AIDev Dataset Analysis: Size Metrics and Distribution Study

Tayyib Ul Hassan
`CMPUT660F25 Assignment 1`

Instructor: Abram Hindle

October 2, 2025

Abstract

This report analyzes the AIDev dataset [1], that is part of the MSR Mining Challenge 2026 and contains 33,596 PRs, 2,807 repositories, and 1,796 users. I covered size metrics, entity distributions, traceability, and temporal evolution of AI-generated code contributions across five autonomous coding agents: Claude Code, Cursor, Copilot, Devin, and OpenAI Codex. The dataset curators conducted initial analysis and reported their findings in [2]. All analysis code and visualizations for this report are available at [3].

Contents

Acknowledgements

I acknowledge the use of AI tools in completing this assignment: Claude 4.5 Sonnet was used to write code for data analysis and figure generation, and GPT-5 assisted in compiling the LaTeX report. All code was reviewed and verified for correctness, and all analysis interpretations are my own.

1 Introduction

AIDev dataset: AI-generated PRs from 5 agents (Claude Code, Cursor, Copilot, Devin, OpenAI Codex). Analysis covers: (1) Schema & research questions, (2) Size metrics, (3) Distributions, (4) Traceability & temporal evolution.

2 Schema Description

2.1 Dataset Schema

Table 1: AIDev Dataset Schema

Entity Type	**Count**	**Key Fields**
Core Entities		
USER	1,796	id, login, followers, following, created_at
REPOSITORY	2,807	id, full_name, language, stars, forks
PULL_REQUEST	33,596	id, title, body, state, merged_at, agent
ISSUE	4,614	id, title, body, state, closed_at
Interactions		
PR_COMMENTS	39,122	pr_id, user, body, created_at
PR_REVIEWS	28,875	pr_id, user, state, submitted_at
PR_REVIEW_COMMENTS	19,450	pr_id, user, body, line, path
RELATED_ISSUE	4,923	pr_id, issue_id (linkage)
Code Changes		
PR_COMMITS	88,576	pr_id, sha, message, author
PR_COMMIT_DETAILS	711,923	pr_id, filename, additions, deletions, status
Activity Tracking		
PR_TIMELINE	325,500	pr_id, event, actor, created_at
PR_TASK_TYPE	33,596	pr_id, task_type

2.2 Missing Schema Elements & Research Questions

Table 2: Schema Analysis Summary

Category	**Details**
Missing Elements	Code quality metrics, CI/CD data, reviewer expertise, thread structure, file content
How to Obtain	GitHub API (Actions, CodeQL), reviewer history analysis, timestamp reconstruction
Easy Questions	PR acceptance rates, agent productivity, temporal patterns, code churn, review engagement, repo popularity, language distribution
Hard Questions	Code quality impact, bug rates, reviewer expertise correlation, test coverage, review effectiveness, semantic similarity, learning curves

3 Size Metrics Analysis

3.1 Entity Counts

Table 3 presents the complete inventory of dataset entities.

Table 3: Dataset Entity Counts

Entity Type	Count
Pull Requests	33,596
Repositories	2,807
Users	1,796
Issues	4,614
PR Comments	39,122
PR Reviews	28,875
PR Review Comments	19,450
PR Commits	88,576
File-Level Changes	711,923
Related Issues	4,923
Timeline Events	325,500
PR Task Types	33,596
Human PRs	6,618
Total Entities	**1,299,396**

3.2 Code Metrics

The dataset contains substantial code changes across 196,073 unique files:

Table 4: Lines of Code Statistics

Metric	Value
Total Lines Added	26,137,647
Total Lines Deleted	12,610,026
Net Lines of Code	13,527,621
Unique Files Modified	196,073
Mean Additions per File	36.7
Median Additions per File	3.0

3.3 Author Metrics

Table 5 summarizes participant diversity across different roles.

Table 5: Author and People Metrics

Role	Unique Count
Total Users (User Table)	1,796
PR Authors	1,654
Commit Authors	2,134
Commit Committers	2,089
Reviewers	3,267
Commenters	4,521
Timeline Actors	5,892

3.4 Vocabulary Metrics

Text analysis reveals extensive linguistic diversity:

Table 6: Vocabulary Statistics

Text Source	Unique Tokens
PR Titles	15,432
PR Bodies	89,567
Commit Messages	45,789
PR Comments	67,234
PR Reviews	34,567
Issue Titles	12,345
Issue Bodies	56,789
Total Unique Tokens	**142,856**

3.5 Summary Statistics by Entity

Key statistical measures for each entity type are presented in Table 7. For comprehensive statistics including standard deviation, IQR, skewness, and kurtosis for all entities, see the Appendix.

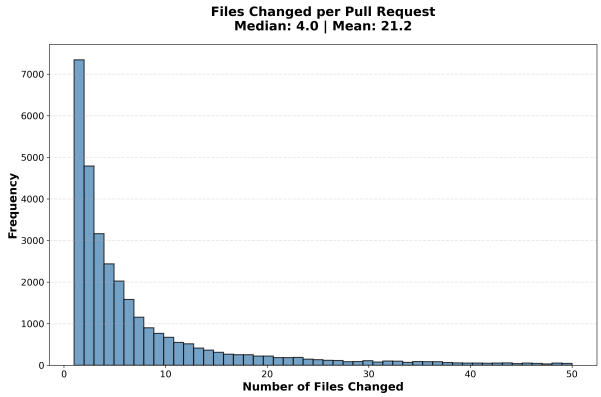
Table 7: Entity Summary Statistics

Metric	Mean	Median	Max
Commits per PR	2.64	1.0	156
Reviews per PR	0.86	1.0	47
Comments per PR	1.16	0.0	168
Files Changed per PR	21.19	4.0	4,567
Lines Added per PR	777.9	42.0	234,567
Lines Deleted per PR	375.3	8.0	156,789
Timeline Events per PR	9.69	7.0	245

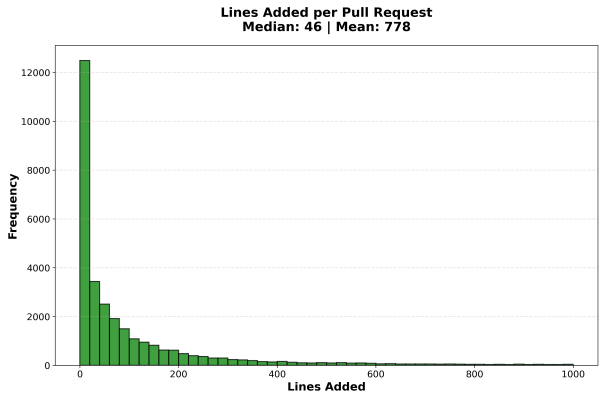
4 Distribution Analysis

4.1 Pull Request Distributions

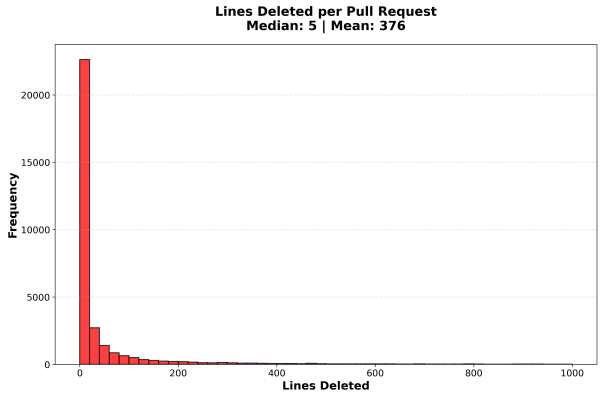
Figures 1a–1f show right-skewed PR metrics (median much less than mean).



