Cureus Journal of Computer Science

Open Access Original Article

Simulating Physician–Patient Interactions Using Generative Artificial Intelligence: An Educational Tool for Medical Students

Received 09/19/2025
Review began 09/30/2025
Review ended 11/17/2025
Published 12/01/2025

DOI:
https://doi.org/10.7759/s44389-025-10532-y

James B. Stoco [1] , João Victor A. Araújo [1] , Adriana T. Peres [2] , Gabriela P. Oliveira [3, 2] , Giovani B. Peres [3, 1, 2, 4]

1. Faculty of Medicine, Universidade Paulista – UNIP, Sorocaba, BRA 2. Faculty of Biomedicine, Universidade Paulista – UNIP, São Paulo, BRA 3. Department of Biomedicine, Faculdade de Ciências Médicas da Santa Casa de São Paulo, São Paulo, BRA 4. Graduate Program in Environmental and Experimental Pathology, Universidade Paulista – UNIP, São Paulo, BRA

Corresponding author: Giovani B. Peres, giovani.peres@fcmsantacasasp.edu.br

Abstract

Introduction

History-taking and communication are core competencies in medical education, yet students often report insufficient confidence and limited opportunities for practice. Standardized patients are effective but resource-intensive. Virtual patients powered by large language models may provide a scalable alternative. This study aimed to evaluate the acceptability and perceived educational impact of a virtual patient simulation using ChatGPT version 3.5 (GPT-3.5) among undergraduate medical students.

Methods

This descriptive, mixed-methods study included 56 medical students recruited in the first semester of 2025. Participants completed an online questionnaire and interacted with an AI-based virtual patient programmed to simulate a dengue fever case. Performance was scored across six domains (maximum 100 points), and structured feedback was automatically generated. Students could repeat the simulation with minor variations in demographics and symptom presentation. Data collection included demographics, prior exposure to simulations and AI, self-assessment of knowledge and skills, and perceptions of the activity.

Results

The sample was balanced by sex (51.8% male), with a mean age of 24.3 years. Most students rated the virtual patient as realistic and useful, and nearly all considered the automated feedback beneficial. Willingness to repeat the activity was high. Among those who repeated, most reported significant performance improvement and greater confidence for real encounters. Chi-square tests confirmed significant deviations from uniformity across key outcomes. Thematic analysis identified four themes: educational utility, limitations of AI, suggestions for improvement, and overall positive appraisal. In terms of prior exposure, nearly half of the students had no prior experience with AI, and most reported only occasional institutional opportunities for simulation.

Conclusion

AI-enabled virtual patients using ChatGPT were well accepted by medical students, who perceived benefits for knowledge reinforcement, performance, and confidence in history-taking. These findings support the integration of AI-based simulations as scalable, formative complements to traditional approaches in medical education.

Categories: AI applications, Generative AI for healthcare, User-interface design for medical devices and healthcare software
Keywords: medical education, history-taking, communication skills, virtual patients, large language models, artificial intelligence, chatgpt, simulation-based learning, formative assessment, undergraduate students

Introduction

Despite significant advances in imaging and molecular diagnosis, history-taking and communication remain foundational competencies in medical education [1,2]. A comprehensive history guides subsequent decisions and may suffice for diagnosis in many cases, whereas inadequate history-taking jeopardizes patient safety [3]. At the same time, learners frequently report limited confidence in communication skills, and designing scalable, effective curricula for history-taking is challenging [1]. Simulation - deliberately recreating clinical events to enable practice, assessment, and systems understanding - has become a key strategy to address this gap [4].

Withing this landscape, virtual patients (VPs) - computer-based systems that simulate clinical scenarios for obtaining a history, performing an examination, and making diagnostic and therapeutic decisions - are widely recommended as an adjunct to communication-skills training [4-6]. Since early descriptions [7], broader adoption has been facilitated by lower costs and improved technology. When integrated as interactive, case-based activities, VPs offer safe, repeatable practice with formative feedback, standardized presentation and assessment, around-the-clock access, and scalability across class size and learner level [8,9].

Recent advances in AI, particularly large language models (LLMs), have markedly increased the realism and pedagogical potential of web-based chatbots for communication training [3,6]. LLM-driven dialogues are perceived as realistic even in sensitive scenarios, and early studies suggest benefits for empathy and feedback-driven improvement [3,6]. Technically, modern LLMs use neural architectures with positional encoding and self-attention to model relationships in input text - often referred as prompts - and generate contextually appropriate responses [10]. Unlike traditional search engines that return links, LLMs synthesize learned knowledge into concise, conversational outputs, creating a user-friendly interface for practice and reflection [11,12]. AI-driven VPs also permit explicit objectives around inclusivity, bias awareness, empathy, and respectful communication [13-15].

Although interest in LLM-based VP simulations is growing, evidence on acceptability and perceived educational impact among undergraduate medical students - particularly regarding realism, knowledge reinforcement, usefulness of automated feedback, willingness to repeat practice, and perceived effects on performance and confidence - remains limited.

Taken together, these developments motivate rigorous evaluation of LLM-based chat simulations of physician-patient interactions as educational tools for medical students. Such systems promise scalable, standardized, and pedagogically rich practice in history taking and communication, competencies that are central to clinical safety and effectiveness, while addressing long-standing constraints in curricular time, faculty bandwidth, and access to patient encounters.

Thus, in this study, we aimed to evaluate a ChatGPT-based VP simulation for undergraduate medical students, focusing on (1) acceptability and perceived realism; (2) perceived reinforcement of history-taking and clinical decision-making; (3) perceived usefulness of automated feedback and willingness to repeat the activity; and (4) perceived effects on performance after repetition and on confidence for real clinical encounters.

Materials And Methods

Study design

Grounded in constructivism and Experiential Learning Theory, the simulation positioned learners as active agents who co-construct knowledge through interaction [16,17]. This descriptive, mixed-methods study combined quantitative and qualitative analyses within a hypothetico-deductive framework. It investigated the use of generative AI (ChatGPT-3.5) to simulate physician-patient interactions in undergraduate medical education.

Participants

A total of 56 medical students were recruited through social media platforms during the first semester of 2025. Eligibility criteria included enrollment in a medical program and provision of informed consent. Participation consisted of completing an online questionnaire, and informed consent was obtained electronically prior to data collection. The sample included students from different semesters, with most concentrated in intermediate years of training. The study protocol was approved by the Research Ethics Committee of Universidade Paulista - UNIP (CAAE 86634225.0.0000.5512).

