

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 question here is narrow and worth asking. An LLM receives a fictional layperson persona and a fictional illness vignette, and must return one action code from a closed set of ten (A0 watchful waiting through A9 call an ambulance), together with a timing code, an escalation code, one to three reason tags, and a confidence integer (§3.2, action codebook in §3.1). Two prompt formulations are compared: P1 states the persona as a narrative paragraph, P2 as a key-value rubric, with the task, the safety rules, and the output schema held identical (§3.3). Everything runs on gemini-2.5-flash through Vertex AI in what the author calls the Antigravity environment (§3.5). Four claims follow. First, that within-batch paired comparison understates prompt sensitivity: the Pilot gave 0 mismatches on all 40 pairs (§5.3), while isolated-prompt replications on 24 pairs ranged from 0.0% in r1 to 45.8% in r3 (§5.4, Cochran $Q=17.17$, $df=2$, $p=0.00019$). Second, that sensitivity depends on the scenario slice: 8.3% (2/24) in the severity=1 non-red-flag slice against 41.7% (10/24) in the red-flag slice, Fisher exact odds ratio 0.127, $p=0.01734$ (§5.5). Third, that a decoding-controlled follow-up keeps prompt mismatch at 78.7% to 82.1% under both temperature 0 and default sampling, with JSD(P1,P2) of 0.148 to 0.196 bits (Table 1), while near-deterministic decoding raises modal agreement in three of four slice-by-prompt cells (Table 2). Fourth, that a stronger rubric moves the action distribution by JSD 0.130 bits and drops mean urgency by 1.29 points, 3.54 to 2.25 (§5.8).

Several things about this manuscript deserve credit before I start objecting. The Pilot result is genuinely instructive, and the author reads it the right way round: a perfect 0/40 prompt agreement is treated as a

warning about the design rather than as a finding about prompts. I checked how surprising it is. §5.2 puts pilot run-to-run agreement for `action_primary` at 0.617, so if P1 and P2 draws were independent, the chance of 40 consecutive agreements is at most $0.617^{40} = 4.1 \times 10^{-9}$, and since agreement with $n=3$ repeats is upward-biased, the true figure is smaller still. The coupling explanation is therefore well motivated. The arithmetic I could recheck holds: $Q=17.17$ on 2 df gives $p = e^{-8.585} = 0.000187$, matching the reported 0.00019; the Fisher odds ratio $(2/22)/(10/14) = 0.1273$ matches; the stress-test means reconcile, since $3.54 \times 24 \approx 85$ and $2.25 \times 24 = 54$, and $(85 - 54)/24 = 1.29$. The author also ran a new experiment in response to an earlier review round instead of arguing back, and labelled the triage comparison in §5.11 as illustrative rather than validation. That is a serious piece of work by one person. My objections are that the headline number of the new experiment is not interpretable as stated, that the decoding control demonstrably did not take, and that the central prompt-versus-decoding comparison is computed at two different aggregation levels.

2. Whether the conclusion holds on the data

2.1 There is no same-prompt baseline, so the 78.7–82.1% cannot be read as a prompt effect. This is my main objection. Table 1 reports the P1-versus-P2 mismatch rate for `action_primary`, and the abstract carries it forward as "prompt sensitivity remained high under both decoding regimes." But mismatch between two conditions is only attributable to the condition if you know the mismatch rate *within* a condition. The paper never reports P1 against P1 across independent runs. It can be bounded from the paper's own numbers. Table 2, row `audit01_base24 / P1`, gives mean modal agreement under `det_t0` of 0.558. For a categorical variable, the chance that two independent draws from the same distribution agree is $\sum p_i^2$, and $\sum p_i^2 \leq \max_i p_i$, which the modal share estimates from above. So same-prompt mismatch under `det_t0` in that slice is at least $1 - 0.558 = 0.442$. Against a cross-prompt figure of 82.1% (Table 1, row `audit01_base24 / det_t0`), the part that prompt framing could possibly explain is at most 37.9 points, and that is an upper bound built on a conservative inequality. Under `stoch_default`, the picture is worse: Table 2 gives modal agreement 0.394, so same-prompt mismatch is at least 0.606, against a cross-prompt 81.9% (Table 1). In the red-flag slice, `det_t0` agreement of 0.529 (P1) and 0.551 (P2) bounds same-prompt mismatch at 0.471 and 0.449 respectively, against 78.7% cross-prompt. The gap that remains may well be real, but it is a fraction of what the abstract reports, and the author has the raw repeats needed to compute the exact comparison.