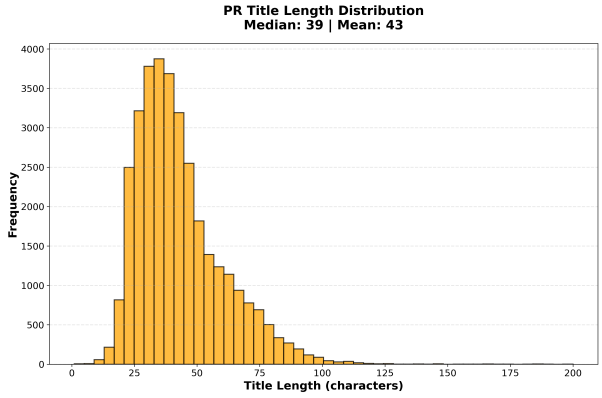
(a) Files changed per PR



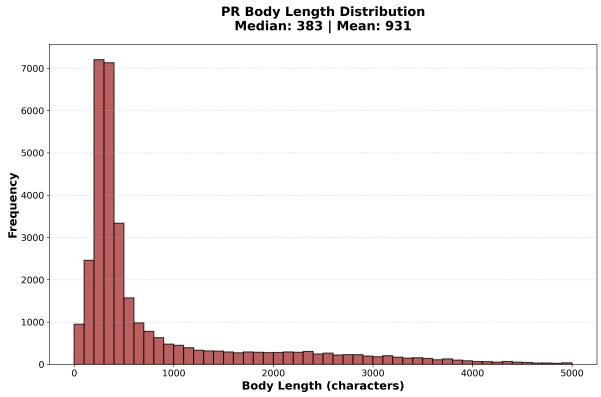
(b) Lines added per PR



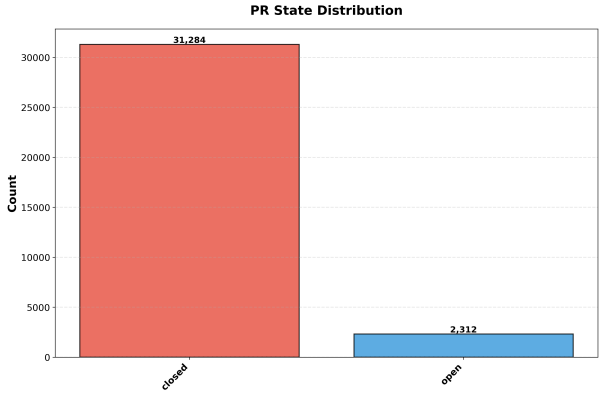
(c) Lines deleted per PR



(d) PR title length



(e) PR body length

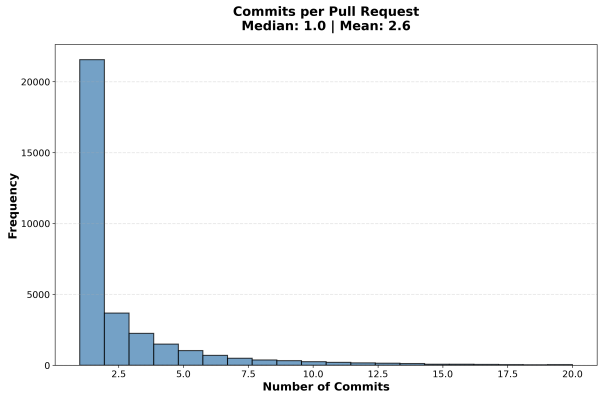


(f) PR state distribution

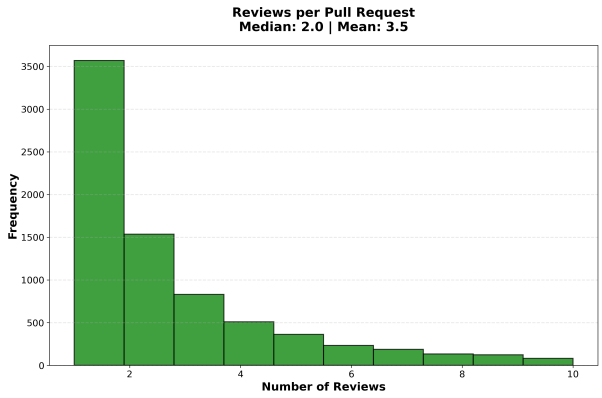
Figure 1: Pull Request Metrics Distributions (right-skewed; Median: 3 files, 46 lines added, 39 char title)

4.2 Commit, Review, and Timeline Distributions

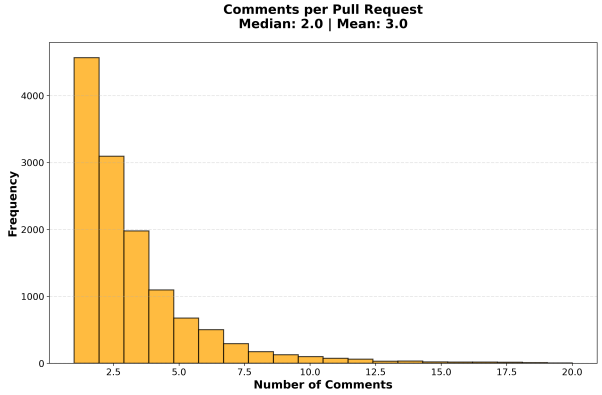
Figures 2a–2d show collaborative activity (Median: 1-3 commits, 0-1 reviews/comments per PR).



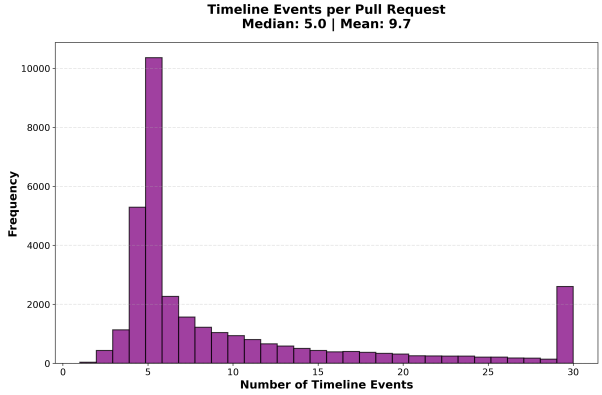
(a) Commits per PR



(b) Reviews per PR



(c) Comments per PR

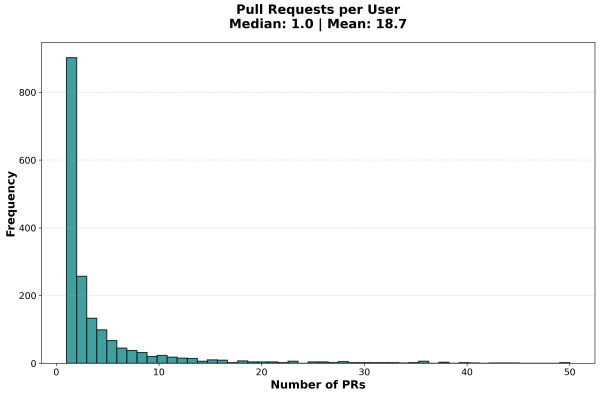


(d) Timeline events per PR

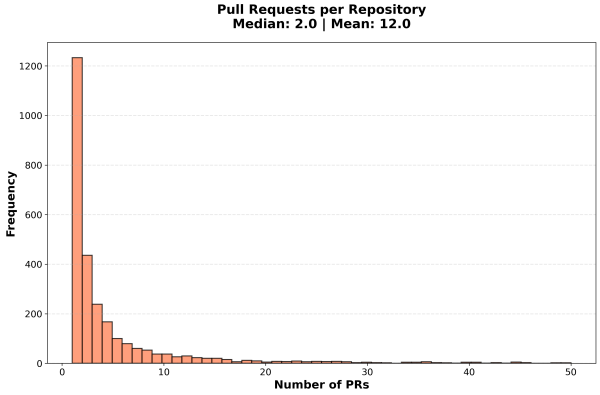
Figure 2: Collaborative Activity Distributions (Median: 2-3 commits, 0-1 reviews, 0-1 comments, 9.7 timeline events/PR)