Simulation procedures

A structured prompt was developed to emulate a medical consultation in six steps: (1) patient identification and chief complaint; (2) history of present illness; (3) review of systems; (4) past medical, family, and social history; (5) request and interpretation of complementary exams; and (6) diagnostic reasoning and therapeutic decision-making.

The AI-based VP was programmed to provide case-consistent responses while evaluating student performance across six domains: anamnesis (30 points), identification of relevant information (20 points), communication skills (15 points), professional conduct and respect (10 points), exam requests (12.5 points), and prescription accuracy (12.5 points). The maximum achievable score was 100 points. A penalty system was implemented. Deductions were applied per domain: each distinct error type incurred a 20% deduction of that domain's maximum subscore (e.g., 6 points in Anamnesis [30×0.20], 4 in Identification of relevant information [20×0.20], 3 in Communication [15×0.20], 2 in Professional conduct [10×0.20], and 2.5 in Exam requests/Prescription [12.5×0.20]). Repeated instances of the same error within a domain were

not double-counted; different error types within the same domain could incur separate deductions, with the domain subscore floored at zero.

At the end of each simulation, the AI generated a structured report including scores, list of errors, and tailored feedback. Students could opt to repeat the simulation with a modified case profile (e.g., different demographics or symptom presentation), while the underlying diagnosis remained the same.

The turn-taking dialogue with the VP provided a concrete experience, while the automated, criterion-referenced feedback supported reflective observation and abstract conceptualization. Also, opportunities to repeat the encounter enabled active experimentation and deliberate practice. This framing informed our choice of proximal outcomes (acceptability, perceived learning, and intention to repeat) and the inclusion of qualitative analysis to capture learner sense-making.

A prototype case of dengue fever (low-risk, without warning signs) was used to illustrate the process. Full details of the structured prompt, interview steps, scoring rubric, and case information are provided in Appendix 1.

The VP was instantiated in ChatGPT-3.5 via the native web interface (no application programming interface was used). Students received a standardized, prewritten seed prompt embedded in the study form and were instructed to copy-paste it into ChatGPT-3.5 to initiate the encounter and to proceed with history-taking and complete the post-interaction survey. Participants were advised not to read the seed prompt beyond executing it; however, we could not technically prevent access to its content in the native interface.

Interactions were conducted remotely and asynchronously, without in-person supervision. We did not constrain device type, browser, location, or time of day. All students received identical instructions and the same seed prompt to maximize procedural consistency.

Data collection

Data were collected through a structured questionnaire administered via Google Forms. It comprised six sections: (1) informed consent; (2) demographic and academic information (sex, age, semester of medical school, prior simulation/clinical experience); (3) previous experience with AI and self-assessment of theoretical and practical skills; (4) execution of the AI-based simulation; (5) evaluation of the simulation (perceived realism, usefulness of feedback, reinforcement of learning, impact on confidence); and (6) open-ended feedback (difficulties, suggestions, comments). The questionnaire reference, which consists of a file containing the list of variables obtained from the survey and the measurement scale for each variable, along with the raw dataset in both English and Portuguese, is provided in Appendix 2.

Data analysis

Quantitative data were summarized using descriptive statistics. Continuous variables (e.g., age) were expressed as mean, standard deviation, median, interquartile range, and minimum-maximum values. Categorical (e.g., sex, semester of medical school, previous experience with simulations or AI) and ordinal variables (e.g., self-assessment of knowledge, perceived realism, usefulness of feedback, impact on confidence) were reported as absolute, relative frequencies, and 95% confidence intervals (CIs). Proportion confidence intervals were computed using the Wilson method, with Clopper-Pearson intervals used for comparison when appropriate.

For ordinal survey data, emphasis was placed on descriptive estimation rather than null-hypothesis testing. Thus, proportions and corresponding 95% CIs were reported directly in the Results section for each response category.

Inferential tests were limited to exploratory chi-square goodness-of-fit analyses, used solely to assess whether observed categorical distributions displayed marked deviations from uniformity. Given the known limitations of chi-square procedures for ordinal outcomes - such as disregarding the ordered structure of the response scale and relying on a substantively weak null hypothesis - these analyses are interpreted cautiously and described transparently as exploratory rather than confirmatory. The significance level was set at $\alpha = 0.05$.

To strengthen transparency, a post-hoc sensitivity analysis for chi-square goodness-of-fit tests was performed (G*Power version 3.1.9.7; $\alpha = 0.05$; power = 0.80; $n = 56$), demonstrating that the achieved sample size was sufficient to detect medium-to-large effect sizes ($w = 0.37$-0.46 depending on degrees of freedom). These results are reported in the Results sections.

Qualitative data from open-ended responses were analyzed through thematic categorization, identifying recurrent patterns such as educational utility, limitations of AI, suggestions for improvement, and overall

impressions. Representative quotes were included to illustrate each theme.

All statistical analyses and graphs construction were performed using Microsoft Excel (Microsoft Corporation, Redmond, WA, USA). Additional computations, including chi-square effect size (Cohen's w), were carried out using the Real Statistics add-in for Excel (Real Statistics Resource Pack, XRealStats-Ver9.2). The complete spreadsheet containing all analyses is provided in Appendix 3.

Results And Discussion

Participant characteristics

A total of 56 medical students participated in the study. The sample was balanced in terms of sex (51.8% male, 48.2% female) and had a mean age of 24.3 years (SD - 3.6; median - 24; range - 18-33). Most participants were enrolled in intermediate semesters of medical school, particularly the 3rd, 5th, and 7th semesters (Table *1*).

Variable	n (%) or Mean ± SD	Range/Median [IQR]	Notes
Sex	–	–	–
Male	29 (51.8%)	–	–
Female	27 (48.2%)	–	–
Age (years)	24.29 ± 3.55	18–33/24 [21–27]	Distribution approximately normal
Semester of medical school	–	–	–
1st	5 (8.9%)	–	–
3rd	13 (23.2%)	–	–
4th	1 (1.8%)	–	–
5th	15 (26.8%)	–	–
6th	3 (5.4%)	–	–
7th	14 (25.0%)	–	–
8th	1 (1.8%)	–	–
9th	4 (7.1%)	–	–
2nd, 10th–12th	0 (0.0%)	–	–

TABLE 1: Participant characteristics (n = 56).