2.2 The near-deterministic configuration is not near-deterministic, and the paper does not say so. §3.4 specifies `det_t0` as temperature 0.0, `top_p` 1.0, `top_k` 20, `candidate_count` 1, `max_output_tokens` 512, fixed seed 17. Under those settings, ten repeats of the same single-turn request should return the same action code, so per-pair modal agreement should sit near 1.0. Table 2 reports 0.558, 0.571, 0.529, and 0.551. That is an important negative result, and the manuscript passes over it: §5.7 says only that `det_t0` "increased mean agreement relative to `stoch_default` in three of four slice×prompt settings." Improvement over 0.394 is not the relevant yardstick when the target is 1.0. Either the configuration was not applied to the requests, or the serving stack is non-deterministic at temperature 0 despite the seed, which is plausible and which §6 gestures at when it cites the provider's "best-effort" wording. Whichever it is, it is a finding, and it also weakens the paper's own recommendation to report a near-deterministic operating point as a reproducible instrument check. An instrument that disagrees with itself 44% of the time is not a reproducibility anchor.

2.3 The prompt-versus-decoding comparison uses two different aggregation levels. §5.7 concludes that "decoding parameters therefore appear to modulate variance and repeatability more strongly than the mean output distribution, while prompt framing shifts the mean distribution directly." The evidence is that `det_t0` versus `stoch_default` gives JSD of 0.018 to 0.073 bits, while P1 versus P2 gives 0.148 to 0.196 bits (Table 1). The same sentence states the asymmetry: the decoding divergences are computed "when aggregated across runs," the prompt divergences "within a run." Plug-in JSD estimated from 24 observations is not comparable with plug-in JSD estimated from roughly 240. The estimator is biased upward, and the bias falls with sample size, so the smaller-sample quantity is inflated relative to the larger-sample one by construction. This is the paper's cleanest-looking result, and I do not think it survives in its present form.

2.4 The size of that bias is of the same order as the reported divergences. Here is the arithmetic, with its assumption stated. For two distributions close to a common r , JSD in nats is approximately $(1/8)\sum(p_i - q_i)^2/r_i$. With independent samples of size n from the same r , $E[(\hat{p}_i - \hat{q}_i)^2] = 2r_i(1-r_i)/n$, so the expected plug-in JSD under the null is about $(k-1)/(4n)$ nats, where k is the number of occupied action codes. At $n=24$ that is 0.135 bits if all ten codes are used, 0.060 bits at five, 0.030 bits at three. The reported within-run prompt divergences are 0.148 to 0.196 bits. I cannot tell how many codes are actually occupied in each slice, because the action distributions live in Figure 2, Figure 3, and Figure 7, and as rendered on Qeios those figures are not legible to me. If the red-flag slice concentrates on A8 and A9, as §5.11 suggests, k is small and the bias is minor. If the non-red-flag slice spreads across A0 through A7, the null

expectation is close to the whole reported value. A permutation test settles this in a few lines of code, and the data are already on disk. The same question applies to the 0.130 bits of §5.8, computed on 24 items per condition.