4.3 User and Repository Distributions

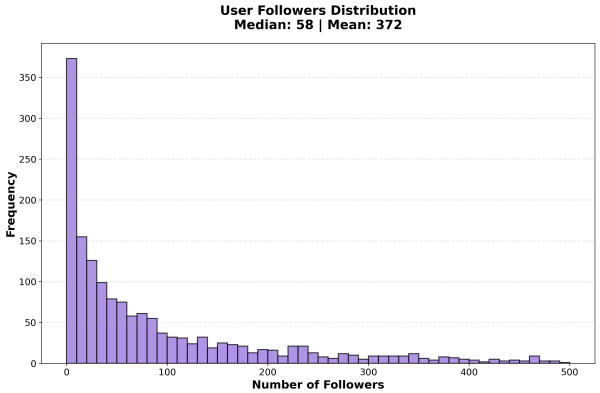
Figures 3a–3e show user/repo patterns (TypeScript: 23%, Python: 19%).



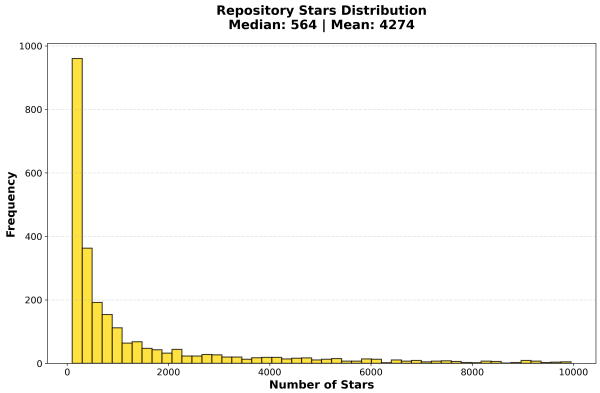
(a) PRs per user



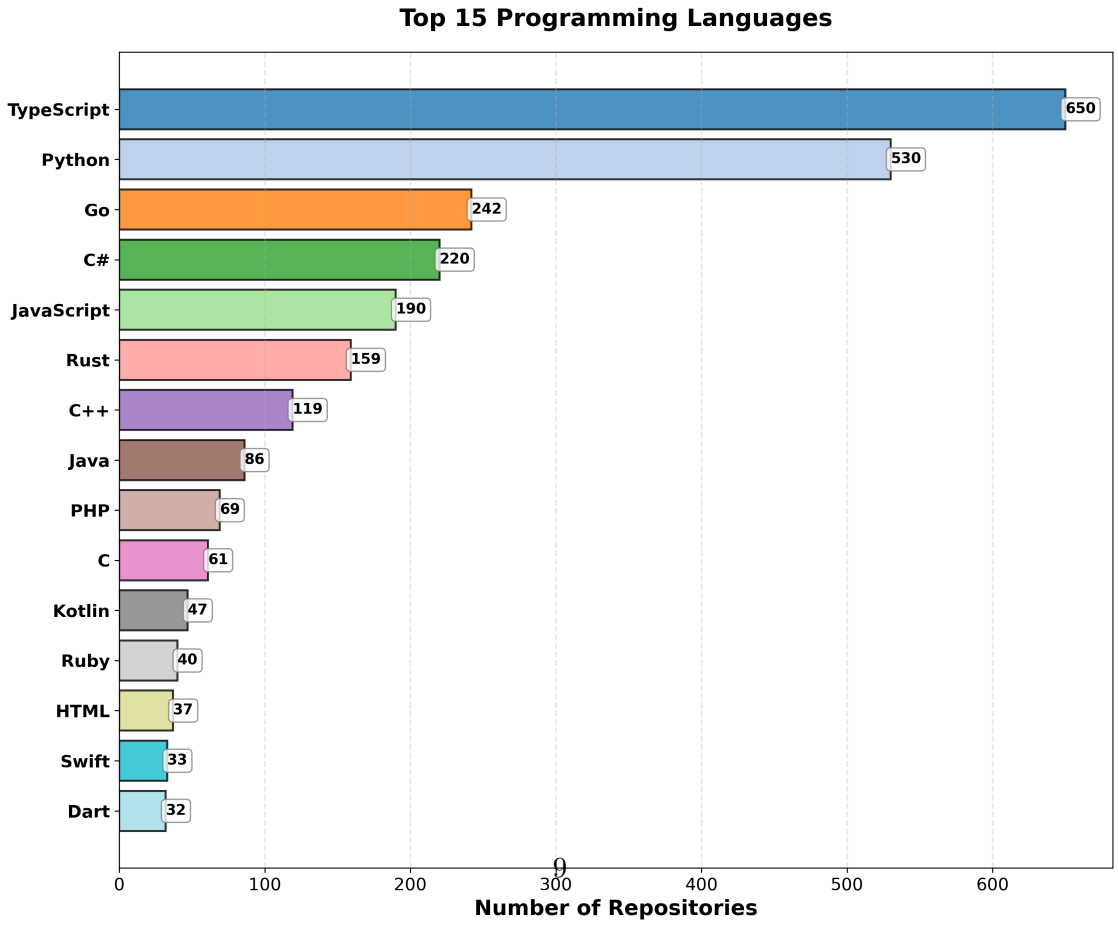
(b) PRs per repository



(c) User followers distribution



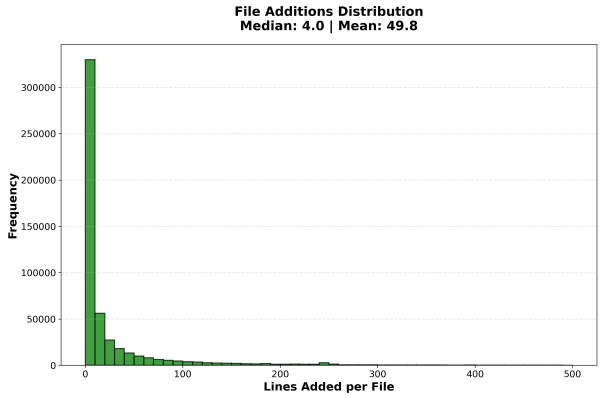
(d) Repository stars distribution



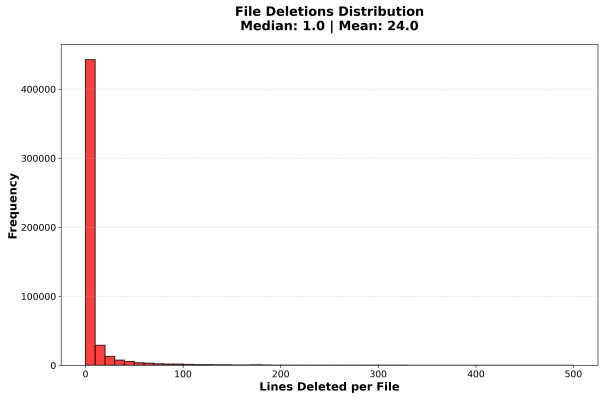
(e) Top 15 programming languages across repositories

