Author–Reviewer Endpoint Matching

Anonymous Author(s)

Abstract

Some cross-agent code-review pipelines send a patch to a reviewer on a different endpoint and ask that reviewer to refute it. We tested both choices in a controlled 2 × 2 study. Of 120 allocated programming tasks, 113 yielded eligible clean changes and 85 also yielded verified single-fault twins. In the first and only repetition, we crossed endpoint matching with neutral or refutation-first framing across 680 reviews. Six blinded software engineers labeled all 618 scoring-window findings at Stage 1 and all 456 twin findings at Stage 2; a seventh resolver settled disagreements. Detection ranged from 95.3 to 98.8 percent across cells, while false blocking of clean siblings ranged from 1.2 to 5.9 percent. The pooled other-minus-matched effects were −0.5 percentage points for detection and +3.0 for false blocking; the corresponding framing effects were −1.7 and +0.6 points. All pooled intervals included zero. In a post hoc comparison on the same 339 clean-review cells, false blocking was 3.24 percent under human labels, 2.36 under per-item automated adjudicator A, and 6.19 under batched adjudicator C. On 334 common twin-review cells, target detection was 97.60, 95.81, and 97.60 percent, respectively. The comparison was selected after outcome access. The C protocol was recorded before any C output, but the resulting A/C analysis remains post hoc. Because A and C differed in model, client, batching, and shared context, the analysis compares complete adjudication protocols. The study was not preregistered, and the estimates are descriptive.

CCS Concepts

• **Software and its engineering** → **Software verification and validation**; • **Computing methodologies** → *Multi-agent systems.*

Keywords

agentic software development, code review, seeded defects, human adjudication, LLM-as-a-judge, adversarial review

1 Introduction

Code review separates authorship from criticism, but its outcomes depend on reviewer behavior and on the criteria used to sustain a finding [1, 2]. Agentic development systems inherit this measurement problem. A production harness routes an agent-authored change to a human operator only after a reviewer from another model lineage, prompted as a refuter, has tried to break it [4]. Its published examples were selected operational episodes, not a controlled defect-rate study. They left two hypotheses untested: an endpoint may share the blind spots of code it authored, and refutation-first framing may surface defects that neutral review misses.

We isolate those routing choices using 85 verified clean changes paired with single-fault twins, following the traditions of differential testing and mutation testing [7, 12]. Clean siblings supply the false-blocking baseline. We completed target cards before reading or adjudicating any finding and used them to define defect detection. Human raters supplied the primary labels, so neither evaluated endpoint determined its own outcome.

We analyzed 680 reviews from the first and only repetition. The pooled, author-stratified relation effects were −0.5 points on target detection and +3.0 on clean false blocking; the corresponding framing effects were −1.7 and +0.6 points. Every pooled interval included zero, and the author-stratum interaction intervals were wide.

We also asked how the measurement instrument affected the result. On identical clean-review cells, false-blocking rates were 3.24 percent under human labels, 2.36 under per-item adjudicator A, and 6.19 under batched adjudicator C. On common twin-review cells, target detection was 97.60, 95.81, and 97.60 percent.

We report a seeded-twin evaluation that pairs each fault with its clean sibling, records the target in advance, labels every finding with human raters, and accounts for attrition. The analysis covers one review repetition, one endpoint pair, and the 85 verified families that survived construction. We also compare the human labels with two automated adjudication protocols. We did not measure machine-gate yield separately.

2 Study Design

2.1 Endpoints, corpus, and estimand

The two commercial configurations each served as author and reviewer. Endpoint A was `gpt-5.6-luna` through Codex CLI 0.144.0 and endpoint B was `claude-opus-4-8` through Claude Code 2.1.195. *Matched* means reviewer and author endpoint labels are equal; *other* means the second endpoint. This comparison applies only to these two endpoint configurations. It does not isolate a general effect of independence, vendor diversity, model family, or self-review. Within each author stratum, changing the reviewer also changes the entire endpoint setup. Equal weighting gives the A- and B-authored strata equal influence but cannot remove this confounding.

