Can You Spot the Virtual Patient? Expert Review, Turing Test, and Linguistic–Semantic Analysis

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Large language models (LLMs) are increasingly used to simulate complex social and cognitive tasks, yet the behavioral regularities and heuristics they employ remain underexplored. In this study, we investigate GPT-4o's behavioral patterns when performing the cognitively demanding task of Virtual patients (VPs) in clinical interviews. VPs offer a promising alternative to traditional tools, but their conversational realism remains underexplored. Using 44 structured illness-script prompts spanning 17 clinical categories, we analyze the model's output through expert review, Turing-style discrimination testing, linguistic profiling, and semantic similarity analysis. 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. BioClinicalBERT-based semantic similarity scores averaged 0.871 (response-level) and 0.842 (transcript-level), indicating strong alignment. This behavioral characterization contributes to understanding how LLMs generalize to cognitively complex, open-ended interaction tasks and provides a reproducible evaluation framework for studying model behaviors in socially and domain-specific contexts.

1 Introduction

Clinical communication is a fundamental competency in medical education, typically taught alongside history taking, physical examination, and clinical reasoning [2]. Simulated environments play a critical role in developing these skills by allowing learners to engage with simulated patients in realistic scenarios [6]. Standardized patients (SPs), trained actors who portray clinical cases, are widely used but costly and difficult to scale [15].

Recent advances in LLM enable scalable, AI-driven VPs that can simulate rich clinical dialogues while providing flexible, low-risk learning opportunities (see Appendix A for related work). While these systems offer flexible, low-risk learning opportunities, little is known about the underlying cognitive behaviors they exhibit—such as conversational strategies, stylistic biases, and heuristics—when performing complex interactions.

In this study, we propose a framework for examining the cognitive behaviors of GPT-4o-generated VPs during complex, domain-specific dialogue. We treat clinical patient simulation not only as a test of conversational realism, but as a behavioral probe to identify the model's conversational strategies, stylistic biases, and heuristics. Our contributions include:

- **Structured Generation:** Generation of 44 VPs from illness-script prompts spanning 17 categories, enabling controlled analysis of model behavior.

- **Expert Review:** Annotation of 1,094 conversation turns for omissions, hallucinations, and repetitions, providing quantitative measures of accuracy and stability.
- **Turing Test Evaluation:** Behavioral discrimination task testing whether participants can distinguish VP responses from human responses, with and without diagnostic hints.
- **Linguistic and Syntactic Analysis:** Profiling of lexical diversity and part-of-speech distributions to uncover systematic stylistic patterns in VP output.
- **Semantic Similarity Assessment:** Measurement of semantic alignment between VP and human responses at both turn and transcript levels using BioClinicalBERT embeddings.

Together, these analyses form a reproducible framework for studying the cognitive behaviors, biases, and generalization strategies of LLM-based VPs in high-stakes, socially and cognitively demanding interactions.

2 Methodology

An overview of the methodological framework is presented in Appendix C (Figure 2).

2.1 Dataset

The Fareez et al. [10] corpus comprises transcripts of simulated encounters between senior Canadian medical students (“doctors”) and resident physicians (“patients”). For this study, we used a subset of 44 respiratory-case transcripts, totaling 2,139 question–answer pairs, to develop the illness scripts.

2.2 Development of the VP

We generated 44 VPs using OpenAI’s ChatGPT-4o (API defaults: temperature = 1.0, top-p = 1.0, presence penalty = 0, frequency penalty = 0) with a zero-shot prompt containing structured illness scripts outlining chief complaint, patient history, symptoms, and additional relevant clinical information. [24] (see Appendix B, Box 2). Missing fields (e.g., name, age) were marked “unknown to the transcript,” prompting the model to produce plausible responses to preserve realism and avoid default persona reversion [11]. VP dialogues used original physician utterances from source transcripts, with examples in Appendix B.

2.3 Evaluation of GPT-4o Virtual Patients

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

Expert Review. Two clinical educators coded VP responses for omissions (missing expected information), inappropriate repetitions, hallucinations (factually incorrect content), successful turns (accurate and context-appropriate responses), and total conversational turns—dimensions reflecting known challenges in VP realism [12]. Interrater reliability, assessed via intraclass correlation coefficients (ICC) (see Appendix 4 for examples). This evaluation captures not only error prevalence but also behavioral tendencies, such as a bias toward adding plausible details over omitting information.

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 human patients [28, 25]. 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 Appendix B, Box 1). These measures enabled analysis of classification accuracy, task sensitivity, and metacognitive certainty.

Linguistic Analysis. Lexical richness and syntactic structure were analyzed using standard metrics from the `LexicalRichness` Python library, including Type-Token Ratio (TTR), Root Type-Token Ratio (RTTR), Maas index, Measure of Textual Lexical Diversity (MTLD), Hypergeometric Distribution Diversity (HDD), Yule’s K, and Moving Average Type-Token Ratio (MATTR) [7]. Many of the metrics for measuring lexical richness, such as MTLD, Yule’s K, and HDD, relate directly to psycholinguistic models of verbal fluency and lexical access. For example, MTLD has been used in cognitive aging research to show how executive function limits vocabulary diversity [16]. In contrast, Yule’s K measures repetitiveness, which can indicate less cognitive flexibility or a rigid style. Using these metrics on VP outputs offers a measurable way to see how LLMs mimic not only surface language but also cognitive patterns tied to the complexity of conversation [16]. 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 [17]. Additional methodological details are provided in Appendix A.

