Can You Spot the Virtual Patient (VP)?

Reyhaneh Hosseinpourkhoshkbhari REYHANEH.HOSSEINPOUR@UTDALLAS.EDU
Wei-chen Huang WEI-CHEN.HUANG@UTDALLAS.EDU
Suvel Muttreja SXM200218@UTDALLAS.EDU
Richard M. Golden GOLDEN@UTDALLAS.EDU
School of Behavioral and Brain Sciences, University of Texas at Dallas, USA

Abstract

Communication is a critical clinical skill, yet scalable, realistic training tools remain limited. Large language model (LLM)-based virtual patients (VPs) offer a promising alternative to traditional tools, but their conversational realism remains underexplored. In this study, we evaluate the realism of GPT-4o-generated VPs using a multi-method approach: expert review, Turing-style testing, linguistic analysis, and semantic similarity. We generated 44 VPs based on real doctor–patient dialogues. Expert annotations of hallucinations, omissions, and repetitions showed high interrater reliability ($ICC > 0.77$). In a Turing test, participants struggled to distinguish VPs from real patients—classification accuracy fell below chance. Linguistic analysis of 2,000+ dialogue turns revealed that VPs produced formal, lexically consistent responses, while human patients showed more emotional and stylistic variability. Semantic similarity scores averaged 0.871 (response-level) and 0.842 (transcript-level), indicating strong alignment. These findings support the use of LLM-based VPs in communication training and offer insights into realism, trust, and refinement, contributing to the safe and responsible deployment of generative AI in healthcare.

Keywords: Generative AI, medical education, virtual patient, realism evaluation.

Data and Code Availability The Fareez et al. (2022) dataset, which was used to develop the illness scripts, consists of transcripts from simulated conversations between senior Canadian medical students (acting as "doctors") and resident doctors (acting as "patients"). The code is available at https://github.com/Reyhanehrhp7/Virtual-Patients-Using-Open-AI-API-2024.

Institutional Review Board (IRB) The protocol for this study was reviewed and approved by the Institutional Review Board at the University of Texas at Dallas (IRB-24-871). All participants provided informed consent prior to participation. The data used were fully de-identified.

1. Introduction

Clinical communication is a fundamental competency in medical education, typically taught alongside history taking, physical examination, and clinical reasoning (Hosseinpourkhoshkbari and Golden, 2025). Simulated environments play a critical role in developing these skills by allowing learners to engage with simulated patients in realistic scenarios (ten Cate and Durning, 2018). Standardized patients (SPs), trained actors who portray clinical cases, are widely used but costly and difficult to scale (Johnson et al., 2020). Recent advances in LLMs enable scalable, AI-driven virtual patients (VPs). These systems can simulate rich clinical dialogues while providing flexible, low-risk learning opportunities. However, the realism and reliability of LLM-based VPs in communication training remain underexplored.

In this study, we propose a framework for evaluating the realism of GPT-4o-generated VPs. Our contributions include the following:

- **Structured Generation:** We generated 44 VPs using illness-script prompts spanning 17 categories.
- **Expert Review:** Two clinicians annotate 1,094 conversation turns for omissions, hallucinations, and repetitions, achieving high interrater reliability.

. Further details are provided in our longer non-archival version, accepted as a poster at the Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance (Hosseinpourkhoshkbari et al., 2025).

- **Turing Test Evaluation:** We assess whether participants can distinguish VP responses from real patient responses, with and without diagnostic hints.
- **Linguistic and Syntactic Analysis:** We compare lexical diversity and part-of-speech distributions between human and VP responses.
- **Semantic Similarity Assessment:** We use BioClinicalBERT embeddings to quantify alignment between VP and human responses at both turn and transcript levels.

Together, these methods offer a reproducible framework to evaluate LLM-based VPs for communication training in medical education.

2. Methodology

An overview of the methodological framework is presented in Figure 1.

