

Peer Review

Review of: "Simulating Lay Health-Seeking Behavior with LLM Personas and Illness Vignettes: Reproducibility, Prompt Sensitivity, and Slice Dependence"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper is a fully synthetic, non-human-subject methodological audit of using an LLM persona as a "synthetic respondent" for lay health-seeking behaviour. A single model (Section 3.5: **gemini-2.5-flash** via Vertex AI, executed in an "Antigravity" environment) is conditioned on one of 60 synthetic personas and one of 30 synthetic illness vignettes (Section 3.1) and must return a closed-schema JSON object whose principal field is a next-action code A0 to A9. Four robustness properties are examined: distribution plausibility, run-to-run repeatability, prompt sensitivity between a narrative persona prompt (P1) and a rubric persona prompt (P2, Section 3.3), and dependence of these on scenario slice. The headline findings are: in the Pilot (40 pairs, 2 prompts, 3 repeats), mean modal agreement was **0.617** (Section 5.2) while P1 and P2 produced **0 mismatches across all 40 pairs** when run inside the same batch (Section 5.3); in the isolated-prompt Audit01 on 24 pairs, prompt mismatch swung from **0.0% (r1) to 45.8% (r3)** (Section 5.4); mismatch was **8.3% (2/24)** in a severity=1 non-red-flag slice against **41.7% (10/24)** in a red-flag slice (Section 5.5); and in a decoding-controlled follow-up with 8 to 10 repeats, prompt mismatch stayed at **78.7% to 82.1%** with JSD(P1,P2) of **0.148 to 0.196 bits** under both temperature 0 and default sampling (Table 1), while temperature 0 raised per-pair modal agreement in three of four slice by prompt cells (Table 2). A rubric stress test shifted the action distribution by JSD **0.130** and lowered mean urgency from **3.54 to 2.25** (Section 5.8). The conclusion drawn is a practice recommendation: isolate prompt conditions, replicate runs, stratify by slice, and report model identifiers and decoding parameters.

Two things deserve credit before the criticism. First, the central observation in Section 5.3 and the Discussion, that a paired design executed inside a shared generation context can produce a perfect, entirely artefactual agreement and be misread as "prompt formulation does not matter", is a genuine and useful contribution that is under-appreciated in this literature. Second, the reported statistics are internally consistent to a degree that is rare. I reconstructed them independently and every one reconciled: Cochran's Q in Section 5.4 ($Q=17.17$, $df=2$) recomputes exactly from column mismatch totals 0, 2, and 11 with exactly one pair mismatching in both r2 and r3, giving $Q = 412.0/24 = 17.17$ and a chi-square p of 0.000187, matching the printed p=0.00019; the Holm-adjusted exact McNemar values follow exactly, $3 \times 2 \times 2^{-11} = 0.00293$ for r3 against r1, $2 \times 24/2048 = 0.0234$ for r3 against r2, and 0.50 for r1 against r2; the Fisher test in Section 5.5 on the table $[[2,22],[10,14]]$ gives an odds ratio of 0.12727 and a two-sided p of 0.01734, both exactly as printed; and the mean urgency values in Section 5.8 recover as $85/24 = 3.5417$ and $54/24 = 2.25$ with a difference of $31/24 = 1.2917$. The modal-agreement means in Table 2 all resolve to exact fractions over the stated repeat counts ($0.558 = 134/240$, $0.571 = 137/240$, $0.529 = 127/240$, $0.551 = 119/216$, $0.394 = 85/216$, $0.407 = 88/216$, $0.366 = 79/216$, $0.510 = 98/192$), which also confirms that Table 2 uses the run counts printed in Table 1. The author has been careful with the numbers. My objections are about what the numbers are allowed to mean.

