Empirical Software Engineering manuscript No.
(will be inserted by the editor)

Measurability Before Power: Pre-Confirmatory Viability Screening for Sparse LLM Experiments

Thupakula Sai Varun

Abstract Experiments comparing large language model (LLM) interventions on sparse software-engineering tasks can spend most of their budget before establishing that a comparison is measurable. We adapt established multi-endpoint feasibility methodology to sparse LLM author/reviewer experiments, operationalise three viability endpoints, and report a prospectively governed case in which the screen stopped a confirmatory study before its reserved evaluation pool was consumed.

Two development pilots, each with 1,320 model calls and 840 judged outcomes, failed criteria fixed before model calls. Task-level outcomes moved on 4 of 40 and 0 of 40 tasks. A replication-aware movement statistic has a positive, success-rate-dependent reference null even under zero treatment effect; evaluated at each pilot's pooled baseline, observed responsiveness was 0.36 and 0.00 times the pooled-binomial independent-arm reference null. Reviewers used a structural NO ISSUES FOUND channel on 0.0–8.3% of drafts against a prospective 10% trigger, while the planning model was near-separated or failed to fit. The prospective stop rests on the frozen endpoints, not on the reference-null comparison or the post-hoc estimability diagnostic.

External construct-alignment tests and assumption-dependent operating-characteristic simulations are supporting evidence only. The external analysis supplies no positive corroboration after a post-hoc, reviewer-motivated noise null. We contribute an LLM-specific operationalisation of established feasibility methodology and a documented case in which a screen stopped a study before its reserved evidence was spent; we claim no new statistical method, causal result about critique, or external validation.

Keywords Empirical software engineering · Large language models · Feasibility studies · Reproducibility · Software engineering experiments

T. S. Varun
Independent Researcher, Hyderabad, India
E-mail: saivarun1410@gmail.com

1 Introduction

A comparison between two language models on a few hundred tasks looks cheap and is not. The expense is not the arithmetic at the end; it is the generation and judging that must happen first, and which is unrecoverable once spent. Worse, a substantial fraction of such comparisons are not *measurable*: if the intervention does not move task-level outcomes, no sample size rescues the estimate, and a power calculation assuming an effect will size a study for one that cannot be observed.

Clinical research has an apparatus for this. Before a definitive trial, a *feasibility study* asks whether the definitive trial can be run at all, and progression is gated on several endpoints at once, each with a threshold fixed in advance. The apparatus is mature, including the recognition that the progression rule is itself a decision procedure whose false-pass and false-reject rates must be computed rather than assumed (Montgomery et al., 2025).

This paper transfers that apparatus to sparse LLM experiments and reports what happened when we applied it to ourselves. We had designed a confirmatory experiment to decompose the gain from stronger-model code review into information carried by a critique and substitution of a stronger solver. The screen stopped it. Three candidate author/reviewer pairs were closed — two after development pilots, one before any inference. The 248 reserved evaluation tasks were never touched (Figure 1).

A stopped study is rarely reported, so we are precise about what it is worth. It is not evidence about critique, code review, or the models involved. It is evidence that a screening procedure, specified in advance, detected an unmeasurable design and prevented the expenditure — plus a description of the diagnostics that did the detecting, one of which we have not seen operationalised for generative-model experiments.

Contributions.

1. **Adaptation** of established multi-endpoint feasibility and progression methodology (Montgomery et al., 2025) to sparse LLM author/reviewer experiments. *The methodology is not ours; the adaptation is.*
2. **A replication-aware responsiveness diagnostic** (Section 3.1, Figure 2): the task-movement statistic used to judge whether an intervention is measurable has a positive null under finite replication, and must be read against that null rather than against zero.
3. A **prospectively governed case** in which viability criteria stopped a planned confirmatory experiment **before its reserved evaluation pool was consumed**.
4. **Supporting simulations** showing progression-rule operating characteristics are sensitive to plausible outcome-generating assumptions, with no examined rule uniformly dominating. *A statistically sophisticated reader will regard this as an expected consequence of the fact that operating characteristics are functionals of the data-generating process; we agree, and position it as support rather than as a finding.*

