Preprints.org

Article

Not peer-reviewed version

DIALOGUE: A Generative AI–Based Pre–Post Simulation Study to Improve Diagnostic Communication in Medical Students Using Type 2 Diabetes Scenarios

Ricardo Xopan Suárez-García , Quetzal Chavez-Castañeda , Rodrigo Orrico-Pérez , Sebastián Valencia-Marin , Ari Evelyn Castañeda-Ramírez , Efrén Quiñones-Lara , Claudio Adrián Ramos-Cortés , Areli Marlene Gaytán-Gómez , Jonathan Cortés-Rodríguez , Jazel Jarquín-Ramírez , Nallely Guadalupe Aguilar-Marchand , Graciela Valdés-Hernández , Tomás Eduardo Campos-Martínez , Alonso Vilches-Flores , Sonia León-Cabrera , Adolfo René Méndez-Cruz , Brenda Ofelia Jay-Jímenez * , Héctor Iván Saldívar-Cerón *

Posted Date: 10 July 2025

doi: 10.20944/preprints202507.0877.v1

Keywords: Generative AI; medical education; diagnostic communication; ChatGPT; virtual patient; standardized patient; communication training; empathy in diagnosis; formative simulation

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

DIALOGUE: A Generative AI–Based Pre–Post Simulation Study to Improve Diagnostic Communication in Medical Students Using Type 2 Diabetes Scenarios

Ricardo Xopan Suárez-García [1,2], Quetzal Chavez-Castañeda [2] , Rodrigo Orrico-Pérez [2], Sebastián Valencia-Marin [1,2], Ari Evelyn Castañeda-Ramírez [1,2], Efrén Quiñones-Lara [1,2], Claudio Adrián Ramos-Cortés [1,2,3], Areli Marlene Gaytán-Gómez [1,2,3], Jonathan Cortés-Rodríguez [3], Jazel Jarquín-Ramírez [2,4], Nallely Guadalupe Aguilar-Marchand [2,4], Graciela Valdés-Hernández [2,5], Tomás Eduardo Campos-Martínez [2,5], Alonso Vilches-Flores [2,5], Sonia León-Cabrera [2,5,6], Adolfo René Méndez-Cruz [2,7] Brenda Ofelia Jay-Jímenez [2,4,*] and Héctor Iván Saldívar-Cerón [1,2,5,6,*]

[1] Unidad de Remisión de Diabetes Mellitus (URDM), Facultad de Estudios Superiores-Iztacala, Universidad Nacional Autónoma de México, Tlalnepantla, 54090 México

[2] Carrera de Médico Cirujano, Facultad de Estudios Superiores-Iztacala, Universidad Nacional Autónoma de México, Tlalnepantla, 54090 México

[3] Laboratorio de Medicina de la Conservación, Escuela Superior de Medicina, Instituto Politécnico Nacional (IPN), Plan de San Luis y Díaz Mirón, Colonia Casco de Santo Tomás, Miguel Hidalgo, Mexico City, 11350 Mexico

[4] Centro Internacional de Simulación y Entrenamiento en Soporte Vital Iztacala (CISESVI), Facultad de Estudios Superiores-Iztacala, Universidad Nacional Autónoma de México, Tlalnepantla, 54090 México

[5] Academia del Módulo de Sistema Endocrino, Carrera de Médico Cirujano, Facultad de Estudios Superiores Iztacala, Universidad Nacional Autónoma de México (UNAM), Tlalnepantla 54090, Estado de México, México

[6] Unidad de Biomedicina (UBIMED), Facultad de Estudios Superiores Iztacala, Universidad Nacional Autónoma de México, Tlalnepantla 54090, Mexico

[7] Laboratorio de Inmunología (UMF), Facultad de Estudios Superiores Iztacala, Universidad Nacional Autónoma de México, Los Barrios N° 1, Los Reyes Iztacala, Tlalnepantla 54090, Estado de México, Mexico

[8] Colegio de Ciencias y Humanidades, Plantel (I) Azcapotzalco (CCH), Av. Aquiles Serdan No. 2060, Ex-hacienda del Rosario, Azcapotzalco, México City, 02020, México

* Correspondence: mcbrendajay@iztacala.unam.mx (B.O.J.-J.); ivansaldi@iztacala.unam.mx (H.I.S.-C); Tel.: (+52 5547976105) (B.O.J.-J.); Tel.: (+52 55-79-80-15-50) (H.I.S.-C.)

Abstract

Effective diagnostic communication—delivering a diagnosis with clarity, structure, and empathy—remains a challenging competency for many undergraduate medical students. This single-arm pre–post study evaluated a generative artificial-intelligence (GenAI) training module designed to improve diagnostic-communication performance. Thirty clinical-phase students completed two pre-test encounters in which they disclosed a type 2 diabetes mellitus (T2DM) diagnosis to a virtual patient powered by ChatGPT (GPT-4o) and were scored with an eight-domain rubric by blinded raters. They then undertook ten asynchronous GenAI scenarios with automated natural-language feedback, followed seven days later by two post-test consultations with human standardized patients assessed in real time with the same rubric. Mean total performance increased by 36.7 points (95 % CI: 31.4–42.1; p < 0.001), and the proportion of high-performing students rose from 0 % to 70 %. Gains were significant across all domains, most notably in opening the encounter, closure, and diabetes-specific explanation. Multiple regression showed that lower baseline empathy (β = –0.41, *p*= 0.005) and higher digital self-efficacy (β = 0.35, *p*= 0.016) independently predicted greater improvement; gender displayed only a marginal effect. Cluster analysis revealed three learner profiles, with the

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

highest-gain cluster characterised by low empathy and high digital self-efficacy. Inter-rater reliability was excellent (ICC ≈ 0.90). These findings provide empirical evidence that GenAI-mediated practice can produce meaningful, measurable enhancements in diagnostic-communication skills and may serve as a scalable, individualised adjunct to conventional clinical education.

Keywords: generative AI; medical education; diagnostic communication; ChatGPT; virtual patient; standardized patient; communication training; empathy in diagnosis; formative simulation

1. Introduction

Artificial intelligence (AI) is rapidly transforming medical education, with generative AI models like ChatGPT garnering significant attention for their ability to produce human-like dialogue and support learning [1]. ChatGPT and similar large language models (LLMs) can swiftly retrieve and synthesize medical information, potentially accelerating students' learning across various domains [2]. A key advantage of these AI tools is their capacity to tailor educational experiences – they can analyze individual learners' needs and provide personalized feedback, adapting to different learning styles and knowledge gaps [2]. Beyond knowledge delivery, generative AI chatbots offer the novel ability to engage learners in interactive case simulations. Recent perspectives suggest that AI-driven dialogues could bolster clinical problem-solving by simulating patient encounters, thereby fostering students' critical thinking and diagnostic acumen in a realistic, real-time manner [2]. In short, generative AI has opened exciting possibilities for scalable, on-demand clinical training in medical education.