2.1. Dataset

The Fareez et al. (2022) dataset, which was used to develop the illness scripts, consists of transcripts from simulated conversations between senior Canadian medical students (acting as "doctors") and resident doctors (acting as "patients"). For this study, a subset of 44 transcripts specifically related to respiratory cases was used. These transcripts consisted of a total of 2,139 question–answer pairs.

2.2. Development of the VP

We used OpenAI's ChatGPT-4o via the API with default parameters (temperature = 1.0, top-p = 1.0, presence penalty = 0, frequency penalty = 0) to generate 44 VPs. A zero-shot prompting strategy was employed, where the model was given structured illness scripts that defined the task without specific examples (Radford et al., 2019). These scripts, derived from the Fareez et al. (2022) dataset, captured key patient details such as chief complaint, medical and social history, and relevant symptoms. When original transcripts lacked information (e.g., name, age, occupation), fields were labeled "unknown to the transcript." The model was prompted to generate plausible responses for missing information to maintain realism and avoid reverting to a default ChatGPT persona (Grévisse, 2024). VP interactions were initiated using the original physician utterances from the source transcripts.

2.3. Evaluation of GPT-4o Virtual Patients

We evaluated GPT-4o-generated VP using four complementary methods:

Expert Review. Two clinical educators developed and applied a coding scheme to assess the quality of VP responses across five dimensions: omissions (missing expected information), inappropriate repetitions, hallucinations (factually incorrect content), successful turns (accurate and context-appropriate responses), and total conversational turns. These categories reflect known challenges in VP realism (Holderried et al., 2024). Interrater reliability was assessed using intraclass correlation coefficients (ICC).

Turing Test. A 20-item survey (10 human, 10 VP responses) was administered to 50 psychology undergraduates to assess whether VP utterances were distinguishable from real patients (Turing, 1950; Rathi et al., 2024). Participants identified the response source, rated their confidence, and had reaction times logged. A between-subjects design tested the effect of a diagnostic hint (n = 25 per group) (see the hint in Box 1). These measures enabled analysis of classification accuracy, task sensitivity, and metacognitive certainty.

Box 1: HINT provided to human participants

Keep in mind that computer-generated responses tend to be more formal and structured than human responses. For example, the computer will tend to avoid filler words such as "Um" and "Ah". The computer also will tend to avoid repeating words. Additionally, computer responses may sometimes be a bit longer and more detailed compared to those from humans.

Linguistic Analysis. Lexical richness and syntactic structure were analyzed using standard metrics from the `LexicalRichness` Python library, including Type-Token Ratio (TTR), Root TTR (RTTR), and the Maas index (Van Hout and Vermeer, 2007; Tweedie and Baayen, 1998). While TTR measures the ratio of unique words to total words, it is sensitive to text length, especially in short or highly variable responses (Muñoz-Ortiz et al., 2024). RTTR and Maas mitigate this issue by normalizing for response length. To further reduce sensitivity to length and capture deeper vocabulary variation, we included

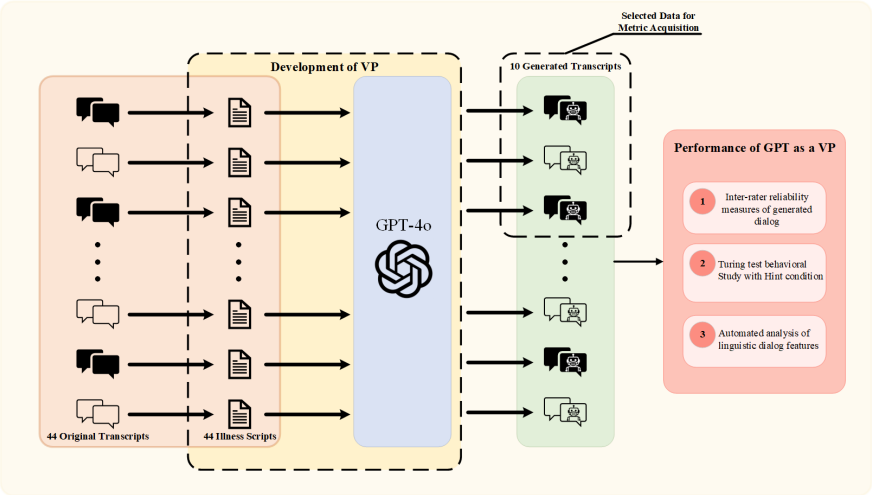

Figure 1: Overview of the VP development and evaluation pipeline.