2. Whether the conclusion holds on the data

2.1 The headline prompt-sensitivity number has no null baseline, and the paper's own Table 2 shows the baseline is large. The abstract and Section 5.7 present "mean action mismatch 0.787 to 0.821" (Table 1, column "Mismatch (action_primary)") as the measure of prompt sensitivity. But mismatch between one P1 run and one P2 run confounds the prompt effect with the model's own run-to-run instability, and Table 2 quantifies that instability directly: under det_t0, row "audit01_base24 (non-red-flag) / P1", per-pair modal agreement is only **0.558**. If an item's response distribution places 0.558 on its modal action, two independent draws from that same distribution agree with probability between $0.558^2 + 0.442^2 = 0.507$ (when the residual mass sits on a single alternative) and roughly 0.38 (when it spreads over three alternatives), so two runs of the *same* prompt would already be expected to mismatch on roughly 49% to 62% of items. Against that, the 82.1% cross-prompt mismatch in Table 1 row "audit01_base24 / det_t0" implies a prompt-attributable excess on the order of 20 to 33 percentage points, not 82. The author already has the data required to settle this: with 10 completed P1 runs and 10 completed P2 runs on the same 24 pairs, the within-prompt cross-run mismatch (P1 run i against P1 run j, and likewise for P2) is

computable at no extra cost and belongs in Table 1 as a column. As printed, the abstract's claim that "prompt sensitivity remained high under both decoding regimes (mean action mismatch 0.787 to 0.821)" is not supported, because nothing in the paper separates the prompt from the noise floor.

The same point can be made from the other side. Table 1 reports JSD(P1,P2) of only **0.148 to 0.196 bits** in the same rows in which mismatch is 78.7% to 82.1%. Near-independent per-item pairing combined with nearly coincident marginals is the signature of within-item stochasticity, not of a systematic prompt-induced shift. The paper notices the low divergence (Section 5.7, "decoding parameters therefore appear to modulate variance and repeatability more strongly than the mean output distribution") but does not draw the implication for its own mismatch statistic.

2.2 A large, unexplained discontinuity between Section 5.4 and Table 1 on the same 24 pairs. Audit01 base24 yields mismatch of 0.0%, 8.3%, and 45.8% (Section 5.4, r1 to r3). The controlled follow-up runs the identical slice, audit01_base24, and reports **82.1%** (Table 1, det_t0) and **81.9%** (Table 1, stoch_default). The same happens on the red-flag slice: **41.7%** in Section 5.5 against **78.7%** and **80.7%** in Table 1. The follow-up therefore reports mismatch roughly double the worst earlier replication on the same items, and the paper nowhere acknowledges, let alone explains, the gap. The most likely explanation is visible in Section 3.4: the follow-up used "per-item isolated mode (one request per persona-scenario item)," whereas Audit01 isolated the prompts from each other but apparently still executed the 24 items inside one run. If so, this is the paper's most important result, a second and larger instance of exactly the coupling effect described in Section 5.3, now at the item level rather than the prompt level: batching items within a request appears to suppress apparent variability far more than batching prompts does. That deserves its own subsection and a direct batched-versus-per-item comparison, not silence.

2.3 The follow-up changes two factors at once, so it does not do what Section 5.7 says it does. Section 5.7 is titled as a decoding-controlled rerun and claims to "disentangle prompt framing effects from stochastic decoding noise". Within Table 1, the decoding contrast is clean because the det_t0 and stoch_default rows differ only in decoding. But every comparison drawn between the follow-up and the earlier experiments is confounded because the follow-up simultaneously changed decoding configuration *and* execution granularity (Section 3.4, per-item isolated mode) *and* top_k (Section 3.5 fixes top_k=20 in both follow-up configurations while the earlier runs used an unrecorded provider default). A clean design needs a third arm: batched execution under det_t0.

2.4 The slice-dependence claim is the weakest of the four headline claims and is contradicted by the controlled data. Section 5.5 compares 8.3% (2/24) against 41.7% (10/24) in single runs and concludes