Communication skills, especially those related to diagnosis, are a core competency for physicians and a critical focus for medical training. Effective doctor–patient communication improves the quality of care, whereas poor communication can undermine history-taking, hinder accurate diagnosis, and negatively affect patient adherence to treatments [3]. Importantly, deficiencies in communication – rather than medical knowledge – are often cited as a leading cause of patient dissatisfaction and complaints [3]. Despite the well-recognized importance of communication, many trainees still feel underprepared in this domain. For example, a recent survey found that while over 70% of senior medical students had received some instruction in "breaking bad news," only about 17% felt adequately prepared to actually deliver serious diagnoses to patients [4]. Nearly all students in that study agreed on the necessity of being well-prepared to communicate difficult news [4], highlighting a gap between current training and learner needs. Clearly, there is an educational imperative to strengthen medical students' diagnostic communication skills – from effective history-taking and explanation of clinical reasoning to empathetic disclosure of diagnoses – through improved teaching methods and practice opportunities.

Simulation-based education has long been an integral strategy for teaching clinical communication and diagnostic reasoning skills in a safe, controlled environment [5]. Traditional simulations often involve standardized patients (actors trained to portray patients), which have been shown to significantly improve learners' diagnostic accuracy, communication abilities, and overall clinical competence [5]. Simulated patient encounters have long been used to develop communication skills in medical training, particularly through Objective Structured Clinical Examinations (OSCEs). These face-to-face simulations are effective but resource-intensive, requiring extensive time, personnel, and coordination, which inherently limits their frequency and scalability [6]. Previous attempts to digitize these interactions—using scripted virtual patients or rule-based chatbots—have yielded mixed results, often hindered by limited interactivity, lack of realism, and poor accessibility [7]. The recent emergence of generative AI models capable of producing coherent, unscripted dialogue opens new possibilities for scalable, on-demand simulations [8]. Unlike traditional systems, LLMs like GPT-4 can simulate dynamic patient behavior, adapt to diverse learner inputs, and provide real-time feedback, potentially transforming how students rehearse complex conversations such as

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

diagnostic delivery [9]. Indeed, the use of virtual patient software has already shown positive educational effects; studies and systematic reviews report that computer-based simulations can enhance students' clinical reasoning and skill acquisition, and are generally well-received by learners [10]. Now, state-of-the-art generative models like ChatGPT enable highly dynamic patient–doctor conversations that were not previously possible. Educators can leverage these AI-driven virtual patients to offer medical trainees essentially unlimited practice of clinical scenarios without the logistical constraints of scheduling actors or specialized labs [10]. Such AI-powered simulations can be tailored to specific learning objectives, provide immediate feedback, and ensure standardized yet responsive patient interactions for every student [10]. In sum, combining time-tested simulation pedagogy with modern AI technology presents a compelling approach to improve diagnostic communication training in a scalable way.

Early experiences with generative AI in medical training are emerging, and the results are promising. They are other pilots studies have explored using ChatGPT or related LLMs as conversational partners for clinical simulations. For instance, Scherr et al. demonstrated that ChatGPT-3.5 could successfully generate interactive clinical case simulations, allowing students to practice forming diagnostic impressions and management plans over a full patient encounter [11]. Notably, the AI-driven cases were able to adapt to learners' inputs and questions, more closely mimicking real-life dialogues than static text vignettes [11]. Other interventions have focused specifically on communication skills. One recent pilot study programmed ChatGPT as a virtual patient in a breaking-bad-news scenario (using the SPIKES protocol) and had second-year medical students conduct a difficult diagnosis discussion via text. The pre–post findings were encouraging: students' self-confidence in communicating with patients – particularly in delivering bad news – increased significantly after the ChatGPT interaction, and their trust in AI as a useful training tool also improved [12]. Qualitative feedback from that study indicated that learners valued the structured practice and felt the AI exercise helped them understand the patient's perspective [12]. Similarly, another experiment using an advanced voice-interactive ChatGPT-4 system as a virtual standardized patient found that medical trainees were satisfied with the realism of the encounter and were enthusiastic about integrating such AI-based practice into their learning [10]. Participants in that study appreciated the uniform and safe environment the AI patient provided, and many reported the experience as a valuable supplement to their clinical education [10]. At the same time, students acknowledged the current limitations of AI simulations – for example, a chatbot cannot yet fully convey human emotions or nuanced non-verbal cues. In the breaking-bad-news simulation, learners emphasized that the ChatGPT exercise was a helpful adjunct but not a replacement for real patient interaction, largely due to the lack of genuine emotional exchange and the predictable nature of AI responses [12]. These early trials underscore both the potential and the challenges of generative AI in communication training. Overall, the evidence to date – though still limited in scope – suggests that AI chatbots can provide meaningful practice that improves learners' confidence and skills, especially when used as part of a blended educational approach [12].

In educational research, pre–post intervention designs are commonly used to evaluate the impact of novel teaching methods on learners' skills and attitudes. By measuring outcomes before and after an intervention in the same cohort, this design allows for an initial assessment of effectiveness without requiring a separate control group. Such an approach has been applied in communication skills training and has demonstrated clear gains. For example, one program that combined didactics and a simulated patient encounter to teach first-year students how to deliver bad news saw the proportion of students reporting confidence in this skill jump from 32% before training to 91% afterward [13]. In a similar vein, the ChatGPT-based breaking-bad-news pilot described above used a pre–post survey to show significant improvement in students' self-rated communication confidence following the AI-facilitated practice session [12]. Although pre–post studies have inherent limitations (such as lack of a parallel control group and potential response-shift bias), they are invaluable for pilot assessments of educational innovations, helping to determine whether an intervention shows promise and warrants further development or more rigorous testing [12,13].

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

In this context, the present study aimed to evaluate the effectiveness of a generative AI–driven simulation tool designed to train diagnostic communication skills in medical students. Through a structured pre–post intervention, we assessed whether repeated simulated conversations with a responsive virtual patient could improve learners' ability to deliver a diagnosis clearly, accurately, and empathetically. Beyond feasibility, we sought to generate robust evidence on learning outcomes and predictive factors associated with greater improvement. By combining performance-based assessment, cluster analysis, and learner feedback, this study contributes to the growing field of AI-enhanced medical education and supports the integration of generative AI as a pedagogical adjunct for clinical communication training [14].

2. Materials and Methods

2.1. Study Design and Overview

The DIALOGUE study (DIagnostic AI Learning through Objective Guided User Experience) was a prospective, single-arm, pre–post intervention designed to evaluate the effectiveness of a GenAI-based training program on diagnostic communication performance in medical students. The study took place between May and June 2025 at the Facultad de Estudios Superiores Iztacala (FESI), Universidad Nacional Autónoma de México (UNAM). Ethics approval was obtained from the institutional committee (ID: CE/FESI/042025/1922), and all participants provided written informed consent. A target sample size of 30–32 students was selected based on feasibility for a pre–post design with blinded evaluation, consistent with prior AI-based educational studies in medical training [12,13]. This number was considered sufficient to detect large within-subject effect sizes in communication outcomes, assuming 80% statistical power and $\alpha = 0.05$, as supported by previous pilot simulations.