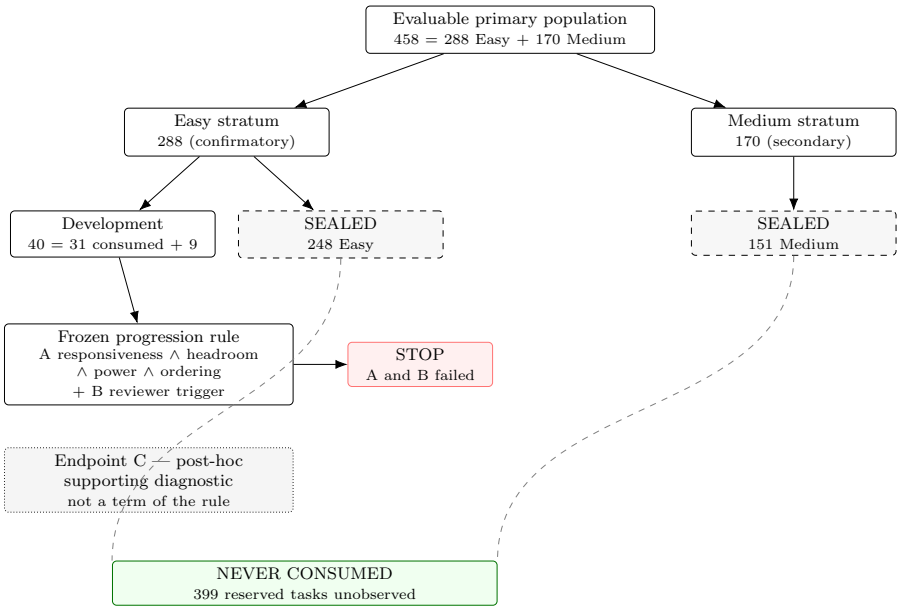

Fig. 1 Viability workflow. The 458-task evaluable population splits into an Easy confirmatory stratum (288 = 40 development + 248 sealed) and a Medium secondary stratum (170, of which 151 remain sealed); 50 tasks had been consumed by earlier screening and pilots. Endpoints A and B are terms of the frozen progression rule; Endpoint C sits off the decision path as a post-hoc supporting diagnostic. Both prospective endpoints failed, so **399** reserved tasks remain unobserved.

5. **Evaluator and provenance procedures**, with an external construct-alignment stress test showing why superficially similar public repair experiments cannot automatically validate the same estimands.

Items 1 and 4 rest on established prior work and we claim no novelty for the underlying ideas.

2 Related Work and the Borrowed Apparatus

Known and adapted, not contributed. Multi-endpoint progression criteria — a feasibility study judged on several endpoints at once, each with a prospectively fixed threshold — and operating-characteristic analysis of the progression rule itself are both established (Montgomery et al., 2025), which also notes that the correspondence between probability-of-success estimates and true power is not obvious. Sparse-data fitting failures of logistic mixed models (separation, non-convergence) are likewise standard. On the substantive side, self-repair in code generation is known to be bottlenecked by the feedback stage rather than the generation stage, with gains that can vanish once repair cost is counted (Olausson et al., 2024); second-pass gains in multi-model pipelines have been decomposed into re-solving, scaffolding and draft content

Table 1 Endpoint definitions.

Endpoint	Construct	Observable statistic	Role	Failure interpretation
A responsiveness	do task-level outcomes move under the intervention?	$\rho = P(d_i \neq 0)$, read against its replication-induced null	support / anti-degeneracy; never treatment evidence	the design cannot measure the target effect, whatever its true value
B reviewer discrimination	does the critic distinguish drafts needing correction from drafts it judges correct?	usage rate of a *structural* NO ISSUES FOUND channel	degeneracy detector	the critic is unconditionally critical; the intervention arm carries no "leave it alone" signal
C estimability	does the planning model fit the pilot data?	convergence, quasi-separation, minimum per-condition event counts	post-hoc in origin; raw diagnostics	a projected power number from a model that did not fit is not a conservative estimate

(Ning et al., 2026); and *systematic overcorrection* in LLM code reviewers is a reported phenomenon whose term we adopt (Jin and Chen, 2026).

Ours. The operationalisation of viability endpoints in this particular sparse LLM author/reviewer setting; the replication-aware responsiveness diagnostic; the prospectively stopped study with its pool intact; the evaluator and provenance practices; and the empirically documented construct-transfer failures of Section 8.

What does not transfer is the *content* of the endpoints. Recruitment, retention and protocol adherence have no analogue here. The failure modes are different in kind: a model may not respond to the intervention at all; a model asked to judge may judge identically every time; and binary outcomes over a few dozen tasks with a few replicates may not support the model used to size the study.

3 LLM-Specific Viability Endpoints

3.1 Endpoint A, and why zero is the wrong reference

This is the element we most want a reader to take away, so we state it slowly.