The test bed is a zero-dependency Node.js reference implementation of the harness protocol. No rater contributed to task authorship or study design. Of 120 allocated task slots, 113 produced an eligible clean patch after one initial attempt and at most three repairs. Seeding then produced 85 accepted twins, 44 A-authored and 41 B-authored. Each accepted twin differed from its clean sibling at one designated causal site, compiled, and failed a hidden reproducer that passed on the clean sibling. Clean siblings passed the full visible and hidden suites. Here, clean denotes the seed-free, suite-passing sibling; it does not claim that no latent or out-of-contract defect remained. The protocol allowed a twin to fail a visible test, and one accepted twin did. A model not otherwise used in the study checked every target card for completeness and relevance to the contract. Fifty-five accepted twins were checked once. After noisy single-sample decisions prompted a protocol amendment, we used the majority decision from three fresh calls to `claude-sonnet-5` for each of the other 30. This procedure did not assess whether the seeded faults resembled production defects. During twin construction, we recorded the causal span, violated outcome, and reproducer. We added the trigger and acceptable propagation points after review execution began but before reading or adjudicating any finding,

using only construction materials. Reviewers did not receive the cards, which were fixed before Stage 2. Seeded faults provide controllable ground truth, not a representative distribution of organic defects [8].

2.2 Review protocol and accounting

Repetition one comprises 680 calls: 85 families, clean and twin variants, and four relation-by-framing cells. The reviewer received the task, acceptance contract, visible tests, submitted patch, and frozen file and call-site context. It received no hidden test, target card, variant label, author identity, relation label, or study hypothesis. Neutral framing requested a code review. Refutation framing requested an attempt to refute correctness using concrete acceptance-contract violations. Prompts, wrappers, execution order, and retained-byte caps were hash-pinned. The first five submitted findings were scored.

All 680 review calls completed. Of these, 672 returned valid schemas and produced 618 scored findings. Families remained in the denominator even when a cell yielded no scoreable finding. In a separate automated- adjudication run, the initial prompt omitted its response schema. We discarded all 121 affected calls, corrected the prompt, and reran the schedule. None of the discarded outputs entered the analysis, and those calls are not part of the 680 review calls. We ran no additional review repetitions and collected no separate machine-gate execution ledger, so this study cannot estimate machine-gate yield.

In the post hoc sensitivity analysis, C was used only to adjudicate findings; it never authored or reviewed a patch. It ran `gpt-5.6-terra` with high reasoning effort through Codex CLI 0.144.0-alpha.4, authenticated with a ChatGPT subscription rather than a platform API. Before collecting any C judgments, we fixed a replacement protocol consisting of twelve new, tool-free batches of at most 100 items. Each C item contained only the adjudication materials applicable to that stage. C received no human or prior automated-adjudicator labels from A or B, no aggregate result, no arm or endpoint-condition label, and no information about prior adjudication outcomes. A judged each item in a separate call, whereas C shared context within each batch. The comparison was chosen after the original outcomes were available. A and C differ in model, client revision, call granularity, batch execution, and shared context, so the analysis compares the two complete adjudication protocols.

2.3 Blinded human adjudication

At Stage 1, each finding received independent blinded labels from two of the six raters. Each pair assessed validity, retention of the reviewer's blocking label, and semantic falsifiability without a target card. After those labels were locked, the same six raters assessed target match and localization for all 456 twin findings at Stage 2. Packages hid the endpoint, arm, framing, variant, family, sibling, and other-rater labels. A separate seventh rater resolved categorical disagreements, and that decision became the final label. The resolver received 26 Stage-1 findings and 17 Stage-2 findings. The final human-label dataset is complete.

The seven paid annotators were independent developers recruited from the author's personal network. The group comprised one junior, three mid-level, and three senior engineers, including the senior resolver; experience ranged from 2 to 15 years with median 6, and all had production JavaScript and Node.js experience. Each communicated only one-to-one with the author and did not know the identities of the other annotators. All were paid the same KRW 100,000 per active hour, rounded to the nearest whole hour, including calibration and final coordination. Before data collection, all seven gave written consent to publication of anonymized agreement statistics and work products. No institutional ethics board reviewed the study.

2.4 Outcomes and analysis status

Detection Y is defined on twins. A cell scores one when at least one finding is both Stage-1 valid and Stage-2 target-matching. Two mentions of the target lacked Stage-1 validity and therefore did not count. False blocking G is defined on clean siblings. A cell scores one when the reviewer classified at least one scored finding as blocking, but adjudication retained none of those findings as both valid and blocking.