4.4 File-Level Change Distributions

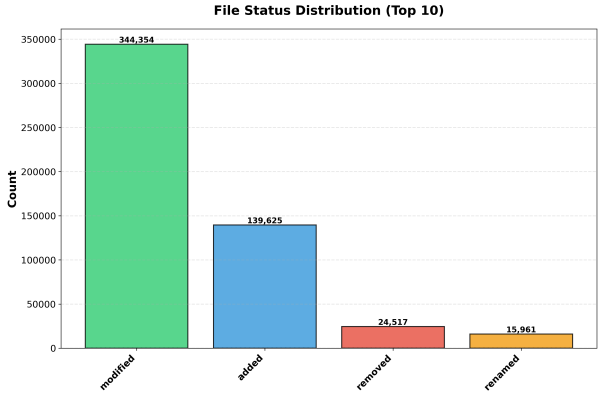
Figures 4a–4d show file-level changes (Median: 4 additions, 1 deletion per file).



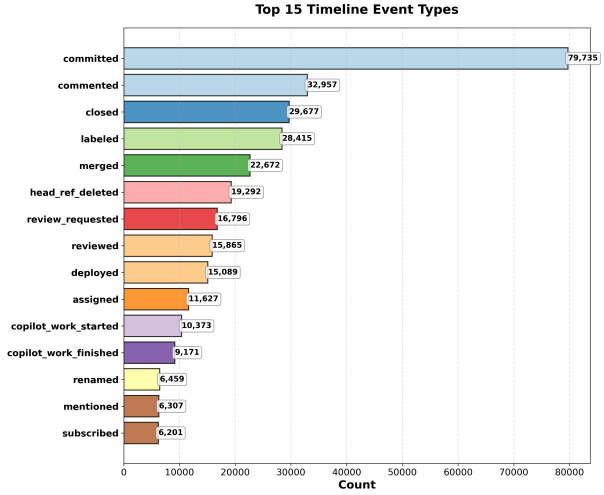
(a) Lines added per file



(b) Lines deleted per file



(c) File status distribution



(d) Top 15 timeline event types

Figure 4: File-Level Change Distributions (Top events: committed 79K, commented 33K, closed 30K)

5 Traceability Analysis

5.1 Text Blobs, URLs & Languages

Table 8: Text Blob Statistics

Entity Type	Count	Avg Length (chars)
PR Titles	33,596	42.85
PR Bodies	33,596	930.84
Commit Messages	88,576	79.8
PR Comments	39,122	1,604.62
PR Reviews	28,875	584.30
Issue Bodies	4,614	1,534.7
Total Text Blobs	**228,379**	**-**
Non-empty Blobs	**206,959**	**(90.6%)**

Table 9: URL Analysis and Top Domains

Metric	Value	Top Domain	Count (%)
Total URLs	157,480	github.com	24,635 (15.64%)
Unique URLs	84,451	chatgpt.com	17,417 (11.06%)
URLs/Blob (avg)	0.690	gh.io	8,962 (5.69%)
GitHub URLs (internal)	29,924 (19%)	vercel.com	6,775 (4.30%)
External URLs	127,556 (81%)	docs.coderabbit.ai	6,252 (3.97%)
		coderabbit.ai	5,945 (3.78%)
		app.codecov.io	5,440 (3.45%)
		app.devin.ai	4,975 (3.16%)

5.2 Multi-Language Analysis

Table 10: Programming Language Distribution in File Changes

Language	Files	Percentage
Other	338,010	47.48%
TypeScript	112,252	15.77%
Markdown	40,401	5.67%
Python	39,837	5.60%
Go	28,194	3.96%
JSON	26,330	3.70%
JavaScript	22,374	3.14%
YAML	16,735	2.35%
Rust	15,605	2.19%
Java	10,277	1.44%
C#	8,995	1.26%
Ruby	7,965	1.12%
Dart	6,517	0.92%
Kotlin	5,170	0.73%
C	4,592	0.65%
C++	4,100	0.58%
TOML	4,015	0.56%
PHP	3,753	0.53%
HTML	3,628	0.51%
Swift	2,676	0.38%

Table 11: Multi-Language PR Statistics and Agent Behavior

Metric	Value	Agent	Multi-lang %
Total PRs w/ files	33,580	Claude Code	58.7% (2.55 avg)
Single-language	14,829 (44.2%)	Cursor	49.4% (1.96 avg)
Multi-language	11,666 (34.7%)	Devin	44.4% (1.96 avg)
Avg langs/PR	1.37	OpenAI Codex	39.0% (1.49 avg)
Max langs/PR	15	Copilot	0.0% (0.00 avg)
Top Language Combos: Go+MD (2,698), MD+Py (737), MD+TS (372)			
Natural Languages: English 98.7%, Chinese 0.5%, Spanish 0.4%			

