Unpacking Rejections in AI-Generated Pull Requests

HWIMIN PARK and CHO WING CHAN

ACM Reference Format:
Hwimin Park and Cho Wing Chan. 2025. Unpacking Rejections in AI-Generated Pull Requests. 1, 1 (November 2025), 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Dataset

The AIDev dataset is an open collection of AI-authored pull requests from GitHub, capturing AI-human software collaborations. It features a curated subset of pull requests from popular projects, enriched with code changes, comments, reviews, commits, and issues from tools like GitHub Copilot, Devin, and Claude Code, supporting research on AI adoption, patch quality, and dynamics.

2 Research Questions

(1) **What common failure patterns cause AI-generated PRs to be rejected?**
This research question identifies prevalent failure patterns, such as inaccuracies, insufficient testing, stylistic issues, and security risks, leading to rejection of AI-generated pull requests. To address it, select rejected AI PRs from the dataset. Then categorize reviewer comments into domains like correctness, tests, style, and security. Next, record patch metrics including size, entropy, and path depth. Conduct frequency analysis to pinpoint dominant issues, then apply clustering or association rules to reveal patterns, such as large untested patches or security folder modifications, which commonly drive rejections.

(2) **How consistent are AI-generated PR descriptions with the actual code changes?**
This question evaluates alignment between AI-generated PR descriptions and code modifications, highlighting discrepancies that erode trust. Steps needed include: extracting keywords from descriptions (such as "fix bug," "refactor"), parsing code diffs for operations like file additions, deletions, or test/source distinctions, and lastly compute similarity via a Keyword-Diff Alignment Score. Identifying mismatches, such as bug-fix claims altering only documentation, and validating via reviewer comments on inaccuracies will also need to be done. These findings will inform strategies for improving AI prompt engineering to enhance descriptive accuracy.

(3) **What early signals predict whether an AI-generated PR will be accepted or rejected?**
This research question uncovers submission-time indicators forecasting AI-generated PR outcomes before reviewer input. Methods include compiling initial features, building models like logistic regression, evaluating with metrics, and using SHAP values to rank signals as key predictors. Such insights can guide automated checks or refinements in AI generation pipelines to boost acceptance rates.

Authors' Contact Information: Hwimin Park; Cho Wing Chan.

ACM XXXX-XXXX/2025/11-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn