Early Warning of High-Cost Agentic Pull Requests via Scenario–Cost Modeling

Anonymous Author(s)

Abstract

Using the AIDev dataset (*snapshot/version: [FILL IN]*), we analyze 33,596 agentic PRs and (RQ1) categorize them into interaction scenarios: S0 (solo agent) 32.94%, S1 (human reviewed) 12.52%, and S2 (human co-edited) 54.55%. We then (RQ2) define a composite cost model spanning review intensity, communication, and iteration, and (RQ3) predict high-cost PRs under a fixed alert-budget policy.

CCS Concepts

• **Software and its engineering** → *Software maintenance tools.*

Keywords

Mining Software Repositories, agentic pull requests, code review, cost modeling, early warning

ACM Reference Format:

Anonymous Author(s). 2025. Early Warning of High-Cost Agentic Pull Requests via Scenario–Cost Modeling. In *Proceedings of MSR '26: Proceedings of the 23rd International Conference on Mining Software Repositories (MSR 2026).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Agentic PRs can reduce developer effort, yet maintainers may face increased review burden and coordination overhead. This paper studies whether simple, early-available signals can provide actionable warnings about high-cost PRs, enabling maintainers to triage limited review resources.

2 Dataset and Experimental Setup

Dataset. We use the AIDev dataset (**snapshot/version: [FILL IN EXACT SNAPSHOT, DATE, OR COMMIT]**). Because the dataset is continuously updated, we report the exact snapshot used for all analyses.

Unit of analysis. Our unit is the PR. We analyze PR metadata, review events, and comments as available in AIDev tables (e.g., `pull_request`, `pr_reviews`, `pr_comments`, `pr_review_comments_v2`).

Reproducibility. We will release a replication package including SQL extraction scripts, analysis notebooks, and figure-generation code (**link/DOI: [FILL IN]**).

MSR 2026, Rio de Janeiro, Brazil

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Scenario distribution by agent (percent within each agent).

Agent	S0	S1	S2
Claude_Code	33.12%	11.76%	55.12%
Copilot	28.71%	41.37%	29.92%
Cursor	50.88%	18.23%	30.89%
Devin	39.65%	31.92%	28.42%
OpenAI_Codex	31.14%	1.25%	67.61%

Table 2: High-cost rate by interaction scenario.

Scenario	*n*	High-cost *n*	Rate
S0 (Solo agent)	11065	549	0.0496
S1 (Human reviewed)	4205	2551	0.6067
S2 (Human co-edited)	18326	3619	0.1975

3 RQ1: Interaction Scenarios of Agentic PRs

Goal. Characterize how agentic PRs differ in human involvement.

Scenario definitions. We label each PR into one of three scenarios: (i) **S0 (Solo agent)**: no human comments/reviews/feedback; merging by humans is allowed. (ii) **S1 (Human reviewed)**: human comments/reviews exist, but no human commits. (iii) **S2 (Human co-edited)**: at least one human-authored commit exists.

Results. Across all agents (N=33,596 PRs), S0 accounts for 11,065 PRs (32.94%), S1 for 4,205 PRs (12.52%), and S2 for 18,326 PRs (54.55%). Scenario distributions vary substantially across agents. For example, Copilot has the highest share of S1 (41.37%), while OpenAI_Codex is dominated by S2 (67.61%) with a very small S1 share (1.25%).

Outputs. We produce a scenario-labeled PR table `pr_scenarios_rq1` used in later RQs.

4 RQ2: Cost Model for Review, Communication, and Iteration

Goal. Quantify PR cost along multiple dimensions relevant to maintainers.

Cost components. We define three cost dimensions: (1) **Review intensity**: *review_count* and *request_changes_count*; (2) **Communication cost**: *comment_count* (including review comments); (3) **Iteration cost**: *post_review_review_count* (number of review rounds after the first review; used as an iteration proxy).

High-cost label. We compute a cost score via log-summed components and label *high_cost* as the top 20% of PRs by cost score. In our dataset, 6,719 out of 33,596 PRs are labeled as high-cost (20.00%).

Table 3: Early-warning performance under fixed alert budgets (repo-level split).