Semantic Similarity. Cosine similarity between VP and human responses was computed using Bio-ClinicalBERT embeddings [1], enabling assessment of semantic similarity in a scalable, annotation-free manner [26]. Higher similarity scores suggest greater similarity with human-like conceptual understanding, coherence, and reasoning, key indicators of cognitively realistic behavior in virtual patients.

3 Results

3.1 Expert evaluation

As shown in Table 1, interrater reliability was at least *good* for all error types, with *excellent* agreement for hallucinations and *good to excellent* for omissions and repetitions. Across 44 VP transcripts, 96.6% of conversation turns were successfully completed, indicating high relevance and coherence. Hallucinations were the most frequent error (2.8%), while 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 tested the effects of *hint* (with vs. without) and *dialog type* (human vs. VP) on classification accuracy. Participants classified human dialogs (M = 81.2%) more accurately than VP dialogs (M = 42%), $F(1, 90) = 8.69$, $p = .004$. Accuracy was higher with a hint (M = 68%) than 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 1 (left), hints improved 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 no-hint (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 tests showed a significant effect of hints for VP dialogs ($p = .0026$) but not for human dialogs ($p = .2861$). VP responses were often misclassified as human (with hints: 51.7%; without: 32.2%), indicating that GPT-4o-generated VPs can produce highly realistic dialogue (see confusion matrix in Appendix C, Table 6).

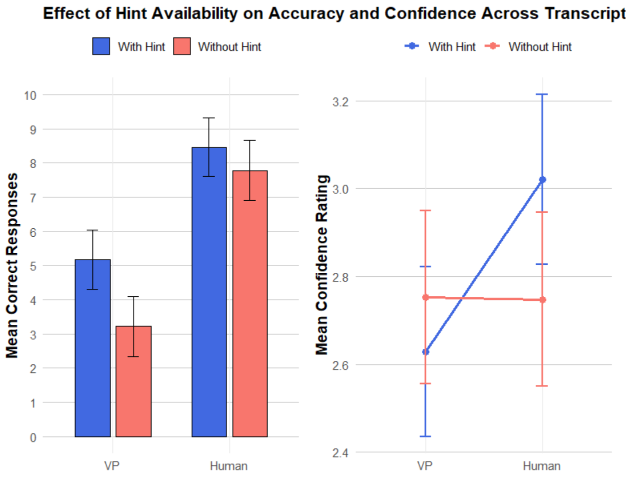

Figure 1: Participants' accuracy and confidence across VP and human transcripts, with and without hints.

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, $F(1,90) = 4.04$, $p = .0474$. The main effect of hint alone was not significant, $F(1,90) = 0.58$, $p = .4483$. With a hint, participants were more confident judging human dialogs than VP dialogs ($p = .0055$), but no such difference appeared without a hint ($p = .9754$) (Figure 1, right). Between groups, VP dialog confidence did not differ significantly ($p = .3795$), while human dialog confidence showed a marginal increase with hints ($p = .0531$).

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]

Response times were log-transformed to reduce skewness. As shown in Table 2, participants without hints responded slightly slower, but differences across hint and dialog type were not significant (all $p > .05$), suggesting response latency did not affect accuracy or confidence. Average VP classification accuracy was below chance ($< 50\%$), indicating VP responses were often indistinguishable from real patients and met the classical Turing Test criterion. Full ANOVA results are in Appendix C (Tables 3–5).

3.3 Linguistic characteristics of VP and human simulated patient dialog turns

Across 44 conversations, we recorded 2,194 turns (question–answer pairs). Human responses were slightly longer (mean = 1.36 sentences) than VP responses (mean = 1.23), though both typically gave single-sentence replies (median = 1) (see Appendix D, Figure 6). VPs tended to over-respond when humans were brief and under-respond when humans were verbose (see Figure 3 in Appendix C).

Kolmogorov–Smirnov tests revealed significant differences in lexical variety ($D = 0.30$, $p < .001$), word count ($D = 0.29$, $p < .001$), and sentence count ($D = 0.06$, $p < .01$). These results suggest that GPT-4o's VP responses, though shorter, exhibit lexical diversity patterns—such as lower repetition (Yule's K) and higher fluency (MTLD) which mirror cognitively efficient and controlled language production, as observed in psycholinguistic studies of verbal ability and aging (Appendix D, Table 7). These results were calculated at the individual-response level (Appendix D, Figure 5).

Part-of-speech comparisons (see Appendix D; Table 8; Figure 4) 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 differences emerged for verbs, auxiliaries, adjectives, conjunctions, or proper nouns ($p > .05$), suggesting similar grammar overall. Fewer interjections and adverbs in VP output point to gaps in emotional and pragmatic realism.

3.4 Semantic Similarity Analysis

Semantic overlap between VP and human responses was measured at two levels:

- **Response level:** Turn-by-turn cosine similarity, averaged per transcript, was 0.871 (SD = 0.13), showing strong local alignment.
- **Transcript level:** Concatenated patient responses yielded a similarity of 0.842 (SD = 0.045), indicating consistent global overlap.

4 Conclusion

This study presents a multi-method framework for examining the behavioral characteristics of GPT-4o-generated VP in clinical communication training. Results indicate strong similarity with human responses in both linguistic and semantic dimensions, suggesting that VPs can simulate key aspects of human-like reasoning and verbal expression. Expert evaluations also rated the quality of VP responses highly, despite occasional hallucinations. However, limitations in domain scope and interaction depth suggest that the cognitive realism of VPs may still be constrained. Future work should explore broader clinical contexts, behavioral cues, and alternative prompting strategies to enhance realism and generalizability.

References

[1] 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.