the Measure of Textual Lexical Diversity (MTLD) and Hypergeometric Distribution Diversity (HDD) (McCarthy and Jarvis, 2010), both of which are designed to handle variable-length texts more robustly. We also computed Yule's K, which quantifies lexical concentration and is unaffected by text length, and the Moving Average TTR (MATTR), which analyzes diversity over sliding word windows (Covington and McFall, 2010). These Metrics were applied to both full transcripts ($n = 44$) and individual utterances ($n = 2{,}194$). Part-of-speech (POS) tagging was used to compare syntactic categories—such as nouns, verbs, and adverbs—across human and VP responses (Kumawat and Jain, 2015).

Semantic Similarity. Cosine similarity between VP and human responses was computed using BioClinicalBERT embeddings, which capture contextual, sentence-level semantics (Alsentzer et al., 2019), enabling assessment of semantic similarity in a scalable, annotation-free manner (Reimers and Gurevych, 2019).

3. Results

3.1. Expert evaluation

As shown in Table 1, interrater reliability was at least *good* for all three error types, hallucinations, omissions, and repetitions, with ICC analyses indicating *excellent* agreement for hallucinations and *good to excellent* agreement for omissions and repetitions. Across the 44 VP transcripts, the average rate of successfully completed conversation turns was 96.6% (range: 90–100%), reflecting the overall relevance and coherence of VP responses. Hallucinations were the most frequent error, occurring in 2.8% of turns on average, whereas omissions (0.31%) and repetitions (0.26%) were rare. While hallucinations occurred more often than other errors, their low absolute frequency indicates that VP responses were generally accurate, contextually appropriate, and reliable.

Metric	ICC	95% CI	Interpretation
Hallucination	0.814	[0.603, 0.920]	Excellent
Omissions	0.783	[0.544, 0.905]	Good–Excellent
Repetition	0.774	[0.528, 0.901]	Good

Table 1: Inter-rater reliability (ICC) for three evaluation metrics.

3.2. Turing test results

A two-way ANOVA examined the effects of *hint* (with vs. without) and *dialog type* (human vs. VP) on participants' accuracy in identifying dialog sources. Participants more accurately classified human dialogs (M = 81.2%) than VP dialogs (M = 42%), $F(1, 90) = 8.69$, $p = .004$. Those who received a hint performed better overall (M = 68%) than those without (M = 55%), $F(1, 90) = 77.34$, $p < .001$. The interaction was not significant, $F(1, 90) = 2.05$, $p = .156$.

As shown in Figure 2 (left), hints improved classification accuracy for both VP (M = 5.17, SE = 0.44, CI [4.29, 6.04]) and human dialogs (M = 8.46, SE = 0.44, CI [7.58, 9.33]), compared to the no-hint condition (VP: M = 3.22, SE = 0.45, CI [2.32, 4.11]; human: M = 7.78, SE = 0.45, CI [6.89, 8.68]). Pairwise comparisons showed that hints significantly improved VP classification ($p = .0026$) but did not affect human dialog accuracy ($p = .2861$). Overall, when participants were provided with hints, overall classification accuracy increased from 55.0% to 68.3%, primarily due to improved VP detection. However, nearly half of VP responses were still judged as human even with hints, suggesting a high degree of realism in VP-generated responses.