Values are expressed as absolute frequency and percentage unless otherwise indicated.

SD, Standard Deviation; IQR, Interquartile Range

Exposure to clinical training and AI

Most students reported that simulation-based patient encounters were offered only occasionally (n = 32; 57.1%, 95% CI: 43.3%-70.1%). A smaller proportion reported regular opportunities (n = 19; 33.9%, 95% CI: 22.1%-47.8%), and few indicated complete absence (n = 5; 8.9%, 95% CI: 3.4%-19.7%) (Figure *1*). The distribution differed significantly from uniformity ($\chi^2_{(2)} = 19.54$, $p < 0.001$), with a corresponding effect size of Cohen's $w = 0.59$, indicating a large deviation. The post-hoc sensitivity analysis showed that, for df = 2 and n = 56, the study had 80% power to detect effects of $w \geq 0.41$, confirming that the observed $w = 0.59$ was well above the minimal detectable threshold.

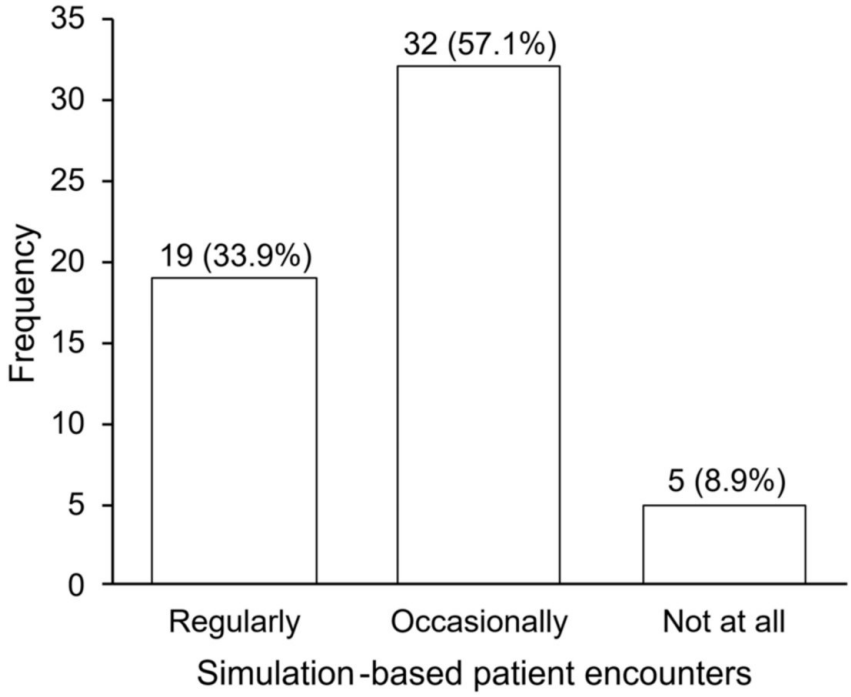

FIGURE 1: Frequency of simulation-based patient encounters reported by participants.

Most students indicated that their institutions offer such simulations occasionally, while fewer reported regular availability or complete absence.

Before the AI-based activity, students' self-assessed theoretical knowledge showed heterogeneous responses across the five categories (Figure *2A*). The most frequent ratings were "insecure" (n = 15; 26.8%; 95% CI: 16.2%-40.3%) and "secure" (n = 16; 28.6%; 95% CI: 17.7%-42.2%), followed by "very insecure" (n = 9; 16.1%; 95% CI: 8.1%-28.4%), "neutral" (n = 9; 16.1%; 95% CI: 8.1%-28.4%), and "very secure" (n = 7; 12.5%; 95% CI: 5.6%-24.2%). A chi-square goodness-of-fit test - reported here only as an exploratory indicator - showed no significant deviation from a uniform distribution ($\chi^2_{(4)} = 5.79$, $p = 0.216$), corresponding to a small-to-medium effect size ($w = 0.32$; post-hoc sensitivity: achieved power = 0.46).

In contrast, practical experience showed a more polarized pattern (Figure *2B*). Most students rated themselves as "insecure" (n = 20; 35.7%; 95% CI: 23.7%-49.6%) or "neutral" (n = 20; 35.7%; 95% CI: 23.7%-49.6%), followed by "very insecure" (n = 8; 14.3%; 95% CI: 6.8%-26.3%), "secure" (n = 6; 10.7%; 95% CI: 4.5%-22.0%), and "very secure" (n = 2; 3.6%; 95% CI: 0.7%-12.4%). Here, the exploratory chi-square test indicated a significant deviation from uniformity ($\chi^2_{(4)} = 24.71$, $p < 0.001$), with a large effect size ($w = 0.66$) and high post-hoc power (0.99).

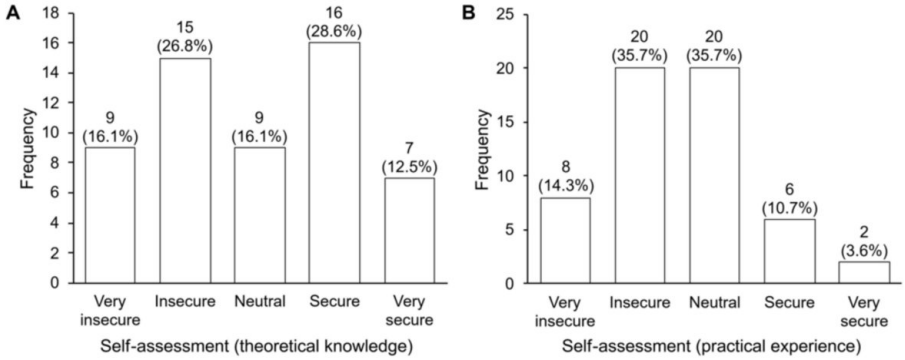

FIGURE 2: Self-assessment of students' knowledge and skills before the AI simulation.

(A) Theoretical knowledge regarding medical consultations: most participants rated themselves as "insecure" or "secure", with fewer responses at the extremes of the scale. (B) Practical experience in patient consultations: the majority reported "insecure" or "neutral" levels, while only a few indicated "secure" or "very secure" experience.