"significantly higher prompt sensitivity in the red-flag slice". The Fisher test treats the 24 pairs as independent units, but the mild slice is 12 scenarios by 2 personas, and the red-flag slice is 6 scenarios by 4 personas (Sections 3.4), so the red-flag result rests on 6 distinct vignettes, and the effective sample is far smaller than 24. More seriously, each slice was measured in a *single* run, and the paper has already established in Section 5.4 that a single run of a fixed slice can land anywhere from 0.0% to 45.8%. The between-run variance the paper itself measured therefore spans the entire between-slice difference it wants to interpret. A red-flag value of 41.7% is inside the range an "r3-like" draw of any slice can produce. And when the two slices are measured properly with 8 to 10 repeats, the effect reverses: Table 1 gives 82.1% and 81.9% for the non-red-flag slice against 78.7% and 80.7% for the red-flag slice, that is, marginally *lower* mismatch in the red-flag slice. The Discussion glosses this as "mismatch remained high in both slices" and never notes the reversal, yet the abstract and title still carry "Slice Dependence" as a headline finding. The slice claim should be either demoted to a hypothesis or re-established with replicated slices under matched execution.

Two further confounds affect Section 5.5. The two slices differ in severity (1 against the red-flag scenarios' severity) and in red-flag status simultaneously, so the comparison cannot attribute the difference to red-flag status; the Discussion concedes this in passing ("severity and vignette ambiguity also matter") but Section 5.5 does not.

2.5 The Pilot repeatability figure is described as "moderate" but is close to chance for a three-option model. Section 5.2 and the abstract report mean modal agreement of 0.617 over 40 pairs with $n=3$ repeats. With $n=3$, the metric has a floor of $1/3$, so the interpretable question is what value would arise if the model chose uniformly among k plausible actions. I computed this exactly by enumeration: for $n=3$, the expected modal agreement is 0.750 for $k=2$, **0.6296** for $k=3$, and 0.5625 for $k=4$. The observed 0.617 sits just below the $k=3$ value. In other words, Pilot repeatability is approximately what would be seen if the model were picking uniformly among about three actions per item. The same exercise applies to Table 2: for $n=10$ repeats, the expected modal agreement under a uniform choice between just two actions is **0.623**, so the `det_t0` value of **0.558** (Table 2, `audit01_base24`, P1) means that at temperature 0.0 with fixed seed 17, the model is *less* self-consistent than a coin flip between two options. That is a striking and publishable finding in its own right, and it materially qualifies the abstract's phrase "near-deterministic decoding improved run-to-run stability", which is true as a relative statement (Table 2, three of four rows) but invites the reader to assume the `det_t0` arm is near-deterministic in outcome. It is not.

2.6 The Section 5.1 plausibility claim rests entirely on figures. Section 4.1 announces a test of whether urgency "increases monotonically" with severity, and Section 5.1 asserts that it does, citing only Figure 2 and Figure 3. No mean urgency per severity level, no confidence intervals, and no trend statistic appear anywhere. The same holds for Section 5.2's claim of "higher repeatability for severity=3", cited to Figure 4 with no numbers, and for the whole of Section 5.9, where the OOV rate is never given as a number anywhere in the manuscript, only as "high OOV rates in Figure 8".

2.7 One statistic does not reconcile. Section 5.8 reports the stress-test Wilcoxon as $W=0.0$ with $p=0.00484$. $W=0$ means every non-zero difference has the same sign, for which the exact two-sided p is 2^{1-n} in n non-zero pairs. Table 3 gives the stress-test mismatch as 41.7%, that is 10 of 24 pairs, and because distinct action codes always have distinct urgency scores under the A0=0 to A9=9 mapping of Section 4.6, the number of non-zero differences must be exactly 10. The exact two-sided p at $n=10$ is therefore **0.001953**, not 0.00484. The printed value instead matches an asymptotic normal approximation, which at $n=10$ without continuity correction gives 0.00506 and moves to roughly 0.0048 once tied difference magnitudes shrink the variance. No integer n reproduces 0.00484 exactly, and I checked n from 6 to 13. The result is unaffected in substance, but the test should be reported as exact (and would then be stronger), with the number of non-zero-difference pairs stated.

2.8 A summary value in Table 3 does not match Table 2. Table 3, row "Controlled follow-up: decoding effect on repeatability", states that `det_t0` raises mean modal agreement "by +0.12 to 0.16 in 3/4 settings". The three significant cells of Table 2 give $0.558 - 0.394 = 0.164$, $0.571 - 0.407 = 0.164$, and $0.529 - 0.366 = 0.163$. The range is +0.163 to +0.164, and the fourth cell gives $0.551 - 0.510 = 0.041$. No pair of Table 2 values produces +0.12. The summary understates its own lower bound.