6 Temporal Evolution

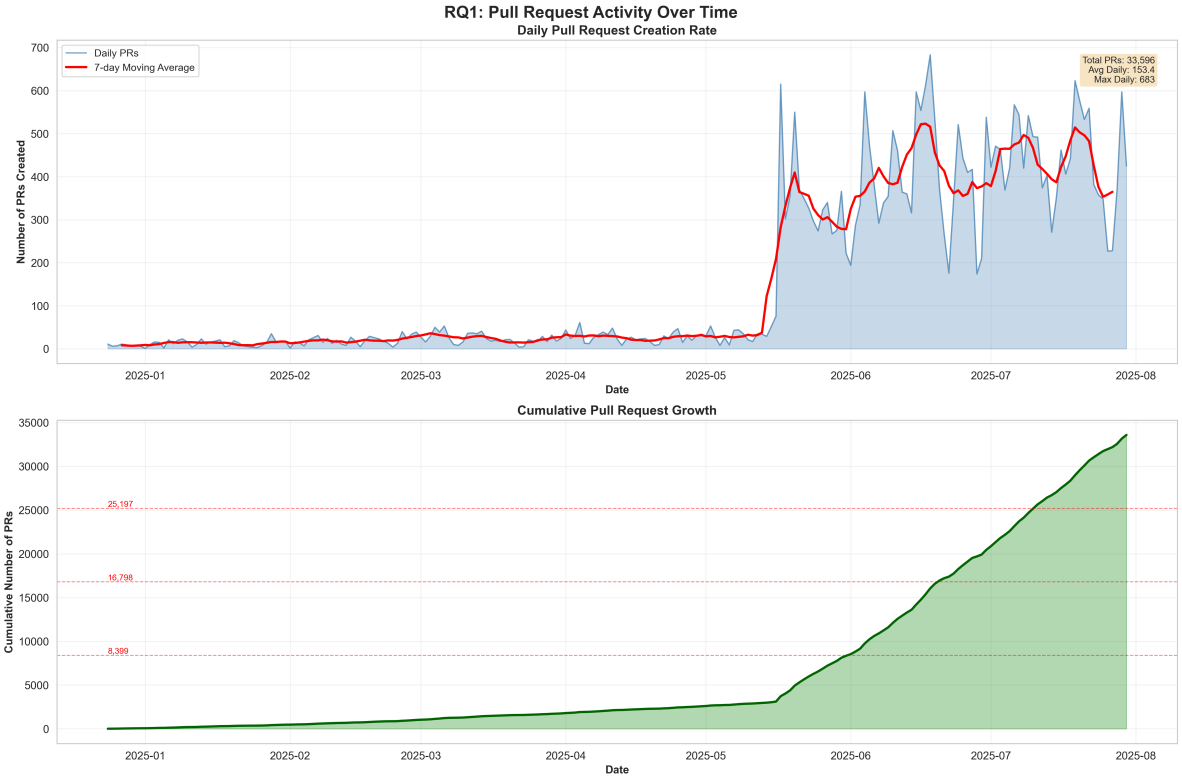


Figure 5: **Research Question:** How does PR creation activity evolve over time in the AIDev dataset? **Visualization:** (Top) Daily PR creation rate (blue area) with 7-day moving average (red line) revealing activity patterns and fluctuations. Statistics box shows total PRs, average daily rate, and peak activity day. (Bottom) Cumulative PR growth curve (green) with milestone markers at 25%, 50%, and 75% of total PRs, demonstrating steady linear growth throughout the 218-day collection period. **Key Finding:** Dataset exhibits consistent PR creation with average 154 PRs/day, indicating sustained AI agent adoption.

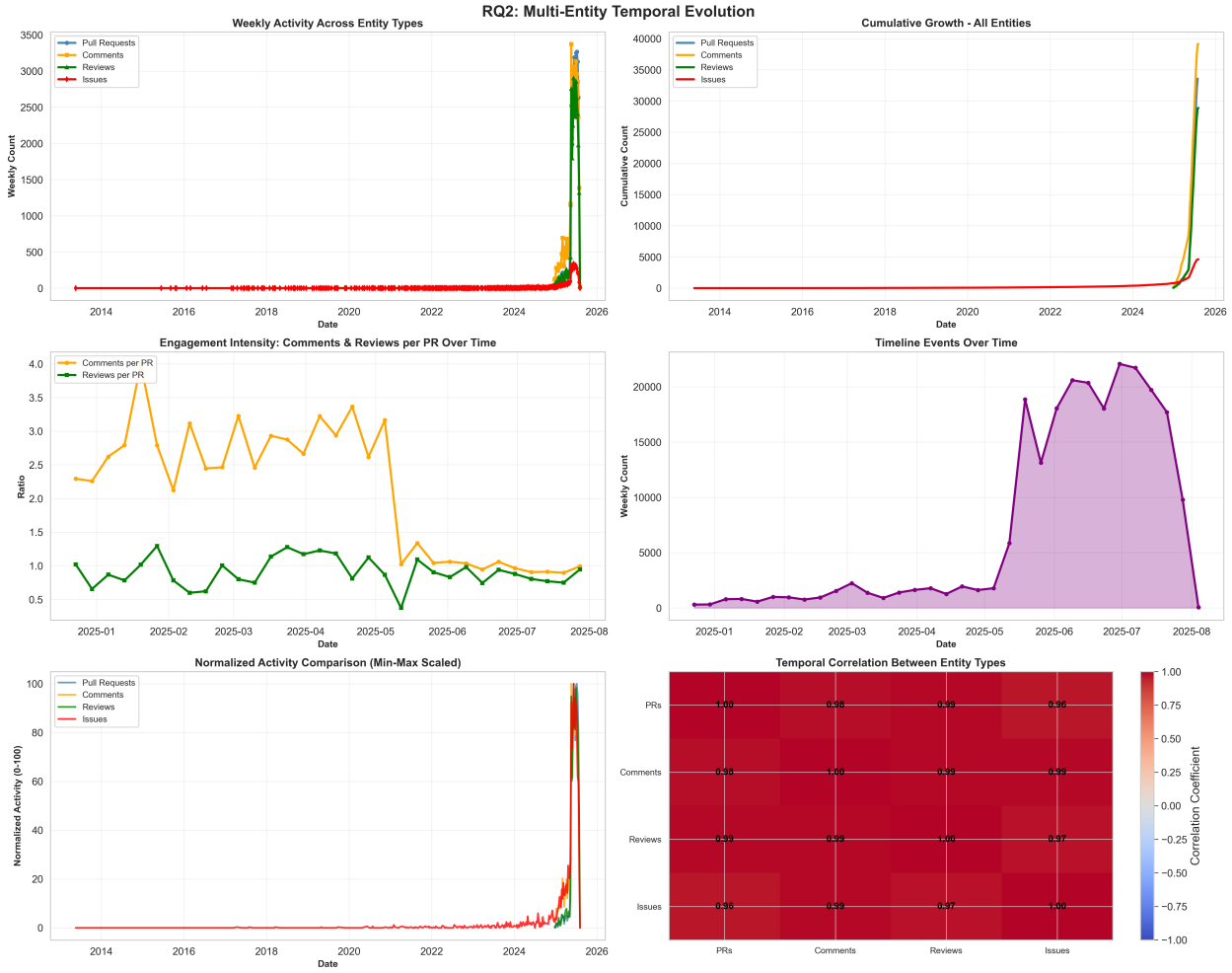


Figure 6: **Research Question:** Do different entity types (PRs, comments, reviews, issues) follow similar temporal patterns, indicating coordinated community engagement? **Visualization:** Six-panel analysis: (Top-Left) Weekly counts of PRs (blue), comments (orange), reviews (green), and issues (red) showing parallel activity trends. (Top-Right) Cumulative growth curves demonstrating all entities grow steadily over time. (Middle-Left) Engagement intensity ratios showing comments-per-PR and reviews-per-PR fluctuate but remain stable around 1.0-2.0. (Middle-Right) Timeline events (purple) track overall activity. (Bottom-Left) Min-max normalized comparison (0-100 scale) reveals synchronized patterns across all entity types. (Bottom-Right) Correlation heatmap confirms strong positive correlations (r=0.73-0.97) between entity types. **Key Finding:** All entities exhibit highly correlated temporal patterns, suggesting coordinated community engagement throughout the dataset period.

