Analyzing DevGPT Dataset

Based on the NAIST DevGPT Dataset

Srinija Dharani
Computer Science
Florida State University
Tallahassee, FL
sd23be@fsu.edu

Roochita Ikkurthy
Computer Science
Florida State University
Tallahassee, FL
ri23b@fsu.edu

Chaitanya Guntupalli
Computer Science
Florida State University
Tallahassee, FL
cg21bb@fsu.edu

ABSTRACT

The current project status report encapsulates our research endeavors centered on ChatGPT's interactions within the developer community, focusing on three pivotal areas. The first area of investigation involves a comprehensive analysis of programming language queries posed to ChatGPT, with an emphasis on identifying temporal shifts in language preferences. This segment is currently in the data collection and extraction phase. The second aspect delves into the variability of conversation lengths with ChatGPT, contingent upon the nature of the issues presented. This involves a methodical categorization of conversation types, a process that is presently being refined. The final facet of our research is concentrated on pinpointing the most prevalent prompts directed at ChatGPT by developers and critically analyzing the AI's response mechanisms. For this, we are in the initial stages of a thorough data mining exercise. The research utilizes sophisticated data analysis tools and adheres strictly to ethical standards in data handling. As we progress, we anticipate generating actionable insights, with subsequent updates planned to chronicle ongoing developments in this multifaceted research initiative.

KEYWORDS

Data Analysis, NLP, NLTK

1 Research Question 1

1.1 Problem Statement

The objective of this research is to ascertain how the length of interactions with ChatGPT fluctuates in relation to the nature of the issues presented. This inquiry is significant as it offers critical insights for developers, illuminating the dynamics between the issue's complexity and the ensuing conversation duration. Understanding these patterns is pivotal for optimizing ChatGPT's efficiency and user experience.

1.2 Dataset Version

For this analysis, we have employed data from Snapshot20231012. This specific version was chosen due to its comprehensive nature, as it integrates data from previous snapshots, thereby streamlining the research process.

1.3 Approach & Tools

The analysis was conducted using Python within a Jupyter Notebook environment. A collaborative approach was adopted, leveraging insights from team discussions. It was revealed that the recent snapshot encompasses data from earlier versions, which significantly reduced the workload and data management complexity. The initial phase involved extracting relevant data from Snapshot20231012, followed by a meticulous categorization of issues. We then calculated the conversation lengths and employed data visualization tools, such as Matplotlib from Python's libraries, to present our findings.

1.4 Results & Implications

The current phase of the project has yielded results indicating the average conversation length for each categorized issue. These results are preliminary and pertain to individual files from Snapshot20231012. Moving forward, the aim is to extend this analysis across all files collectively. Additionally, we plan to refine our issue identification process by incorporating a broader range of keywords. This enhancement is expected to reduce the number of issues categorized as 'others', thereby yielding a more granular and insightful analysis of conversation lengths across a wider spectrum of issues.

1.5 About the Primary Contributor of this Research Question

Name: Chaitanya Guntupalli

FSU ID: cg21bb@fsu.edu

2 Research Question 2

2.1 Problem Statement

The primary objective of this research is to systematically identify and analyze the most frequent prompts that developers pose to ChatGPT. This includes determining the nature, context, and specific content of these queries. Following the identification of these common prompts, the study aims to meticulously examine how ChatGPT addresses these inquiries, i.e., code or no code solutions. The focus will be on evaluating the patterns, accuracy, and effectiveness of ChatGPT's responses in relation to the identified prompts. The outcome of this research will provide insights into the interaction dynamics between developers and ChatGPT, potentially informing improvements in AI response algorithms and enhancing user experience for developers engaging with ChatGPT.

2.2 Dataset Version

For this analysis, we have employed data from Snapshot20231012. This specific version was chosen due to its comprehensive nature, as it integrates data from previous snapshots, thereby streamlining the research process.

2.3 Approach & Tools

In order to identify the most commonly used keywords in prompts that ChatGPT receives, a list of such keywords was generated. Subsequently, a function was developed to summarize the prompt and generate the aforementioned list. The presence of the words in the summary was then checked against the list of keywords. If the words were present in the list, the ListOfCode column of the dataset was examined to determine if it was empty or not. If the column was empty, it indicated that ChatGPT provided a solution without code. Conversely, if the column was not empty, it indicated that ChatGPT provided a solution with code.
The tools used were – Jupyter Notebook, Pandas, NLP, NLTK.

2.4 Results & Implications

It would be premature to draw any conclusions at this time since the analysis has not been completed for all the data files. Once the analysis is finished, graphs will be created to display the results.

1.5 About the Primary Contributor of this Research Question

Name: Srinija Dharani

FSU ID: sd23be@fsu.edu

3 Research Question 3

3.1 Problem Statement

This analysis aims to identify the most occurring programming languages that developers ask ChatGPT to resolve their bugs. More specifically, what mostly occurring programming languages do developers ask ChatGPT to rectify their bugs and solve their issues in?

3.2 Dataset Version

For this analysis, we have employed data from Snapshot20231012. This specific version was chosen due to its comprehensive nature, as it integrates data from previous snapshots, thereby streamlining the research process.

3.3 Approach & Tools

To check if the result is correct, we have first selected a collection of keywords. The objective is to identify the most frequent PL using the terms (bugs, errors, features, and test) that are categorized from the Prompt column. We extracted the data and completed the categorization using the NLTK library. Subsequently, we discovered which programming languages developers asked ChatGPT to support, as well as how many of each category there were. We then created a graphic representation of PL versus categories.
The tools used were – Jupyter Notebook, Pandas, NLP, NLTK, Matplotlib.

3.4 Results & Implications

The results indicate that Python is the most frequently mentioned programming language when developers ask ChatGPT to resolve their coding issues. More analysis will follow in the final stage.

3.5 About the Primary Contributor of this Research Question

Name: Roochita Ikkurthy

FSU ID: ri23b@fsu.edu

4 Conclusion

A proper conclusion cannot be formulated at this time.