2.9 Table 4 does not support the sentence that cites it. Section 5.11 states that the model selected high-urgency actions "especially under P2", but Table 4 shows P1 and P2 both giving A8 for S008 and for S018, with only S010 differing (P1: A8, P2: A9). One of three rows is not "especially".

2.10 Causal language. The manuscript is mostly careful, but three places slip. Section 5.5 says the Fisher result "indicat[es] significantly higher prompt sensitivity in the red-flag slice" from a single-run contrast. Section 5.9 attributes the OOV step-change to "the complexity of simultaneously satisfying multiple constraints", an explanation that cannot be right as stated, because the prompts and constraints were identical in Audit01 r1 (0% OOV) and in the later runs (high OOV); a constant cannot explain a step change. Section 5.7's claim to "disentangle" prompt from decoding effects overstates what a two-cell decoding contrast with no same-prompt baseline can do.

3. Reproducibility

This section is where the paper is simultaneously above average for its field and still short of what its own conclusions demand. It deserves credit for stating the model identifier (Section 3.5: gemini-2.5-flash via Vertex AI), for stating single-turn stateless requests with `candidate_count=1`, for fully specifying the follow-up decoding configurations (`det_t0: temperature 0.0, top_p 1.0, top_k 20, candidate_count 1, max_output_tokens 512, fixed seed 17; stoch_default: temperature 1.0, top_p 0.95, top_k 20, candidate_count 1, max_output_tokens 512, per-run seeds`), for openly disclosing in Section 3.5 that the Pilot, audit, slice, and stress-test runs did *not* have their generation configuration overridden or logged, and for repeating that limitation in Section 7. Few papers in this area disclose that much.

The remaining gaps, in rough order of severity:

No model snapshot and no run dates. Section 3.5 names "gemini-2.5-flash", which is a floating alias, and Section 7 concedes that server-side model revisions were not pinned. The manuscript contains no execution dates anywhere, for any run. This is not a formality here, because the paper's two most interesting irregularities both take the form of a step change between replications on identical inputs: Audit01 r3 is the mismatch outlier at 45.8% against r1's 0.0% (Section 5.4, Figure 5 annotation), and OOV goes from 0% in the Pilot and Audit01 r1 to high in the later runs (Section 5.9, Figure 8 annotation, described in the figure caption as a "step-change"). Two independent metrics stepping at the same point in the run sequence, on the same prompts and the same items, is at least as consistent with a serving-side model or default change between run dates as with sampling variation, and the paper cannot distinguish the two because it did not record either the returned model version string or the calendar dates. The Vertex and Gemini responses carry a model version field; it should have been logged per call and reported per run.

Most results have no recorded decoding configuration at all. Sections 5.1 through 5.6, 5.8, and 5.9, that is the majority of the manuscript, rest on runs whose temperature, top_p, top_k, and seed are unknown. Section 3.5's fallback, treating "documented provider defaults as the best available reference" with citations [18][19] dated 2026, is a reasonable disclosure but is not evidence: it assumes the documented default at citation time equals the effective default at the (unstated) run time, and it assumes the Antigravity layer did not override anything. Those results should be labeled as configuration-unknown wherever they are used, including in Table 3.

The "Antigravity environment" is never defined. Section 3.5 names it in a single clause and never says what it is, whether it injects a system instruction, whether it wraps calls in an agentic loop, whether it applies its own retries or its own generation defaults, or how the prompt text in Section 3.3 maps onto `system_instruction` versus user content. For a study whose central claim is that execution context couples outputs, an undescribed execution layer is a first-order omission.

Seeds. `det_t0` uses "fixed seed 17" (Section 3.5) but `stoch_default` uses unlisted "per-run seeds". List them. Also note that the seed is best-effort on this platform, which the Discussion correctly cites, and the `det_t0` modal agreement of 0.558 in Table 2 is the empirical demonstration that it did not deliver determinism. That fact should be stated as a result, not left implicit.

