

Minh T C Nguyen Computer Science



Prof. Tingting Yu Project Advisor

# Data Cleaning

## A GPT3-based bot that helps you in cleaning data and so on

Computer Science, Machine Learning | Class of 202III

## Data cleaning and analysis are two critical processes in data science, as they involve identifying and correcting errors, inconsistencies, and inaccuracies in a dataset, as well as extracting meaningful insights and knowledge from the data. However, data cleaning and analysis can be timeconsuming and challenging tasks, especially if you don't have great domain knowledge of cleaning data or the data itself, it will cost a lot of time for looking at tools, solutions, pipelines on the

## Solution

Problem

Data Cleaning is an assistant that helps you to provide verbal solution and programming solution to clean data, extract data, and even featureengineering data adopting a fine-tuned GPT3based models, cutting-edge AI technology at current 2023.

#### **Features**

- It is open until it is not open.
- Provide data cleaning, data analyzing, and feature engineering solutions, given a dataset.
- Guarantee almost zeros data leakage: we keep no user's data and reduce the information send to OpenAl.
- A user-friendly chatbot with non-negative responses. Sometimes misunderstanding your question

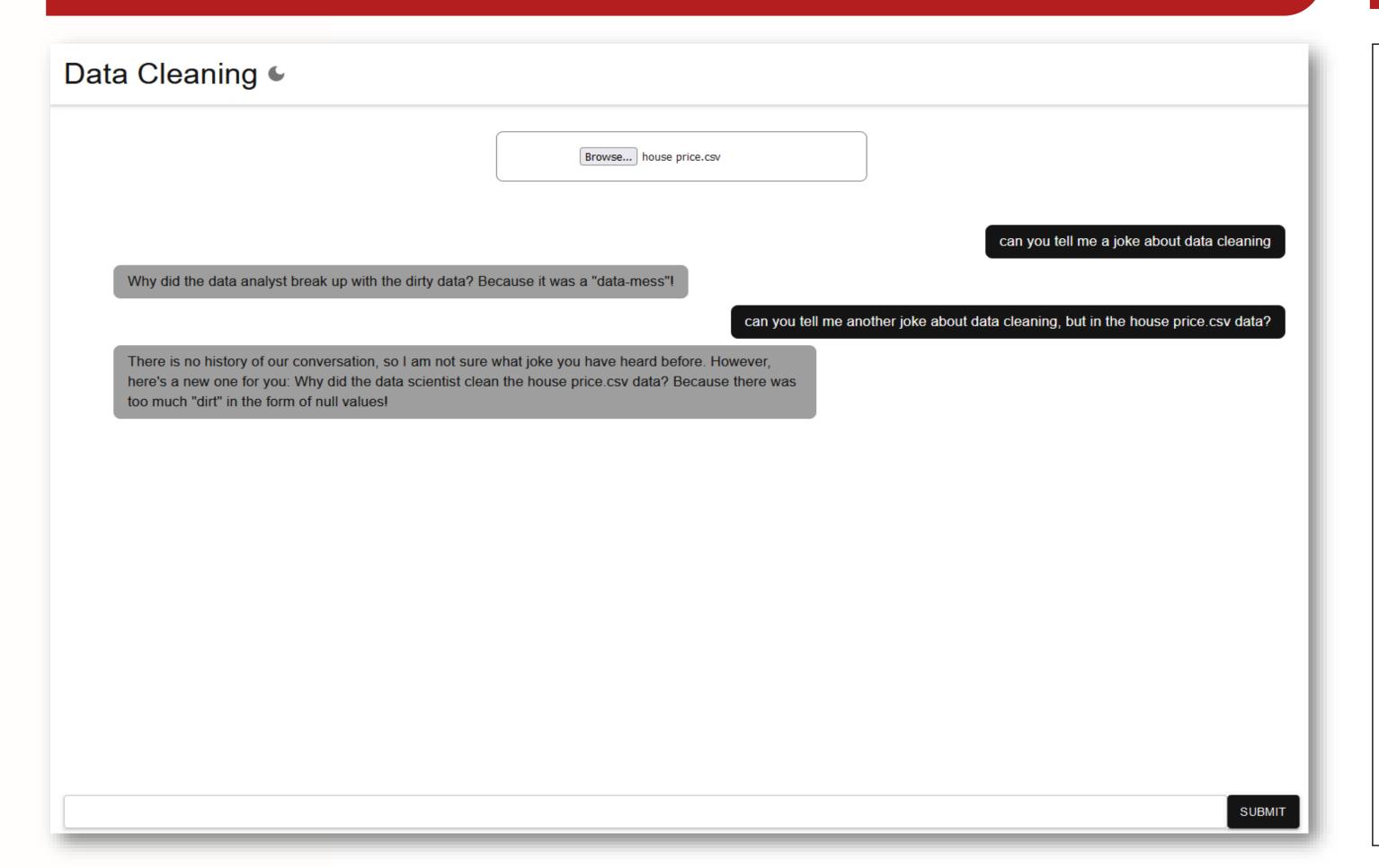
### Reference:

Internet.

C. Liu, L. Zhu, and M. Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. Applied and Computational Harmonic Analysis, 59:85–116, 2022.

# Diagram submit instruction Malware check yets. When did the date occurrent countries have been been all the countries. POST Filter instruction **OpenAI** user message get response

#### Tech Stack Interface





implement and use as assistants for various information

Results

After 39 hours of working including collecting data, fine-tuning

the model, testing the model on some new datasets, planning

discussing with my project advisor, I have deployed a chat user-

Proving that the current Large Language Models(LLMs), such as

GPT-3, have the potential to be valuable tools for businesses to

Challenges

and implementing the website, testing web stability, and

friendly interface that is able to provide solutions for data

cleaning, data analysis, and feature engineering

concerns.

- Availability: Due to the rapid growth of ChatGPT, OpenAI has decided to modify its plans by discontinuing public access to certain models that have been trained on specific tasks, such as GPT-3 Codex. As a result, alternative solutions like GPT-3.5 Turbo have been used, but these models may not be as effective as Code one since it has a smaller amount of code in their dataset. While GPT-3.5 Turbo can be used as a chatbot, it may not be as effective as a code provider compared to GPT-3. Moreover, the API rate limit has been reduced recently comes along with the API outage problem from OpenAI, the bot cannot function well.
- Data Collecting: modern AI contains several major factors including big data. Recent research proves that bias-variance trade-off (U-shaped risk curve) can be disregarded when we overparameterized enough using big data and deep neural network, and allow us to reach the modern optimum (Liu et al., 2022). Big data is important, and if I do have resources to collect a much larger dataset, I could improve the play of the model.
- Budget: using OpenAI and Azure cloud are expensive. If I have enough money to allow users to use my OpenAl API, I could implement reinforcement learning to learn from users' feedback.

## **Future Plans**

- With the release of GPT-4, LLaMA, and Stanford Alpaca. I am joining their waitlists so I could implement one of the models. Especially GPT-4, since it is a multimodal model which can take not only text but other types of inputs.
- Providing users with outputs like plots and stats
- Finding donors and collaborators for my project.