2.5 The step from 45.8% to 82.1% on the same 24 pairs is never reconciled. Audit01 and the controlled follow-up both use audit01_base24. Table 3 lists Audit01 mismatch as r1 0.0%, r2 8.3%, r3 45.8%; Table 1 lists 82.1% and 81.9% on the same pairs and the same prompts. An almost twentyfold change against the median of the earlier runs needs an explanation, and the manuscript offers none. My reading is that the two numbers are different quantities: the earlier one compares modal or single-run outputs paired within a replication, the later one averages item-level mismatch over 8 to 10 independent repeat-pairs, which drags it toward the same-prompt noise floor discussed in 2.1. If that is right, the summary Table 3 puts two incommensurable statistics in adjacent rows under the same label "prompt mismatch (action_primary)," and §4.3 defines mismatch once, in a single sentence, without saying how repeats are paired.

2.6 Slice dependence appears in the title but not in the controlled experiment. The evidence for slice dependence is §5.5: 8.3% in the severity=1 non-red-flag slice against 41.7% in the red-flag slice. Both figures come from runs whose decoding configuration was not recorded, at three repeats, and the comparison is confounded because the two slices differ in severity *and* in red-flag status at once, so the design cannot separate them. The controlled follow-up then dropped the severity=1 slice entirely and reran audit01_base24 (severity=2) against redflag24, where Table 1 gives 82.1% and 78.7% under det_t0: no slice difference at all. §6 handles this gracefully, saying that red-flag status is "one useful stratification but not the only driver," but a title-level claim now rests on the least controlled experiment in the paper while the best-controlled one fails to reproduce it. Rerunning the severity=1 slice under det_t0 would resolve this, and it is 24 pairs.

2.7 Table 4 sits oddly beside Table 1. In §5.11 the modal red-flag action is P1: A8 / P2: A8 for S008, P1: A8 / P2: A9 for S010, P1: A8 / P2: A8 for S018. Two of three displayed scenarios agree across prompts, in the slice where Table 1 reports 78.7% mismatch. That is the signature of the problem in 2.1: at the level of the item mode, the prompts largely agree, while individual draws disagree because the model is unstable within a prompt. The manuscript should draw that distinction itself.

2.8 The urgency scale is not an urgency scale. §4.6 and §5.8 treat A0 through A9 as an ordinal urgency score with A0=0 and A9=9. The codebook in §3.1 does not order that way in its middle. A4 is "consult family/friends", A5 is "search online for information", A3 is "consult a pharmacy/pharmacist". Searching

online is coded as more urgent than speaking to a pharmacist, which no triage model I know of would endorse; A4 and A5 are information-seeking modes rather than urgency levels at all. The reported drop of 1.29 points is driven by exactly these cells, since §5.8 lists A7 5→1, A5 6→3, A2 2→6, A3 1→3. The A5→A3 movement counts as a decrease of two urgency points purely because of where the codes were placed in the list. Either justify the ordering explicitly, or restate the result as a redistribution across categories. A minor bookkeeping point in the same passage: those four listed changes net to -1 item, so at least one more cell moved and is not reported, yet the JSD of 0.130 is computed over the full ten-bin distribution.

2.9 Pairs are treated as independent when they are clustered. audit01_base24 is 12 scenarios × 2 personas and redflag24 is 6 scenarios × 4 personas (§3.4). Every test in §4.6 uses the persona-scenario pair as the analysis unit, which gives n=24 in both slices, but the effective number of independent scenarios is 12 and 6. The Fisher test in §5.5 compares 2/24 against 10/24 as if 48 independent observations were in play. §7 concedes that the tests "do not eliminate all potential dependence structures"; that concession should become a cluster-aware analysis, or at minimum a report of how much of the mismatch variance is between scenarios rather than between personas.

2.10 One statistic does not reconcile. §5.8 reports the paired Wilcoxon test as W=0.0, p=0.00484. With W=0 all non-zero differences share a sign, so the exact two-sided p is 2×2^{-n} : 0.0078 at n=8, 0.0039 at n=9. The reported value falls between and matches neither, so it presumably comes from a normal approximation. On 24 items, a p of that size implies eight or nine pairs carry the whole result.

