Yusen Rong, Yue Wang

Project Proposal – AIDev Dataset

Our project investigates how autonomous coding agents behave within real GitHub workflows by comparing their pull-request outcomes to those of human developers. We focus on three core research questions that require combining multiple dataset features and allow for deeper analytical work rather than simple descriptive summaries.

The first research question examines whether language of the repo affects the adoption and frequency of the usage of AI agents. This requires comparing several databases by repo ID to determine not only language frequency but also compare it to agent use frequency.

The second research question explores how change complexity interacts with author identity to affect merge probability. By linking additions, deletions, and files changed with merge outcomes, we can assess whether agents tend to submit smaller or simpler patches and whether complexity impacts them differently than human contributors. This allows us to analyze not only the structural differences in the types of changes submitted but also how those changes influence acceptance rates across author types.

The third research question analyzes whether agent pull requests attract different levels of human oversight, such as number of reviewers, total review comments, and time to first review, and how these review patterns relate to final merge decisions. This combines collaboration data with merge outcomes to understand whether agent contributions are subjected to additional scrutiny and whether such scrutiny meaningfully predicts PR success.

For RQ1, we will systematically store the frequency of language type for each repo from the repo df, and then parse through the pull request data frame to also store the frequency of ai agents per repo and then compare the results. For RQ 2 and 3, we will load and combine the agent and human pull-request datasets along with the repository table using the HuggingFace `hf://` parquet interface. After unifying both PR sources and creating an explicit author-type variable, we will parse all relevant timestamps, derive merge flags, compute time-to-merge in hours, and join repository metadata to each PR. We will also extract complexity metrics and review-related fields to enable multi-feature analysis. For RQ2, we will analyze distributions of complexity features across author types and fit logistic regression models with author-type and complexity interactions to evaluate their combined effect on merge success. For RQ3, we will summarize review patterns and construct models relating oversight intensity to merge outcomes. Finally, we will extend the analysis by training a Random Forest classifier that incorporates author identity, repository characteristics, complexity, and review signals to predict merge likelihood, using feature importance to understand which factors most strongly influence PR acceptance.

This approach yields a clear, well-structured investigation that satisfies the project requirements: each question combines multiple dataset elements, the methodology demonstrates solid data-wrangling and analytical planning, and the writing remains concise and coherent for a single-page submission.