Missing runs are uncharacterized. Section 3.4 and Section 5.7 say completed repeats ranged from 8 to 10 "due to occasional transient failures", and Table 1 shows the pattern is not random-looking: P2 completed fewer runs than P1 in both red-flag rows (10/9 and 9/8). No failure counts, causes (HTTP error, timeout, safety block, refusal, JSON parse failure), or per-item breakdown are given. If failures concentrated on red-flag items or on one prompt, the red-flag estimates in Table 1 and Table 2 are subject to informative missingness, which matters most in exactly the cell that produced the one null result (Table 2, `redflag24`, P2, $p=0.251$).

Safety settings. None are reported. The task is medical-adjacent, the vignettes include anaphylaxis and severe chest pain (Table 4, S010 and S018), and safety filters can alter or block such completions. Report the safety-setting configuration and the per-call finish reason distribution.

Prompt language. Section 3.3 notes in passing that the excerpts are "translated from the synthetic Japanese inputs". The experiment was therefore run in Japanese, which is not stated in the abstract, in Section 3.1, or in the Limitations. This is material: the narrative-versus-rubric contrast that defines P1 against P2 is a linguistic manipulation, and its magnitude is language-dependent. State it plainly in Section 3.1 and add it to Section 7.

Artefacts. Data Availability points to `supplementary-material.zip`. There is no persistent identifier, no license, no directory manifest, no software environment (Python and library versions, in particular the version used for the Cochran, McNemar, Fisher, and Wilcoxon tests of Section 4.6, given the exact-versus-asymptotic issue in item 2.7), and no statement of whether the archive contains raw API responses or only parsed records. Raw responses matter here specifically because they carry the model version, usage metadata, finish reason, and safety ratings that Section 3.5 says were not retained.

Call accounting. Only the Pilot's output count is given (Section 3.4: 240). Audit01, the slice audits, the stress test, and the follow-up have no totals. From the printed run counts in Table 1, the follow-up alone is $24 \times (10+10+9+9+10+9+9+8) = 1776$ calls. Report the totals so that the cost and scale of replication are knowable.

4. *What specifically to fix*

1. **Table 1:** add two columns giving the same-prompt cross-run mismatch, P1 run i against P1 run j and P2 run i against P2 run j, averaged over all run pairs within each cell, as the null baseline for the printed P1-against-P2 mismatch. Then restate the prompt effect as the excess over that baseline, or as a chance-corrected statistic such as Cohen's kappa. The abstract's "mean action mismatch 0.787 to 0.821" must not stand as a measure of prompt sensitivity without it.
2. **Section 5.7 and Table 1:** explain the discontinuity between Section 5.4 (audit01_base24 mismatch 0.0%, 8.3%, 45.8%) and Table 1 (same slice, 82.1% and 81.9%), and between Section 5.5 (red-flag 41.7%) and Table 1 (redflag24, 78.7% and 80.7%). If the cause is batched execution in Audit01 against per-item isolation in the follow-up, say so explicitly and promote it to a result.
3. **Section 3.4 and Section 5.7:** add a third experimental arm, batched execution under det_t0 on audit01_base24, so that execution granularity and decoding configuration are separated. As designed, the follow-up changes granularity, decoding, and top_k at once relative to the earlier runs.
4. **Table 1:** state how a P1 run is paired with a P2 run when the completed counts differ (rows redflag24 det_t0, 10/9, and redflag24 stoch_default, 9/8). The printed means reconcile exactly with $\min(P1, P2)$ pairings ($78.7\% = 170/216$ over 9 pairings; $80.7\% = 155/192$ over 8), so the rule appears to be "drop the surplus runs," but it is never stated, and the choice of which runs to drop is not reported.
5. **Table 1:** define the bracketed intervals. Level, method (bootstrap, normal across runs, or other), and unit of resampling are all unspecified for both the mismatch and the JSD columns. Note also that intervals computed across runs do not account for clustering of the 24 items across run pairs.
6. **Table 2:** compare modal agreement at matched repeat counts. Modal agreement is mechanically biased by n, and the compared cells use 10 against 9, 10 against 9, 10 against 9, and 9 against 8. By enumeration, the bias is small here; expected modal agreement under a uniform two-way choice is 0.623 at n=10 against 0.6367 at n=9, so roughly 0.014, against a reported effect of 0.164; the finding survives, but subsampling every cell to the common minimum makes the comparison exact and costs nothing.