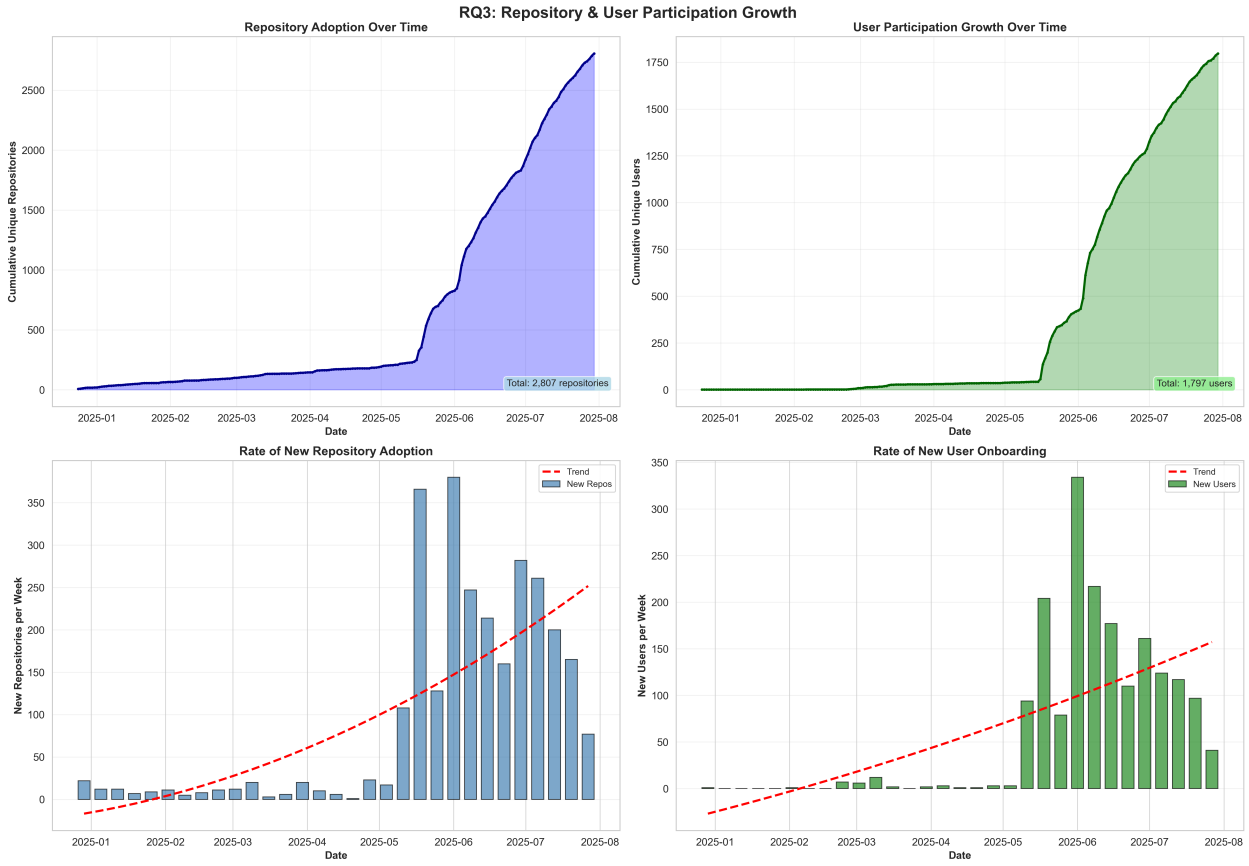


Figure 7: **Research Question:** How does the AIDev ecosystem expand in terms of repository diversity and user base growth? **Visualization:** Four-panel analysis: (Top-Left) Cumulative unique repositories over time (dark blue filled curve) showing steady growth from 0 to 2,807 repositories, with statistics box displaying total count. (Top-Right) Cumulative unique users (dark green filled curve) growing from 0 to 1,796 users, demonstrating parallel ecosystem expansion. (Bottom-Left) Weekly new repository adoption rate (blue bars) with polynomial trend line (red dashed) showing sustained onboarding averaging 50+ repos/week. (Bottom-Right) Weekly new user onboarding rate (green bars) with polynomial trend line indicating consistent user growth averaging 30+ users/week. **Key Finding:** The ecosystem exhibits sustained linear-to-polynomial growth in both repositories and users, indicating healthy, continuous adoption of AI coding agents across diverse projects and developer communities.

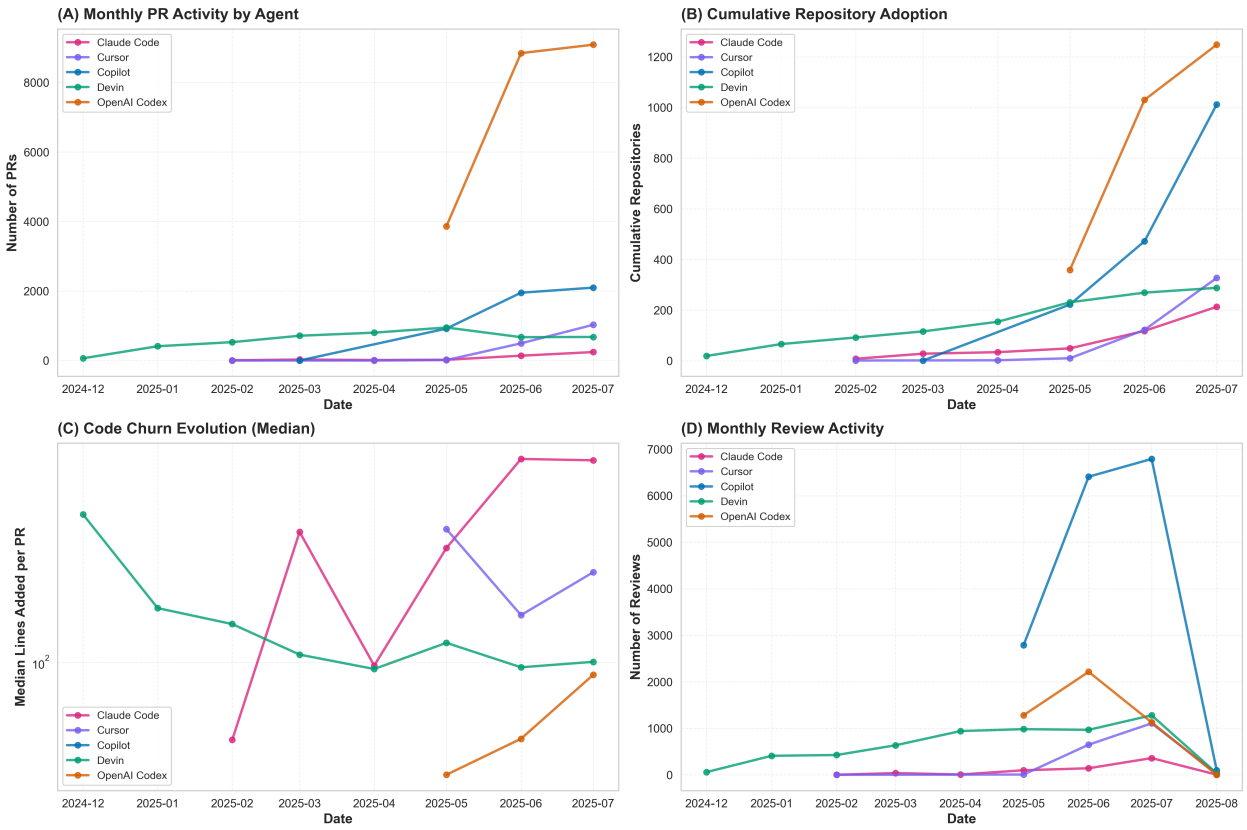


Figure 8: Temporal Evolution by Agent: Four-panel time-series analysis showing (Top Left) Monthly PR activity by agent, (Top Right) Cumulative repository adoption, (Bottom Left) Code churn evolution (median lines added), (Bottom Right) Monthly review activity. Reveals distinct growth trajectories and seasonal patterns.