Let d_i be the task-level difference between the intervention outcome and its paired control for task i, and define $\rho = P(d_i \neq 0)$ — the probability that a task *moves* at all. Because $|d_i| \leq 1$,

$$|E[d]| \;\leq\; \rho, \tag{1}$$

so for a target effect δ^* the condition $\rho \geq \delta^*$ is **necessary**. A 5-percentage-point effect is arithmetically impossible unless at least 5% of tasks move. This makes ρ a **support and anti-degeneracy diagnostic**, and it is never treatment-effect evidence: a high ρ with a mean near zero is a system moving in both directions.

The subtlety is the reference point. Each task's outcome is not observed once but averaged over k sampled generations. Two *identical* systems — a true effect of exactly zero — will still disagree on some replicates by chance, so the observed $\hat{\rho}$ has a **strictly positive expectation under the null** whenever $0 < p < 1$. Treating the two arms as **independent** Binomial$(k, p)/k$ means with a common success probability p, that reference null is

$$\rho_{\text{null}}(p, k) \;=\; 1 - \sum_{j=0}^{k} \left[\binom{k}{j} p^j (1-p)^{k-j} \right]^2 . \tag{2}$$

This is not a function of k alone — it depends on the task success rate p as well. At $k = 3$ it ranges from 0.111 at $p = 0.02$ or 0.98 up to 0.688 at $p = 0.5$, so 0.111 is the *minimum over that whole range*, not a value any particular study should compare itself against. A reference null should therefore be evaluated at the study's own baseline and replicate count (Figure 2) and published alongside the observed value.

Two assumptions, stated rather than buried. This reference null assumes (i) a *homogeneous* success probability across tasks, which we evaluate at the pooled pilot baseline S_{C1}, and (ii) *independent* arms. Neither holds exactly here: task difficulty is plainly heterogeneous, and our downstream arms share a single generated draft, so they are positively correlated. Positive correlation *reduces* the chance of disagreement, so the independent-arm value is an **upper bound** on the true null and the comparison below is conservative in the direction that matters. We therefore call 0.278 and 0.136 a **pooled-binomial independent-arm reference null**, not the exact null of either pilot. A pilot that tests responsiveness against zero is testing against a reference the statistic can never attain, and will read ordinary sampling noise as evidence of responsiveness.

This is not a theorem about LLMs; it is a property of any movement statistic computed over finite replicated draws. We claim only that it matters here and that we have not seen it operationalised for generative-model feasibility screening.

Frozen thresholds. The one-sided 95% Clopper–Pearson (Clopper and Pearson, 1934) lower bound on ρ must reach $\delta^* = 5$ percentage points — the lower bound rather than the point estimate, because the pilot is small and $\hat{\rho}$ is noisy upward. A headroom condition accompanies it: since $G_{\text{info}} \leq 1 - S_{C7}$, the one-sided 95% *upper* bound on the control pass rate must not exceed $1 - \delta^*$.

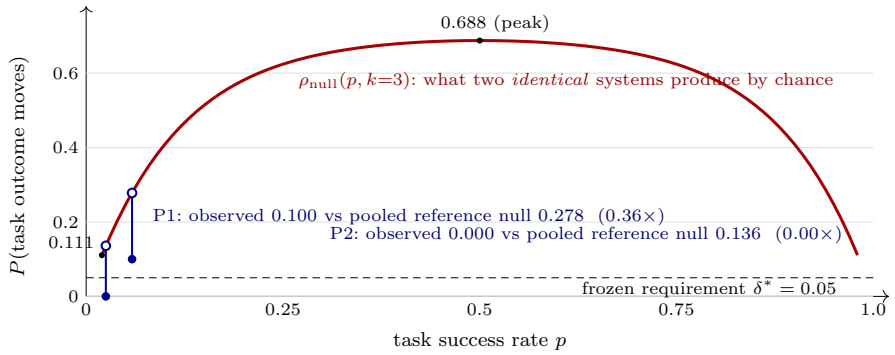

Fig. 2 Endpoint A. The curve is the **pooled-binomial independent-arm reference null** $\rho_{\text{null}}(p, 3)$ — the task-movement rate two independent identical arms produce by chance at $k = 3$, as a function of a common task success rate p. It is lowest (0.111) at the extremes and peaks at 0.688. Each pilot is placed at its pooled baseline: P1 observed 0.100 against a reference null of 0.278, P2 observed 0.000 against 0.136. The reference null is an upper bound under the shared-draft correlation of our design, so these comparisons are conservative. **The stop decision does not depend on them** — it follows from the frozen lower-bound requirement (Section 5).