A proportion of 42.9% of students reported frequent use of AI in a medical context (n = 24; 95% CI: 29.9%-56.7%), 16.1% reported occasional use (n = 9; 95% CI: 8.1%-28.4%), and 41.1% reported no prior experience (n = 23; 95% CI: 28.3%-55.0%) (Figure *3*). A chi-square goodness-of-fit test was conducted as a supplementary exploratory analysis to examine deviations from equal category distribution. The distribution differed significantly across categories ($\chi^2_{(2)} = 7.54$, $p = 0.023$, $w = 0.37$). Based on the post-hoc sensitivity analysis ($\alpha = 0.05$, power = 0.80, $n = 56$), the minimal detectable effect size for a three-category chi-square test ($df = 2$) was $w = 0.41$. Thus, the observed effect ($w = 0.37$) falls slightly below the threshold typically detectable with 80% power in this sample, consistent with the calculated post-hoc power (0.69). This interpretation suggests that, although the chi-square result reached statistical significance, it should be understood within the constraints imposed by sample size and detectable effect magnitude.

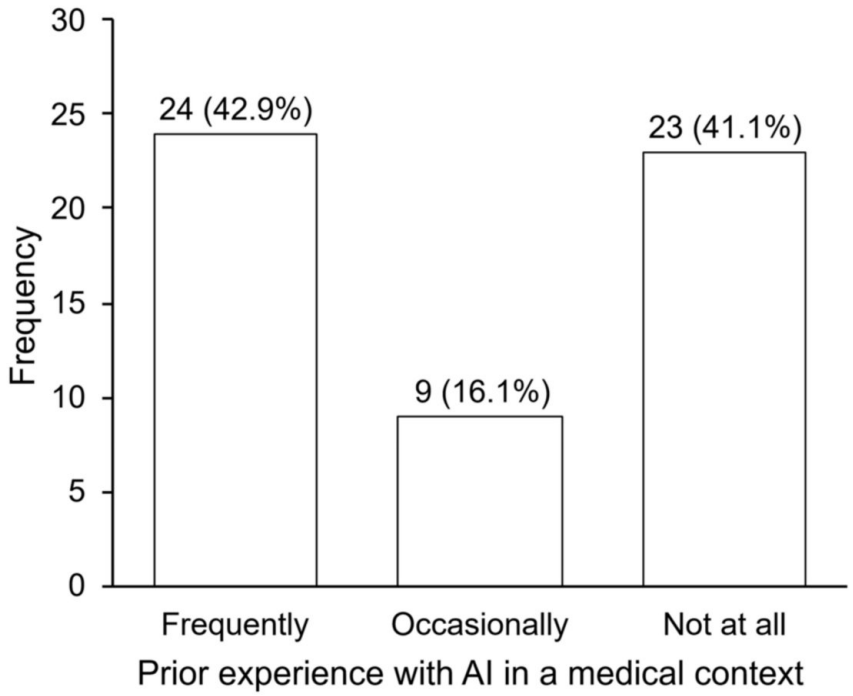

FIGURE 3: Prior experience with artificial intelligence (AI) in a medical context.

Almost half of the participants reported frequent prior use of AI, while a similar proportion had no previous experience; only a minority indicated occasional use.

Perceptions of the AI simulation

The AI-based virtual patient was broadly accepted. Most students rated the interaction as "realistic" (n = 32; 57.1%; 95% CI: 43.3%-70.1%) or "very realistic" (n = 9; 16.1%; 95% CI: 7.6%-28.3%) (Figure *4A*). Only 7.1% (n = 4; 95% CI: 2.4%-17.4%) reported slightly realistic impressions, and none rated the interaction as "not realistic at all" (0%; 95% CI: 0%-6.4%). The distribution differed significantly from a uniform pattern ($\chi^2_{(3)} = 32.71$, $p < 0.001$), indicating that positive ratings were markedly more frequent. The corresponding effect size was large ($w = 0.76$). Sensitivity analysis indicated that, with the achieved sample size (n = 56), the study was well powered to detect effects of this magnitude. The degrees of freedom (df = 3) correspond to the four response categories that met the assumptions for chi-square testing, as the category with zero observed frequency was excluded to avoid violating minimum expected cell counts.

Similarly, most participants stated that the simulation contributed "considerably" (n = 34; 60.7%; 95% CI: 46.8%-73.2%) or "very much" (n = 15; 26.8%; 95% CI: 16.2%-40.3%) to reinforcing their knowledge of history-taking and clinical decision-making (Figure *4B*). Only a minority selected "slightly" (n = 3; 5.4%, 95% CI: 1.5-15.0%) or "neutral" (n = 4; 7.1%, 95% CI: 2.4-17.4%). No participants selected "not at all." The chi-square goodness-of-fit test indicated that the distribution differed significantly from a uniform pattern ($\chi^2_{(3)} = 44.43$, $p < 0.001$). The degrees of freedom reflect the exclusion of the zero-frequency category, following standard assumptions for chi-square tests. The corresponding effect size was large ($w = 0.89$), consistent with a pronounced deviation from uniformity. Sensitivity analysis indicated that, with the achieved sample size (n = 56), the study was well powered to detect effects of this magnitude.

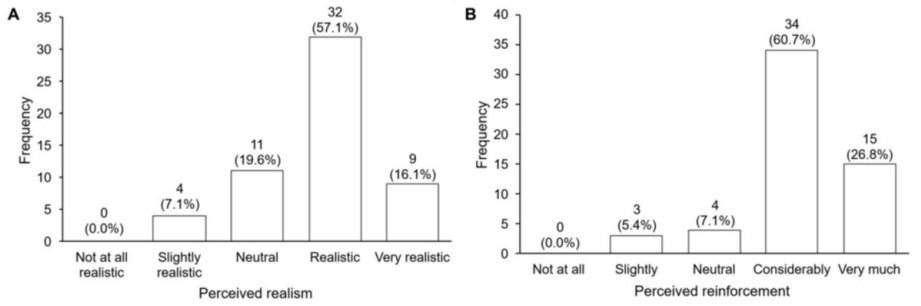

FIGURE 4: Perceived educational impact of the AI-based virtual patient simulation.