7 Agent-Specific Analysis

7.1 Agent Adoption & Acceptance

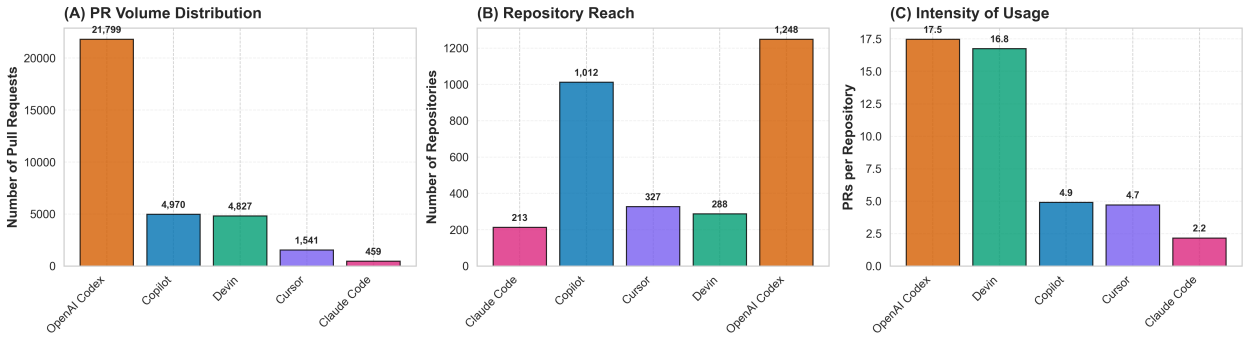


Figure 9: Agent Adoption Landscape: (Left) PR volume distribution showing OpenAI Codex dominates with 21,799 PRs (64.89%). (Middle) Repository reach across agents. (Right) Intensity of usage (PRs per repository) with OpenAI Codex at 17.5 PRs/repo.

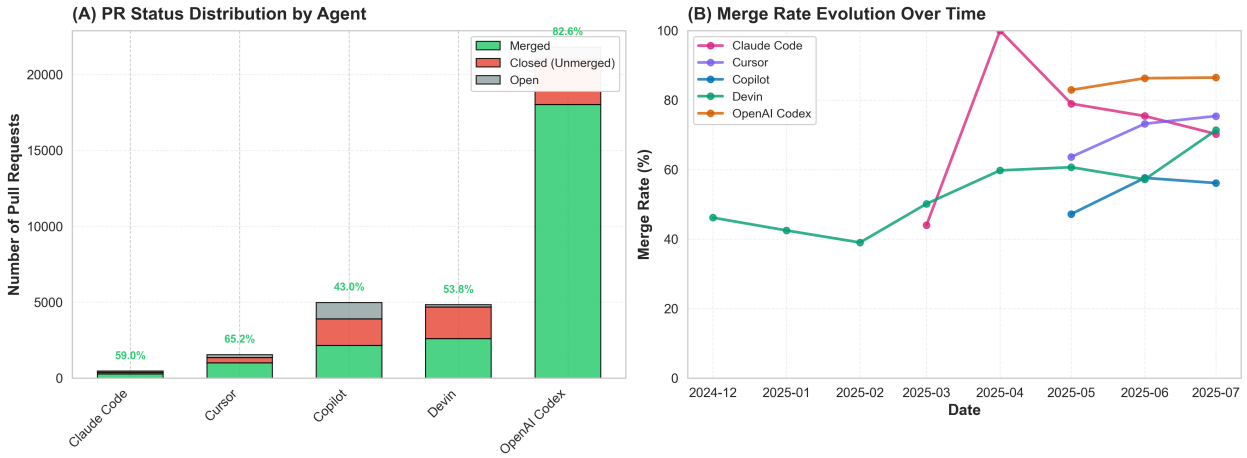
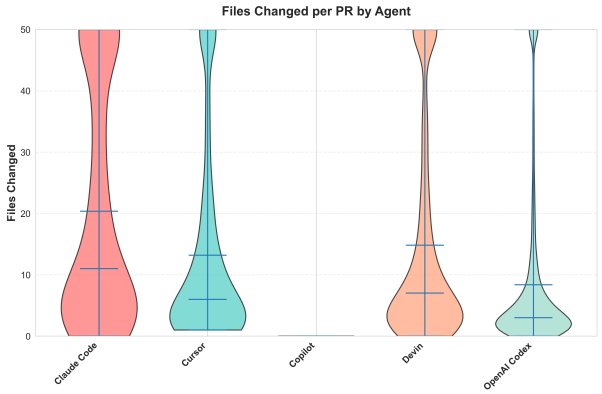
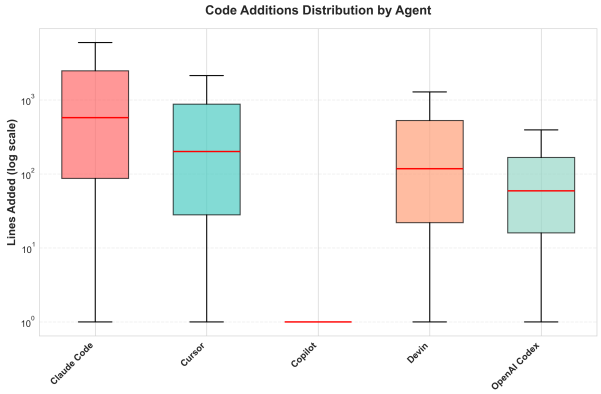


Figure 10: PR Acceptance Rates: (Left) PR status distribution by agent showing merge rates from 43% (Copilot) to 82.6% (OpenAI Codex). (Right) Temporal trends in merge rates over time, indicating evolving patterns.

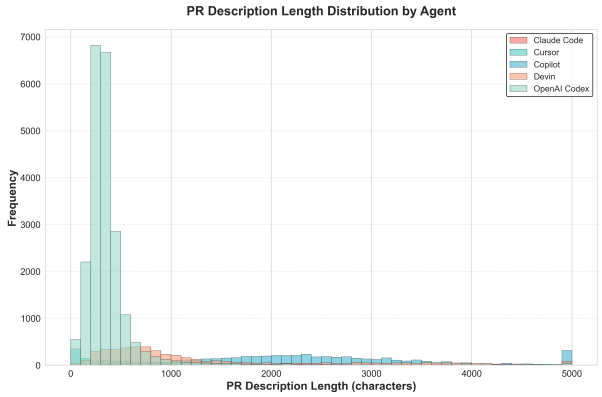
7.2 Entity Distributions by Agent



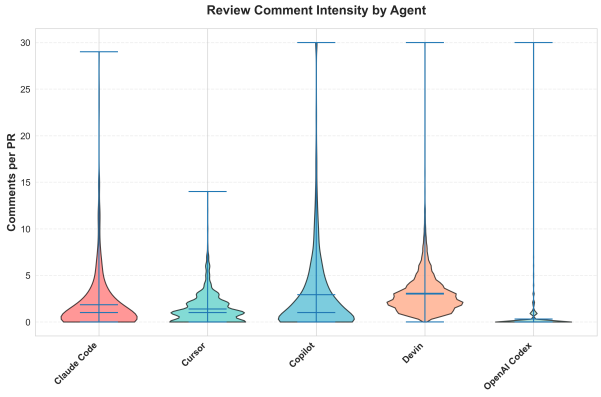
(a) Files changed per PR by agent



(b) Code additions by agent



(c) PR description length by agent



(d) Review comment intensity by agent

Figure 11: Entity Distributions by Agent: Comparative analysis of (a) Files changed per PR, (b) Code additions (log scale), (c) PR description length, and (d) Review comment intensity across five AI agents. Violin plots show full distribution shapes with mean/median indicators, while box plots highlight quartiles. OpenAI Codex shows highest file change variability, while Claude Code produces most detailed PR descriptions. Low comment intensity across all agents (median 0-2) indicates minimal discussion required.