3.2 Endpoint B — reviewer discrimination

The critique prompt must permit an explicit NO ISSUES FOUND verdict, so the intervention arm can carry a "leave it alone" signal exactly as the control arm can. The endpoint is that channel's usage rate; the frozen degeneracy trigger fires below 10%. We insist on a *structural* channel rather than post-hoc semantic classification of free-form prose, for a reason that generalises: without an explicit channel a non-discriminating reviewer is **invisible** — you observe critiques, and cannot observe the absence of discrimination.

3.3 Endpoint C — planning-model estimability

Fit diagnostics for the mandated random-intercept logistic model. **This endpoint is post-hoc in origin**, formulated after observing fit behaviour in the first pilot, and is marked as such wherever it appears. Where no defensible universal threshold exists we report raw diagnostics rather than a binary verdict.

Endpoint C was not part of the frozen progression rule and did not determine the stop. The frozen rule's viability conjunction comprises author responsiveness, control-arm headroom, projected power and capability ordering; estimability is not one of its terms. Endpoint C is reported as a **post-hoc supporting planning diagnostic**, and the prospective stop decision would have been the same without it.

Table 2 Pilot outcomes and gate decisions.

	S_{C1}	S_{C2}	S_{C3}	S_{C4}	S_{C7}	S_{C8}	G_{info}	Δ_{sub}	Decision
P1	0.0583	0.2917	0.3000	0.1000	0.0583	0.0333	+0.0417	+0.0083	**STOP**
P2	0.0250	0.1333	0.1000	0.0250	0.0250	0.0250	0.0000	−0.0333	**STOP**

4 Prospective Study Design

The intended confirmatory experiment compared a weaker *author* A with a stronger *reviewer* B on contamination-controlled C++ competitive-programming tasks: $C1$ A alone; $C2$ B alone; $C3$ A's draft rewritten by B; $C4$ A's draft, B's critique, A *repairs*; $C7$ A's draft, a task-information-free revision instruction, A repairs; $C8$ A's draft, a critique from a *different* task, A repairs. $C1$ is generated once and shared downstream.

Two estimands were pre-specified: $G_{\text{info}} = S_{C4} - S_{C7}$, the incremental effect of task-relevant critique over a task-information-free control *with the author writing the final solution in both arms*; and $\Delta_{\text{sub}} = S_{C3} - S_{C2}$, routing a draft through the stronger model versus using it as the solver. Design sensitivity was fixed at **5 percentage points** — a design target, *not* a smallest-effect-of-interest — with 80% target power.

Admissibility enforced a contamination cutoff against the author's training date, leaving a frozen **evaluable primary population of 458 tasks — 288 Easy and 170 Medium**. Fifty had already been consumed by earlier screening and pilots (31 Easy, 19 Medium). The **Easy** stratum was designated the confirmatory population and split before any inference into a **40-task development set** (the 31 consumed Easy tasks plus 9 drawn from the 257 untouched Easy tasks under a frozen seed) and the remaining **248 sealed confirmatory tasks**; 288 = 40+248. The **Medium** stratum was redesignated a secondary generalisation boundary and never run confirmatorily, leaving its **151** untouched tasks sealed as well. **399 tasks are therefore reserved and unobserved.** Both reserved pools are sealed by an allow-list guard that denies generation, judging, cache inspection and result merging for any reserved id, and fails closed if its manifests cannot be verified.

5 Development Pilot Results

Two pilots, each **1,320 model calls and 840 judged outcomes** (Table 2). Both are *development evidence only* and are not reinterpretable as confirmatory results.

P1: author `qwen2.5-coder-32b-instruct`, served by a third-party host that does not document serving precision, so replication on raw weights may not reproduce these numbers exactly → reviewer `gpt-5.1-2025-11-13`. P2: author `gpt-4o-mini-2024-07-18` → reviewer `gpt-4.1-mini-2025-04-14`. The reviewer-solo arm exceeds the author baseline in both, so the intended capability ordering held.

Diagnostic	P1	P2	Interpretation
Author baseline S_{C1}	0.0583	0.0250	
Reviewer-solo S_{C2}	0.2917	0.1333	capability ordering held
A task movement $\hat{\rho}$	0.100	0.000	**FAIL** — both
pooled reference null $\rho_{\text{null}}(S_{C1}, 3)$	0.278	0.136	baseline-specific
ratio to that null	0.36×	0.00×	at or below the null
B NO ISSUES rate	0.0% / 8.3%	2.5% / 0.0%	**FAIL** (all $< 10\%$)
issues per draft	5.49–9.03	5.49–9.03	unconditionally critical
C planning fit *(post-hoc)*	near separation	near separation	*not a rule term*
Decision (on A and B)	**STOP**	**STOP**	pool preserved

Fig. 3 Feasibility profiles for both pilots. Endpoints A and B are prospective and carried the stop decision; Endpoint C is a post-hoc supporting diagnostic, shown separated from them.

