Qeios

Peer Review

Review of: "Balancing Safety and Educational Availability in a Large Language Model-Based Virtual Patient for Medical Interview Training: Robustness Evaluation Under Direct and Indirect Instructional Contamination"

Iliano Fasolino[1]

1. Department of Computer Science, University of Milan, Italy

This paper asks a question that robustness work in my own area tends to skip: whether a virtual patient that stays safe under contamination is still usable for practice. Counting refusal and lost information recovery as failures, next to leakage and harm, is a good evaluation choice, and the design is clean: five fixed cases, a fixed 10-turn protocol, six matched conditions with 100 episodes each, and a primary metric that drops the scripted greeting so the score reflects questioning. The indirect contamination result in Table 2, with slot F1 excl. init falling from 0.901 to 0.489 and refusal rising to 0.047, is striking. I read the paper from the computer science side, robustness and evaluation validity, and leave clinical and educational judgments to reviewers with that background.

Sulabha Asha has already raised the single model, the unexplained 0.927 under direct contamination, case validation, the wording of the defense conclusion, and the 0.978 appendix baseline, and the author has replied to each with planned revisions. I do not repeat them; where I touch on them, I add a check the existing logs could answer.

My main concern is that each channel is tested at a single level. Table 1 and Appendix 2 show one representative string per channel, and the paper does not say how many templates were used (Data Availability mentions "attack templates" in Multimedia Appendix 1); please list every template used per channel. Nor does it vary how strong the note is, where it sits, or what share of turns is contaminated.

doi.org/10.32388/L165N4

With one level per channel, the paper has a single data point for each, not a dose response, so 0.489 shows that this note hurts this model in this setup and says nothing about how degradation scales. In batched mode, I also could not tell whether the note is appended once per episode or after each of the 10 questions, which changes the effective dose by a factor of ten. A small curve would help: contaminate 1, 3, 5, and 10 of the turns, and vary how authoritative the note sounds.

My second concern is what "indirect" means here. Both perturbations are appended after the learner turn, and no retrieval step is described; the indirect note differs from the direct suffix mainly in its label ("Reference note:") and its quality assurance framing. The Discussion concludes that "the contamination channel matters." I think the safer reading is that framing matters within one input channel. Placing the note in a separate context field, as a retrieval pipeline would, is what would test the channel claim.

Third, what do the failed answers look like? Under indirect contamination, the refusal rate is 0.047 and the clarification rate is 0.000, yet slot F1 excl. init falls by 45.7% relative to clean (0.901 to 0.489), so refusals account for little of the loss. The note asks for the hidden diagnosis, yet no forbidden leakage or harm was detected. Were the failed answers out of role, misaligned with the numbered questions, generic, or phrased in a way the evidence patterns miss? Leakage, contradiction, and harm are "deterministic counts" (Appendix 3) with no reported sensitivity check, so zero events means zero rule matches. Asha's summary already called the slot scoring unvalidated; a hand-labelled sample of indirect episodes would check the safety counters and the F1 drop together. The same logs could test the author's reply to Asha on the 0.927 result: if the model really shifted into a terser extraction mode, the mean answer length per condition should show it.

Fourth, the sanitizer. It targets "explicit note delimiters (e.g., 'Reference note:')", "ignore previous instructions", and "you are not a patient", which are the markers in the representative strings of Appendix 2, so for those strings, recovery to 0.903 is expected by construction. If other templates were used, please report whether each carried one of these markers. The revised wording promised in the reply ("demonstrated the potential to restore performance") fixes the tone Asha objected to, but the evaluation still tests only the attack the defense was written against. A paraphrased note with no delimiter would show what survives an attacker who knows the rules. The paper does not say whether the strings were in English or Japanese, which matters for a string matcher. There is no clean plus defense condition, so the sanitizer's cost on benign questions is not measured; Table A1 uses heuristic risk scores, not this sanitizer.

Statistically, the 100 episodes are 5 cases x 20 repeats at temperature 0.2, and Wald intervals treat them as independent. The effective unit is closer to five cases, and the Wald interval behaves poorly near zero, as the lower bound of -0.001 for refusal under direct contamination shows. Per-case results with a case-clustered bootstrap would be safer, and separately, Wilson rather than Wald intervals for rates near zero. The standard deviations that Statistical Analysis promises do not appear in Table 2. No model snapshot or run date is given, although Appendix 5 calls gpt-4o-mini a "fixed model". In Table A1, the baseline F1 is 0.978 and the baseline refusal is 0.000 in every row, indirect injection included, against a refusal rate of 0.047 for indirect contamination in Table 2. The reply says the replayed logs were "primarily a subset of clean, successful interactions"; if so, the indirect rows are not replays of the primary indirect episodes, and the appendix should say what was replayed. Finally, Methods says the full prompt template is in the appendix, while Appendix 1 gives "representative prompt fragments".

Of these requests, the dose curve and the paraphrased attack matter most to me because they decide whether 0.489 and 0.903 hold beyond the strings shown.

Declarations

Potential competing interests: No potential competing interests to declare.