AIDev: Studying AI Coding Agents on GitHub

Hao Li, Haoxiang Zhang, Ahmed E. Hassan
Queen's University, Canada
{hao.li, haoxiang.zhang}@queensu.ca, ahmed@cs.queensu.ca

***Abstract*—AI coding agents are rapidly transforming software engineering by performing tasks such as feature development, debugging, and testing. Despite their growing impact, the research community lacks a comprehensive dataset capturing how these agents are used in real-world projects. To address this gap, we introduce AIDev, a large-scale dataset focused on agent-authored pull requests (Agentic-PRs) in real-world GitHub repositories. AIDev aggregates 932,791 Agentic-PRs produced by five agents: OpenAI Codex, Devin, GitHub Copilot, Cursor, and Claude Code. These PRs span 116,211 repositories and involve 72,189 developers. In addition, AIDev includes a curated subset of 33,596 Agentic-PRs from 2,807 repositories with over 100 stars, providing further information such as comments, reviews, commits, and related issues. This dataset offers a foundation for future research on AI adoption, developer productivity, and human-AI collaboration in the new era of software engineering.**

I. High-Level Overview

The vision of AI Teammates [1] and recent evidence of their adoption in practice [2] signal a major transition in software engineering (SE). Coding Agents are increasingly acting as AI Teammates that participate in core development workflows. They now contribute thousands of pull requests (PRs)[1] daily, becoming routine actors in collaborative software development. This shift marks the emergence of SE 3.0 [1], where human-AI collaboration is deeply integrated into real-world projects.

Figure 1 illustrates how a Coding Agent operates within a real GitHub workflow. In this example, the agent (GitHub Copilot) is assigned an issue, generates a code patch, and submits a PR with a detailed description. A human reviewer provides feedback, which the agent addresses in a follow-up commit and reply.[2] This interaction showcases the emerging dynamics of human-AI collaboration in software development, where Coding Agents not only contribute code but also remain engaged in the review process.

To support systematic study of this paradigm shift, we introduce AIDev, a large-scale dataset of Agentic-PRs from real-world GitHub projects. AIDev comprises 932,791 Agentic-PRs authored by five agents: OpenAI Codex, Devin, GitHub Copilot, Cursor, and Claude Code, across 116,211 repositories involving 72,189 developers (dataset cutoff: August 1, 2025). Each PR is linked to its corresponding repository and developer, along with additional metadata. For deeper analysis, we curated a subset of 33,596 Agentic-PRs from 2,807 repositories with more than 100 GitHub stars. This enriched subset provides review comments, commit-level diffs, event timelines, and related issues. AIDev enables research on adoption, quality, review, and risks of Coding Agents.

II. Internal Structure

Table I summarizes the major tables, their sizes, and the kinds of artifacts they contain. The AIDev dataset provides PR-level, repository-level, and developer-level metadata. The curated subset of repositories with more than 100 GitHub stars provides enriched artifacts such as inline review comments, commit-level diffs, linked issues, and full PR timelines. To support further analysis, the dataset also includes automated

[1]We refer to PRs authored by Coding Agents as Agentic-PRs.

[2]Not all Coding Agents currently support addressing review comments.

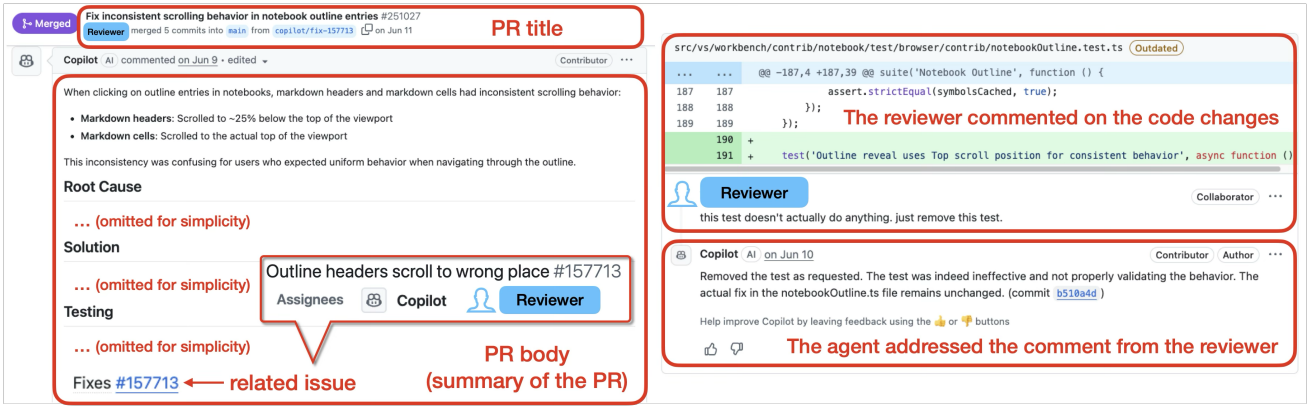

Fig. 1. Example of an Agentic-PR on GitHub. The Coding Agent (GitHub Copilot) authored a pull request, received feedback from a human reviewer, and addressed the comment in a follow-up commit.