2.2. Participants

Eligible participants were undergraduate medical students from the FESI-UNAM, enrolled in the clinical phase of their MD program (semesters 2 to 7). Convenience sampling was used to recruit students from May to June 2025. Inclusion criteria included: (a) enrollment in clinical semesters, (b) completion of at least one core clerkship, (c) no prior formal training in diagnostic communication, and (d) no participation in prior pilot studies involving AI tools. Baseline characteristics—including age, gender, academic semester, GPA, prior use of generative AI tools, digital self-efficacy, empathy (Jefferson Scale of Empathy–Student version), and previous clinical exposure to real patients—were collected via structured questionnaires. The MD program at FES-Iztacala spans six years, including preclinical and clinical phases. The clinical phase begins in year 3 and progresses through structured clerkships and simulation-based training modules."

2.3. Study Phases

The study was conducted in three sequential phases: (1) a pre-test diagnostic communication simulation, (2) an asynchronous remote AI-based training intervention, and (3) a post-test diagnostic consultation with a human SP.

2.3.1. Pre-Test Phase

In the pre-test phase, each participant completed two simulated diagnostic encounters, where they were required to communicate a diagnosis of T2DM to a virtual patient powered by a generative AI model (ChatGPT, GPT-4o-mini, release 2025-05-15; temperature = 0.7, max_tokens = 1000). Prior to each simulation, participants were provided with a five-minute written clinical case summary. Interactions were conducted individually, in real time, through audio-based dialogue and without scripted prompts. Each performance was independently evaluated by two blinded clinical raters using a validated 8-domain diagnostic communication rubric (see Section 2.4). All clinical scenarios

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

were co-designed by the research team and faculty members with expertise in endocrinology and medical education. Each case involved a newly diagnosed adult with type 2 diabetes mellitus and incorporated psychosocial variables of moderate complexity, such as denial, family concern, or emotional distress. Both scenarios presented a diagnosis of T2DM but differed slightly in psychosocial framing—Scenario 1 involved a middle-aged patient with minimal emotional reaction, while Scenario 2 portrayed a younger adult with significant concern about long-term complications and lifestyle impact. All participants completed both scenarios in randomized order to mitigate sequencing bias and ensure balanced exposure.

2.3.2. Educational Intervention: AI-Based Training

Following the pre-test, participants completed an asynchronous two-part educational intervention delivered remotely:

Module 1: Prompt Engineering Workshop — A 20-minute instructional session introducing the principles of effective prompt construction for clinical simulations using generative AI tools.

Module 2: Diagnostic Communication Workshop — A 40-minute training video focused on core elements of patient-centered diagnostic delivery, including empathy, emotional regulation, clarity of explanation, and strategies for communicating a T2DM diagnosis.

After completing both modules, each student conducted ten independent simulated diagnostic conversations with ChatGPT acting as a virtual patient. Each scenario was designed with increasing psychosocial complexity and targeted specific communication competencies (e.g., adapting medical language to patient understanding, managing emotional responses, or addressing denial). Simulations were performed in separate conversation threads. At the conclusion of each interaction, participants entered the prompt “FEEDBACK” to trigger an automated reflection by the AI, providing personalized formative guidance in natural language. All simulation links were submitted to the study coordinator and archived for subsequent qualitative analysis.

2.3.3. Post-Test Phase

Seven days after completing the AI-based training, each participant engaged in two live diagnostic consultations with human standardized patients (SPs), each representing one of the two original clinical scenarios. Thus, every participant was delivered two diagnoses during the post-test phase. The encounters lasted approximately 5–7 minutes each and were conducted in person. Performances were assessed in real time by a third clinical evaluator, blinded to the participant’s prior exposure to the AI intervention. Scoring was based on the same 8-domain rubric used in the pre-test phase to ensure consistency across evaluations. Post-test evaluations were carried out by clinical raters who were licensed physicians with 5 to 15 years of clinical experience and formal teaching appointments in the MD program at FES-Iztacala. All raters underwent structured calibration using a standardized scoring manual and were blinded to participants’ pre-test scores and AI training exposure.

2.4. Evaluation Instruments

The primary outcome was the improvement in diagnostic communication competency, assessed using a modified version of the Kalamazoo Essential Elements Communication Checklist (Adapted), originally developed by the American Academy on Communication in Healthcare [15]. The rubric was adapted for T2DM diagnostic disclosure and included eight domains: (1) Opening and rapport, (2) Patient-centered history, (3) Empathic listening, (4) Clarity of explanation, (5) Emotional containment, (6) Lay language use, (7) Shared decision-making, and (8) Professionalism. Each domain was scored on a 5-point Likert scale (1 = poor, 5 = excellent). Rubric descriptors were refined through pilot testing with five students not included in the main analysis. Clinical evaluators received structured calibration training using a standardized scoring manual. Inter-rater discrepancies ≥2 points were resolved through consensus discussion or adjudication by a third blinded reviewer.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

2.5. Data Collection and Analysis

Rubric scores and self-reported measures (confidence and empathy) were collected manually using standardized paper-based forms and subsequently transcribed into a digital spreadsheet. Pre- and post-test scores were paired for each participant using anonymized identifiers. All statistical analyses were performed using R version 4.5.1 (R Foundation for Statistical Computing, Vienna, Austria). The Shapiro–Wilk test was employed to assess the normality of score distributions. Depending on the results, either paired t-tests or Wilcoxon signed-rank tests were applied to compare pre- and post-intervention scores at both the total and domain-specific levels. Effect sizes were calculated using Cohen's d for parametric tests or r for non-parametric comparisons. Inter-rater reliability was evaluated using intraclass correlation coefficients (ICCs; two-way random effects model, absolute agreement).

2.6. Qualitative Analysis of AI Feedback

AI-generated feedback from the training phase was analyzed through inductive thematic analysis. Two independent researchers manually coded the feedback using Microsoft Excel (Microsoft Corporation, Redmond, WA, USA), following an iterative, line-by-line approach. Codes were then grouped into themes through constant comparison. Discrepancies in coding were resolved through discussion until consensus was reached. The process aimed to identify core patterns in learners' reflections and AI-generated suggestions.

2.7. Use of Generative AI in the Study

Generative AI was utilized in two distinct capacities within the study: (1) ChatGPT (GPT-4o-mini; OpenAI, San Francisco, CA, USA) served as a virtual simulated patient during the training phase; and (2) the same model was prompted to generate natural language feedback following each simulated consultation. No model fine-tuning or external training was conducted. Safety settings, token limits, and API parameters were configured in accordance with OpenAI's safety guidelines (version 2.4). Generative AI was not used in the writing, editing, or translation of the manuscript text.

2.8. Statistical Analysis

All statistical analyses were conducted using R version 4.5.1 (R Foundation for Statistical Computing, Vienna, Austria). A two-tailed p-value < 0.05 was considered statistically significant. Baseline characteristics were summarized using descriptive statistics. Normality of continuous variables was assessed via the Shapiro–Wilk test. Paired comparisons between pre- and post-intervention scores (total and domain-specific) were performed using paired t-tests or Wilcoxon signed-rank tests, with effect sizes reported as Cohen's d or rank-based r, respectively. Improvement scores (Δ-scores) were calculated as the difference between post-test and pre-test rubric totals. To identify predictors of improvement, we fitted multiple linear regression models including baseline empathy, self-efficacy, GPA, gender, prior AI use, and other relevant covariates. Model selection followed stepwise procedures using the Akaike Information Criterion (AIC), and multicollinearity was assessed using variance inflation factors (VIF). Model assumptions were verified via residual plots. Cluster analysis (k-means, k = 3) was performed on standardized baseline variables to identify latent learner profiles. Clusters were compared on Δ-scores using ANOVA or Kruskal–Wallis tests with Tukey or Dunn's post-hoc corrections. Correlation analyses (Pearson or Spearman) were used to explore associations between baseline traits (e.g., empathy, motivation) and performance outcomes. Rubric reliability was assessed through Cronbach's alpha (internal consistency) and intraclass correlation coefficients (ICC, two-way random effects, absolute agreement) for inter-rater agreement. Sensitivity analyses excluded participants with high baseline scores or incomplete training engagement. No data imputation was performed.