[2] Leen Aper, Wemke Veldhuijzen, Tim Dornan, Monica van de Ridder, Sebastiaan Koole, Anselme Derese, and Jan Reniers. "should i prioritize medical problem solving or attentive listening?": The dilemmas and challenges that medical students experience when learning to conduct consultations. *Patient education and counseling*, 98(1):77–84, 2015.

[3] Alexandra Aster, Sophia Viktoria Ragaller, Tobias Raupach, and Ambra Marx. Chatgpt as a virtual patient: written empathic expressions during medical history taking. *Medical Science Educator*, pages 1–10, 2025.

[4] Alexander Borg, Benjamin Jobs, Viking Huss, Cidem Gentline, Fabricio Espinosa, Mini Ruiz, Samuel Edelbring, Carina Georg, Gabriel Skantze, and Ioannis Parodis. Enhancing clinical reasoning skills for medical students: a qualitative comparison of llm-powered social robotic versus computer-based virtual patients within rheumatology. *Rheumatology International*, 44 (12):3041–3051, 2024.

[5] Alexander Borg, Ioannis Parodis, and Gabriel Skantze. Creating virtual patients using robots and large language models: a preliminary study with medical students. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 273–277, 2024.

[6] Th. J. ten Cate and Steven J Durning. Approaches to assessing the clinical reasoning of preclinical students. *Principles and practice of case-based clinical reasoning education: a method for Preclinical Students*, pages 65–72, 2018.

[7] 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.

[8] Douglass Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. A practical part-of-speech tagger. In *Third conference on applied natural language processing*, pages 133–140, 1992.

[9] M Divyapushpalakshmi and R Ramalakshmi. An efficient sentimental analysis using hybrid deep learning and optimization technique for twitter using parts of speech (pos) tagging. *International Journal of Speech Technology*, 24(2):329–339, 2021.

[10] 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.

[11] 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.

[12] 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.

[13] Bahar Irfan, Sanna Kuoppamäki, Aida Hosseini, and Gabriel Skantze. Between reality and delusion: challenges of applying large language models to companion robots for open-domain dialogues with older adults. *Autonomous Robots*, 49(1):9, 2025.

[14] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[15] 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.

[16] Susan Kemper and Aaron Sumner. The structure of verbal abilities in young and older adults. *Psychology and aging*, 16(2):312, 2001.

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

[18] Yanzeng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. Leveraging large language model as simulated patients for clinical education. *arXiv preprint arXiv:2404.13066*, 2024.

[19] David Malvern, Brian Richards, Ngoni Chipere, and Pilar Durán. *Lexical diversity and language development*. Springer, 2004.

[20] Gonzalo Martínez, José Alberto Hernández, Javier Conde, Pedro Reviriego, and Elena Merino-Gómez. Beware of words: Evaluating the lexical diversity of conversational llms using chatgpt as case study. *ACM Transactions on Intelligent Systems and Technology*, 2024.

[21] 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.

[22] 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.

[23] Mose Park, Yunjin Choi, and Jong-June Jeon. Does a large language model really speak in human-like language? *arXiv preprint arXiv:2501.01273*, 2025.

[24] 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.

[25] 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.

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

[27] Neil Sardesai, Paolo Russo, Jonathan Martin, and Anand Sardesai. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. *Postgraduate medical journal*, 100(1182):237–241, 2024.

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

[29] 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.

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

[31] Yongjin Yi and Kyong-Jee Kim. The feasibility of using generative artificial intelligence for history taking in virtual patients. *BMC Research Notes*, 18(1):80, 2025.

[32] Yongxiang Yuan, Jieyu He, Fang Wang, Yaping Li, Chaxiang Guan, and Canhua Jiang. Ai agent as a simulated patient for history-taking training in clinical clerkship: an example in stomatology. *Global Medical Education*, 2025.

Appendix A

Related Work

Recent advancements in LLMs have facilitated the development of VPs in medical education. For example, [12] used structured prompts based on illness scripts covering five categories and assessed realism through script consistency, medical plausibility, and student feedback. They found that ChatGPT can simulate a VP experience with mostly plausible responses and a generally positive user experience, though occasional implausible information was noted. [3] used short, unstructured prompts that were symptom-focused. The realism of their developed VP was evaluated using a short post-interaction questionnaire measuring students' perceived autonomy during history taking (e.g., freedom of choice and task relevance) and their prior experience using ChatGPT. [4] applied a detailed, structured prompt, including patient description and prior dialog turns; while students preferred this format over non-interactive alternatives, delays and repetitive responses reduced perceived realism. [31] used prompts with limited patient detail and relied on expert ratings for response quality across five dimensions (relevance, accuracy, fluency, succinctness, and impersonation), reporting some plausible answers. [32] introduced a VP system but did not specify the prompt design used to generate responses. Their evaluation focused on usability before and after optimization, showing improvements in personality and user experience, yet clinical appropriateness and accuracy across diverse populations were not assessed. [12], [18], and [3] assessed VP performance through human judgments, focusing on characteristics such as empathy, appropriateness of responses, and consistency with predefined patient profiles. [32] conducted structured pre- and post-optimization usability testing with medical students using standardized questionnaires; however, their evaluation did not cover the clinical appropriateness, freedom from bias, or medical accuracy of the responses across diverse populations. [11] conducted a pilot study involving 15 medical students, collecting participant feedback on the authenticity of the interactions. Similarly, [4] analyzed student perceptions of authenticity, providing insights into how users interpret the realism and communication quality of VP interactions. [27] used a structured prompt with a custom patient story that included medical history, family background, and details about the planned anesthesia and surgery. They also conducted qualitative evaluations through a survey, reporting that while VPs generated human-like responses, these often lacked the complexity of real patient interactions. They also noted that the responses were too formal and lengthy, reducing perceived realism. [31] used prompts with limited patient information and did not focus on interactive dialogue. Instead, they evaluated the quality of VP responses using expert reviewers. The tool included five items—relevance, accuracy, fluency, succinctness, and the VP's ability to impersonate a patient—rated on a five-point Likert scale. [31] showed the VP responses were generally plausible and only a few instances of implausible answers generated by the VP. While these studies report promising results in learner engagement and VP realism, several challenges remain. One challenge is the variability in prompt design across studies, which ranges from brief clinical scenarios to highly detailed ones, and few offer guidance on optimal prompt design for consistency and realism ([12, 18, 3]). Consequently, response delays caused by lengthy prompts can disrupt the natural flow of VP interactions ([13, 5, 11]). In addition, VP often breaks character and reverts to its default assistant persona, particularly when students fail to provide clear clinical cues ([11]). "Hallucinations" (i.e., generating confusing or meaningless responses) continue to pose a threat to learning outcomes ([14]). Furthermore, most current evaluations rely on human judgments focused on attributes such as empathy, response appropriateness, and alignment with predefined patient profiles, often limited to a single clinical case and short-term usability, while overlooking the long-term educational impact.