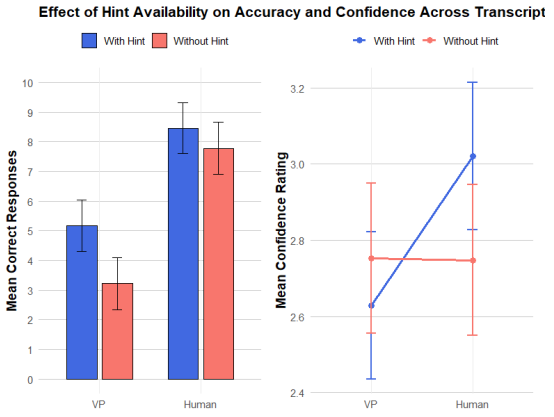

Figure 2: Participants' accuracy and confidence across VP and human transcripts, with and without hints.

A second two-way ANOVA on confidence ratings showed a main effect of dialog type, $F(1, 90) = 4.04$, $p = .0475$, and an interaction with hint condition, $F(1, 90) = 4.04$, $p = .0474$. However, the main effect of the hint condition alone was not significant, $F(1, 90) = 0.58$, $p = .4483$. With a hint, participants were more confident judging human dialogs than VP dialogs ($p = .0055$); no such difference appeared without a hint ($p = .9754$) (see Figure 2, right). Between-group comparisons showed no significant difference in confidence for VP dialogs ($p = .3795$), but there was a marginal trend toward higher confidence in human dialog classification when a hint was provided ($p = .0531$).

Response times were log-transformed to reduce skewness. As shown in Table 2, participants in the no-hint group responded slightly slower, but differences across hint and dialog type were not significant (all $p > .05$). Thus, response latency likely did not confound accuracy or confidence. Notably, average classification accuracy for VP responses remained below chance (< 50%), suggesting they were often indistinguishable from real patients—meeting the classical Turing Test criterion.

Table 2: Log-Time analysis by dialog type and hint condition. SE = standard error; CI = confidence interval; $t(90)$ = t-statistic with 90 degrees of freedom. No significant differences were observed.

Hint	Dialog	Log-Time	SE	95% CI
With	VP	3.13	0.164	[2.80, 3.45]
Without	VP	3.36	0.168	[3.03, 3.69]
With	Human	3.05	0.164	[2.72, 3.37]
Without	Human	3.23	0.168	[2.89, 3.56]

3.3. Linguistic characteristics of VP and human simulated patient dialog turns

Across 44 conversations, we recorded 2,194 turns (question–answer pairs). As shown in Figure 3, human responses were slightly longer (mean = 1.36 sentences) than VP responses (mean = 1.23), though both typically gave single-sentence replies (median = 1). VPs tended to over-respond when humans were brief and under-respond when humans were verbose.

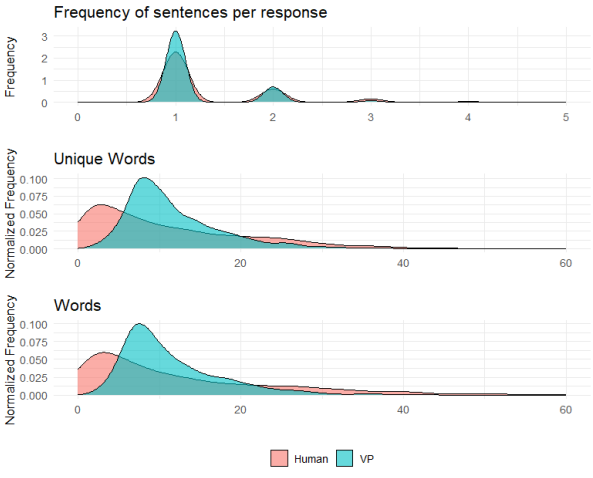

Figure 3: Distributions of sentence count, unique words, and total word count per turn.

Kolmogorov–Smirnov tests revealed significant differences in lexical variety ($D = 0.30$, $p < .001$), word count ($D = 0.29$, $p < .001$), and sentence count ($D = 0.06$, $p < .01$). Despite using fewer words overall, VP responses were more lexically diverse. Lower Yule's K scores and higher MTLD, MATTR, CTTR, and HDD values indicate less repetition and more efficient vocabulary. These results held at the individual-response level.

