WHY DO YOU FAIL ME MR. BOT?

2025/10/28

Amaan Ahmed, Asim Waheed, Youssef Souati

CS846 – Project Proposal

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

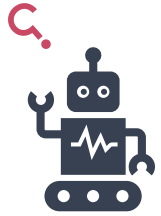
Analyzing Pull-Request Failures in the AIDev Dataset

AI Coding Agents now author hundreds of thousands of PRs

But many PRs fail to be merged or are reverted

Our goal:

Understanding *why*



UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

Why study PR failures?

PR acceptance = natural test of trust in AI teammates

Prior work[1] showed agent PRs accepted far less than humans

Recent data suggests trend reversed

Understanding failures → Improving human-AI workflows

[1] Wyrich, M., Ghit, R., Haller, T., & Müller, C. (2021). Bots Don't Mind Waiting, Do They? Comparing the Interaction With Automatically and Manually Created Pull Requests. *2021 IEEE/ACM Third International Workshop on Bots in Software Engineering (BotSE)*, 6–10. https://doi.org/10.1109/BotSE52550.2021.00009

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

The AIDev Dataset

- 900,000+ agent-authored PRs across 100,000+ repos
- Metadata on PRs, reviews, comments and timelines
- Rich comparisons between agents and humans
- Useful to study PR failure across authors and contexts

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

Research Questions

RQ1: In what contexts are agentic PRs more likely to succeed than human PRs?

RQ2: What factors lead to rejection of agentic PRs?

RQ3: What practices make agentic PRs more successful?

RQ1 (Where) → RQ2 (Why) → RQ3 (How) → Best Practices

EARLY RESULTS

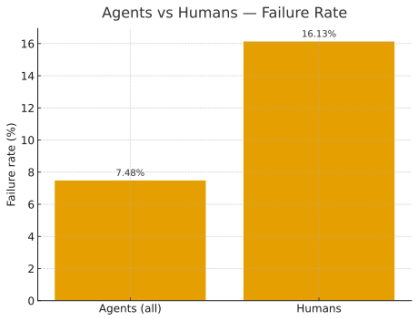
Agentic PRs fail less often

Failure Rate

- Agents: 7.48%
- Humans: 16.13%

Questions:

- Are agentic PRs more relevant?
- Difference in open vs. closed source?
- Is there a difference in code quality?

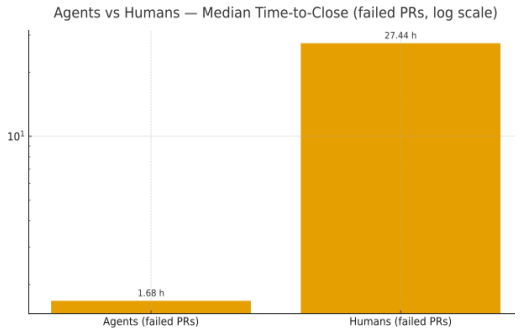


Agentic PRs are Rejected Much Faster

- Failed agentic PRs close in 1.68 hours
  - Humans in 27.4 hours
- AI work is reviewed faster

Questions:

- Bias or something else?
- Pipelines that use agents more efficient?

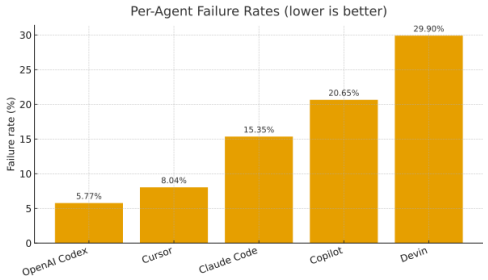


Some agents perform better

- Codex has lowest failure rate (5.77%)
- Devin has the highest failure rate (29.9%)

Questions:

- What kind of repos for each agent?
- Reasons of failure for each agent?



Using LLMs to label PR rejection reasons

Natural question:

- Can PR rejection reasons be labelled by an LLM?

Relevance?

- Agent trained to submit a PR
- LLM trained to review PR → helps Agent
- Rejection reasons useful for reviewing

Initial Experiment

Steps:

1. Extract failed PRs
2. Create context files for each PR
3. Curate prompt to LLM

Currently done manually

```
You are labeling failure causes of pull requests.

Context:
# paste contextn here

Task:
1) Assign zero or more labels from this set:
["test_failure","build_setup","style_lint","api_mismatch","logic_bug",
 "flaky_timeout","security_policy","insufficient_context","other"]
2) Provide 1–3 short evidence quotes (exact phrases) from the CI or reviews.
3) Give confidence ∈ {low, medium, high}.

Output EXACTLY this JSON:
{"labels":[...],
 "evidence":["...", "..."],
 "confidence":"low|medium|high"}
```


Initial Experiment

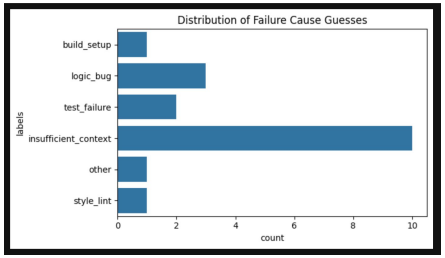
Randomly picked 15 rejected PRs

In general:

- Confidence was low
- Most PRs had insufficient context

Next steps:

- Extract more information for each PR



Next Steps

Further analysis:

- Categorize PRs by size, repo activity, and complexity (RQ1)
- Identify rejection reasons – e.g., redundancy, inactivity, merge conflicts (RQ2)
- Extract success patterns -> guidelines for future agents (RQ3)

What could go wrong? (Threats to Validity)

- Labelling rejection reasons is quite difficult and varied[1,2]
  - Defining labels beforehand leads to bias
- Differentiating between abandoned and rejected PRs
- Repo size or activity may bias results
- In general: controlling for confounding variables

[1] Zhang, X., Yu, Y., Gousios, G., & Rastogi, A. (2023). Pull Request Decisions Explained: An Empirical Overview. *IEEE Transactions on Software Engineering, 49*(2), 849–871. https://doi.org/10.1109/TSE.2022.3160598
[2] Gottigundala, T., Seresathien, S., & da Silva, B. (2021). Qualitatively Analyzing PR Rejection Reasons from Conversations in Open-Source Projects. *2021 IEEE/ACM 13th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*, 109–112. https://doi.org/10.1109/CHASE52884.2021.00021

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

DISCUSSION

Why might agentic PRs appear more successful?

Are faster rejections a sign of bias or efficiency?

How could we fairly evaluate PR quality?

What is the least amount of contextual information we need?

What analyses would ✨YOU✨ like to see?