CS846 Project Proposal

Why Do You Fail Mr. Bot?

Amaan Ahmed Asim Waheed Youssef Souati

1 Introduction

AI has become deeply embedded in modern software engineering. Developers now routinely rely on AI tools for tasks ranging from general problem solving [2] and code completion to even delegating entire issues to autonomous coding agents [3, 4]. However, despite their growing presence, the effectiveness and reliability of these agents remain unclear.

Recent work has shown that autonomous agents, such as Codex, Devin, Copilot, Cursor, and Claude Code, are now responsible for hundreds of thousands of pull requests (PRs) on GitHub [4]. Yet, many of these PRs fail to be merged or are quickly reverted, suggesting that AI teammates still struggle to produce contributions that meet human standards of quality and trust. Understanding *why* these failures occur is essential for improving future generations of AI-assisted development workflows.

In this proposed work, we analyze the primary causes of failed Agentic PRs using the AIDev dataset [4], which captures over 900,000 agent-authored pull requests across more than 100,000 repositories. Our analysis will address the following research questions:

1. Are agent-authored PRs more or less likely to succeed than human-authored PRs?

2. What factors lead to the rejection of agent-authored PRs?

3. How can we identify practices that maximize the success of agent-authored PRs?

2 Prior Work

The reasons behind PR rejection have long been studied in software engineering. Many rejections stem not from technical flaws but from process- or context-related issues. For example, redundancies are common in fork-based development [5], and many PRs are rejected because they duplicate existing changes [6]. Other studies note that rejection factors vary with project size, activity, and community dynamics [1]. Overall, rejection often depends more on *context*, such as timing, coordination, or relevance, than on code quality itself.

Zhang et al. find that the single strongest predictor of PR acceptance is whether the author is a core team member [8], emphasizing the social dimension of PR evaluation. While these insights explain human-authored PRs, it remains unclear whether the same dynamics hold for agentic PRs, where authorship, trust, and accountability differ.

3 Initial Analysis

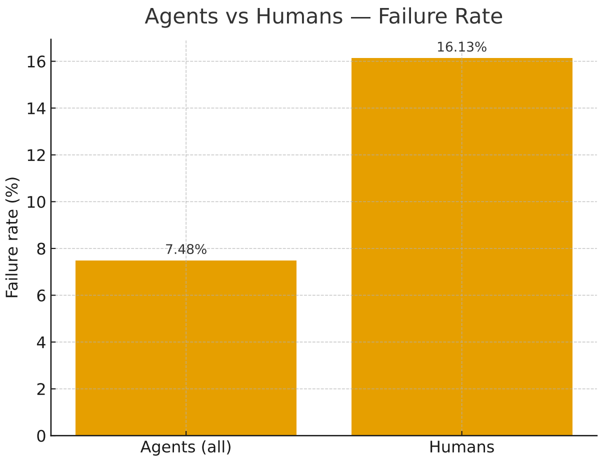

Figure 1: Human-authored PRs are more likely to be rejected than agent-authored PRs based on this dataset.

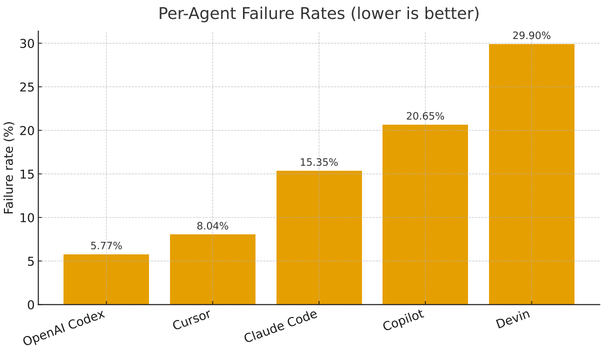

Figure 2: Initial comparison of the failure rates of different agents.

Our initial analysis reveals a surprising trend: agent-authored pull requests fail significantly less often than human-authored ones (Figure 1). Across all agents, the average failure rate is only 7.48%, compared to 16.13% for humans, reversing earlier findings. For example, Wyrich et al. found that in 2021, human pull requests were accepted in 72.53% of cases, while bot-authored ones were merged only 37.38% of the time [7]. Our data suggest that this gap has not only closed but flipped, hinting at a shift that needs to be studied.

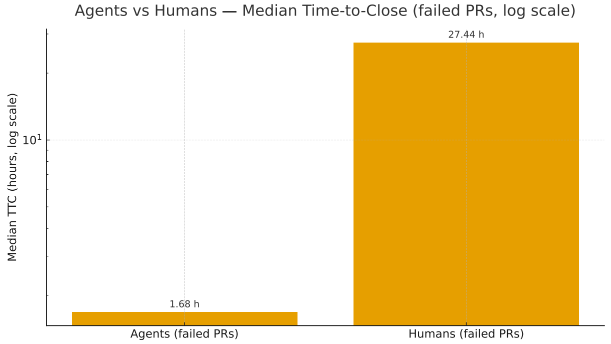

Figure 3: Agent-authored PRs are more likely to be rejected faster than human-authored PRs.

We define a failed PR as one whose state is closed and whose merged_at field is null. We also compute the time-to-close (*TTC*) for failed PRs as the elapsed time between created_at and closed_at, measured in hours. As shown in Figure 3, agent-authored PRs are typically closed much faster than human-authored ones, with median closure times of 1.68 hours versus 27.44 hours, respectively. This sharp difference suggests that reviewers treat agentic PRs differently.