Endpoint A failed. Task-level outcomes moved on **4 of 40** tasks in P1 and **0 of 40** in P2. At 4/40 the one-sided 95% lower bound on ρ does not reach the frozen 5-point requirement; at 0/40 it is zero by construction. **That frozen lower-bound requirement is what the gate applies, and it is what the stop rests on.** As supporting context — not as a gate term — the pooled-binomial independent-arm reference null of Section 3.1, evaluated at each pilot's pooled baseline, is 0.278 for P1 (baseline 0.0583) and 0.136 for P2 (0.0250), placing the observed values at 0.36× and 0.00× of it. Because that reference null is an upper bound under our shared-draft correlation, the comparison is conservative. **This is a measurability failure, not a finding about critique.** P1's seven discordant replicate observations and their transition counts are recorded as exploratory only; seven observations in a pilot that failed its screen support no inference and we draw none.

Endpoint B failed, independently. Across the four reviewer×pilot cells the structural NO ISSUES FOUND channel was used on 0.0%, 8.3%, 2.5% and 0.0% of drafts — every cell below the frozen 10% trigger — while reviewers flagged 5.49 to 9.03 issues per draft across 480 stored critiques. The reviewers were not silent; they were **unconditionally critical**, which is exactly the non-discrimination this endpoint detects, and is consistent with reported systematic overcorrection (Jin and Chen, 2026). We do not claim the flagged issues were semantically wrong — that would require adjudication we did not perform; the endpoint concerns the *distribution of verdicts*, not their content. The result rests on a single critique prompt, and since prompt elaboration is reported to amplify over-criticism (Jin and Chen, 2026), prompt and model effects are **not separable in our data**.

Endpoint C failed. The mandated planning model fitted P1 near separation ($b_0 \approx -10.52$). Figure 3 summarises both pilots against all three endpoints.

6 Why the Confirmatory Study Was Stopped

Three failures on three axes from two pilots — though only two bear on the prospective decision. **The stop rests on endpoints A and B**; Endpoint C is post-hoc in origin, was not a term of the frozen viability conjunction, and would not have changed the outcome. A and B are moreover *independent*: an author that does not move under the intervention (A) and a reviewer that never abstains (B) are distinct defects with distinct remedies, and either alone makes the planned estimand unmeasurable. That they are separable is a property of *our two systems*, not a general claim — with $n = 2$ we can say the axes were separable in our data and no more.

The rule was fixed before inference and applied as written: the pilots did not progress, and the confirmatory pool was not consumed. A third candidate pair was closed *before* any inference on cost and capability-evidence grounds.

248 confirmatory tasks and 151 harder tasks remain unobserved. Had the screen been skipped, they would have been spent on a confirmatory experiment that did not satisfy the pre-specified viability conditions required to estimate its planned primary contrast. That is the empirical claim of this paper, and it is a claim about a decision procedure, not about language models.

7 Planning and Operating-Characteristic Sensitivity

Supporting analysis, reported separately from the prospective decision criteria.

Having failed viability, we asked a narrower question: *if* viability had held, how would the progression rule have behaved? A two-stage precision protocol was frozen inside the simulation script before it ran — a coarse stage at 30 pilots per family-cell, then a refinement rule selecting cells for a rerun at 60 pilots with 200 outer-bootstrap refits. The rule selected 19 cells; all 19 were run, under three outcome families: a fitted random-intercept family and two *stress families* (a sparse responsive/inert mixture, and random-slope heterogeneity) that are prospectively specified but not validated against data and are not claims about the true data-generating process.

Across the 19 canonical cells, median-rule false-pass spans 0.067–0.983 and the conservative 10th-percentile rule spans 0.000–0.783 (Figure 4); median-rule false-reject spans 0.000–0.383 and the conservative rule 0.000–1.000. The conservative rule bought lower false-pass in **19 of 19** cells and paid for it with strictly higher false-reject in **17 of 19**. In the two remaining cells both false-reject rates are 0.000, so the conservative rule is weakly dominant there — but that rests entirely on a tie at zero, each zero carrying the interval $[0.000, 0.060]$ at 60 pilots, so neither case is distinguishable from Monte-Carlo noise. **No examined rule uniformly dominates.** Rule selection is a bet on which departure from the planning model is most plausible, and a 40-task pilot cannot settle it, because the misspecification that matters is invisible at that size.