7. **Table 2 and Section 5.7:** report the absolute interpretation of the `det_t0` agreement values, not only the relative gain. State that 0.558 (`audit01_base24`, P1, `det_t0`) and 0.529 (`redflag24`, P1, `det_t0`) are below the 0.623 expected at $n=10$ from a uniform choice between two actions, that is, temperature 0.0 with fixed seed 17 did not yield near-deterministic behavior. Adjust the abstract sentence "near-deterministic decoding improved run-to-run stability in three of four settings" accordingly.
8. **Table 3, row "Controlled follow-up: decoding effect on repeatability":** correct "+0.12 to 0.16" to the values actually implied by Table 2, +0.163 to +0.164 across the three significant cells (and +0.041 in `redflag24` / P2).
9. **Section 5.8:** report the number of non-zero-difference pairs for the Wilcoxon test and use the exact test. With $W=0.0$ and, by the 41.7% mismatch in Table 3, exactly 10 non-zero differences, the exact two-sided p is 0.001953; the printed $p=0.00484$ corresponds to an asymptotic approximation. State which implementation and which method argument were used, here and for the tests in Section 4.6.
10. **Section 5.8:** give the full baseline and P2_strong action distributions over all ten codes. The text lists only A7 (5 to 1), A5 (6 to 3), A2 (2 to 6), and A3 (1 to 3), which account for 14 of 24 baseline and 13 of 24 strong-rubric items; the remaining codes must sum to 13 urgency points over 10 items at baseline and 11 points over 11 items after, and readers cannot verify the JSD of 0.130 without them.
11. **Section 5.5 and abstract:** either replicate both slices with 8 to 10 repeats under matched per-item execution, or demote the slice-dependence claim. As it stands, both slice values come from single runs while Section 5.4 shows single-run mismatch on a fixed slice ranging from 0.0% to 45.8%, and the replicated data in Table 1 shows the non-red-flag slice at 82.1% and 81.9% against the red-flag slice at 78.7% and 80.7%, a reversal of the claimed direction that the Discussion does not mention.
12. **Section 5.5:** address clustering and confounding. The mild slice is 12 scenarios by 2 personas, and the red-flag slice is 6 scenarios by 4 personas (Section 3.4), so Fisher's test on 24 pairs per slice overstates the effective sample, most severely in the red-flag slice with 6 distinct vignettes. Report a scenario-clustered analysis, and state that severity and red-flag status vary together, so the comparison cannot isolate red-flag status.
13. **Section 5.1, Figures 2 and 3:** replace the assertion of monotonic urgency with numbers, mean urgency per severity level (1, 2, 3) and per `red_flag` value with intervals, plus the trend statistic that Section 4.1 promises. Do the same for Section 5.2's claim of "higher repeatability for severity=3," currently supported only by Figure 4.
14. **Section 5.9, Figure 8:** give the OOV rate as a number per run and prompt in a table. At present, no OOV value appears anywhere in the text. Additionally, revise the causal explanation: prompt

complexity is constant across Audit01 r1 (0% OOV) and the later runs (high OOV), so it cannot account for a step change on identical prompts and items. Test the alternative that the serving-side model changed between runs.