3. Reproducibility

This is better than most LLM papers I read, and the gaps that remain are specific. The model identifier is real (gemini-2.5-flash via Vertex AI, §3.5), the follow-up decoding parameters are complete down to max_output_tokens and the seed value 17, repeat counts are reported honestly as a range with the reason (§3.4), and personas, scenarios, and codebook are named as files with a supplementary archive attached.

The initial experiments are not reproducible and the paper says so. §3.5 states that for the Pilot, the audits, and the stress test, the generation configuration was not overridden and platform-level request logs were not retained, so provider defaults are used as a reference. That covers §5.1 through §5.6 and §5.8, which is most of the results and all of the slice-dependence evidence. The frank admission is to the author's credit; the consequence is that those numbers cannot be regenerated by anyone, including the author.

No model snapshot. `gemini-2.5-flash` is a floating alias, and §7 acknowledges that server-side revisions are not pinned. Since the experiments span two phases separated by a review round, the initial and follow-up runs may not have hit the same weights, which is an alternative explanation for the discontinuity in 2.5 and worth naming as such. I could find no run dates, which would at least bracket the served version.

The execution environment is undefined. "The Antigravity environment" appears in §3.5 with no citation, version, or description. This matters more than it looks: the Pilot's within-batch coupling, which is the paper's headline methodological warning, is a property of how requests were issued, and a reader cannot evaluate the coupling hypothesis without knowing what sat between the script and the API. §3.5 asserts "single-turn, stateless requests," but §5.3 shows that in the Pilot they were not effectively independent. The wrapper code, not just the prompt templates, is the object that needs publishing here.

Prompts are shown in translation only. §3.3 gives abbreviated P1 and P2 excerpts and notes the content was "translated from the synthetic Japanese inputs." The actual strings sent to the model were Japanese. Prompt language is not a formatting detail for a persona-conditioning study, and it is not discussed anywhere in §6 or §7. The exact original templates belong in an appendix, with the translations beside them.

Seeds, logs, and the archive. `stoch_default` uses "per-run seeds" (§3.4) whose values are not listed, so those runs cannot be repeated even in the best case. The supplementary archive has no version, no DOI, and no license. For a paper whose subject is reproducibility, an archived deposit with a fixed identifier would match its own recommendations. Nothing is said about the failure rate behind "occasional transient failures" or how failed repeats were handled, which matters when the denominator varies between 8 and 10.

Results that exist only as pictures. Figures 2, 3, 4, 5, 6, and 8 are not legible as rendered on Qeios, and several carry results that appear nowhere as numbers in the text. §5.9 is the worst case: it calls reason-tag OOV drift "a critical failure mode," describes a step change from 0% to "high," and gives not one rate, referring the reader to Figure 8. The same applies to the severity gradient of §5.1 and the per-field mismatch rates of Figure 5, which §4.3 says were computed for `action_time` and `escalation_if_worse` as well. The confidence field is collected in the schema (§3.2) and never analyzed at all.