3. Results

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

3.1. Participant Flow and Baseline Characteristics

3.1.1. Participant Flow

Of the 41 medical students assessed for eligibility, 32 met the inclusion criteria and agreed to participate (Figure 1). All 32 students completed the baseline questionnaire and both pre-test diagnostic consultations with an AI-based virtual patient. Subsequently, all enrolled participants engaged in and completed the asynchronous AI remote training module. However, two students (6.3%) were lost to follow-up and did not attend the scheduled human standardized patient (SP) post-test. Therefore, 30 participants (93.8%) completed the full study protocol and were included in both the intention-to-treat (ITT) and per-protocol analyses.

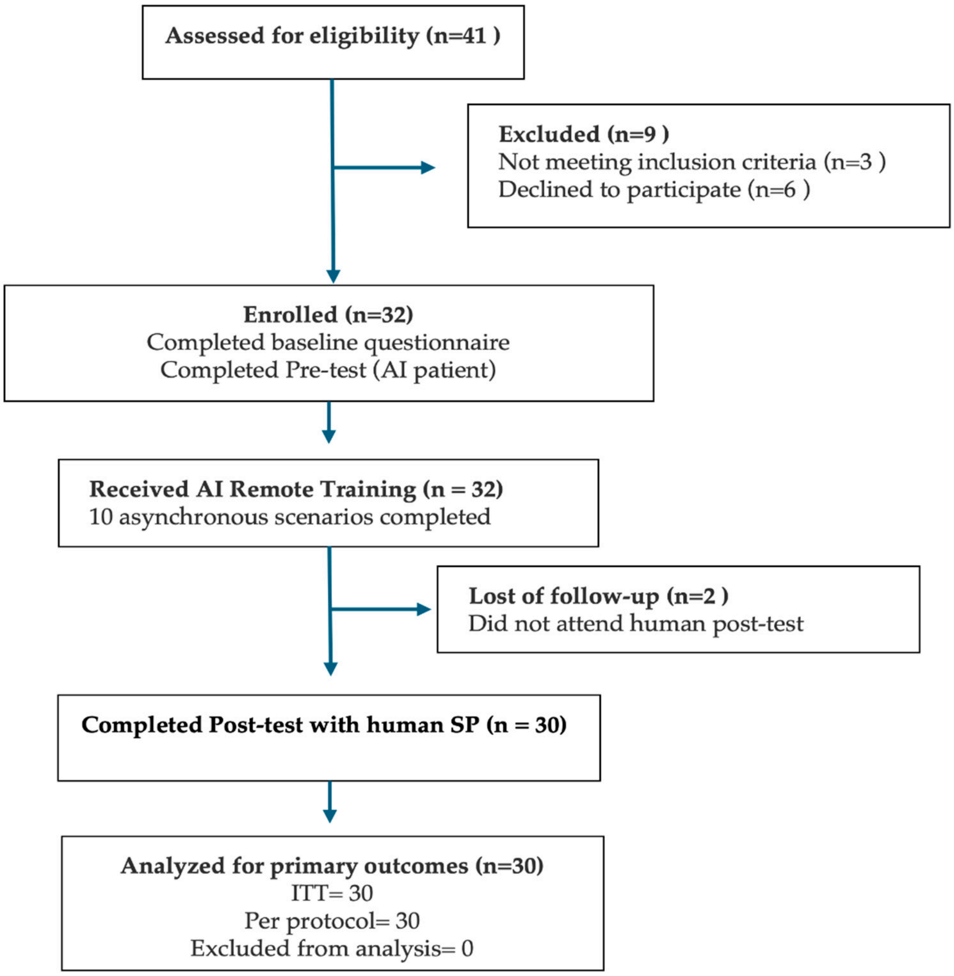
Assessed for eligibility (n=41)
Excluded (n=9)
Not meeting inclusion criteria (n=3)
Declined to participate (n=6)
Enrolled (n=32)
Completed baseline questionnaire
Completed Pre-test (AI patient)
Received AI Remote Training (n = 32)
10 asynchronous scenarios completed
Lost of follow-up (n=2)
Did not attend human post-test
Completed Post-test with human SP (n = 30)
Analyzed for primary outcomes (n=30)
ITT= 30
Per protocol= 30
Excluded from analysis= 0

Figure 1. Participant flow through the DIALOGUE pre–post study, based on the 2025 CONSORT diagram. The figure outlines the number of students assessed for eligibility, enrolled, exposed to the AI-based remote training intervention, and completing the SP post-test. Final analyses were conducted on all participants who completed the protocol [16].

3.1.2. Baseline Characteristics

Baseline characteristics for the enrolled cohort (n = 32) are summarized in Table 1. The mean age was 21.1 ± 4.1 years (range: 19–43), with 22 participants identifying as female (73.3%) and 8 as male (26.7%). Most students were in the fourth clinical semester (63.3%), with smaller proportions in semester 2 (13.3%), semester 6 (6.7%), and semester 7 (16.7%). No participant had previously received formal instruction in diagnostic communication, and none had completed an Objective Structured Clinical Examination (OSCE). Over two-thirds of students (70%) reported having ≥10 prior

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

interactions with ChatGPT or similar large language models (LLMs), and 40% had previous experience communicating clinical findings to real patients. Digital self-efficacy was rated as high or very high by 43.3%, medium by 53.3%, and low by 3.3% of students. Laptops or desktop computers were the most frequently used devices for simulation (73.3%), followed by tablets and smartphones. Internet quality was self-rated as "good" (56.7%), "medium" (33.3%), or "poor" (10%). The median self-reported motivation to participate was 9 on a 10-point scale. Confidence in core communication skills varied: highest for explaining laboratory results (mean = 3.3 ± 0.7), and lowest for performing teach-back (mean = 2.3 ± 0.8). The mean total score on the Jefferson Scale of Empathy was 116 ± 15.

Table 1. Baseline sociodemographic, academic, and digital profile of enrolled medical students (n = 30.