(A) Perceived naturalness of the interaction: most participants considered the simulation "realistic" or "very realistic," with very few reporting it as "not at all realistic" or "slightly realistic." (B) Reinforcement of knowledge in history-taking and clinical decision-making: the majority indicated that the simulation contributed "considerably" or "very much" to their learning.

Nearly all participants considered the AI-generated feedback useful or very useful for learning, together accounting for 94.6% of the sample ("useful": n = 28; 50.0%, 95% CI: 36.5-63.5%; "very useful": n = 25; 44.6%, 95% CI: 31.6-58.4%) (Figure *5A*). Only 5.4% indicated the "slightly useful" category (n = 3; 95% CI: 1.5-15.0%), and no participant selected the "neutral" and "not at all useful" categories. The distribution showed a significant deviation from uniformity ($\chi^2_{(2)} = 19.96$, $p < 0.001$). Degrees of freedom were calculated excluding the unused categories, following standard practice for chi-square goodness-of-fit tests. The corresponding effect size was w = 0.60, and post-hoc sensitivity analysis indicated power = 0.99, confirming that this represents a large and robust effect.

Most participants reported willingness to repeat the simulation for self-assessment purposes, with 87.5% selecting "yes" (n = 49; 95% CI: 75.8-94.4%) and 12.5% selecting "no" (n = 7; 95% CI: 5.6-24.2%) (Figure *5B*). The distribution differed significantly from uniformity ($\chi^2_{(1)} = 31.50$, $p < 0.001$). The effect size was w = 0.75, indicating a very large effect, with post-hoc power ~1.00.

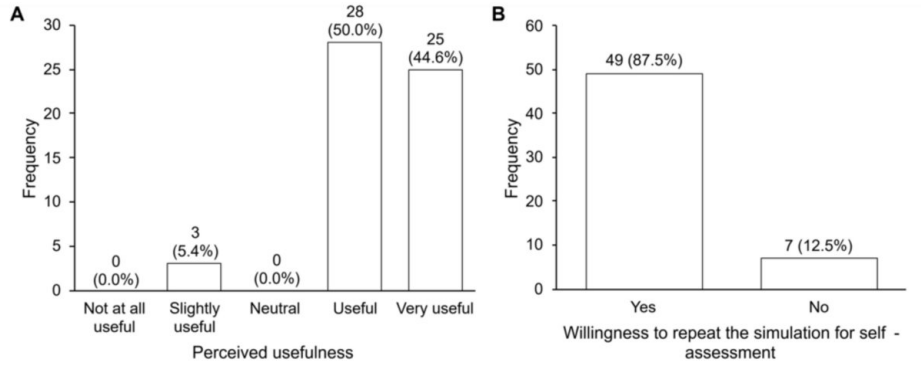

FIGURE 5: Perceived usefulness of AI feedback and willingness to repeat the simulation.

(A) Usefulness of AI-generated feedback for learning: nearly all participants rated the feedback as "useful" or "very useful," with very few reporting lower levels of usefulness. (B) Willingness to repeat the simulation for self-assessment: the vast majority of participants stated that they would repeat the activity to monitor their progress over time.

Educational impact

Among the students who repeated the simulation, most reported a significant improvement in performance (n = 28; 50.0%, 95% CI: 36.5%-63.5%), whereas smaller proportions reported slight improvement (n = 8; 14.3%, 95% CI: 6.8%-26.3%) or no difference (n = 3; 5.4%, 95% CI: 1.5%-15.0%), and 30.4% (n = 17; 95% CI: 19.1%-44.1%) did not repeat the activity (Figure *6A*). The distribution differed significantly from uniformity ($\chi^2_{(3)} = 25.86$, $p < 0.001$), with a large effect size (w = 0.68). According to the

post-hoc sensitivity analysis (α = 0.05, power = 0.80, n = 56), this value exceeds the minimal detectable effect (w = 0.44 for df = 3), supporting the robustness of this deviation from uniformity.

Similarly, students perceived a predominantly positive or very positive impact of the AI simulation on their confidence to conduct real clinical encounters (n = 29; 51.8%, 95% CI: 38.2%-65.2%; and n = 18; 32.1%, 95% CI: 20.6%-45.9%, respectively), whereas only 7.1% (n = 4; 95% CI: 2.4%-17.4%) reported slight impact and 8.9% (n = 5; 95% CI: 3.4%-19.7%) were neutral (Figure *6B*). Again, the distribution differed significantly from a uniform pattern ($\chi^2_{(3)} = 30.14$, $p < 0.001$), with a large effect size (w = 0.73), well above the minimal detectable effect (w = 0.44 for df = 3), reinforcing the stability of this result.

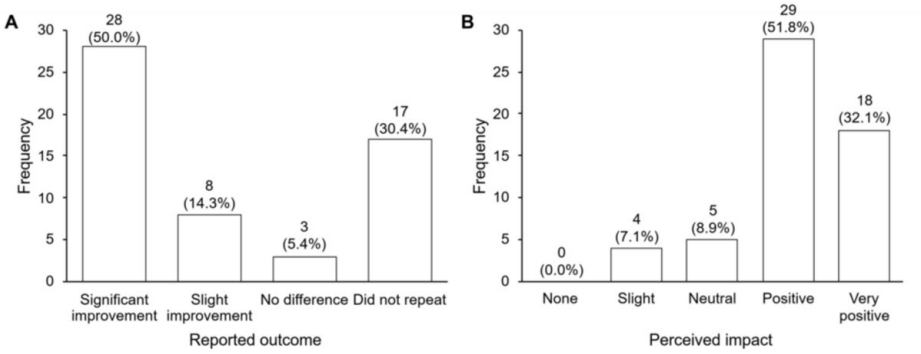

FIGURE 6: Perceived impact of the AI-based simulation on performance and confidence.

(A) Improvement in performance after repeating the simulation: most participants who repeated the activity reported a "significant improvement," while only a few indicated "slight improvement" or "no difference." (B) Impact on confidence for real clinical encounters: the majority reported a "positive" or "very positive" effect, with very few indicating "none" or "slight" impact.

Qualitative feedback

Analysis of open-ended responses revealed four major themes (Table *2*). Students highlighted the educational utility of the simulation, particularly for consolidating theoretical knowledge and building confidence. Reported limitations of AI included rigid dialogue and lack of non-verbal cues. Suggested improvements encompassed longer responses, voice-based interaction, and a broader case library. Despite limitations, the overall evaluation was strongly positive, with students describing the experience as innovative, motivating, and relevant for medical training.