Contrasts are other-minus-matched, refutation-minus-neutral, and their interaction. Each of 10,000 bootstrap draws resamples complete families with replacement within the 44 A-authored and 41 B-authored strata, retains all four cells, and averages the two stratum estimates 50:50. The interval endpoints are the 251st and 9,750th ordered draws. We used 2026071103 as the random seed for reproduction. The study was not preregistered, and we specified no formal decision threshold, so Y and G are primary descriptive outcomes.

For Human/A/C comparisons, we retained only items or cells with computable values from all three sources; we did not impute values across sources. The reported rates and contrasts are unweighted descriptive margins. We retained C's batch identifiers and did not count judgments from a shared batch as independent model calls.

2.5 Generative AI assistance

Beyond the model calls evaluated in the study, generative AI agents accessed through Codex and Claude Code assisted with design ideation, implementation of study code and experimental materials, analysis and verification code, and manuscript drafting and editing. Agent-generated planning documents were not treated as preregistration. The sole human author is responsible for the submitted study and its claims; no AI system is listed as an author.

3 Results

3.1 Attrition and primary outcomes

Of the 120 allocated slots, 60 belonged to each author endpoint. Seven clean patches failed the eligibility criteria, leaving 57 A-authored and 56 B-authored patches. Seeding rejected another 28, leaving 44 A-authored and 41 B-authored twins. Twelve of the 28 seeding losses occurred on the engine surface, and only six O6 cleanup and concurrency families survived. Our estimates therefore apply to the 85 families that passed every construction check, rather than all 120 allocated tasks.

Table 1 shows near-ceiling detection in every cell and low false blocking. Every pooled interval in Table 2 includes zero. With no

Table 1: Human-adjudicated cell means in percent. Y is target detection on twins and G is false blocking on clean siblings. Each cell contains all 85 verified families.

	Matched		Other	
Outcome	Neutral	Refut.	Neutral	Refut.
Y detection	97.6	97.6	98.8	95.3
G false block	2.4	1.2	3.5	5.9

prespecified equivalence margin, these estimates neither demonstrate equality nor exclude policy-relevant differences. The point estimates ran opposite to both initial hypotheses: relation and framing effects on detection were slightly negative, whereas their false-blocking effects were positive but imprecise. Differences between the author strata also show the confounding introduced by endpoint configuration. The relation estimate for Y was −3.4 points on A-authored changes and +2.4 on B-authored changes; for G, it was 0.0 and +6.1. The wide interaction intervals, extending from −15.9 points for Y to +17.1 for G, preclude an equivalence claim within either stratum.

3.2 Rater agreement and judge sensitivity

Initial-rater raw agreement was 96.3 percent on validity, 96.8 on blocking retention, and 99.4 on falsifiability across 618 Stage-1 findings. Because the rater pair varied by finding, nominal Krippendorff's α was .616, .874, and .597, respectively [10]. Gwet's AC1 was .959, .957, and .993 [3]. Across 456 Stage-2 findings, agreement was 98.0 percent for target match (α = .920, AC1=.974) and 96.3 for three-way localization (α = .870, AC1=.956). Resolution, rather than majority voting or an automated tie-break, produced the final labels.

After correction, A returned usable labels for 615 of 618 Stage-1 findings and 450 of 456 Stage-2 findings. C returned usable labels for all 1,074 findings across twelve batches, without retries or tool calls. The incomplete automated-adjudication run using endpoint B is retained as archival evidence. It returned 616 usable Stage-1 labels and 268 usable Stage-2 labels, leaving 188 Stage-2 items missing. Those items correspond to 188 initial requests and 376 failed retries, for 564 unsuccessful CLI attempts whose service responses reported HTTP 429 and "weekly limit." The run stopped because the service returned a weekly quota limit. The study used no outcome-contingent stopping rule.

Table 3 shows high agreement on both stages, although C departs more from the human labels than A does on the three Stage-1 axes. At the review-cell level, Y and its contrasts are nearly unchanged between human and C labels, while the absolute level and contrasts of G are more sensitive. The G relation contrast remains positive but shrinks under C, and its framing contrast changes sign.