15. **Section 3.5 and a new reporting table:** pin and report the model snapshot, the model version string returned in each response, and the calendar date of every run (Pilot, Audit01 r1 to r3, both slice audits, stress test, and each follow-up condition). This is required to adjudicate item 14 and the Audit01 r3 outlier in Section 5.4.
16. **Section 3.5:** describe the Antigravity environment: what it is, whether it injects a system prompt or tool scaffolding, whether it applies its own retries or generation defaults, and how the P1 and P2 text of Section 3.3 is split between system instruction and user content.
17. **Section 3.4 and Section 5.7:** report per-condition failure counts and causes behind "completed repeats ranged from 8 to 10," and show that missingness is unrelated to slice, prompt, and item. Table 1 shows P2 short by one run in both red-flag rows, and the only non-significant cell in Table 2 is redflag24 / P2.
18. **Section 3.2 and Section 3.5:** report safety-setting configuration and the distribution of finish reasons, particularly for the red-flag vignettes summarized in Table 4 (S010 severe dyspnoea with chest pain, S018 facial and lip swelling with throat discomfort).
19. **Section 3.1 and Section 3.3:** state prominently that the personas, vignettes, and prompts were written in Japanese and that Section 3.3 shows translations, and add the language dependence of the narrative-versus-rubric contrast to Section 7.
20. **Section 4.3 against Section 5.4 and Section 5.7:** report `action_time` and `escalation_if_worse` mismatch numerically. Both are defined as metrics in Section 4.3 and neither appears as a number anywhere; Figure 5 is the only trace of them.
21. **Section 3.2 against the Results:** the confidence field (integer 0 to 100) is collected and never analyzed. Either report it, ideally as calibration of confidence against per-pair modal agreement from Table 2, which would be a genuinely informative addition, or remove it from the schema description.
22. **Section 5.6:** test the P1 against P2 repeatability difference (0.625 against 0.764, the same 24 pairs) rather than only stating it. A paired Wilcoxon is available, and if the rubric prompt is reliably more repeatable than the narrative prompt, that is a practical result worth stating in the Conclusion.
23. **Section 5.3:** state the unit behind "0 mismatches" precisely. The Pilot had 3 repeats per prompt, so it is unclear whether the comparison is 40 pair-level comparisons or 120 run-level comparisons, and

whether zero mismatches held in every one of the three runs.

24. **Section 5.11 and Table 4:** either remove "especially under P2" or support it. Table 4 shows P1 and P2 agreeing on A8 for S008 and S018, with a difference only at S010 (A8 against A9).
25. **Data Availability:** deposit the archive under a persistent identifier with an explicit license, a manifest, the software environment used for Section 4.6's tests, and the raw API responses including model version, usage metadata, finish reason, and safety ratings, rather than parsed outputs alone.
26. **Section 3.4:** state how the 40 Pilot pairs were sampled from the 60 personas by 30 scenarios grid of Section 3.1, and how the audit and slice pairs were selected within their strata.

5. Recommendation

Major revisions. The study addresses a real and under-examined methodological problem, the Section 5.3 within-batch coupling observation is a genuine contribution, the reporting of decoding parameters in the controlled follow-up is better than the norm for this literature, and the internal arithmetic is unusually sound: I reconstructed Cochran's Q, the Holm-adjusted McNemar values, the Fisher odds ratio and p, the urgency means, and every modal-agreement mean in Table 2, and all of them reconciled exactly. The revisions I am asking for are therefore not about the rigor of execution but about the inferential step from numbers to claims. Three of them are blocking. First, the headline prompt-sensitivity figure in Table 1 and the abstract has no same-prompt null baseline, while Table 2's `det_t0` modal agreement of 0.558 shows that the baseline is large; the excess attributable to prompt framing is on the order of 20 to 33 points, not 79 to 82. Second, the roughly twofold discontinuity between Section 5.4 and Section 5.5 on one hand and Table 1 on the other, on identical slices, is unexplained and is probably the paper's most interesting finding. Third, the slice-dependence claim in the title and abstract is built on single-run measurements whose run-level variance the paper itself shows to be as large as the effect, and the replicated data in Table 1 point the other way. All three are answerable from data the author already holds, which is why this is a major revision rather than a rejection. The reproducibility items, in particular pinning the model snapshot and recording run dates (item 15), also bear directly on whether the Audit01 r3 mismatch outlier and the OOV step change in Section 5.9 are sampling effects or provider-side drift, and resolving that would strengthen rather than weaken the paper's central message.

Declarations

Potential competing interests: No potential competing interests to declare.