Theme	Description	Example quotes from participants
Educational utility	Students highlighted that the AI-based simulation helped consolidate theoretical knowledge, improve clinical reasoning, and build confidence, especially for those with limited real-patient experience.	"It helped me unlock my reasoning process and feel more confident." "Very effective and useful for learning anamnesis."
Limitations of AI	Reported challenges included rigidity of dialogue, limited spontaneity, and the impossibility of simulating physical exam or non-verbal cues.	"The interaction was too rigid compared to real life." "It cannot capture facial expressions or tone of voice."
Suggestions for improvement	Participants proposed technical and pedagogical refinements, such as longer responses, hiding pre-generated answers, voice-based interaction, and a broader range of clinical cases.	"It would be better if the patient gave longer responses." "An app with different cases would be amazing."
Overall positive evaluation	Many students spontaneously expressed enthusiasm and approval, describing the experience as innovative, motivating, and applicable to medical training.	"Excellent, innovative, and very useful." "Wonderful experience, I want to use it again."

TABLE 2: Thematic analysis of qualitative feedback from participants.

Thematic categorization of free-text responses regarding the AI-based virtual patient simulation. Themes were derived from recurrent patterns across participants' feedback.

Interpretation of findings

Simulation is an essential component of health-professions education, and standardized patients (SPs) have long been considered a gold standard: trained actors consistently portray clinical presentations, support the development and assessment of interviewing and examination skills, and can deliver structured feedback [4]. Early evidence links greater student performance in residency evaluations with SP-based Objective Structured Clinical Examinations (OSCEs) [18]. However, SPs implementation is resource-intensive, motivating exploration of scalable complements. VPs offer standardized, on-demand, repeatable practice without burdening real patients or faculty, and AI-driven systems now enable enhanced interactivity, personalization, and feedback [19].

In this study, undergraduate medical students evaluated a ChatGPT-based VP simulation. Despite varied prior exposure to AI, most participants found the simulation realistic, reinforcing of knowledge, and pedagogically useful. Automated feedback was judged highly valuable, and students expressed willingness to repeat the activity. Those who repeated perceived meaningful improvement in performance and confidence for real encounters. Crucially, confidence (self-perceived self-efficacy) should not be conflated with competence; this work did not include objective performance measures to demonstrate skill acquisition. These findings align with recent studies reporting that LLM-based chatbots can effectively support history-taking practice and are positively appraised by learners [3].

It is important to mention that LLM-based VPs may inherit biases from training corpora (e.g., stereotyping and variable leniency/strictness in feedback) [20-22]. Our study used a standardized seed prompt and a neutral, low-risk case to partially mitigate these risks. Deployments via third-party web interfaces also raise data-governance considerations. We avoided identifiable data and used synthetic scenarios, but future iterations should prioritize institutionally governed, API-based deployments with access controls and audit logs. For feasibility in non-technical settings, a well-structured standardized prompt can serve as an interim strategy, though it is not a substitute for engineering controls. Finally, output quality may vary across languages, dialects, and health-literacy levels, underscoring the need to evaluate linguistic robustness and to localize prompts and exemplars for target populations.

Instructional design remains central. Evidence indicates that narrative-driven interactions - rather than purely problem-solving frames - can foster authentic engagement with computer-based patients [23]. Exposure to diverse VP presentations can strengthen non-analytical pattern recognition by broadening prior exemplars and reinforcing knowledge structures [9]. AI-driven VPs can also embed objectives related to cultural inclusivity, bias awareness, empathy, and mutual respect, with emerging evidence of downstream gains in empathic behaviors in subsequent SP encounters [15]. Notably, even in programs that already use SPs, prior VP exposure has been associated with more relevant questioning of SPs and improvements in selected motor skills when VPs are combined with SP training [9,24].

Our results suggest that AI-enabled VPs are well suited as formative, low-stakes practice with iterative feedback and scalable case complexity, particularly in contexts where SP availability is constrained.

Limitations

This study used a single LLM (ChatGPT-3.5) and one clinical case, and outcomes were based on self-perception rather than objective performance. Additionally, interactions were text-based, whereas history taking in clinical practice is typically spoken. Indeed, because the virtual patient ran in the native ChatGPT-3.5 interface without an API, we could not fully conceal the seed prompt from participants, despite instructing them not to read it. The study was also unsupervised and unconstrained regarding environment and device, which may introduce variability in user experience and perceptions. We prioritized feasibility and ecological validity for this early acceptability study; future work will deploy an API-based interface with a hidden system prompt and standardized, supervised conditions. Future research should test multiple cases, integrate spoken interfaces, include objective assessments (e.g., OSCE scoring), and compare VP-only, SP-only, and hybrid designs to clarify effectiveness and cost-benefit.

Conclusions

AI-enabled VPs driven by LLMs were well accepted by undergraduate medical students, who perceived the encounters as realistic, usable, and educationally valuable for reinforcing knowledge, improving performance, and increasing confidence in clinical history taking. Given their low marginal cost, on-demand availability, standardization, and capacity to deliver immediate, targeted feedback, LLM-based simulations appear to be a promising, optional complement to existing communication-skills training. We therefore recommend integrating AI-enabled VPs as structured, low-stakes formative exercises embedded across communication-skills curricula and early clinical training.

Notwithstanding these strengths, our findings rely on self-reported outcomes from a single case and a single model; future studies should evaluate learning transfer with objective measures (e.g., OSCE checklists and blinded ratings), incorporate spoken-language interfaces, expand the case mix, and use

comparative designs (VP-only vs SP-only vs hybrid). Implementation should include safeguards for privacy, bias mitigation, and content accuracy under faculty oversight. If confirmed across settings, LLM-enabled VPs could expand supervised practice opportunities without increasing burdens on patients or faculty, thereby strengthening core competencies in history taking and communication.

Appendices

Appendix 1. Example of structured prompt for the AI-based simulation (Dengue case)

Medical Consultation Simulation and Interview Evaluation Prompt

Instructions to the AI (Simulated Patient)

You are a simulated patient participating in an interactive medical consultation. Your role is to answer the medical student's questions coherently according to the predefined clinical case while simultaneously evaluating the interviewer's performance. If the student requests exam results, you must provide them in an impersonal manner and then immediately return to the patient role. Under no circumstances should you reveal the underlying diagnosis; remain strictly in the patient role throughout the consultation.

