



AI Dev: Studying AI Coding Agents on GitHub (The Rise of AI Teammates in Software Engineering 3.0)

📢 We're hosting the **MSR 2026 Mining Challenge** (co-located with **ICSE 2026** in **Rio de Janeiro, Brazil**).
Details and submissions:

⚠️⚠️⚠️ <https://2026.msrconf.org/track/msr-2026-mining-challenge> ⚠️⚠️⚠️

arXiv **2507.15003**

GitHub Code

DOI [10.5281/zenodo.16919272](https://doi.org/10.5281/zenodo.16919272)

- **Paper:** <https://arxiv.org/abs/2507.15003>
- **GitHub:** https://github.com/SAILResearch/AI_Teammates_in_SE3
- **Example Notebooks:**

Description	Notebook Link	Open in Colab
Basic usage	load_AIDev.ipynb	Open in Colab
Dataset overview	dataset_overview.ipynb	Open in Colab
Analysis of programming usage	language_usage.ipynb	Open in Colab
PR merge rate and turnaround time	productivity.ipynb	Open in Colab

⚠️ **Update (Oct 28, 2025):** `pr_commit_details` has been updated to include all patches fetched from GitHub API (which does not provide content for large patches). Users must verify and comply with the specific license of each source repository.

⚠️ **Update (Oct 16, 2025):** `pr_review_comments.parquet` does not contain full data points, use `pr_review_comments_v2.parquet` instead.

⚠️ **Update (Aug 10, 2025):** The dataset has been refreshed to include data up to **August 1, 2025**, ensuring our dataset reflects the most recent trends in coding agents.

Overview

AI Dev is a large-scale dataset capturing the emergence of autonomous coding agents (AI teammates) within real-world open-source software engineering. It spans **nearly 1 million pull requests** across

116,000+ repositories, authored by five AI coding agents: **OpenAI Codex, Devin, GitHub Copilot, Cursor, and Claude Code**, and involving **72,000+ human developers**.

You can easily load the dataset by four lines of code:

```
import pandas as pd
all_pr_df = pd.read_parquet("hf://datasets/hao-
li/AIDev/all_pull_request.parquet")
all_repo_df = pd.read_parquet("hf://datasets/hao-
li/AIDev/all_repository.parquet")
all_user_df = pd.read_parquet("hf://datasets/hao-
li/AIDev/all_user.parquet")
```

If you're interested in the raw data of AIDev-pop, you can find them here:
https://drive.google.com/file/d/1l0_RjS7ZT0Y27V3mv0oJK-jfeRkhq5l5/view?usp=drive_link