Linguistic Features of Conversations

Third, we explore various linguistic patterns between VP and human responses. While the performance of conversational LLMs has been evaluated from various perspectives, much less attention has been given to analyzing the linguistic features of the text they generate and how closely they resemble actual human conversations ([20, 23, 22]). This is surprising, considering that LLMs are language models, and understanding how they use language and whether their language use resembles that of humans is essential. Lexical diversity refers to how varied a person's vocabulary is within a given response. One of the most common ways to measure lexical diversity is through the Type-Token Ratio (TTR), which looks at the number of unique words used compared to the total number of words ([19]). However, TTR may not provide a reliable measure when the responses are too short, too long, or differ greatly in length ([22]). To reduce this sensitivity to length, we also calculate metrics like Root TTR (RTTR) and the Maas index ([30, 29]). These build on the basic TTR formula but adjust for the total number of words, making them more reliable when comparing responses of different lengths. Measure of Textual Lexical Diversity (MTLD) and Hypergeometric Distribution Diversity (HDD) are designed to handle variations in text length more effectively and provide a stronger sense of vocabulary richness, especially in longer or more complex responses ([21]). Additionally, we use Yule's K, a well-established measure that captures how concentrated or repetitive the vocabulary is. It's particularly valuable because it is not affected by the length of the text. Lastly, we calculate Moving Average TTR (MATTR), which works by analyzing the text in blocks or sequences rather than all at once ([7]). Part-of-speech (POS) tagging, also known as grammatical tagging, is a fundamental task in natural language processing (NLP) that involves assigning syntactic labels to each word in a sentence based on its context ([8, 9, 17]). These tags typically include categories such as nouns, verbs, adjectives, adverbs, and other grammatical classes. In this study, we use POS tagging to compare responses from human-simulated patients and VPs, offering an additional linguistic perspective on how their language patterns differ.

Appendix B

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.

Box 2: Excerpt from an Example Ilness Script

You are going to play the role of a medical patient.

The patient details are as follows:
Name: Unknown to the transcript.
Age: Unknown to the transcript.
Chief Complaint: Cough that brings up gunk.
Symptoms: Green, sometimes yellow, gunk; couple Kleenexes worth of sputum.
Duration: Past few years.
Nature and rating of pain: Coughing is always pretty bad, worse with exertion like walking up stairs.
Pain/symptom progression: Cough is getting worse now, worse in the morning. Deep breaths make it worse.
Additional symptoms: Low energy, feeling tired, slightly swollen, possible weight gain, trouble breathing with exertion, frequent urination at night, poor sleep.
Medical History: High blood pressure, C-section for birth of daughter, hospitalized for a few days after that.
Medications: Lisinopril.
Immunizations: Unknown to the transcript.
Family History: Father had a heart attack at age 78.
Alcohol, tobacco, marijuana, other drugs: No recreational drugs; three to four glasses of wine per week (5–6 oz per glass); smokes cigarettes (1–2 packs/day for the last 40 years).
Living Conditions: Lives in an apartment with husband and daughter.
Recent Travel: No recent travel.
Occupation: Worker at local grocery store.
Other: Uses one pillow. Lost five pounds in the last few months. Recently tested negative for tuberculosis. Has a decent diet but does not exercise.
Answer with the main complaint initially. Respond with other symptoms only when asked or when relevant in the conversation, maintain conversational language.
If a response slot says "unknown to the patient", respond in a way that indicates the patient doesn’t know and has no information about it.
If a response slot says "the patient is unsure", respond in a way that indicates uncertainty and lack of surety.
If a response slot says "absent in the patient", respond with a response indicating such.
If a response slot says "unknown to the transcript", make up a relevant response.

Box 3: A complete example of Doctor-VP Interaction

User: I was wondering if you could tell me what brought you in.

ChatGPT: I've been having a lot of trouble with coughing and shortness of breath lately. The cough has been there for the past five years, but it's gotten a lot worse in the last three days.

User: OK, when did the cough start?

ChatGPT: The cough started about five years ago.

User: OK, did anything happen 3 days ago that has brought the cough on?

ChatGPT: No, nothing specific happened three days ago. The cough just seemed to get a lot worse all of a sudden.

User: OK, is there any particular time of day that the cough is worse, like morning, night?

ChatGPT: Yes, the cough is definitely worse in the morning.

