DATA 542 Project – Milestone 1: Research Plan

HAOZHONG JI, University of British Columbia Okanagan, Canada
YIN-WEN TSAI, University of British Columbia Okanagan, Canada

1 Research Questions and Dataset

We use the AIDev dataset. It is about pull requests on GitHub that use AI coding agents. The dataset has information on pull requests, repositories, users, reviews, and commits. In this project, we only study pull requests that involve AI agents. We also keep only pull requests with valid times. Their final status must be clear. So they are either merged or closed.

We will answer three research questions (RQs). Each RQ uses more than one feature.

- **RQ1:** Do different AI coding agents get used in different types of repositories? We compare repositories with different popularity and activity.
- **RQ2:** How do the size and type of AI pull requests relate to review work and merge results, when we group pull requests into simple size and file-type buckets?
- **RQ3:** For AI-agent pull requests, how do simple human–AI collaboration patterns (based on follow-up commits) relate to merge rate and time-to-merge?

2 Planned Methods

For all RQs, we will load the pull request table and the repository table. We will join them by repository ID. We will keep only AI-related pull requests with clear outcomes. We will create features for our analysis. These include agent type, repository stars, and repository activity. We will also use commit-level tables and user information to detect human and agent commits inside each pull request.

RQ1 (adoption patterns). We will group repositories by star counts. For example, low, medium, and high stars. We will count how many AI pull requests each agent makes in each group. We will also look at changes over time, such as by month. We will show the results with line charts and bar charts.

RQ2 (characteristics and outcomes). For each pull request, we will measure its size. We use lines added, lines deleted, and files changed. We will label file types using file paths. For example, code, tests, docs, or config files. We will record review comments, merge outcome, and time-to-merge. We will then group pull requests into simple size buckets (for example, small, medium, large) and by main file type. For each group, we will summarize merge rate, median time-to-merge, and average number of review comments. We will compare these groups using tables, bar charts, and box plots. If time allows, we may run a very simple logistic regression as an exploratory check, but our main focus will be on data wrangling and descriptive analysis.

RQ3 (collaboration patterns). For RQ3, we will use commit-level tables joined with the user table to detect human and agent commits inside each pull request. In AIDev, every pull request is opened by an AI agent, so we study teamwork only through follow-up commits. For each pull request, we will count human and agent commits after the first agent commit and assign a simple pattern label, such as “agent-only” (no human commits) or “agent-then-human” (at least one human follow-up commit). For each pattern, we will summarize merge rate, median time-to-merge, and average review comments and compare the patterns with tables and a few bar or box plots.

Authors’ Contact Information: Haozhong Ji, jhzzs@student.ubc.ca, University of British Columbia Okanagan, Kelowna, Canada; Yin-Wen Tsai, ywtsai@student.ubc.ca, University of British Columbia Okanagan, Kelowna, Canada.