Variable	Category/units	N (%) or Mean ± SD
Age	Years	21.1 ± 4.1
Sex	Female	22 (73.33%)
	Male	8 (26.66 %)
Clinical semester	2	4 (13.33%)
	4	19 (63.33%)
	6	2 (6.66%)
	7	5 (16.66%)
Cumulative GPA	0-10	8.31 ± 0.50
Prior formal course in diagnostic communication	Yes	0 (0%)
Prior ECOE completed	Median (IQR)	0 (0%)
Prior experience with real patients	Yes	12 (40.00%)
Prior ChatGPT/LLM use	≥10 interactions	21 (70%)
	<10 interactions	9 (30%)
	Never	1 (3.33%)
Digital self-efficacy †	Very high / high	13 (43.33%)
	Medium	16 (53.33%)
	Low	1 (3.33%)
Device most used for simulation	Laptop / PC	22 (73.33%)
	Tablet	4 (13.33%)
	Smartphone	4 (13.33%)
Self-rated internet quality	Good	17 (56.66%)
	Medium	10 (32.25%)
	Poor	3 (10%)
Motivation score (1–10)	0-10	9
Self-confidence score, 1–5	Explain lab results	3.3 ± 0.7
	Explain DM criteria	3.0 ± 0.8
	Convey bad news empathetically	3.2 ± 0.8
	Teach-back	2.3 ± 0.8
	Anxiety reduction	2.4 ± 0.7
Jefferson Scale of Empathy, total (20–140)	-	116 ± 15

† Self-efficacy categories derived from a 5-point Likert scale (Very low = 1, Very high = 5).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

3.2. Baseline Diagnostic-Communication Performance

At baseline, the mean total rubric score was 48.83 ± 9.65 for Scenario 1 and 51.10 ± 9.82 for Scenario 2, showing a modest but statistically significant difference (*p*= 0.05). This suggests that some participants may have experienced procedural gains or growing familiarity with the simulation format between the two baseline encounters. Despite this overall difference, no significant variation was observed at the domain level (all *p* >0.05), indicating general consistency across the eight communication domains. However, item-level comparisons revealed significant differences in a few key subskills. Participants scored higher in Scenario 2 when assessed on their ability to assess the daily-life impact of diabetes (Item 4.3, *p*= 0.02), to check patient understanding through teach-back (Item 5.3, *p*= 0.01), and to close the consultation with empathy and follow-up planning (Item 7.3, *p*= 0.02). Additionally, the ability to provide a final summary (Item 7.1) approached significance (*p*= 0.05), suggesting a mild improvement in closure behavior during the second encounter. These differences likely reflect initial learning effects or adjustment to the AI-simulated environment rather than true intervention effects, given that no formal training had yet occurred. Across both scenarios, most participants demonstrated limited baseline communication proficiency. Based on total scores, 63% (19/30) of students were classified as low performers, 33% (10/30) as intermediate, and none as high performers. The lowest baseline scores were consistently observed in the subdomains of teach-back and goal negotiation. In contrast, active listening (Item 3.2) received the highest mean score (4.00 ± 0.18 in Scenario 1). However, this rating should be interpreted with caution, as it may reflect the structural nature of AI-mediated conversations. Unlike human interactions—where overlaps and interruptions are common—dialogues with ChatGPT occur in a turn-based format, which may have inadvertently facilitated uninterrupted responses and inflated perceived listening quality. A radar plot comparing mean domain scores between the two pre-test scenarios (Figure 2) visually confirms the similarity in participants' communication profiles, with only minor fluctuations between domains. A detailed breakdown of scores by item and scenario is provided in Table 2.

Table 2. Baseline diagnostic-communication scores obtained during the two pre-test scenarios (n = 30).

Item	Domain / Skill (abridged)	Scenario 1 Mean ± SD	Scenario 2 Mean ± SD	*p*
1	Relationship	3.10 ± 0.58	3.15 ± 0.55	0.50
1.1	Eye contact & open posture	3.86 ± 0.73	3.83 ± 0.79	0.74
1.2	Friendly body language	3.46 ± 0.86	3.56 ± 0.72	0.41
1.3	Self-introduction & role	2.00 ± 1.16	2.06 ± 1.20	0.70
2	Opening	1.81 ± 0.71	1.92 ± 0.69	0.22
2.1	Greeting & ID verification	1.73 ± 0.63	1.63 ± 0.71	0.18
2.2	State's purpose of visit	1.90 ± 0.84	2.13 ± 0.89	0.06
2.3	Explores patient expectations	1.80 ± 0.98	2.00 ± 0.94	0.22
3	Information Gathering	2.71 ± 0.64	2.58 ± 0.69	0.12
3.1	Uses open questions	2.00 ± 1.01	2.30 ± 0.99	0.76
3.2	Listens without interrupting	4.00 ± 0.18	3.90 ± 0.54	0.32
3.3	Summarizes to confirm	2.13 ± 1.04	1.83 ±1.11	0.08
4	Patient perspective	1.91 ± 0.85	2.00 ± 0.84	0.36
4.1	Elicits beliefs about illness	1.80 ± 1.06	1.70 ± 1.02	0.57
4.2	Explores worries/concerns	2.13 ± 0.97	1.96 ± 1.03	0.23
4.3	Assesses daily-life impact	1.8 ± 0.92	2.33 ± 1.12	0.02
5	Information sharing	1.76 ± 0.76	2.05 ± 0.83	0.18
5.1	Explains diagnosis in plain language	1.80 ± 1.06	1.70 ± 1.02	0.57
5.2	Uses visual aids/examples	2.13 ± 0.97	1.96 ± 1.03	0.23
5.3	Checks understanding (teach-back)	1.80 ± 0.92	2.33 ± 1.12	0.01
6	Plan negotiation	2.03 ± 0.65	2.03 ± 0.53	1.00
6.1	Discusses therapeutic options	2.36 ± 0.85	2.10 ± 0.99	0.07
6.2	Involves patient in decisions	2.06 ± 0.90	2.16 ± 0.74	0.44
6.3	Negotiates realistic goals	1.66 ± 0.80	1.83 ± 0.83	0.20
7	Closure	1.67 ± 0.62	2.01 ± 0.59	0.06

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

7.1	Provides final summary	1.63 ± 0.71	1.86 ± 0.81	0.05
7.2	Checks for residual questions	1.53 ± 0.93	1.86 ± 0.93	0.07
7.3	Closes with empathy & follow-up	1.86 ± 0.97	2.30 ± 0.83	0.02
8	Diabetes specific	1.88 ± 0.56	1.93 ± 0.56	0.73
8.1	Explains lab results (HbA1c)	1.80 ± 0.71	1.90 ± 0.71	0.50
8.2	Provides initial plan & alleviates anxiety	1,96 ± 0.71	1.90 ± 0.75	0.57
Total	Overall score (23–115)	48.83 ± 9.65	51.1 ± 9.82	0.05

aPaired-samples t-test; Shapiro–Wilk confirmed normality for all difference distributions. Significant differences ($p < 0.05$) observed in Items 4.3, 5.3, and 7.3 (paired t-tests).