User: OK, and are you bringing up any sputum?

ChatGPT: Yes, I'm bringing up sputum. It used to be clear, but now it's yellow and sometimes green. And I'm bringing up more sputum than before.

User: OK. With your cough prior to three days ago were you bringing up sputum with that?

ChatGPT: Yes, I was bringing up sputum before, but it was clear back then and there was less of it compared to now.

User: OK, and are you bringing up more sputum now, the same or less?

ChatGPT: I'm bringing up more sputum now than before.

User: OK, have you noticed any blood?

ChatGPT: No, I haven't noticed any blood in the sputum.

User: OK, and you mentioned you've been you've had a wheeze, when you do you notice that?

ChatGPT: I mostly notice the wheeze at night.

User: OK, um and what types of activities have you had shortness of breath?

ChatGPT: I've been getting short of breath just from walking and doing day-to-day things.

User: OK. And have you been experiencing any other symptoms besides the cough, wheeze and shortness of breath?

ChatGPT: Yes, I've also been feeling fatigued, though I sleep normally. Sometimes I get a dull chest pain from coughing so frequently.

User: OK. So a bit more tired. How has your sleep been?

ChatGPT: My sleep has been normal, despite feeling more fatigued during the day.

User: OK, have you had any ear pain or eye pain?

ChatGPT: No, I haven't had any ear pain or eye pain.

User: Any discharge from either, either your ears or eyes?

ChatGPT: No, I haven't had any discharge from my ears or eyes.

User: OK. Have you had a runny nose?

ChatGPT: No, I haven't had a runny nose.

User: Or felt congested?

ChatGPT: No, I haven't felt congested either.

User: OK, have you had a sore throat?

ChatGPT: No, I haven't had a sore throat.

User: Or any difficulty swallowing?

ChatGPT: No, I haven't had any difficulty swallowing either.

User: Changes to your sense of smell or taste?

ChatGPT: No, I haven't noticed any changes to my sense of smell or taste.

User: OK, have you had any changes in appetite?

ChatGPT: No, my appetite has been normal.

User: Alright, and have you had any weight loss over the last several months?

ChatGPT: No, I haven't had any weight loss over the last several months.

User: OK. How about have you had any fevers or chills?

ChatGPT: No, I haven't had any fevers or chills.

User: Nights, how about night sweats?

ChatGPT: No, I haven't had any night sweats.

User: Have you had any chest pain?

ChatGPT: Yes, I've had some dull chest pain, but it seems to be from coughing frequently.

User: OK, do you get chest pain when you exert yourself physically?

ChatGPT: No, the chest pain doesn't seem to get worse with physical exertion; it mostly happens because of the frequent coughing.

User: OK, so it's just after the coughing?

ChatGPT: Yes, the chest pain is mostly just from coughing so much.

User: OK and how would you describe the character of the chest pain? Some people will say sharp, dull, achy.

ChatGPT: I'd describe the chest pain as dull.

Box 4: Conversation Cont.

User: OK, and have you had any like heart palpitations, feeling like your heart's racing?
ChatGPT: No, I haven't experienced any heart palpitations or felt like my heart is racing.
User: And how about any episodes where you felt like you were going to faint or, yeah, pass out?
ChatGPT: No, I haven't had any episodes where I felt like I was going to faint or pass out. **User**: OK. Uhm and have you had any swelling in your feet, ankles, hands?
ChatGPT: No, I haven't noticed any swelling in my feet, ankles, or hands.
User: OK. And then have you had any urinary problems?
ChatGPT: No, I haven't had any urinary problems.
User: Any bowel problems?
ChatGPT: No, I haven't had any bowel problems either.
User: OK, and then any joint pains or muscle aches?
ChatGPT: No, I haven't had any joint pains or muscle aches.
User: And how about any rashes anywhere?
ChatGPT: No, I haven't had any rashes.
User: OK, and then, sorry, have you had any nausea or vomiting?
ChatGPT: No, I haven't had any nausea or vomiting.
User: OK. In terms of your past medical history, have you been diagnosed with any conditions before?
ChatGPT: Yes, I have been diagnosed with high blood pressure and diabetes.
User: I see. Are you on any medications for those?
ChatGPT: Yes, I'm taking a beta blocker for my high blood pressure and metformin for my diabetes. I also take calcium and vitamin D supplements.
User: OK, uh, do you, are you taking any other medications?
ChatGPT: No, just the ones I mentioned: the beta blocker, metformin, calcium, and vitamin D.
User: Any surgeries in the past?
ChatGPT: No, I haven't had any surgeries in the past.
User: Any allergies to medications?
ChatGPT: No, I don't have any allergies to medications.
User: OK. Are your immunizations up to date?
ChatGPT: Yes, my immunizations are up to date, but I haven't gotten the pneumococcal vaccine.
User: Yeah, you're you're right. So for um, in Canada or anybody or in Ontario, at least, anybody who's older than 65, it's recommended to get this vaccine to protect against certain types of bacterial infections type which can cause pneumonia. So it would be good something that you can set up with your family doctor.
ChatGPT: That sounds like a good idea. I'll make sure to talk to my family doctor about getting the pneumococcal vaccine.
User: Yeah, and so are you, so do you smoke cigarettes?
ChatGPT: I used to smoke a pack a day for 45 years, but I stopped two years ago.
User: OK, and how much do you smoke or how much did you use to smoke?
ChatGPT: I used to smoke a pack a day.
User: OK, that's great that you were able to stop. Do you drink alcohol?
ChatGPT: Yes, I drink alcohol occasionally, about every two weeks.
User: OK, and any triggers that you can think of you were exposed to either at work or home like chemicals or things that you were exposed to that you had to breathe, that you breathe in?
ChatGPT: No, I haven't been exposed to any triggers like chemicals at work or home.
User: OK, have you traveled anywhere?
ChatGPT: No, I haven't traveled anywhere recently.
User: Alright, and have you been near anybody who's been sick or has similar symptoms?
ChatGPT: No, I haven't been near anyone who is sick or has similar symptoms.
User: Alright, and anybody in the family have a heart or lung condition?
ChatGPT: No, but there is liver cancer in my family.
User: Any cancers in the family?
ChatGPT: Yes, there's liver cancer in my family.
User: OK, and, I think that.. OK, and that was all I wanted to ask on history. Was there anything else that you wanted to tell me today?
ChatGPT: No, I think we covered everything. Thanks for asking.
User: Yeah, certainly so umm, it, it sounds like it could be maybe a viral infection that I brought on this cough and worsen the sputum. This could be a bacterial infection, it could be a COPD exacerbation and also one of the other things that we would have to consider right now is Covid as well and so we will do a PCR nasal swab for Covid and a chest Xx-ray and some other investigations as well and uhm, we will try to get to the bottom of what's causing your your shortness of breath.
ChatGPT: That sounds good. Thank you for explaining. I appreciate you looking into this and helping me figure out what's going on.