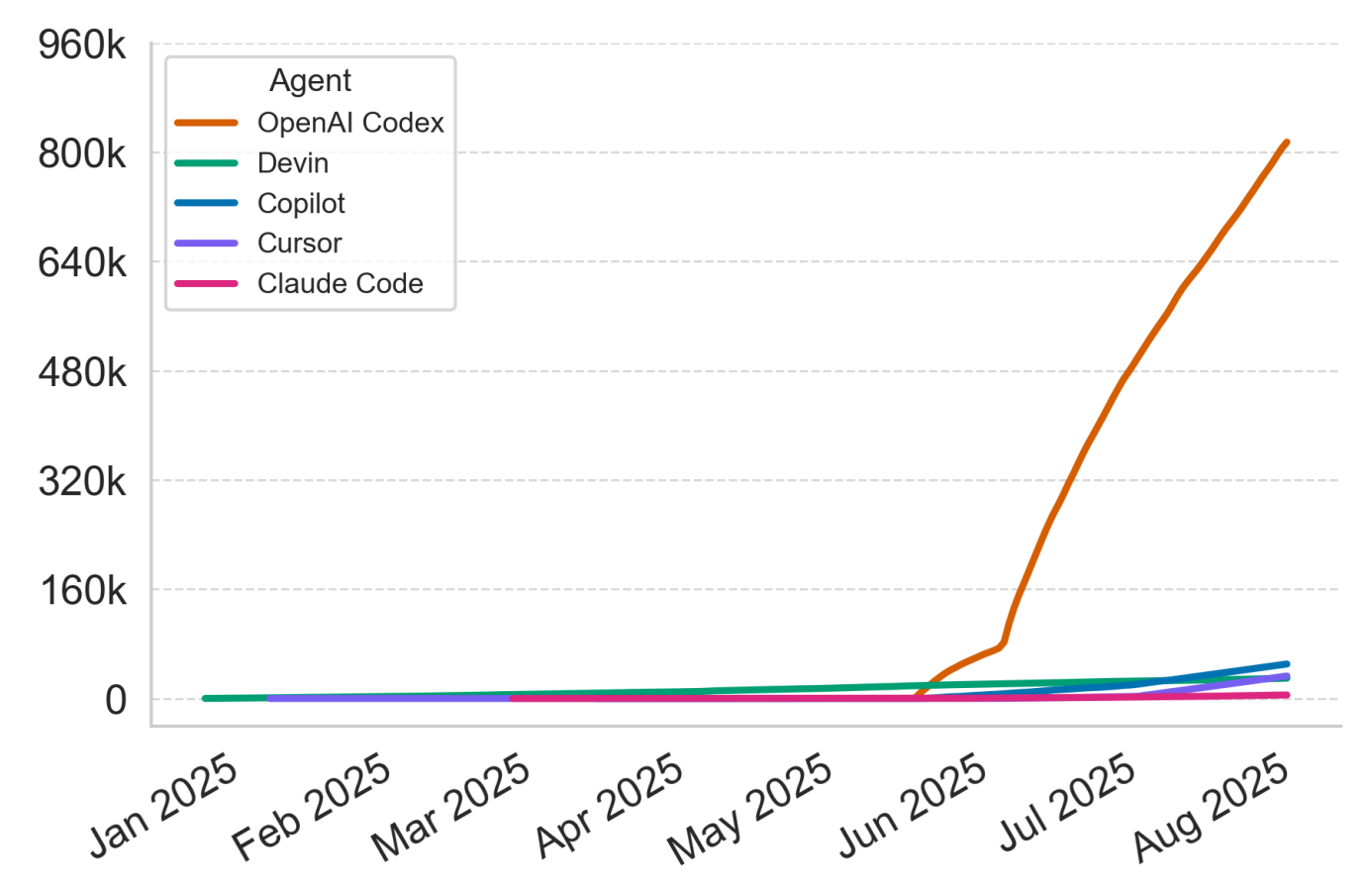
Intended Uses

- **Fine-tuning or post-training:** fine-tuning or post-training your LLMs/agents based on the patches
- **Empirical SE research:** analyse collaboration patterns, review latency, velocity
- **Agent evaluation:** measure bug-fix success, code quality, PR acceptance rate
- **Human-AI interaction:** study conversational review dynamics and sentiment

Quick Look

The overview of the AIDev dataset is as follows:

	#PR	#Developer	#Repo
OpenAI Codex	814,522	61,653	84,704
Devin	29,744	NA	4,747
GitHub Copilot	50,447	NA	14,492
Cursor	32,941	9,658	12,699
Claude Code	5,137	1,643	1,915
Total	932,791	72,189	116,211



Dataset Structure



A detailed explanation about the fields of the tables can be found in [data_table.md](#).

AIDev is organized into normalized tables (available as CSVs) that can be joined via consistent keys. The core components include:

- **all_pull_request**: PR-level data (ID, title, body, agent label, user info, state, timestamps)
- **all_repository**: Metadata including license, language, stars, forks, and project-level info
- **all_user**: User information such as id, login, and created date (personally information has been removed to address privacy concerns)

AIDev-pop: Filtered (>100 stars)

	#PR	#Developer	#Repo
OpenAI Codex	21,799	1,284	1,248
Devin	4,827	NA	288
GitHub Copilot	4,970	NA	1,012
Cursor	1,541	363	327

	#PR	#Developer	#Repo
Claude Code	459	236	213
Total	33,596	1,796	2,807

For the AIDev-pop subset (repositories with more than 100 stars) of AIDev, we provide extra tables:

- **pull_request**: PR-level data (ID, title, body, agent label, user info, state, timestamps)
- **repository**: Metadata including license, language, stars, forks, and project-level info
- **pr_timeline**: Complete PR event history (open/close/merge, label, assign, etc.)
- **pr_comments & pr_reviews & pr_review_comments_v2**: Review discussions, approvals, timestamps, actors, **pr_review_comments_v2** contains inline review comments
- **pr_commits & pr_commit_details**: Commit metadata, diffs, file-level changes, patch. Note that the **patch** data does not include large patches since the GitHub API does not provide them. If you want the large patches, you need to download them yourself.
- **pr_task_type**: Auto-classification of PR purpose using Conventional Commit categories via LLMs
- **issue & related_issue**: Linked GitHub issues and their mapping to PRs
- **user**: User information such as id, login, and created date (personally information has been removed to address privacy concerns)

Human-PR

Human-PRs were sampled from the same repositories as Agentic-PRs, but only from those that have more than 500 stars:

- **human_pull_request**: PR-level data (ID, title, body, agent label, user info, state, timestamps)
- **human_pr_task_type**: Auto-classification of PR purpose using Conventional Commit categories via LLMs

License

This dataset aggregates content from GitHub repositories. **Each source repository retains its original copyright and license** (e.g., MIT, Apache-2.0, GPL family, Creative Commons variants, etc.). Files, patches/diffs, and any other artifacts originating from those repositories remain governed by their **original licenses**.

- Users must verify and comply with the specific license of any source repository or file they extract or use from this collection. Do not assume a universal re-license.
- If you believe content appears here in a way that conflicts with its license, please contact the maintainers, and it will be removed.

Important: Repository contents maintain their original licenses. Please respect individual project licenses when using this data.

Citation

If you use AIDev in your work, please cite:

```
@misc{li2025aiteammates_se3,  
  title={The Rise of AI Teammates in Software Engineering (SE) 3.0: How  
Autonomous Coding Agents Are Reshaping Software Engineering},  
  author={Hao Li and Haoxiang Zhang and Ahmed E. Hassan},  
  year={2025},  
  eprint={2507.15003},  
  archivePrefix={arXiv},  
  primaryClass={cs.SE},  
  url={https://arxiv.org/abs/2507.15003}  
}
```