to make informed decisions about their health and wellbeing.