Appendix C

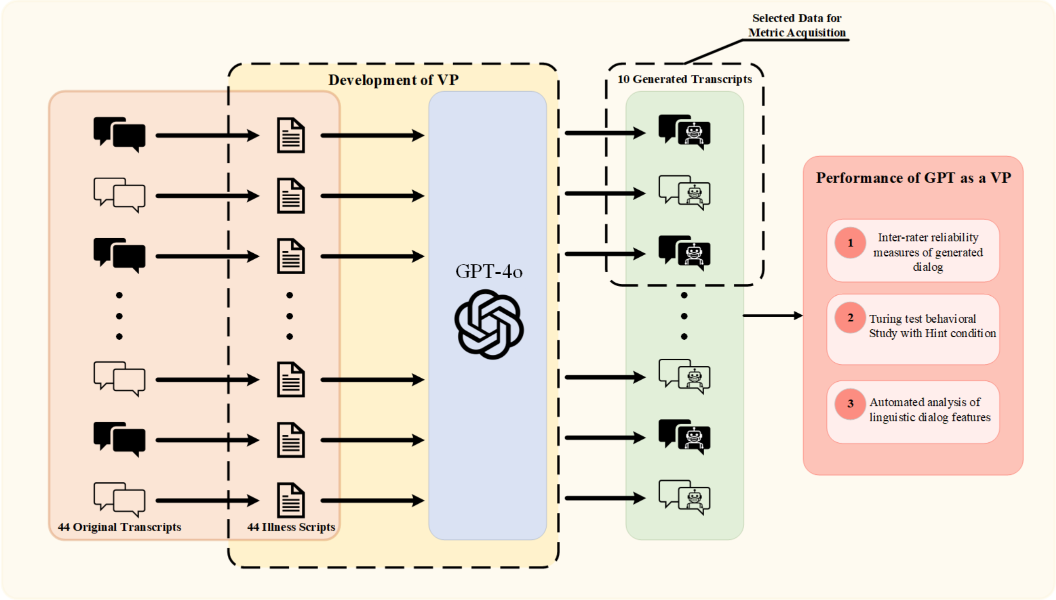
Selected Data for Metric Acquisition
Development of VP
10 Generated Transcripts
GPT-4o
Performance of GPT as a VP
1 Inter-rater reliability measures of generated dialog
2 Turing test behavioral Study with Hint condition
3 Automated analysis of linguistic dialog features
44 Original Transcripts
44 Illness Scripts

Figure 2: Overview of the VP development and evaluation pipeline.

Table 3: Estimated Marginal Means and Pairwise Comparisons of Confidence Ratings by Hint Condition and Dialog Type

Hint Condition	**Dialog Type**	**Mean Confidence**	**SE**	**95% CI**
With Hint	VP-Generated	2.63	0.0974	[2.44, 2.82]
Without Hint	VP-Generated	2.75	0.0995	[2.55, 2.95]
With Hint	Human-Generated	3.02	0.0974	[2.83, 3.21]
Without Hint	Human-Generated	2.75	0.0995	[2.55, 2.95]

Table 4: Confidence Ratings by Human Evaluators in the Turing Test

Comparison	**Mean Difference**	**SE**	**t(90)**	***p*-value**
With Hint: Human vs. VP	-0.39	0.138	-2.84	.0055 *
Without Hint: Human vs. VP	0.004	0.141	0.03	.9754
VP: With vs. Without Hint	-0.12	0.139	-0.88	.3795
Human: With vs. Without Hint	0.27	0.139	1.96	.0531 †

Note. CI = Confidence Interval. ** $p < .01$. † Trend-level difference ($p < .10$).