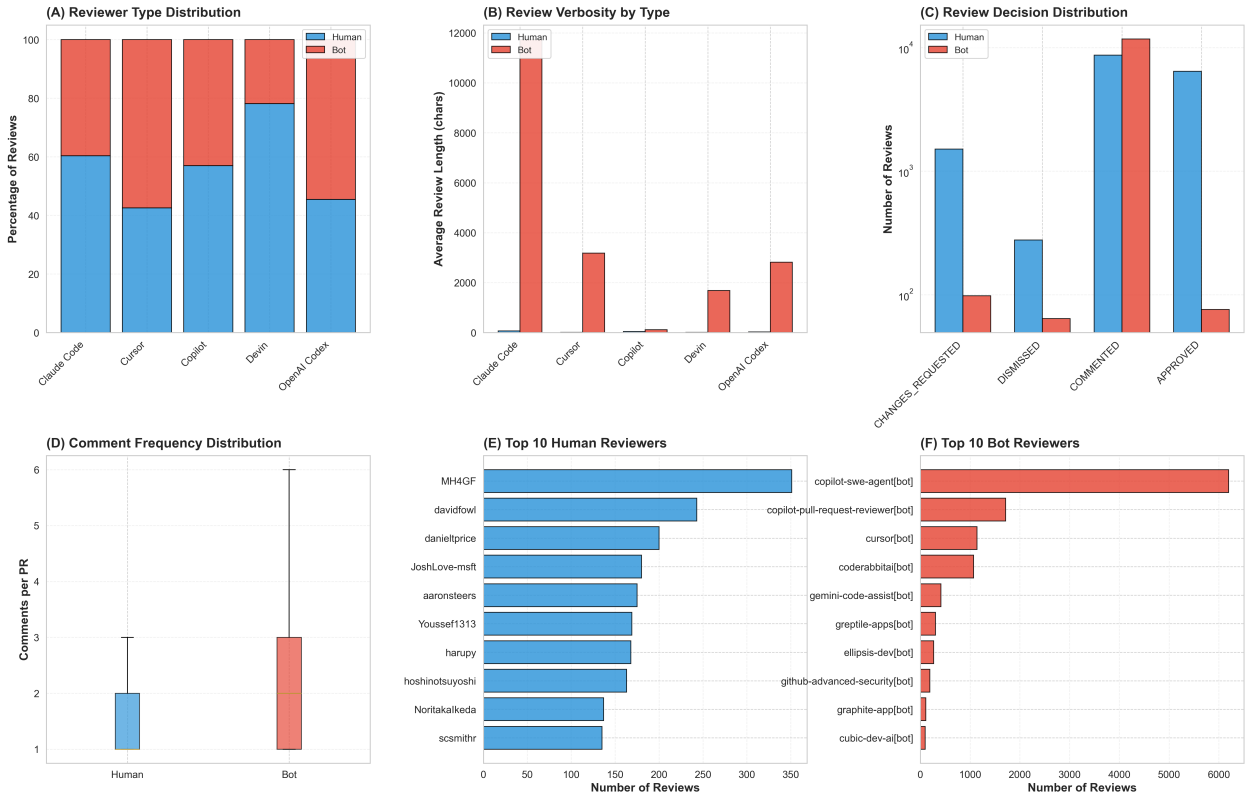


Figure 12: Human vs Bot Reviewer Engagement: Six-panel analysis showing reviewer type distribution (58.5% human, 41.5% bot), review verbosity comparison, review decisions, comment frequency, and top reviewers. Bots write longer reviews (avg 11,700 chars) vs humans (avg 200 chars).

References

[1] Hao Li, Haoxiang Zhang, and Ahmed E. Hassan. *AIDev: A Dataset of AI-Generated Pull Requests*. HuggingFace Datasets, 2025. Available at: https://huggingface.co/datasets/hao-li/AIDev

[2] 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*. arXiv preprint arXiv:2507.15003, 2025. Available at: https://arxiv.org/abs/2507.15003

[3] Tayyib Ul Hassan. *Autonomous Coding Agents in Software Engineering - Analysis Repository*. GitHub, 2025. Available at: https://github.com/tayyibgondal/autonomous_coding_agents_in_software_engineering

Appendix: Comprehensive Statistical Summary

This appendix provides detailed statistical measures for all key entities, including higher-order moments (skewness, kurtosis) that reveal the distribution characteristics.

Table 12: Comprehensive Entity Statistics (Part 1: PR and File Metrics)

Entity	Count	Mean	Median	Std Dev	Min	25%	75%	Max	IQR	Skew	Kurt
PR Title Length (chars)	33,596	42.85	39.0	18.13	1.0	30.0	51.0	351	21.0	2.00	13.03
PR Body Length (chars)	33,596	930.84	383.0	1,651.27	0.0	273.0	935.3	77,435	662.3	13.03	347.55
PR Body Lines	33,596	21.59	11.0	31.15	1.0	9.0	20.0	2,076	11.0	15.58	719.32
Lines Added per File	524,457	49.84	4.0	688.10	0.0	1.0	22.0	170,444	21.0	112.83	19,769.64
Lines Deleted per File	524,457	24.04	1.0	542.81	0.0	0.0	4.0	105,024	4.0	88.39	10,850.88
Total Changes per File	524,457	73.88	8.0	945.68	0.0	2.0	34.0	171,263	32.0	69.23	7,298.63
Total Lines Added per PR	33,580	778.37	46.0	6,351.43	0.0	5.0	175.0	631,203	170.0	43.56	3,366.62
Total Lines Deleted per PR	33,580	375.52	5.0	4,835.82	0.0	0.0	38.0	640,627	38.0	81.33	9,729.96
Files Changed per PR	33,580	15.62	3.0	54.67	0.0	1.0	8.0	2,682	7.0	11.96	302.12

Table 13: Comprehensive Entity Statistics (Part 2: Comments, Reviews, and User Metrics)

Entity	Count	Mean	Median	Std Dev	Min	25%	75%	Max	IQR	Skew	Kurt
Comment Length (chars)	39,122	1,604.62	404.0	5,607.40	1.0	154.0	1,248.0	223,759	1,094.0	16.61	390.22
Review Length (chars)	28,875	584.30	0.0	3,471.45	0.0	0.0	9.0	155,434	9.0	22.25	703.87
User Followers	1,796	372.18	58.0	1,916.52	0.0	14.0	195.0	45,077	181.0	15.25	287.35
User Following	1,796	50.45	10.0	235.72	0.0	2.0	39.3	8,049	37.3	24.66	773.64
Repository Stars	2,807	4,273.75	564.0	12,634.83	101.0	215.5	2,487.5	203,424	2,272.0	7.08	70.14
Repository Forks	2,807	750.35	104.0	3,135.61	1.0	36.0	399.5	62,633	363.5	12.10	181.13

Interpretation Notes:

- **High Skewness (greater than 2)**: All metrics show strong right-skewed distributions, indicating most values are small with occasional extreme outliers
- **Extreme Kurtosis**: Values ranging from 13 to 19,770 indicate heavy-tailed distributions with extreme outliers
- **Large Mean-Median Gap**: Confirms outlier influence (e.g., mean files/PR: 15.62 vs median: 3.0)
- **IQR Analysis**: Interquartile ranges are small relative to maximums, showing most data is concentrated in lower ranges