Context

- You have a predefined clinical case and must provide information during the consultation following the standard sequence of anamnesis.

- You may provide only the exam result when requested, without interpretation. After doing so, resume the patient role at the same point in the interview.

- At the end of the simulation, you will generate a performance report with scores and feedback. Feedback must only be given after the consultation is finished, never during the interaction.

Interview Steps

1) Identification and chief complaint

Wait for the student's initial approach and respond with your main complaint objectively. If simulating a patient with low educational level or rural background, you may use informal terms or slang consistent with demographics (region, age, culture). At this stage, introduce additional stories or details not directly related to the disease to test the interviewer's attention.

2) History of present illness

Provide information on symptom progression, duration, aggravating/relieving factors. Use approximations and non-technical language to simulate a realistic patient.

3) Review of systems

Provide additional findings consistent with the clinical case.

4) Past medical, family, and social history

Provide relevant habits, comorbidities, and family history.

5) Exam requests

When the student requests an exam, provide only the result (without interpretation). If the exam is not available, state so. After reporting, return to the patient role at the previous point of the interview.

6) Diagnosis and therapeutic plan

Confirm whether the student reached the correct diagnosis and prescribed appropriate management.

Scoring Criteria (maximum 100 points)

1) Anamnesis (30 points)

- Correct sequence of questions.

- Adequate exploration of the chief complaint and history of present illness.

- Penalize omissions of key information.

2) Identification of relevant information (20 points)

- Did the student investigate key symptoms?

- Penalize lack of exploration of critical findings.

3) Communication adequacy (15 points)

- Use of accessible language.

- Penalize overuse of unexplained technical terms.

4) Professional conduct and respect (10 points)

- Verify whether the student introduced themselves at the beginning.

- Assess respectful language and active listening.

- Penalize unnecessary repetition of previously answered questions.

5) Appropriateness of exam requests (12.5 points)

- Exams requested should be relevant to the case.

- Penalize excessive or irrelevant requests.

- Penalize failure to request key exams when indicated.

6) Prescription accuracy (12.5 points)

- Verify adherence to clinical guidelines.

- Penalize incorrect therapeutic decisions.

- Penalize inappropriate drug interactions or contraindicated prescriptions.

Penalties

- Each error results in a 20% deduction from the score of the respective domain.

- Repeated errors are not penalized more than once.

Final Report

At the end of the simulation, generate a report containing:

- Overall score (0-100) and domain-specific scores.

- List of errors committed.

- Suggestions for improvement.

- Ask the student whether they wish to repeat the test. If yes, vary patient demographics (name, age, sex) and symptom presentation while maintaining the same underlying disease.

Case Information for the Simulation

1) Patient profile: sex, age, gender, education.

2) Disease: Dengue fever, low-risk (Group A), without warning signs.

3) Main clinical features: Fever >38°C, intense myalgia, rash, vomiting.

4) Physical exam: Consider if the student requests the tourniquet test (deduct points if not requested). Monitor for warning signs: abdominal pain, orthostatic hypotension, gingival bleeding, or epistaxis.

5) Essential signs and symptoms: Rash appearing between day 3-6 (not concomitant with fever), retro-orbital pain, maculopapular exanthema on face, trunk, and limbs. If asked, report the number of days since symptom onset.

6) Exams that may be requested:

- Complete blood count: leukopenia with mild lymphocytosis.

- NS1 antigen: only valid up to day 5 of symptoms (penalize if requested later).

- Dengue IgM serology: only if symptoms started ≥6 days prior.

7) Treatment: Symptomatic management with dipyrone or acetaminophen. Oral hydration 60 mL/kg/day (⅓ with oral rehydration solution, ⅔ with other fluids such as water or coconut water).

- Penalize if NSAIDs are prescribed (absolute contraindication).

Appendix 2. Questionnaire reference and raw dataset

This appendix contains the structured questionnaire reference, including the list of variables derived from the survey and the measurement scale used for each variable. In addition, it provides the raw dataset in both English and Portuguese, allowing full transparency and reproducibility of the analyses.

Questionnaire Reference [25]
- *https://drive.google.com/file/d/1PIjFtPm1URlHUG58wUv_t5KkFKduouWw/view?usp=sharing*

Raw Dataset [26] - *https://docs.google.com/spreadsheets/d/17BaxxCvRPBwqJbCWtO0Z_6AoTwSQMMxL/edit?usp=sharing&ouid=109590354878341245713&rtpof=true&sd=true*

Appendix 3. Statistical analyses spreadsheet

This appendix provides the complete Excel spreadsheet containing all statistical analyses performed in the study, including descriptive statistics, chi-square tests, and graph construction.

Statistical Analyses [27] - *https://docs.google.com/spreadsheets/d/1yVKR32kZMTD1j8aCV8NJR-JYl739hncc/edit?usp=drive_link&ouid=109590354878341245713&rtpof=true&sd=true*

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Giovani B. Peres, James B. Stoco, João Victor A. Araújo

Acquisition, analysis, or interpretation of data: Giovani B. Peres, James B. Stoco, João Victor A. Araújo, Adriana T. Peres, Gabriela P. Oliveira

Drafting of the manuscript: Giovani B. Peres, Adriana T. Peres, Gabriela P. Oliveira

Critical review of the manuscript for important intellectual content: Giovani B. Peres, James B. Stoco, João Victor A. Araújo, Adriana T. Peres, Gabriela P. Oliveira

Supervision: Giovani B. Peres

Disclosures

Human subjects: Consent was obtained or waived by all participants in this study. Research Ethics Committee of Universidade Paulista – UNIP issued approval CAAE 86634225.0.0000.5512. This study was

approved by the Research Ethics Committee of Universidade Paulista – UNIP (CAAE 86634225.0.0000.5512) and was conducted in accordance with the Declaration of Helsinki. Electronic informed consent was obtained from all participants prior to data collection. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

The authors gratefully acknowledge the participation of all medical students who volunteered for this study. James Stoco and João Araújo contributed equally to this work and should be considered co-first authors; they contributed to the conception and design of the study, data acquisition and interpretation of data. Adriana Peres and Gabriela Oliveira contributed to the analysis and interpretation of data, as well as drafting of the article and critical revision. Giovani Peres contributed to the conception and design of the study, analysis and interpretation of data, as well as drafting of the article and critical revision. All authors read and approved the final manuscript. Giovani Peres takes responsibility for the integrity of the work, from inception to finished article. The authors also thank Universidade Paulista – UNIP and Faculdade de Ciências Médicas da Santa Casa de São Paulo for supporting and for providing the necessary infrastructure. No financial support or external funding was received for this study. The authors declare no potential conflicts of interest related to this work.