Table 5: Summary of ANOVA Results for Correct Responses, Confidence Level, and Log-Transformed Response Time. Significance levels: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

Variable	**Source**	**Sum Sq**	**Df**	**Mean Sq**	**F**	**p-value**
Correct Responses	Group	40.50	1	40.50	8.689	0.0041 **
	Transcripts	360.20	1	360.20	77.34	0.001 ***
	Group × Transcripts	9.50	1	9.50	2.045	0.1562
	Residuals	419.10	90	4.66		
Confidence Level	Group	0.132	1	0.132	0.580	0.4483
	Transcripts	0.920	1	0.920	4.039	0.0475 *
	Group × Transcripts	0.921	1	0.921	4.042	0.0474 *
	Residuals	20.50	90	0.228		
Log Time	Group	1.00	1	1.00	1.549	0.217
	Transcripts	0.27	1	0.27	0.412	0.523
	Group × Transcripts	0.01	1	0.01	0.022	0.881
	Residuals	58.24	90	0.647		

Table 6: Turing Test Confusion Matrices With and Without Hint

(a) With Hint

	Human	VP	Total
Judged Human	203	116	319
Judged VP	37	124	161
Total	240	240	480

(b) Without Hint

	Human	VP	Total
Judged Human	179	156	335
Judged VP	51	74	125
Total	230	230	460

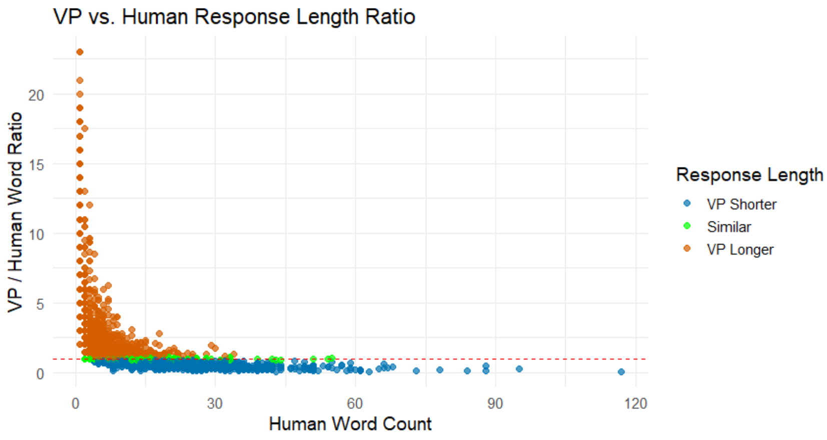

Figure 3: VP vs. Human Response length Ration

	Total Words	**Unique Words**	**TTR**	**RTTR**	**MTLD**	**HDD**	**MATTR**
Human	30,371	1,559	0.0513	8.95	40.33	0.798	0.817
VP	27,517	1,660	0.0603	10.01	55.96	0.825	0.825

Table 7: Lexical richness metrics comparing human and VP responses.

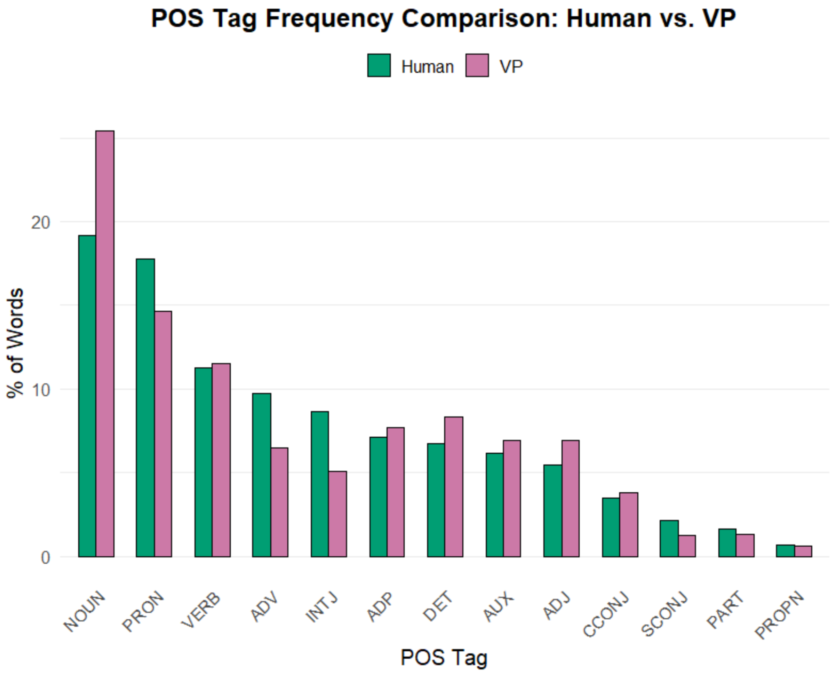

Figure 4: Percentage differences in the use of POS category for VP responses in comparison to human simulated patients.

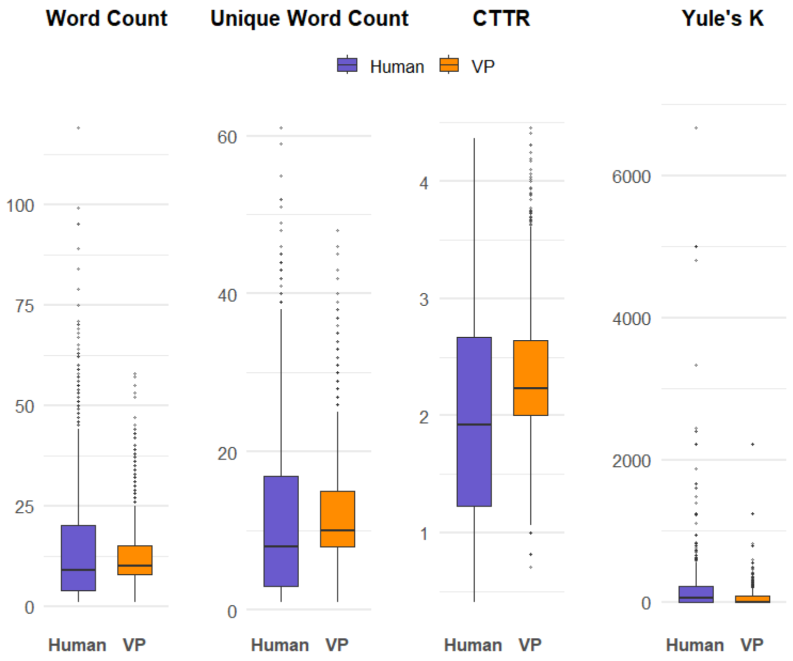

Figure 5: Lexical metric distributions computed per response for Human and VP answers.