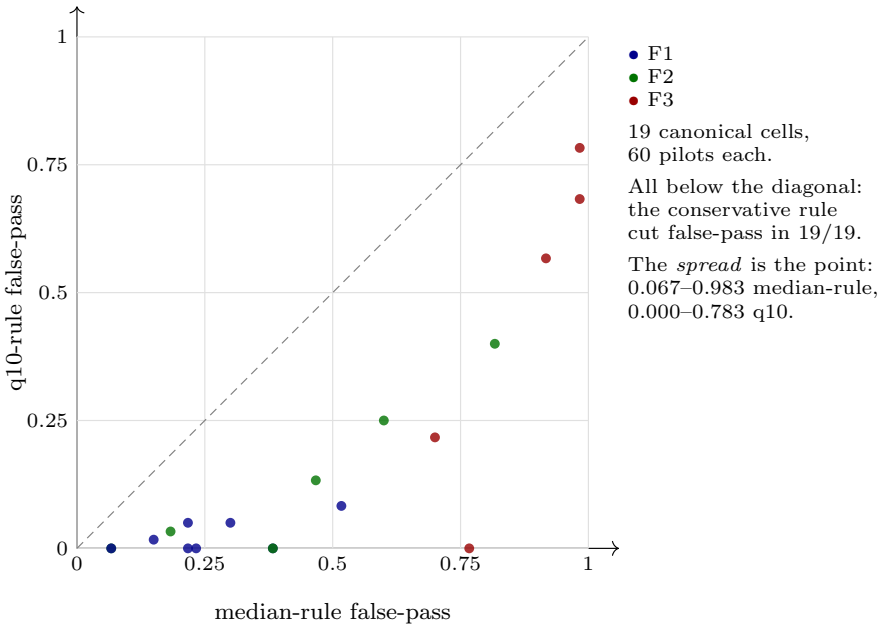

Fig. 4 Canonical operating characteristics across 19 cells and three outcome families.

The refinement stage changed nothing: comparing all 19 cells × 4 rates between the 30-pilot and 60-pilot runs by Fisher exact test gives **0 of 76 comparisons at** $p < 0.05$, against ≈ 3.8 expected by chance. That is a null result for a pre-specified step, reported as one — we ran it because it was frozen before the first-stage results existed, and declining a pre-specified step because its outputs disappointed is the practice this paper argues against.

8 External Stress Test

External construct-alignment stress test; no positive corroboration is claimed.

We pre-registered a protocol to test whether our constructs transfer to independent published artifacts, froze the endpoint definitions and analysis code, and passed a 23-check quality gate against synthetic fixtures *before* opening any outcome file. Table 3 summarises the outcomes.

The public release of Xiang et al. (2026) also differs from its paper's headline sample (137 tasks against 116), so we analyse the artifact *as released* and describe nothing as reproducing a published headline.

The noise null deserves a sentence on method, because it is the analysis we least wanted to run. Holding each condition's effect fixed at its observed value and re-randomising only the 25% partition — preserving the shared-task dependence exactly rather than treating the eight conditions as independent — the observed instability is what finite-sample variation produces. The direction

Table 3 External stress-test outcomes.

Artifact / analysis	Outcome	Limitation established
Self-repair study (Olausson et al., 2024)	repairs attempted *only on failing completions* — 4,287/4,287 failing carry repairs, 0/3,913 passing do	effect one-sided by construction; zero negative transitions is an artefact; best-of-10, not single repair. **Cannot validate endpoint A**
Cross-model review (Xiang et al., 2026)	reviewer emits a *code artifact*, evaluated as such	contrast is structurally substitution, not critique information; neither artifact instantiates our headline estimand
Baseline-as-proxy claim (ours)	higher baseline inverse-ordered with responsiveness in 31 of 31 within-stratum pairs	our earlier claim has external *counter-evidence*; **withdrawn**, not softened
Pilot/held-out instability (ours)	5 of 8 conditions disagreed on effect direction, worst 10/20	under a **post-hoc, reviewer-motivated** noise null frozen before execution, observed 5 against a *null mean of 5.22*, $P(\geq 5) = 0.80$; **withdrawn**
Endpoint B	no screened artifact exposes a structural clean-verdict channel	**not externally evaluable**

holds under all three nulls specified in that frozen protocol ($P = 0.80$, 0.65, 0.97), and two independent implementations agree with zero disagreements. It is labelled a **post-hoc, reviewer-motivated sensitivity analysis**; it was specified after outcomes were read, and frozen and hashed before it was executed.

