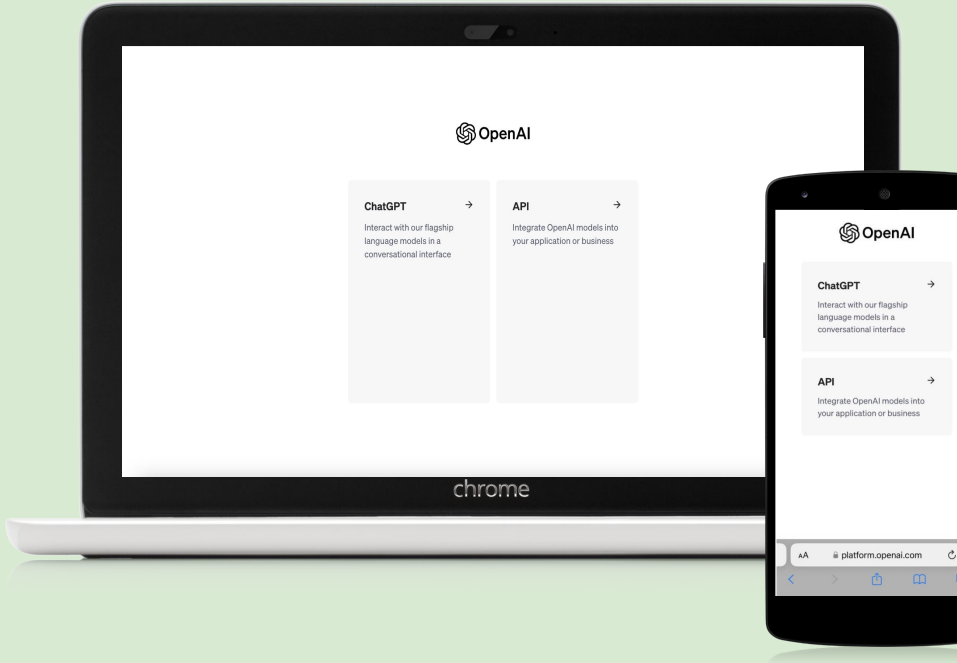
OpenAI ChatGPT API

Solving Classification Problems Using AI



Outline

The Problem

The task of accurately classifying data is a fundamental challenge in various domains, especially in medical diagnostics. In the context of liver disease prediction, traditional machine learning models often face limitations in achieving high classification accuracy. Despite extensive preprocessing, feature engineering, and model optimization, these models may still yield unsatisfactory results..

The Dataset

Source : Kaggle
Size : 410 rows with labels and 12 attributes.
Description : A collection of 410 patient records, each meticulously labeled with 12 attributes. This dataset serves as a valuable resource for conducting research and analysis within the domain of liver disease classification. The primary objective of this dataset is to facilitate the development and evaluation of machine learning models and algorithms. Researchers, data scientists, and healthcare professionals can leverage this dataset to create predictive models that accurately classify patients into categories such as "positive" or "negative" for liver disease, based on the provided attributes.

Link : https://www.kaggle.com/competitions/classifying-liver-disease-patients

Traditional Models & their Accuracies

Model Name	Accuracy	F1-Score
Decision Tree	65.8%	75.8%
XG Boost	68.2%	78.6%
Support Vector Classifier	54.4%	58.8%
Logistic Regression	73.1%	83.2%

Traditional Models & their Accuracies After Hyper-Parameter Tuning

Model Name	Accuracy	F1-Score
Logistic Regression	68.2%	79.3%
Decision Tree	65.04%	77%
Support Vector Classifier	63.4%	75.9%
Random Forest	67.4%	79.9%

Problems??

- Limited Accuracy
- Potential Bias in Classifications
- Lack of Context Awareness.
- While these algorithms have been in use, they've encountered certain challenges that must be addressed

Solution Proposal

Leverage Large Language Models

- Large Language Models (LLMs) represent a breakthrough in artificial intelligence, with their ability to understand, generate, and manipulate human language.
- By using LLMs like GPT-3.5 Turbo, we tap into their immense language processing capabilities to enhance the accuracy of complex tasks such as classifying patients for liver diseases.
- LLMs can analyze vast amounts of textual data, recognize patterns, and make predictions with a high degree of accuracy.