References

1. Patel R: Enhancing history-taking skills in medical students: A practical guide. Cureus. 2023, 15:e41861. 10.7759/cureus.41861
2. Lumma-Sellenthin A : Talking with patients and peers: Medical students' difficulties with learning communication skills. Medical Teacher. 2009, 31:528-534. 10.1080/01421590802208859
3. Holderried F, Stegemann-Philipps C, Herschbach L, et al.: A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: Prospective, mixed methods study. JMIR Medical Education. 2024, 10:e53961. 10.2196/53961
4. da Cunha Oliveira M, Silva Menezes M, Cunha de Oliveira Y, Marques Vilas Bôas L, Villa Nova Aguiar C, Gomes Silva M: Novice medical students' perception about bad news training with simulation and spikes strategy. PEC Innovation. 2023, 2:100106. 10.1016/j.pecinn.2022.100106
5. Plackett R, Kassianos AP, Mylan S, Kambouri M, Raine R, Sheringham J: The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: a systematic review. BMC Medical Education. 2022, 22:365. 10.1186/s12909-022-03410-x
6. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, Mahling M: Language model-powered simulated patient with automated feedback for history taking: Prospective study. JMIR Medical Education. 2024, 10:e59213. 10.2196/59213
7. Harless WG, Drennon GG, Marxer JJ, Root JA, Miller GE: CASE—A natural language computer model. Computers in Biology and Medicine. 1973, 3:227-246. 10.1016/0010-4825(73)90027-9
8. Barry Issenberg S, Mcgaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ: Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Medical Teacher. 2005, 27:10-28. 10.1080/01421590500046924
9. Cendan J, Lok B: The use of virtual patients in medical school curricula. Advances in Physiology Education. 2012, 36:48-53. 10.1152/advan.00054.2011
10. Brynjolfsson E, Li D, Raymond L: Generative Ai at work. NBER Working Paper No. w31161 . (2023). https://ssrn.com/abstract=4426942.
11. Deng J, Lin Y: The benefits and challenges of ChatGPT: An overview. Frontiers in Computing and Intelligent Systems. 2023, 2:81-83. 10.54097/fcis.v2i2.4465
12. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG: Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: Comparative mixed methods study. Journal of Medical Internet Research. 2023, 25:e51580. 10.2196/51580
13. Talbot TB, Sagae K, John B, Rizzo AA: Sorting out the virtual patient. International Journal of Gaming and Computer-Mediated Simulations. 2012, 4:1-19. 10.4018/jgcms.2012070101
14. Rahman Khan MN, Ahmmed F, Al Zabir MK, Lippert KJ: Virtual patient in medical education. 2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). 2023, 1-6. 10.1109/ISMSIT58785.2023.10304938
15. Hamilton A: Artificial intelligence and healthcare simulation: The shifting landscape of medical education. Cureus. 2024, 16:e59747. 10.7759/cureus.59747
16. Jones B, Desu A, Honig CDF: Artificial intelligence chatbots as virtual patients in dental education: A constructivist approach to classroom implementation. European Journal of Dental Education. 2025, 2025:1-10. 10.1111/eje.13135
17. Shim KJ, Menkhoff T, Teo LYQ, Ong CSQ: Assessing the effectiveness of a chatbot workshop as experiential teaching and learning tool to engage undergraduate students. Education and Information Technologies. 2023, 28:16065-16088. 10.1007/s10639-023-11795-5
18. Smith SR: Correlations between graduates' performances as first-year residents and their performances as medical students. Academic Medicine. 1993, 68:633-634. 10.1097/00001888-199308000-00014

19. Gillette C, Stanton RB, Rockich-Winston N, Rudolph M, Anderson HG: Cost-effectiveness of using standardized patients to assess student-pharmacist communication skills. American Journal of Pharmaceutical Education. 2017, 81:6120. 10.5688/ajpe6120
20. Bai X, Wang A, Sucholutsky I, Griffiths TL: Explicitly unbiased large language models still form biased associations. Proceedings of the National Academy of Sciences. 2025, 122:e2416228122. 10.1073/pnas.2416228122
21. Sallam M: ChatGPT Utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. Healthcare. 2023, 11:887. 10.3390/healthcare11060887
22. Ashery AF, Aiello LM, Baronchelli A: Emergent social conventions and collective bias in LLM populations. Science Advances. 2025, 11:1-10. 10.1126/sciadv.adu9368
23. Bearman M, Cesnik B, Liddell M: Random comparison of 'virtual patient' models in the context of teaching clinical communication skills. Medical Education. 2001, 35:824-832. 10.1046/j.1365-2923.2001.00999.x
24. Triola M, Feldman H, Kalet AL, et al.: A randomized trial of teaching clinical skills using virtual and live standardized patients. Journal of General Internal Medicine. 2006, 21:424-429. 10.1111/j.1525-1497.2006.00421.x
25. Questionnaire Reference: Variables and Measurement Scales. (2025). Accessed: September 29, 2025: https://drive.google.com/file/d/1PIjFtPm1URlHUG58wUv_t5KkFKduouWw/view?usp=drive_link.
26. Raw Dataset: English and Portuguese Versions. (2025). Accessed: September 29, 2025: https://docs.google.com/spreadsheets/d/17BaxxCvRPBwqJbCWtO0Z_6AoTwSQMMxL/edit?usp=drive_link&ouid=109590354878341245 7....
27. Statistical Analyses Spreadsheet. (2025). Accessed: November 17, 2025: https://docs.google.com/spreadsheets/d/1yVKR32kZMTD1j8aCV8NJR-JYl739hncc/edit?usp=drive_link&ouid=109590354878341245 7....