TABLE I
OVERVIEW OF THE AIDEV DATASET

	Table	# Records	Content
Core Metadata	all_pull_request	932,791	Pull request metadata (title, body, agent, state, timestamps, repository, user)
	all_repository	116,211	Repository-level metadata (name, license, language, URL, stars, forks)
	all_user	72,189	Developer/user metadata (login, followers, creation date)
The following is a subset of PRs from repositories with more than 100 GitHub stars			
Core Metadata	pull_request	33,596	Same fields as all_pull_request, restricted to curated subset
	repository	2,807	Same fields as all_repository, restricted to curated subset
	user	1,796	Same fields as all_user, restricted to curated subset
Comments & Reviews	pr_comments	39,122	Discussion-style comments on PRs (author, body, timestamp)
	pr_reviews	28,875	Review verdicts (e.g., approve or request changes) with metadata (author, body, timestamp)
	pr_review_comments	19,450	Inline code review comments with file-level context (path, diff hunk, timestamp)
Commits & Diffs	pr_commits	88,576	Commits linked to PRs with metadata (SHA, author, message)
	pr_commit_details	711,923	File-level commit diffs including additions, deletions, and patches
Issues & Events	related_issue	4,923	Mapping between PRs and related issues
	issue	4,614	GitHub issues related to PRs (title, body, state, user, timestamps)
	pr_timeline	325,500	Full PR event history (e.g., committed, closed, merged, labeled, reviewed)
Annotation	pr_task_type	33,596	Automated classification of PR purpose (Conventional Commits categories, GPT-based)

annotations of PR purpose (e.g., bug fix, feature, documentation), following the Conventional Commits categories. The complete relational schema is available in our dataset on Hugging Face and Zenodo (see Section IV).

III. POTENTIAL RESEARCH QUESTIONS

The AIDEV dataset opens avenues for a wide range of research opportunities into the role of Coding Agents in SE. We outline example research questions below:

1) Adoption and Practices:

a) Who adopts Coding Agents on GitHub (e.g., newcomers vs. experienced developers)? How do adoption patterns vary across repositories and ecosystems?

b) What practices (e.g., PR size, task type, and commit granularity) correlate with the quality of Agentic-PRs? How can these practices inform concrete guidelines for developers to work with Agentic-PRs?

2) Code Patch Characteristics:

a) How do Agentic-PRs change code (e.g., additions, deletions, files touched)? How consistent are their descriptions with the actual code changes?

b) To what extent do Agentic-PRs introduce original code versus reusing existing snippets? What are the implications for maintainability?

3) Testing Behavior:

a) How frequently do Coding Agents contribute tests? What types (e.g., unit, integration, end-to-end) are most common? What is the test-to-code churn ratio across ecosystems?

b) When tests are missing in initial Agentic-PRs, do developers intervene to ensure reliable software testing (via follow-up commits or related PRs)?

4) Review Dynamics:

a) What aspects of Agentic-PRs (e.g., correctness, style, security, testing) receive the most attention during review?

b) To what extent do Coding Agents address review comments? Which comment types are challenging for agents to resolve?

5) Failures Patterns and Risks:

a) What common failure patterns and code quality issues appear in Agentic-PRs? Why do they occur? How can we leverage these insights to reduce failure rates, optimize human-AI collaboration, and improve AI model training that prioritizes learning from mistakes?

b) How well can early signals (e.g., PR description, touched paths, and patch characteristics) predict Agentic-PRs rejection or review effort?

c) How frequently do Agentic-PRs introduce or mitigate security vulnerabilities?

IV. HOW TO ACCESS (LINKS)

The AIDEV dataset is available for download on Hugging Face and Zenodo. On Hugging Face, the dataset can be explored interactively through the "Data Studio" interface, which supports in-browser SQL queries. For reproducibility and ease of use, we also provide example Jupyter notebooks with ready-to-use Google Colab links in our GitHub repository. These notebooks demonstrate how to download, filter, and analyze the dataset. The related links are provided below:

- Hugging Face: https://huggingface.co/datasets/hao-li/AIDev
- Zenodo: https://doi.org/10.5281/zenodo.16919051
- GitHub: https://github.com/SAILResearch/AI_Teammates_in_SE3

REFERENCES

[1] A. E. Hassan, G. A. Oliva, D. Lin, B. Chen, and Z. M. Jiang, "Towards AI-Native Software Engineering (SE 3.0): A Vision and a Challenge Roadmap," 2024.

[2] H. Li, H. Zhang, and A. E. Hassan, "The Rise of AI Teammates in Software Engineering (SE) 3.0: How Autonomous Coding Agents Are Reshaping Software Engineering," 2025.