4. What specifically to fix

1. **Add the same-prompt mismatch baseline to Table 1.** For each slice and each decoding configuration, compute P1-against-P1 and P2-against-P2 mismatch over independent repeat pairs, put those columns next to the P1-against-P2 column, and report the difference as the prompt effect. The repeats already exist. Until this is in the table, the abstract's "mean action mismatch 0.787–0.821" should not be described as prompt sensitivity, and my bound from Table 2 (same-prompt mismatch of at least 0.442 under `det_t0` in `audit01_base24`) suggests the corrected effect will be much smaller.
2. **§5.7 and Table 2: confront the fact that `det_t0` produced modal agreement of 0.558, 0.571, 0.529, and 0.551 rather than values near 1.0.** Report it as a result about the serving layer, check whether the configuration actually reached the API by logging the echoed request, and adjust the recommendation in §8 that a near-deterministic operating point supports "reproducible instrument checks".
3. **Recompute the JSD comparison in §5.7 at one aggregation level.** Either both quantities within a run or both pooled across runs, with the n used stated in the caption of Table 1.
4. **Give every JSD a null.** Permute labels within the pooled sample and report the permutation distribution for the 0.130 of §5.8 and the 0.148 to 0.196 of Table 1. My back-of-envelope null at $n=24$ is roughly $(k-1)/(4n)$ nats, which is 0.135 bits if all ten codes are occupied, so this is not a formality. Report the number of occupied codes per slice while you are there.
5. **§4.3: define the mismatch statistic precisely, including how repeats are paired.** Then reconcile the Audit01 row of Table 3 ($r3$ 45.8%) with the Table 1 rows on the same 24 pairs (82.1%, 81.9%), or relabel them as different measures. Right now, Table 3 presents both under one heading.
6. **Rerun the severity=1 non-red-flag slice under `det_t0` and `stoch_default`** so that the slice-dependence claim in the title rests on the controlled protocol rather than on §5.5, whose runs have no recorded decoding parameters. While revising §5.5, note that the two slices differ in severity and red-flag status simultaneously, so the contrast cannot isolate red-flag status.
7. **§5.8 and §4.6: justify the A0–A9 ordinal mapping or drop it.** A5 (search online) ranks above A3 (pharmacist) and A4 (family/friends) in the codebook of §3.1, and the reported -1.29 shift is built on movement through exactly those codes. If the ordering cannot be defended, report the redistribution and the JSD and leave the mean out.

8. **Publish the numbers behind Figure 8.** §5.9 makes a strong claim with no rate in the text. A short table of OOV percentage by run and prompt, with the denominator, would carry the section. Same treatment for Figure 5 (per-field mismatch, including `action_time` and `escalation_if_worse`) and Figure 4.
9. **Cluster-aware inference in §4.6.** With 12 scenarios in `audit01_base24` and 6 in `redflag24`, report scenario-clustered tests or a mixed model with a scenario random effect, and say what the effective sample size is.
10. **§5.3: rule out duplication before attributing 0/40 to autoregressive coupling.** Report whether the Pilot's identical outputs extended to `action_time`, `escalation_if_worse`, `reason_tags`, and the integer confidence. Exact agreement on a 0–100 confidence value across all 40 pairs would point at a pipeline bug rather than in-context anchoring, and the two have different implications for the paper's central warning.
11. **§3.5: describe the Antigravity environment, cite it, version it, release the wrapper.**
12. **Add an appendix with the original Japanese P1 and P2 templates in full, beside the translations,** plus the `stoch_default` seed values, the run dates, and the handling of incomplete repeats. Deposit the archive under a persistent identifier.
13. **§5.8: state the Wilcoxon variant and the count of non-zero differences.** $W=0.0$ with $p=0.00484$ matches neither $2 \times 2^{-8} = 0.0078$ nor $2 \times 2^{-9} = 0.0039$.
14. **Table 1: say what the bracketed intervals are.** "82.1% [77.2–87.0]" does not specify bootstrap, across-run range, or binomial, nor whether the resampling unit is the run or the pair.

5. Recommendation

Major revisions.

I am not asking for a different paper. The design is sound in outline, the pilot warning about within-batch pairing is a real contribution, the schema-drift observation is practically useful, and the author's willingness to run a new controlled experiment rather than defend the old one is the right instinct. What blocks acceptance is that the headline result, the 78.7–82.1% mismatch of Table 1, is reported without the within-condition comparison that would make it a measurement of prompt sensitivity rather than of model instability, and that the supporting JSD contrast in §5.7 is computed at two different sample sizes. Both are fixable from data already on disk, without new API calls. The temperature-0 behavior in Table 2 needs an honest reckoning, and the slice-dependence claim in the title needs the controlled rerun it

never received. If items 1 through 6 come back with the numbers, I expect the conclusions to survive in weakened and more defensible form, and the paper would then be a useful reference for anyone building synthetic-responder pipelines in health contexts.

Declarations

Potential competing interests: No potential competing interests to declare.