a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

DATA 542 Project (Group work)

Team size: 2-3 students.

Dataset:

You have worked with the DevGPT dataset in one of the classes. The dataset repository is here: https://github.com/NAIST-SE/DevGPT. Alternatively, you can download the dataset from Zenodo: https://zenodo.org/records/8304091.

The project is open-ended. You can select your own research questions or choose the ones from below and explore them using the dataset. These RQs are given from the MSR Mining Challenge 2024.

Suggested RQs:

1. What types of issues (bugs, feature requests, theoretical questions, etc.) do developers most commonly present to ChatGPT?
2. Can we identify patterns in the prompts developers use when interacting with ChatGPT, and do these patterns correlate with the success of issue resolution?
3. What is the typical structure of conversations between developers and ChatGPT? How many turns does it take on average to reach a conclusion?
4. In instances where developers have incorporated the code provided by ChatGPT into their projects, to what extent do they modify this code prior to use, and what are the common types of modifications made?
5. How does the code generated by ChatGPT for a given query compare to code that could be found for the same query on the internet (e.g., on Stack Overflow)?
6. What types of quality issues (for example, as identified by linters) are common in the code generated by ChatGPT?
7. How accurately can we predict the length of a conversation with ChatGPT based on the initial prompt and context provided?
8. Can we reliably predict whether a developer's issue will be resolved based on the initial conversation with ChatGPT?
9. If developers were to rerun their prompts with ChatGPT now and/or with different settings, would they obtain the same results?

a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

Your tasks

1) Select or define three research questions. **The questions require combining at least two of the files from the dataset.**
2) Define a methodology to answer the questions and implement them. Release your code on GH, and provide any instructions required to access them. The code should be well written and documented.
3) Write a short report (2-3 pages) about the methodology to answer each research question, as well as the obtained results and interpretations.

Code reuse

Should you decide to reuse code (written by you or others or GenAI) or the papers published on this dataset:

- indicate its copyright and how your usage does not violate it,
- give proper credits,
- how did you use the published papers, e.g., their code is used, their RQs is adopted, the results is compared with them.

Submission

To complete the assignment, submit to Canvas a concise pdf report that includes a link to your GitHub repository.

Grading rubric

	Weights	**Subtotals**
Report		40
Research questions and methodology	10	
Obtained results	10	

Interpretation of the results	10	
Concise and clear writing	10	
Code static		20
Follow conventions	10	
Comments	10	
Code execution		20
Syntax-error free, runs	10	
Takes reasonable time to run (so if you need to optimize your code or for loops, do so)	10	
Adhering the project requirements		20
Using at least two files	10	
Code/ideas reuse statements	10	
Total	100	100