Part-of-speech comparisons showed humans used more nouns, pronouns, interjections, and adverbs ($p < .001$), while VPs used more adpositions ($p < .001$). Particles were also more common in human speech ($p = .045$). No significant differences were found in verbs, auxiliaries, adjectives, conjunctions, or proper nouns ($p > .05$), suggesting grammatical similarity overall. However, reduced interjections and adverbs in VP output highlight gaps in emotional and pragmatic realism.

3.4. Semantic Similarity Analysis

Semantic overlap between VP and human responses was assessed using cosine similarity of BioClinicalBERT embeddings at two level:

- **Response level.** Cosine similarity was computed turn-by-turn and averaged per transcript. The mean similarity was 0.871 (SD = 0.13), indicating strong local alignment.
- **Transcript level.** All patient responses were concatenated, and similarity was computed between full VP and human transcripts. The mean was 0.842 (SD = 0.045), reflecting consistent global semantic overlap.

4. Conclusion

This study presents a multi-method evaluation framework for assessing the realism of GPT-4o-generated VP in clinical communication training. Findings indicate strong linguistic and semantic alignment with human responses and high expert-rated quality, despite occasional hallucinations. While results highlight strong linguistic and semantic alignment with human responses, limitations in domain scope and interaction depth remain. Future work should explore broader clinical contexts, behavioral cues, and alternative prompting strategies to enhance realism and generalizability.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Michael A Covington and Joe D McFall. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100, 2010.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1): 313, 2022.

Christian Grévisse. Raspatient pi: A low-cost customizable llm-based virtual standardized patient simulator. In *International Conference on Applied Informatics*, pages 125–137. Springer, 2024.

Friederike Holderried, Christian Stegemann-Philipps, Lea Herschbach, Julia-Astrid Moldt, Andrew Nevins, Jan Griewatz, Martin Holderried, Anne Herrmann-Werner, Teresa Festl-Wietek, Moritz Mahling, et al. A generative pretrained transformer (gpt)–powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. *JMIR medical education*, 10(1):e53961, 2024.

Reyhaneh Hosseinpourkhoshkbari and Richard M. Golden. Clinical reasoning assessment methods: Current simulated environment practice and future prospects using AI and Psychometrics. *Available at SSRN 5400917*, 2025.

Reyhaneh Hosseinpourkhoshkbari, Wei chen Huang, Suvel Muttreja, and Richard M. Golden. Can you spot the virtual patient? Expert Review, Turing Test, and Linguistic–Semantic Analysis. In *First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models*, 2025. URL https://openreview.net/forum?id=uWiOA7p4Wt. 18 pages.

Kelly V Johnson, Allison L Scott, and Lisa Franks. Impact of standardized patients on first semester nursing students self-confidence, satisfaction, and

communication in a simulated clinical case. *SAGE open nursing*, 6:2377960820930153, 2020.

Deepika Kumawat and Vinesh Jain. Pos tagging approaches: A comparison. *International Journal of Computer Applications*, 118(6), 2015.

Philip M McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392, 2010.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ishika Rathi, Sydney Taylor, Benjamin K Bergen, and Cameron R Jones. Gpt-4 is judged more human than humans in displaced and inverted turing tests. *arXiv preprint arXiv:2407.08853*, 2024.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Olle ten Cate and Steven J Durning. Approaches to assessing the clinical reasoning of preclinical students. *Principles and practice of case-based clinical reasoning education*, page 65, 2018.

Alan M Turing. *Computing machinery and intelligence*. Springer, 1950.

Fiona J Tweedie and R Harald Baayen. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32: 323–352, 1998.

RWNM Van Hout and AR Vermeer. *Comparing measures of lexical richness*. Cambridge University Press Cambridge, 2007.