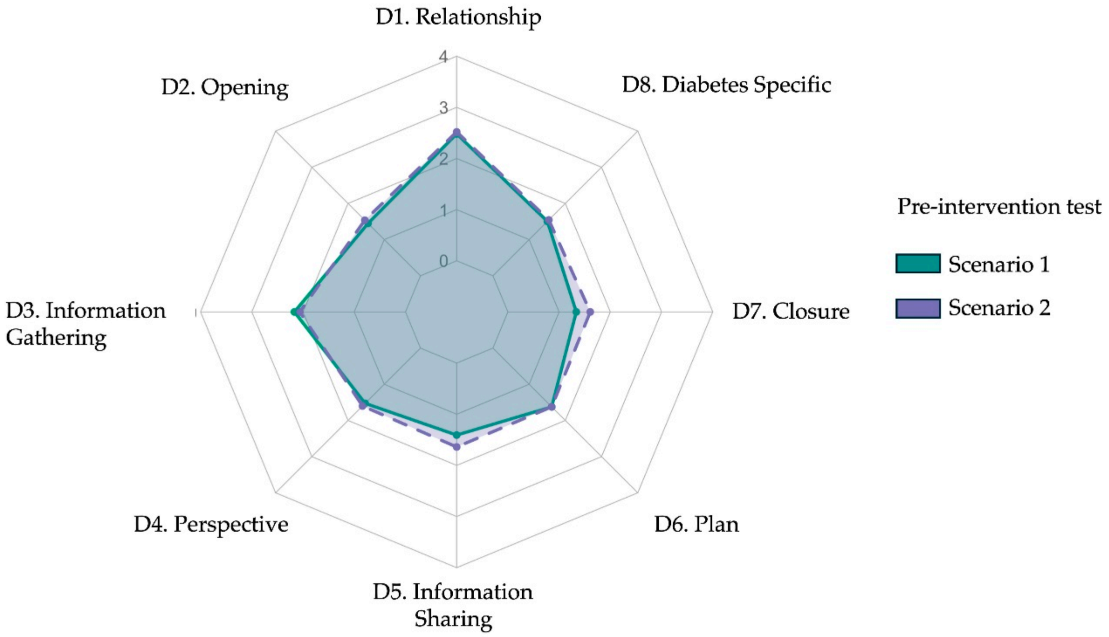

Figure 2. Radar plot comparing mean baseline diagnostic communication scores across eight rubric domains for Scenario 1 and Scenario 2. Each axis represents one domain from the adapted Kalamazoo rubric: (D1) Relationship and rapport, (D2) Opening the encounter, (D3) Information gathering, (D4) Exploring patient perspective, (D5) Information sharing, (D6) Plan negotiation, (D7) Closure, and (D8) Diabetes-specific explanation. Colored lines indicate mean Likert-scale scores (range 1–5): green for Scenario 1 and purple for Scenario 2. Pre-intervention profiles were broadly similar across domains.

3.3. Post-Intervention Diagnostic Performance and Pre–Post Comparison

3.3.1. Post-Test Performance Across Scenarios

After completing the AI-based intervention, participants underwent two live diagnostic consultations with standardized patients. The mean total rubric scores were 84.5 ± 17.8 for Scenario 1 and 88.9 ± 17.4 for Scenario 2 ($p = 0.08$), with no statistically significant difference between scenarios. As shown in Table 3, domain-level analyses revealed consistent performance across both encounters. Slight variations emerged in patient greeting (D2.1, $p = 0.01$), active listening (D3.2, $p = 0.01$), and diabetes-specific plan explanation (D8.2, $p = 0.02$), but overall post-test profiles were comparable. The spider plot in Figure 4 illustrates the near-overlapping domain scores between scenarios, suggesting internal consistency and performance stability across post-test cases.

Table 3. Baseline diagnostic-communication scores obtained during the two post-test scenarios (N = 30). No statistically significant differences were observed between Scenario 1 and Scenario 2 for any rubric item (paired t-tests, all $p > 0.05$).

Item	**Domain / Skill (abridged)**	**Scenario 1**	**Scenario 2**	***p***
		Mean ± SD	**Mean ± SD**	

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

1	Relationship	4.05 ± 0.63	4.15 ± 0.67	0.36
1.1	Eye contact & open posture	4.15 ± 0.54	4.26 ± 0.86	0.21
1.2	Friendly body language	4.26 ± 0.55	4.26 ± 0.62	1.00
1.3	Self-introduction & role	3.85 ± 0.84	3.81 ± 0.86	0.82
2	Opening	3.71 ± 0.64	3.86 ± 0.81	0.25
2.1	Greeting & ID verification	3.81 ± 0.71	4.35 ± 0.84	0.01
2.2	State's purpose of visit	3.68 ± 0.78	3.93 ± 0.96	0.11
2.3	Explores patient expectations	3.63 ± 0.81	3.30 ± 1.04	0.16
3	Information Gathering	3.86 ± 0.60	4.01 ± 0.68	0.19
3.1	Uses open questions	3.88 ± 0.67	3.96 ± 0.00	0.56
3.2	Listens without interrupting	4.21 ± 0.31	4.43 ± 0.40	0.01
3.3	Summarizes to confirm	3.50 ± 1.09	3.63 ± 1.12	0.08
4	Patient perspective	3.43 ± 0.62	3.61 ± 0.54	0.31
4.1	Elicits beliefs about illness	3.21 ± 0.78	3.15 ± 0.82	0.79
4.2	Explores worries/concerns	3.91 ± 0.57	3.60 ± 0.53	0.11
4.3	Assesses daily-life impact	3.50 ± 0.75	3.76 ± 0.52	0.15
5	Information sharing	3.43 ± 1.05	3.61 ± 1.04	0.31
5.1	Explains diagnosis in plain language	3.21 ± 1.09	3.15 ± 0.97	0.79
5.2	Uses visual aids/examples	3.91 ± 1.16	3.60 ± 1.10	0.11
5.3	Checks understanding (teach-back)	3.50 ± 1.23	3.76 ± 1.16	0.15
6	Plan negotiation	3.41 ± 0.95	3.55 ± 1.00	0.34
6.1	Discusses therapeutic options	3.53 ± 097	3.70 ± 0.90	0.42
6.2	Involves patient in decisions	3.43 ± 1.11	3.55 ± 1.16	0.37
6.3	Negotiates realistic goals	3.26 ± 1.15	3.40 ± 1.10	0.47
7	Closure	3.53 ± 1.11	3.74 ± 0.99	0.14
7.1	Provides final summary	3.30 ± 1.24	3.46 ± 0.93	0.36
7.2	Checks for residual questions	3.62 ± 0.97	3.75 ± 1.02	0.41
7.3	Closes with empathy & follow-up	3.68 ± 1.28	4.01 ± 1.10	0.06
8	Diabetes specific	3.80 ± 0.81	4.12 ± 0.79	0.03
8.1	Explains lab results (HbA1c)	3.81 ± 0.80	4.05 ± 0.71	0.12
8.2	Provides initial plan & alleviates anxiety	3.78 ± 0.97	4.20 ± 0.74	0.02
Total	Overall score (23–115)	84.46 ± 17.75	88.93 ± 17.37	0.08

aPaired-samples t-test; Shapiro–Wilk confirmed normality for all difference distributions.

