Qeios

Peer Review

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

Hwon Heo[1]

1. Asan Medical Center, Seoul, South Korea

Summary

This manuscript offers a timely and valuable methodological audit of LLM-based behavioral simulations. The slice-aware analysis—**particularly the focus on red-flag versus non-red-flag scenarios**—is a notable strength. However, the current experimental design makes it difficult to fully support the paper's main conclusions. The lack of strict control over the generation configuration—specifically stochastic settings (e.g., temperature, top-p) and platform-level behaviors like context caching—makes it impossible to confidently isolate the true source of the observed variability. Moreover, three runs per condition are insufficient to robustly estimate stability. **While the authors raise intriguing hypotheses, establishing them requires tighter experimental controls and a deeper engagement with the existing literature.**

Comment

The authors present a meaningful evaluation of how well LLMs can simulate lay health-seeking behavior. The structured action codebook and the distinction between red-flag and non-red-flag scenarios are excellent methodological choices. That said, I have several concerns regarding the experimental design that currently limit how the results can be interpreted.

My primary concern lies in the lack of clarity and control over the model's generation configuration. Since the study's central objective is to evaluate prompt sensitivity and run-to-run variability, the

 doi.org/10.32388/CO4PG1

parameters that dictate model output must be strictly controlled and reported. For a model like Gemini 2.5 Flash, this means explicit management of stochastic variables (e.g., temperature, top-p) and platform-level states. The authors acknowledge that Vertex AI defaults (temperature 1.0, topP 0.95) were likely used, but without locking these down or recording the exact logs, we cannot know whether the observed variance stems genuinely from prompt framing or simply from the model's inherent decoding noise. To disentangle these effects, the model should be evaluated under both a near-deterministic setting (e.g., temperature = 0) and a default stochastic setting.

Additionally, relying on only three repetitions (runs) per condition is too low to reliably estimate run-to-run variability. Robustly characterizing the stability of these effects requires a higher number of repeats.

The manuscript also needs to be better positioned within the broader, emerging literature on using LLMs as synthetic respondents. Studies by Tjuatja et al. (2024)[1], Strachan et al. (2024)[2], and Kosinski (2024)[3] have already investigated how LLMs diverge from human distributions and respond to prompt variations. The authors need to explicitly map out how their methodology and findings build upon or differ from this prior work.

Regarding the hypothesis that within-batch prompt comparisons may underestimate prompt sensitivity—this is a very interesting point. However, the current experiments do not isolate the actual mechanism at play. To substantiate this claim, the study shouldn't just point to future work; it needs to actively control for batching procedures, context reuse, or implicit platform-level caching to prove what is actually driving the perfect agreement seen in the pilot.

Finally, for any methodological audit, full transparency is non-negotiable. **Reproducibility requires exact model versions, decoding parameters, and any platform-specific settings to be fully documented.**

Overall, this study asks excellent methodological questions, but the current experimental controls are not quite there yet. Addressing the confounders around generation parameters, increasing the run counts, and clarifying the paper's contribution relative to existing literature will be necessary steps for revision.

References

1. ^Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, et al. (2024). Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. doi:10.1162/tacl_a_00685.
2. ^James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, et al. (2024). Testing theory of mind in large language models and humans. Nat Hum Behav, vol. 8 (7), 1285-1295. doi:10.1038/s41562-024-01882-z.
3. ^Michal Kosinski. (2024). Evaluating large language models in theory of mind tasks. Proc. Natl. Acad. Sci. U. S.A., vol. 121 (45). doi:10.1073/pnas.2405460121.

Declarations

Potential competing interests: No potential competing interests to declare.