Alert budget	k	AUC	Precision@k	Recall@k
Top-10%	1301	0.831	0.694	0.727
Top-20%	2601	0.831	0.409	0.855

Table 4: High-cost rate by agent and scenario (cell shows rate with sample size).

Agent	S0 (Solo)	S1 (Human reviewed)	S2 (Human co-edited)
Claude_Code	0.06 (n=152)	0.33 (n=54)	0.49 (n=253)
Copilot	0.01 (n=1427)	0.64 (n=2056)	0.81 (n=1487)
Cursor	0.16 (n=784)	0.50 (n=281)	0.61 (n=476)
Devin	0.13 (n=1914)	0.65 (n=1541)	0.65 (n=1372)
OpenAI_Codex	0.02 (n=6788)	0.26 (n=273)	0.08 (n=14738)

5 RQ3: Early Warning of High-Cost PRs

Goal. Predict whether an incoming PR will be high-cost using only early-available signals.

Features and split. We train a logistic regression classifier with categorical early signals: *agent*, *scenario label* (RQ1), and PR *state*. We evaluate using a **repository-level split** to reduce within-repo leakage and report AUC.

Budget-based alerting (Top-k). Because the feature space is coarse-grained and yields tied risk scores, probability thresholding can produce unstable alert volumes. We instead use a fixed alert-budget policy: flag only the Top-k highest-risk PRs in each batch (ties broken deterministically by PR id), matching practical review-resource constraints.

Results. The model achieves AUC= 0.831. Under a Top-10% alert budget (k = 1301), it attains Precision@10%= 0.694 and Recall@10%= 0.727 (F1= 0.710). Under a Top-20% budget (k = 2601), it achieves Precision@20%= 0.409 and Recall@20%= 0.855 (F1= 0.553), demonstrating a clear precision–coverage trade-off (Table 3).

Interpretability. Agent-level high-cost rates remain well separated after accounting for uncertainty. For example, Copilot has a high-cost rate of 0.5119 (Wilson 95% CI [0.4980, 0.5258]) and Devin 0.4411 ([0.4271, 0.4551]), substantially higher than OpenAI_Codex 0.0617 ([0.0585, 0.0649]), supporting agent identity and scenario labels as actionable early-warning features.

Agent × scenario heterogeneity. High-cost risk is highly heterogeneous across the intersection of agent and scenario. For example, Copilot exhibits extremely high high-cost rates in S2 (0.81, n=1487) and S1 (0.64, n=2056), but remains very low in S0 (0.01, n=1427). In contrast, OpenAI_Codex shows a moderate high-cost rate in S1 (0.26, n=273) but a much lower rate in S2 (0.08, n=14738), suggesting that the "human-reviewed" workflow is disproportionately associated with high-cost PRs for certain agents. These patterns support scenario label and agent identity as actionable early-warning signals.

Figure 1: High-cost rate by agent with Wilson 95% confidence intervals.

6 Discussion and Implications

Our findings suggest maintainers can control alert noise versus coverage by choosing an alert budget. A Top-10% policy yields high precision (fewer false alarms), while Top-20% captures most high-cost PRs. The strong concentration of high-cost PRs in human-reviewed workflows indicates that collaboration mode and agent identity are meaningful early signals for triage.

7 Ethical Implications

We analyze publicly available repository artifacts and report aggregate results. We avoid releasing any personally identifying information (PII) and do not attempt to deanonymize users. Automated warnings may influence maintainers' attention; therefore, alerts should be used as decision support rather than as automated rejection signals, and should be periodically audited for unintended bias across projects or contributors.

8 Threats to Validity

Construct validity. Our cost model uses observable proxies (review rounds, comments, request-changes) and may not capture all forms of effort (e.g., offline discussion).

Internal validity. Some event logs may be incomplete; we mitigate this by relying on stable tables for review/comment counts and by using repository-level splits.

External validity. Results are specific to the AIDev snapshot and the studied repositories; agent distributions are imbalanced, and small-sample cells should be interpreted cautiously.

9 Conclusion

We propose a scenario–cost framework for agentic PRs and show that early-available categorical signals can provide strong high-cost risk ranking (AUC= 0.831) and practical Top-k early warning under fixed alert budgets. Future work will incorporate richer early PR features (e.g., code-change characteristics) and study downstream outcomes such as acceptance and turnaround time.

References