3.3.2. Pre–Post Intervention Gains

When comparing pre- and post-intervention results, participants demonstrated substantial and statistically significant improvements across all rubric domains. The total score increased from 49.96 ± 9.72 to 86.70 ± 17.56 (Δ = 36.74, 95% CI: 31.39 to 42.09, $p < 0.001$; Cohen's d = 2.58), representing a very large effect size. As shown in Table 4, every domain showed meaningful gains, with the largest improvements observed in "Opening" (Δ = 1.92, d = 2.68), "Closure" (Δ = 1.79, d = 2.07), and "Diabetes-specific explanation" (Δ = 2.06, d = 2.95). The most improved subitems included teach-back (Item 5.3, Δ = 2.12), goal negotiation (6.3, Δ = 1.58), and empathetic closure (7.3, Δ = 1.77), all with large effect sizes. Figure 4 presents a radar plot overlaying pre- and post-intervention domain means, visually confirming significant gains across all communication areas. In parallel, Figure 5 displays a violin-box plot contrasting the total rubric score distribution before and after training, highlighting both the upward shift in central tendency and reduced score dispersion after the intervention. Together, these findings indicate that the generative AI–based training module produced broad and meaningful enhancements in students' diagnostic communication competencies, spanning both structural and affective elements of clinical dialogue.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

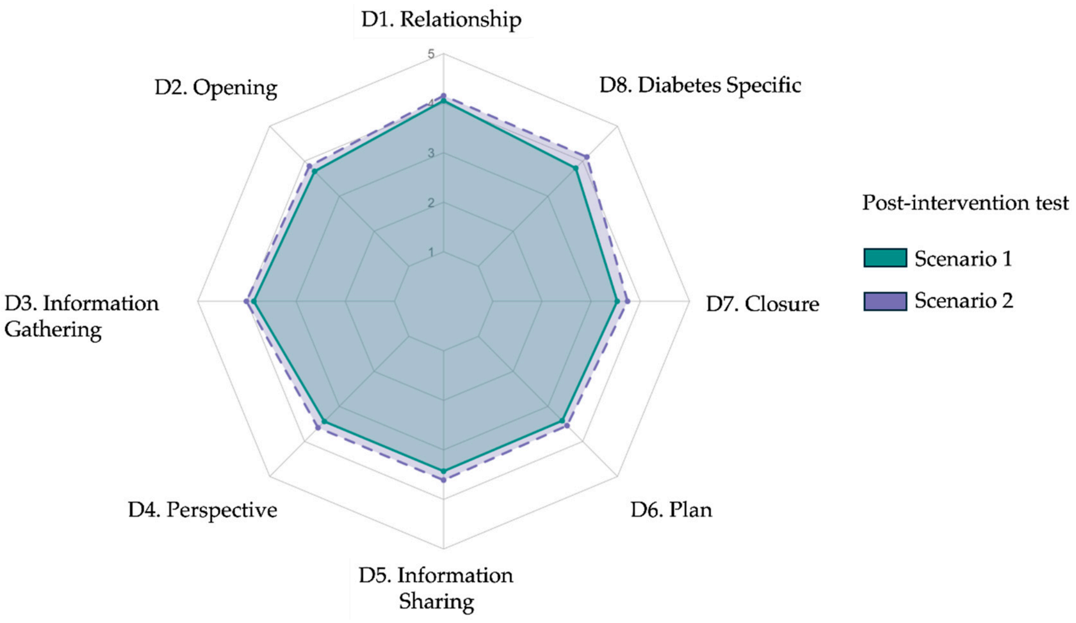

Figure 3. Radar plot comparing mean post-intervention diagnostic communication scores across eight rubric domains for Scenario 1 and Scenario 2. Each axis represents one domain from the adapted Kalamazoo rubric: (D1) Relationship and rapport, (D2) Opening the encounter, (D3) Information gathering, (D4) Exploring patient perspective, (D5) Information sharing, (D6) Plan negotiation, (D7) Closure, and (D8) Diabetes-specific explanation. Colored lines indicate mean Likert-scale scores (range 1–5): green for Scenario 1 and purple for Scenario 2. Post-intervention profiles show consistent improvement across domains.

Table 4. Pre–post comparison of diagnostic communication scores across rubric domains (N = 30). Mean scores, mean differences (Δ), 95% confidence intervals, effect sizes (Cohen’s d), and p-values are shown for each rubric item comparing pre- and post-intervention performance. All differences were statistically significant (p < 0.001, paired t-tests).

Item	Domain / Skill (abridged)	Pre-intervention Mean ± SD	Post-intervention Mean ± SD	Δ Mean	95% CI (Δ)	Cohen's d	*p*
1	Relationship	3.13 ± 0.56	4.10 ± 0.65	0.97	0.74-1.20	1.59	0.001
1.1	Eye contact & open posture	3.85 ± 0.75	4.20 ± 0.60	0.35	0.09-0.61	0.51	0.004
1.2	Friendly body language	3.51 ± 0.79	4.26 ± 0.58	0.75	0.49 -1.01	1.08	0.001
1.3	Self-introduction & role	2.03 ± 1.19	3.83 ± 0.84	1.80	1.41-2.19	1.74	0.001
2	Opening	1.86 ± 0.70	3.78 ± 0.73	1.92	1.65-2.19	2.68	0.001
2.1	Greeting & ID verification	1.68 ± 0.67	4.08 ± 0.83	2.40	2.12-2.68	3.18	0.001
2.2	State’s purpose of visit	2.01 ± 0.87	3.80 ± 0.87	1.79	1.46-2.12	2.05	0.001
2.3	Explores patient expectations	1.90 ± 0.96	3.46 ± 0.94	1.56	1.20-1.92	1.64	0.001
3	Information Gathering	2.65 ± 0.56	3.93 ± 0.64	1.28	1.05-1.51	2.12	0.001
3.1	Uses open questions	2.01 ± 0.99	3.92 ± 0.68	1.91	1.59-2.23	2.24	0.001
3.2	Listens without interrupting	3.95 ± 0.38	4.32 ± 0.37	0.37	0.23-0.51	0.98	0.001
3.3	Summarizes to confirm	1.98 ± 1.11	3.56 ± 1.09	1.58	1.17-1.99	1.43	0.001
4	Patient perspective	1.95 ± 0.56	3.52 ± 1.04	1.57	1.25-1.89	1.87	0.001
4.1	Elicits beliefs about illness	1.75 ± 1.03	3.18 ± 1.13	1.43	1.02-1.84	1.32	0.001
4.2	Explores worries/concerns	2.05 ± 0.99	3.75 ± 1.13	1.70	1.30-2.10	1.60	0.001
4.3	Assesses daily-life impact	2.06 ± 1.05	3.63 ± 1.18	1.57	1.15-1.99	1.40	0.001
5	Information sharing	1.91 ± 0.56	3.57 ± 0.92	1.66	1.37-1.95	2.17	0.001
5.1	Explains diagnosis in plain language	2.28 ± 1.09	3.54 ± 1.05	1.26	0.86-1.66	1.17	0.001
5.2	Uses visual aids/examples	1.91 ± 0.99	3.06 ± 1.33	1.15	0.71-1.59	0.98	0.001
5.3	Checks understanding (teach-back)	1.53 ± 0.92	3.65 ± 1.10	2.12	1.74-2.50	2.09	0.001