> The external stress test was informative primarily through **falsification and construct-alignment failures** rather than positive validation. Our classification of the external evidence is **weak**. What it demonstrates is that artifacts which look topically similar can be structurally incompatible with the estimand they appear to address.

9 Evaluator and Provenance Validity

These instruments are *descriptive*; none is a result.

Evaluator oracle. An oracle validated 445 tasks by byte-exact answer replay, plus three at a second tier, over a 448-task set — establishing that the harness reproduced known-correct outcomes before it was trusted to label unknown ones. A three-execution timing policy fired on 18.9% and 11.1% of sources across the two pilots and *changed no verdict.*

Benchmark structure and admissibility. Within one pinned archive revision we identified five confirmed structural inconsistencies; this is scoped to that revision and generalises neither to the benchmark as a project nor to benchmarks at large. A sensitivity check on the cutoff itself: moving the frozen contamination boundary one month later would shrink the *admissible* primary Easy+Medium pool from 461 to 403 tasks. That 403 is a counterfactual about cutoff drift, not the population this study used, which is the 458 evaluable tasks of Section 4. We do *not* conclude that contamination-controlled evaluation prices out capable models.

A fabricated dataset, found and corrected. A judging stage once "completed" 840 sources in about 90 seconds and reproduced a previous experiment's outcomes exactly. The cause was resume identity: the harness keyed resumable work on `(task_id, seed, condition)` and omitted `source_sha256`, so a new experiment's sources collided with the previous ledger and *inherited its verdicts without executing*, producing a complete, internally consistent, entirely fabricated dataset. That ledger is permanently quarantined; resume identity is now the full causal key, and a regression suite covering six properties passes 7/7 while importing the real modules — including the frozen defective module as a control that must still reproduce the defect — rather than mocking the module under test.

A concurrency defect in our own simulation. Two processes wrote the same simulation ledger and the last flush silently replaced the other's records for two cells. The count-based resume check could not detect it: it validated record *contents*, which both writers satisfied, and never established exclusive ownership. Rather than adjudicate between two observed ledgers after seeing both, all 19 cells were re-run once, in a single writer, under a pinned and recorded environment with append-only writes and an exclusive lock. Every simulation number here comes from that canonical ledger; the superseded lineages are quarantined. The canonical run is statistically indistinguishable from both — 0 of 76 rate comparisons reach $p < 0.05$ — and two reported counts moved, which we report rather than absorb.

Three lessons generalise, and only three:

1. Resumable harnesses must key on the full causal identity of the unit of work.
2. Implausible speed is a defect signal.
3. Integrity checks that validate record contents are insufficient if they do not also guarantee exclusive writer ownership.

10 Practical Recommendations for Sparse Replicated LLM-SE Experiments

The following recommendations are scoped to sparse, replicated LLM author/reviewer experiments of the kind studied here. They do not establish general rules for LLM experimentation or software engineering at large.

1. **Screen measurability before power.** Power projection presumes an estimable effect; it cannot detect the absence of one.
2. **Report the diagnostic's null.** A responsiveness statistic over replicated outcomes has a positive null that depends on the outcome rate as well as the replicate count. Publish a reference null at the pilot's own baseline and k, with its independence and homogeneity assumptions stated, or the diagnostic will confirm whatever it is asked to confirm.
3. **Give judging models a structural abstention channel, and measure its use.** Without an explicit channel, non-discrimination is invisible.
4. **Check that the planning model fits before trusting what it projects.**
5. **Freeze the refinement steps too, and run the ones you froze.**
6. **Verify construct alignment structurally, not topically, before importing external evidence.**
7. **Compute the null for your own favourable result before a referee does.** A directional finding without a null is the failure mode most likely to survive internal review, precisely because it is welcome.

11 Limitations

The screen's specificity is untested, and this is the most important limitation. Both pilots failed, so every threshold was exercised in one direction only. A screen that stops everything is worthless, and we cannot demonstrate that these thresholds do not also stop viable studies. That demonstration requires a system that passes the screen and a confirmatory study that then estimates the effect; we have neither.

Two development pilots, both failing, on one difficulty stratum of one benchmark, in C++ competitive programming. **No confirmatory causal study was run**, and the paper contains no confirmatory causal evidence.