Approach : Fine-Tuning LLM

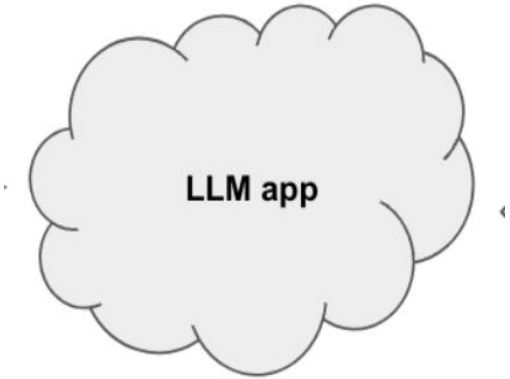
- Fine-Tune existing LLM models **eg** : gpt-3.5-turbo, davinci-002, babbage-002
- Gpt-3.5-turbo is the best model out there for fine tuning.
- Train the model with our data in the expected format by openai
- Expected formats are json, jsonl
- Can choose prompt completion format or chat completion format.
- Use Chat completion format when working with single dataset or with multiple data sources.

Single Dataset

Choose the data set or Data source

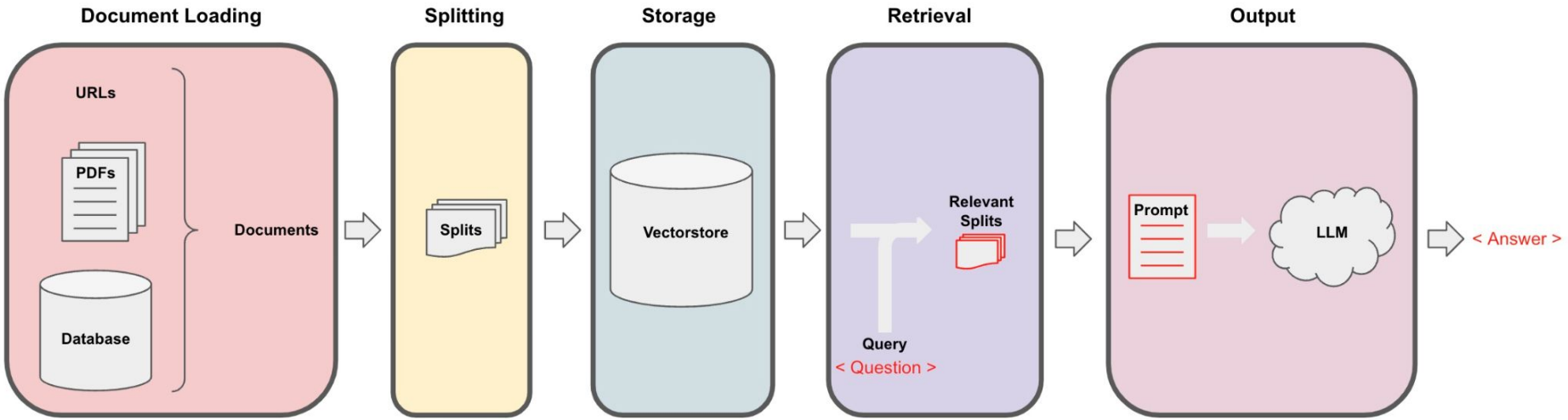
LangChain

Prepare the data pipelines using Langchain framework



Train the Large language model

Multiple Data Sources



Code Walkthrough

Current Limitations

- Unable to send large input data to the OpenAI API due to limitation in the number of tokens that can be sent per API call. Hence, we were able to send only few rows of data with one API call which limited our training ability.

- Can divide data into segments and send them separately to the API but it would be time consuming if our test dataset has millions of records.

How to overcome?

We are trying to use Langchain’s ability to preprocess large amounts of data and store in the form of documents which will in turn be stored on vector index stores. This would make data retrieval fairly easy and efficient which can be provided as input to OpenAI API.

References

Open AI Fine Tuning: https://platform.openai.com/docs/guides/fine-tuning

Chat Completion API : https://platform.openai.com/docs/guides/gpt/chat-completions-api

RAG pipeline : https://python.langchain.com/docs/use_cases/question_answering/

Open Ai pricing : https://openai.com/pricing

Questions?