4 Interpretation and Threats

On this corpus, neither routing choice showed a consistent advantage. Detection was already 95–99 percent, which limited possible gains, although the intervals remain compatible with meaningful losses. The estimated increases in clean false blocking were likewise uncertain. These results cover only first-pass review. The experiment did not measure later production-gate stages such as falsifiable-finding requirements, repair and reverification, deterministic checks, or final human approval.

Here, target detection was stable across human and batched-C labels, but false blocking moved from 2.36 percent under per-item A to 6.19 percent under C, with the human rate between them. Such sensitivity is consistent with known LLM-as-judge variation [14]. Automated review evaluations should name the full adjudication protocol, publish missingness, and anchor a subset to independent human labels.

Several limitations narrow the result. The study was not preregistered, and version control is not a substitute for preregistration. We ran only one repetition and chose the Human/A/C comparison after examining the outcomes. Although we fixed C's replacement protocol before collecting its judgments, comparing A with C still changes the model, client revision, call granularity, and shared context together. C also used only seven Stage-1 and five Stage-2 batches, so judgments within a batch may be dependent. The common-cell analyses also inherit selection from A's small number of missing adjudications; the reported margins therefore describe the complete Human/A/C overlap rather than all 340 cells. The code came from a single public repository that the models may have encountered during training. Seeded single faults may not represent organic or interacting faults [8], and attrition left relatively few engine and O6 cases. Within each author stratum, reviewer endpoint is confounded with vendor and capability.

We completed two target-card fields after review execution began but before opening any review output. This preserved outcome blinding but offered less protection than recording the cards before execution under independent access control. Recruiting paid experts from the author's personal network may introduce selection effects and demand characteristics, despite rater-to-rater anonymity, arm blinding, independent labels, and fixed packages. No institutional ethics board reviewed the study.

5 Related Work

Multi-agent review pipelines [13] and critic models [11] report what reviewers find; debate-style setups put opposed agents before a judge [6]. This study instead holds the patch and context fixed, pairs seeded twins with clean siblings, and uses complete human adjudication. Its ground-truth construction follows differential and mutation testing [7, 8, 12]. The routing question inherits intuitions from code inspection [2] and N-version independence [9], while self-correction fragility [5] motivates an external reviewer. Within this corpus, we found no clear advantage from switching from the matched to the other endpoint or from adding refutation framing. The automated sensitivity further shows that reported outcome levels can vary with the adjudicator and its call protocol.

6 Conclusion

Across 680 reviews of 85 verified seeded twins, every pooled interval for endpoint matching and refutation framing included zero for both human-adjudicated target detection and clean false blocking. These estimates apply to the studied corpus and endpoint pair. On common cells, false blocking was 3.24 percent under human labels, 2.36 under per-item A, and 6.19 under batched C, while target

Table 2: Family-paired contrasts in percentage points with central 95 percent family-resampling intervals. Pooled estimates give the A-authored and B-authored strata equal total weight.

Outcome	Contrast	Pooled	A-authored	B-authored
Y	relation	−0.5 [−2.8, 1.9]	−3.4 [−8.0, 0.0]	+2.4 [0.0, 6.1]
Y	framing	−1.7 [−4.5, 0.7]	−3.4 [−8.0, 0.0]	0.0 [−3.7, 3.7]
Y	interaction	−3.4 [−8.3, 1.3]	−6.8 [−15.9, 0.0]	0.0 [−7.3, 7.3]
G	relation	+3.0 [−0.6, 7.4]	0.0 [−3.4, 3.4]	+6.1 [−1.2, 14.6]
G	framing	+0.6 [−2.9, 4.0]	0.0 [−5.7, 4.5]	+1.2 [−2.4, 4.9]
G	interaction	+3.7 [−1.1, 8.6]	0.0 [−6.8, 6.8]	+7.3 [0.0, 17.1]

Table 3: Exploratory Human/A/C sensitivity. Panel A gives percent rates and raw agreement on 615 Stage-1 and 450 Stage-2 findings. Panels B and C give unweighted margins on 339 clean and 334 twin cells.