Table 8: POS Tag Frequency Counts for Human and VP Responses. **ADJ** = adjective, **ADP** = adposition, **ADV** = adverb, **AUX** = auxiliary, **CCONJ** = coordinating conjunction, **DET** = determiner, **INTJ** = interjection, **NOUN** = noun, **PART** = particle, **PRON** = pronoun, **PROPN** = proper noun, **SCONJ** = subordinating conjunction, **VERB** = verb.

Group	**ADJ**	**ADP**	**ADV**	**AUX**	**CCONJ**	**DET**	**INTJ**	**NOUN**	**PART**	**PRON**	**PROPN**	**SCONJ**	**VERB**
Human	1674	2190	2982	1899	1082	2073	2651	5878	496	5460	213	663	3451
VP	1915	2296	2481	1909	1056	1995	2126	7020	370	4044	181	346	3188

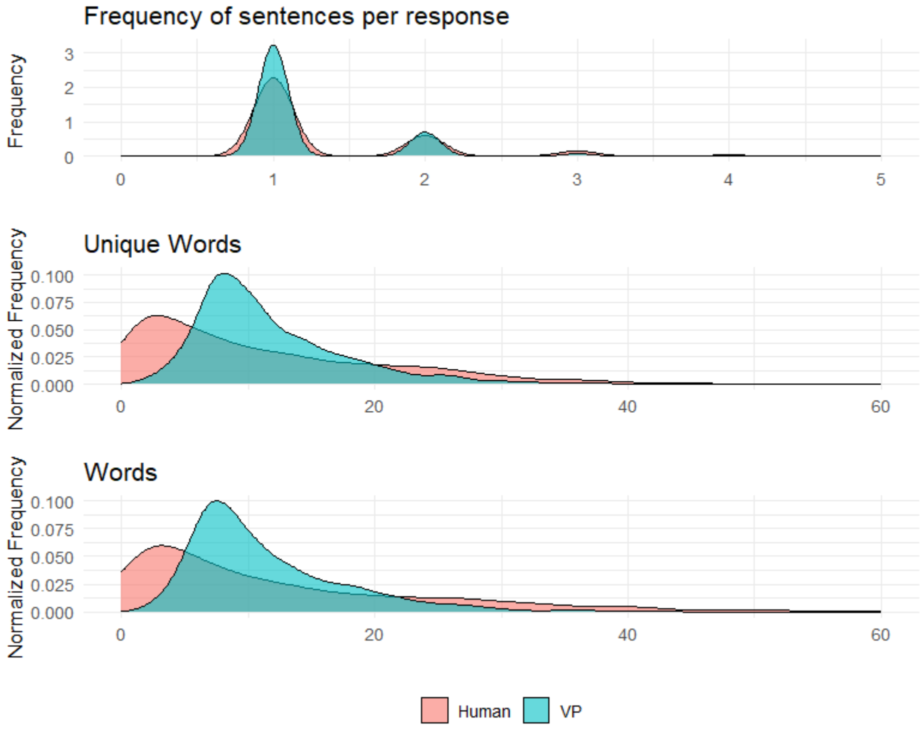

Figure 6: Distributions of sentence count, unique words, and total word count per turn.

Appendix D

Turns with Omissions

This metric refers to the number of conversations where ChatGPT-4o omits important information from the illness script that is necessary for making an accurate diagnosis.

Illness Script:

> The patient has no family history of lung conditions. Their mother had a heart attack at the age of 56. The patient has not traveled recently.

Example of Conversation:

> D: You are OK. Any family history of lung conditions?
> P: No lung condition.
> D: But how about heart?
> P: Yeah, it's just my mom had a heart attack when she was 56.

Example of a Turn with an Omission:

> D: You are OK. Any family history of lung conditions?
> P: No lung condition.
> D: But how about heart?
> P: No heart issues.

Turns with Hallucinations (Could Influence Diagnosis)

Hallucinations in this context mean instances where ChatGPT-4o provides false, fabricated, or misleading information relative to the illness script, which could lead to incorrect diagnoses or inappropriate treatment plans.

Illness Script:

> The patient has no family history of lung conditions. Their mother had a heart attack at the age of 56. The patient has not traveled recently.

Example of Conversation:

> D: You are OK. Any family history of lung conditions?
> P: No lung condition.
> D: But how about heart?
> P: Yeah, it's just my mom had a heart attack when she was 56.

Example of a Turn with a Hallucination:

> D: But how about heart?
> P: Yeah, it's my mom and dad had a heart attack.

Turns with Inappropriate Repetitions

This refers to the number of conversation turns where ChatGPT-4o repeats information or questions inappropriately in a way that is not human-like.

Example of a Turn with a Repetition:

> D: Can you describe the nature of your pain?
> P: The pain is sharp and localized to my lower abdomen. It started suddenly two days ago.
> D: Have you experienced any nausea or vomiting?
> P: Yes, I've had nausea and vomiting. Additionally, the pain is still sharp and localized to my lower abdomen, and it hasn't changed in intensity since it started suddenly two days ago.