Endpoint asymmetry. Endpoint A is prospective and externally examined on one usable artifact. Endpoint B is prospective but *has no external validation* — no screened public artifact exposes a structural clean-verdict channel — and rests on a single critique prompt, so prompt and model effects are not separable. Endpoint C is *post-hoc in origin*.

External evidence is weak and, after the noise null of Section 8, contributes no positive support. Sign agreement was read across 13 cells $\times$ 20 seeds with no multiplicity correction, so the worst case is descriptive.

Progression-rule behaviour depends on assumed outcome-generating families. Two of three families are stress families; the heterogeneity scale is specific to the simulated model; results are for one frozen design grid; other families could order the rules differently. Bit-exact replay of the simulation is tied to the interpreter's string-hash family.

Model versioning. Two of the three model identifiers used in the pilots are dated, immutable snapshots; the third is served by a host that does not document serving precision. Any future confirmatory work in this line re-

quires immutable, callable model snapshots for both roles, and we found that requirement to be in tension with cost.

We do not claim this framework is universally optimal, or optimal at all. It is one adaptation, applied once, that stopped one study.

12 Conclusion

We adapted multi-endpoint feasibility methodology to sparse LLM experiments, operationalised three LLM-specific viability endpoints — including a responsiveness diagnostic that must be read against its replication-induced null rather than against zero — and applied them prospectively to a planned confirmatory study of critique information in cross-model code review. The study did not progress. Two development pilots failed the responsiveness and reviewer-discrimination endpoints independently, the planning model failed to fit, and the reserved 248-task pool was never consumed.

We then tested our own constructs against independent artifacts and found the strongest external result we had was compatible with finite-sample noise, and that the nearest published artifact is structurally incompatible with the estimand it appeared to address. Both are reported against our own interest.

What we offer is not a result about language models. It is a screening procedure, its LLM-specific diagnostics, and a documented case in which the procedure stopped a study before it spent its evidence — including the part where the procedure was turned on our own favourable finding. A prospectively specified critique-transfer experiment remains reserved for future work with adequate independent-task scale and reproducibly versioned model access.

Declarations

Funding. This work was self-funded by the author. No external funding was received.

Competing interests. The author declares no competing interests.

Ethics approval and consent. Not applicable. This study involved no human participants or animals. Consent to participate and consent for publication are not applicable.

Author contribution. Thupakula Sai Varun is the sole author and is responsible for the study design, formal analysis, interpretation, study progression or stopping decisions, and scientific conclusions.

Generative AI use. ChatGPT and Claude assisted with code and simulation development and manuscript development. The author independently performed the analyses and interpretations, made all study progression or stopping decisions, verified the final manuscript and its cited evidence, and remains fully responsible for the accuracy, originality, code, analyses, claims, citations, and conclusions. No generative AI system is listed as an author.

Data and code availability. The frozen analysis code, canonical simulation ledger, derived results, checksum manifests, and reproduction instructions

supporting the claims are supplied to the journal as supplementary material for peer review. The confirmatory task pools remain sealed and unobserved; no task outcomes exist to share. Third-party raw artifacts are not redistributed because they remain subject to their upstream licences; the supplementary material records upstream commit hashes, checksums, and retrieval instructions. The evaluator harness is not redistributed, and its x86-64 Linux portability boundary is documented in the supplementary material. Any public archival release will be limited to materials the author has the right to distribute and will follow the journal's artifact and supplementary-material policy.

References

Clopper CJ, Pearson ES (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26(4):404–413, DOI 10.1093/biomet/26.4.404

Jin H, Chen H (2026) Are LLMs reliable code reviewers? systematic overcorrection in requirement conformance judgement. Automated Software Engineering DOI 10.1007/s10515-026-00638-5, 2603.00539

Montgomery RN, Bodde AE, Vidoni ED (2025) Jointly assessing multiple endpoints in pilot and feasibility studies. Statistical Methods in Medical Research 34(3), DOI 10.1177/09622802241311219

Ning J, Li X, Yu C (2026) Revision or re-solving? decomposing second-pass gains in multi-llm pipelines. In: Conference on Language Modeling (COLM), DOI 10.48550/arXiv.2604.01029, 2604.01029

Olausson TX, Inala JP, Wang C, Gao J, Solar-Lezama A (2024) Is self-repair a silver bullet for code generation? In: International Conference on Learning Representations (ICLR), 2306.09896

Xiang Z, Zhang Y, Zhang Y, Xu H (2026) Cross-model LLM code review: Should you use Claude to review Codex or vice versa? In: Agentic Software Engineering Workshop at KDD, DOI 10.48550/arXiv.2607.21656, 2607.21656