A. Finding-level labels

Axis	Positive H/A/C	Agreement A–H/C–H/A–C
S1 validity	94.63/95.45/91.06	97.24/94.15/93.01
S1 blocking	85.04/85.20/83.41	97.89/96.75/95.61
S1 falsifiability	99.02/97.56/95.28	97.89/94.63/94.80
S2 target match	85.11/83.11/85.56	97.56/99.56/97.11

B. G on 339 of 340 clean cells

Judge	Count (%)	Relation	Framing
Human	11/339 (3.24)	+2.93	+0.57
A per-item	8/339 (2.36)	+2.35	+1.17
C batched	21/339 (6.19)	+0.55	−0.63

C. Y on 334 of 340 twin cells

Judge	Count (%)	Relation	Framing
Human	326/334 (97.60)	−0.03	−2.34
A per-item	320/334 (95.81)	−0.05	−2.30
C batched	326/334 (97.60)	−0.03	−2.34

detection was 97.60, 95.81, and 97.60 percent. These shifts show that the adjudicator and its call protocol are part of the measurement instrument.

Artifact Availability

The anonymized artifact at https://anonymous.4open.science/r/refute-agenticdev26-artifact/ includes the redacted corpus, clean and twin patches, target cards, review and adjudication records, human-label provenance, analysis code, attrition accounting, and records of protocol changes and execution incidents. It documents the 121 discarded adjudication attempts and preserves the incomplete automated-adjudication ledger for endpoint B, in which 188 missing Stage-2 items correspond to 564 unsuccessful CLI attempts reporting HTTP 429. For endpoint C, the artifact includes the amendment fixed before collecting C's judgments, item and batch schedules, all twelve call records, the complete 1,074-row item ledger, and the resulting analysis. Target cards, seed-site records, and clean–twin diffs identify the seeded edits. Hidden-test sources are withheld, but their SHA-256 commitments are included.

References

[1] Alberto Bacchelli and Christian Bird. 2013. Expectations, Outcomes, and Challenges of Modern Code Review. In *Proceedings of the 35th International Conference on Software Engineering (ICSE)*. 712–721. doi:10.1109/ICSE.2013.6606617

[2] Michael E. Fagan. 1976. Design and Code Inspections to Reduce Errors in Program Development. *IBM Systems Journal* 15, 3 (1976), 182–211. doi:10.1147/sj.153.0182

[3] Kilem Li Gwet. 2008. Computing Inter-Rater Reliability and Its Variance in the Presence of High Agreement. *Brit. J. Math. Statist. Psych.* 61, 1 (2008), 29–48. doi:10.1348/000711006X126600

[4] Harness 2026. Anonymized Production Orchestration Harness. URL withheld for double-anonymous review. Public architecture and anonymized run traces of the harness. An anonymized mirror of the cited pages ships in the artifact bundle.

[5] Jie Huang, Xinyun Chen, Swaroop Mishra, et al. 2024. Large Language Models Cannot Self-Correct Reasoning Yet. In *International Conference on Learning Representations (ICLR)*.

[6] Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI Safety via Debate. arXiv:1805.00899

[7] Yue Jia and Mark Harman. 2011. An Analysis and Survey of the Development of Mutation Testing. *IEEE Transactions on Software Engineering* 37, 5 (2011), 649–678. doi:10.1109/TSE.2010.62

[8] René Just, Darioush Jalali, Laura Inozemtseva, Michael D. Ernst, Reid Holmes, and Gordon Fraser. 2014. Are Mutants a Valid Substitute for Real Faults in Software Testing?. In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE)*. 654–665. doi:10.1145/2635868.2635929

[9] John C. Knight and Nancy G. Leveson. 1986. An Experimental Evaluation of the Assumption of Independence in Multiversion Programming. *IEEE Transactions on Software Engineering* SE-12, 1 (1986), 96–109. doi:10.1109/TSE.1986.6312924

[10] Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology* (4 ed.). SAGE Publications.

[11] Nat McAleese, Rai Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. 2024. LLM Critics Help Catch LLM Bugs. arXiv:2407.00215

[12] William M. McKeeman. 1998. Differential Testing for Software. *Digital Technical Journal* 10, 1 (1998), 100–107.

[13] Xunzhu Tang, Kisub Kim, Yewei Song, et al. 2024. CodeAgent: Autonomous Communicative Agents for Code Review. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://aclanthology.org/2024.emnlp-main.632/

[14] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.