When broken down by agent (Figure 2), we find substantial variation: OpenAI Codex has the lowest failure rate (5.77%), while Devin's is nearly six times higher (29.90%). These disparities imply that not all agents are perform equally. Future analysis will examine whether these differences stem from contextual factors such as project size, PR complexity, or the scope of the change.

4 Research Questions

RQ1: Are agent-authored PRs more or less likely to succeed than human-authored PRs?

To analyze this beyond the single plot in figure 1, we would need to first categorize PRs based on their size, context, and any additional information we can find. We hypothesize that PRs submitted by bots are simpler, require fewer changes, and are initiated by a developer knowing that it would get accepted.

RQ2: What factors lead to the rejection of agent-authored PRs?

We will borrow analysis techniques from the prior work mentioned above, including potential reasons for PR rejection. We hypothesize that most failed PRs would, as shown in prior work, have little to do with the quality of code itself, but other contextual issues.

RQ3: How can we identify practices that maximize the success of agent-authored PRs?

Using the results obtained in the previous research questions, we wish to identify factors that lead to successful PRs. We envision using AI to further analyze this, with the hope of coming up with a process that allows agents to maximize the chances of successful PRs.

5 Threats to Validity

Our analysis is subject to several threats to validity. First, the labeling of agent-authored PRs relies on dataset heuristics (e.g., branch names and bot identifiers), which may misclassify some human contributions. Second, our definition of PR failure (state = closed and merged_at = null) may conflate abandoned with explicitly rejected PRs. However, we will look for additional features that may be used to highlight this difference. Finally, differences in repository activity, popularity, or review policies may influence failure rates independently of agent quality. To circumvent this, we may need additional data on the repositories mentioned in the dataset.

6 Milestones

RQ1: Answering this RQ will form the basis of the rest of our analysis. By November 10, we aim to finalize the descriptive statistics comparing agent- and human-authored PRs, including merge and failure rates, as well as time-to-close distributions. This milestone establishes a baseline understanding of whether agentic contributions behave differently from human ones.

RQ2: By November 20, we will investigate the contextual and structural factors that drive PR rejection. This includes analyzing variables such as repository size, project activity, PR length, and the presence of review comments. The goal is to identify which characteristics correlate most strongly with failed agent-authored PRs.

RQ3 and Final Report: By November 30, we will synthesize our findings to highlight practices that improve the success of agent-authored PRs, supported by examples from the dataset. The final report, integrating results from all three RQs, will be completed and submitted by **December 1**.

References

[1] Tanay Gottigundala, Siriwan Sereesathien, and prefix=da useprefix=true family=Silva, given=Bruno. Qualitatively Analyzing PR Rejection Reasons from Conversations in Open-Source Projects. In *2021 IEEE/ACM 13th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*, pages 109–112. doi: 10.1109/CHASE52884.2021.00021. URL https://ieeexplore.ieee.org/document/9463312.

[2] Huizi Hao, Kazi Amit Hasan, Hong Qin, Marcos Macedo, Yuan Tian, Steven H. H. Ding, and Ahmed E. Hassan. An empirical study on developers' shared conversations with ChatGPT in GitHub pull requests and issues. 29(6):150. ISSN 1573-7616. doi: 10.1007/s10664-024-10540-x. URL https://doi.org/10.1007/s10664-024-10540-x.

[3] Ahmed E. Hassan, Gustavo A. Oliva, Dayi Lin, Boyuan Chen, Zhen Ming, and Jiang. Towards AI-Native Software Engineering (SE 3.0): A Vision and a Challenge Roadmap. URL http://arxiv.org/abs/2410.06107.

[4] Hao Li, Haoxiang Zhang, and Ahmed E. Hassan. The Rise of AI Teammates in Software Engineering (SE) 3.0: How Autonomous Coding Agents Are Reshaping Software Engineering. URL http://arxiv.org/abs/2507.15003.

[5] Luyao Ren, Shurui Zhou, Christian Kästner, and Andrzej Wasowski. Identifying Redundancies in Fork-based Development. In *2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 230–241. doi: 10.1109/SANER.2019.8668023. URL https://ieeexplore.ieee.org/document/8668023.

[6] Qingye Wang, Xin Xia, David Lo, and Shanping Li. Why is my code change abandoned? 110:108–120. ISSN 0950-5849. doi: 10.1016/j.infsof.2019.02.007. URL https://www.sciencedirect.com/science/article/pii/S0950584919300424.

[7] Marvin Wyrich, Raoul Ghit, Tobias Haller, and Christian Müller. Bots Don't Mind Waiting, Do They? Comparing the Interaction With Automatically and Manually Created Pull Requests. In *2021 IEEE/ACM Third International Workshop on Bots in Software Engineering (BotSE)*, pages 6–10. doi: 10.1109/BotSE52550.2021.00009. URL https://ieeexplore.ieee.org/document/9474402.

[8] Xunhui Zhang, Yue Yu, Georgios Gousios, and Ayushi Rastogi. Pull Request Decisions Explained: An Empirical Overview. 49(2):849–871. ISSN 1939-3520. doi: 10.1109/TSE.2022.3165056. URL https://ieeexplore.ieee.org/document/9749844.