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

6	Plan negotiation	2.03 ± 0.56	3.48 ± 0.97	1.45	1.15-1.75	1.83	0.001
6.1	Discusses therapeutic options	2.23 ± 092	3.61 ± 1.00	1.38	1.02-1.74	1.43	0.001
6.2	Involves patient in decisions	2.11 ± 0.82	3.49 ± 1.13	1.38	1.01-1.75	1.39	0.001
6.3	Negotiates realistic goals	1.75 ± 0.81	3.33 ± 1.11	1.58	1.21-1.95	1.62	0.001
7	Closure	1.84 ± 0.62	3.63 ± 1.05	1.79	1.46-2.12	2.07	0.001
7.1	Provides final summary	1.75 ± 0.77	3.38 ± 1.24	1.63	1.24-2.02	1.57	0.001
7.2	Checks for residual questions	1.70 ± 0.94	3.68 ± 0.99	1.98	1.62-2.34	2.05	0.001
7.3	Closes with empathy & follow-up	2.08 ± 0.92	3.85 ± 1.19	1.77	1.37-2.17	1.66	0.001
8	DM specific	1.90 ± 0.56	3.96 ± 0.81	2.06	1.80-2.32	2.95	0.001
8.1	Explains lab results (HbA1c)	1.85 ± 0.70	3.93 ± 0.88	2.08	1.78-2.38	2.61	0.001
8.2	Provides initial plan & alleviates anxiety	1.93 ± 0.73	3.99 ± 0.89	2.06	1.75-2.37	2.53	0.001
Total	Overall score (23–115)	49.96 ± 9.72	86.70 ± 17.56	36.74	31.39-42.09	2.58	0.001

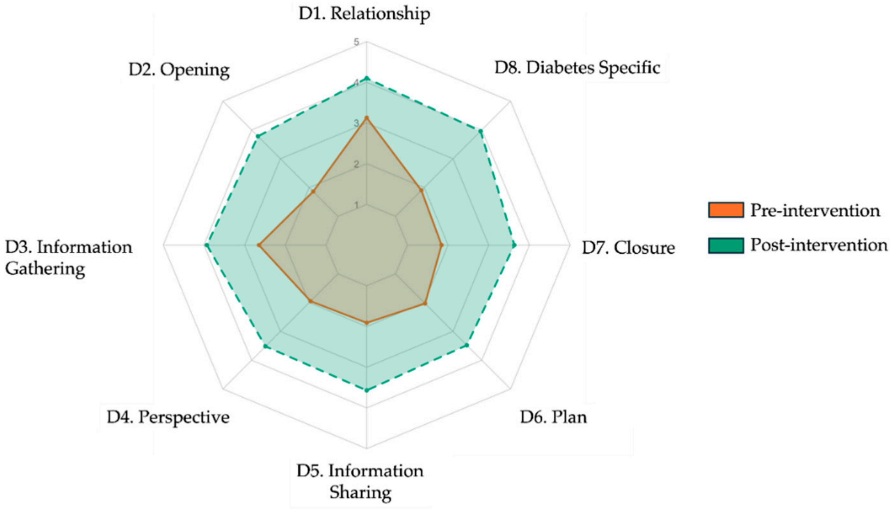

Figure 4. Radar plot comparing mean post-intervention diagnostic communication scores across eight rubric domains for Scenario 1 and Scenario 2. Each axis represents one domain from the adapted Kalamazoo rubric: (D1) Relationship and rapport, (D2) Opening the encounter, (D3) Information gathering, (D4) Exploring patient perspective, (D5) Information sharing, (D6) Plan negotiation, (D7) Closure, and (D8) Diabetes-specific explanation. Colored lines indicate mean Likert-scale scores (range 1–5): orange for Scenario 1 and green for Scenario 2. Post-intervention profiles were broadly consistent across scenarios.

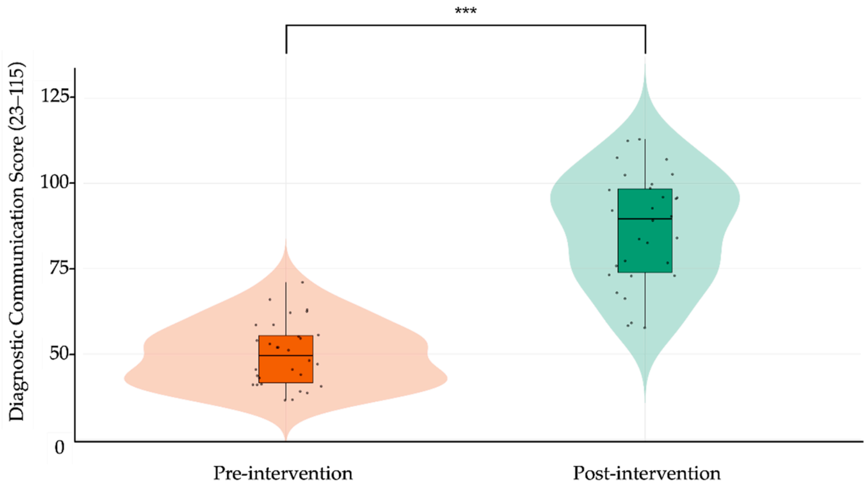

Figure 5. Violin plot comparing pre- and post-intervention diagnostic communication scores. Each dot represents an individual participant's total score (range: 23–115), derived from the adapted Kalamazoo rubric. Violin plots illustrate the distribution and density of scores; overlaid boxplots show the median (horizontal bar)

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

and interquartile range (box). Vertical lines indicate score range (minimum–maximum). Orange denotes pre-intervention scores; green denotes post-intervention scores. Asterisks (***) indicate statistically significant improvement after the intervention (paired-samples t-test, $p < 0.001$).

3.4. Predictors of Improvement

To identify baseline factors associated with the magnitude of improvement in diagnostic-communication scores, we fitted a multiple linear regression model with the Δ-score (post − pre total rubric score) as the dependent variable. Candidate predictors were age, gender, school term, GPA, previous interaction with real patients (yes/no), baseline empathy (Jefferson Scale total score), digital self-efficacy (1–5 scale), self-reported motivation (1–10 scale), and prior use of ChatGPT or other LLMs (≥ 10 interactions vs. < 10). The overall model was statistically significant (R^2= 0.43, adjusted R^2= 0.37; F(6, 23) = 4.13, p= 0.004), explaining approximately 37% of the variance in Δ-scores. Two variables emerged as independent predictors: baseline empathy was negatively associated with improvement (β = −0.41, SE = 0.13; 95% CI: −0.68 to −0.14; p = 0.005; Figure 6E), indicating that students with lower initial empathy scores tended to achieve larger gains, likely due to a ceiling effect among highly empathic learners. Conversely, digital self-efficacy was positively associated with improvement (β = 0.35, SE = 0.14; 95% CI: 0.07–0.63; p = 0.016; Figure 6D), suggesting that greater confidence with digital tools enhanced the effectiveness of GenAI training. Gender showed a marginal effect (p = 0.044; Figure 6G) in the bivariate analysis, with female students demonstrating higher median Δ-scores; however, this difference was no longer significant after adjusting for other covariates. No significant associations were observed for age (Figure 6A), school term (Figure 6B), GPA (Figure 6C), motivation (Figure 6F), prior patient contact (Figure 6H), or previous LLM use. Overall, these results indicate that learners entering the training with lower empathy yet stronger digital self-efficacy derived the greatest incremental benefit from GenAI-mediated communication practice, highlighting the combined influence of emotional baseline and technological readiness on learning outcomes. Figure 6 visualizes these relationships through scatterplots (panels A–F) and boxplots (panels G–H), underscoring the heterogeneity of individual learning trajectories.

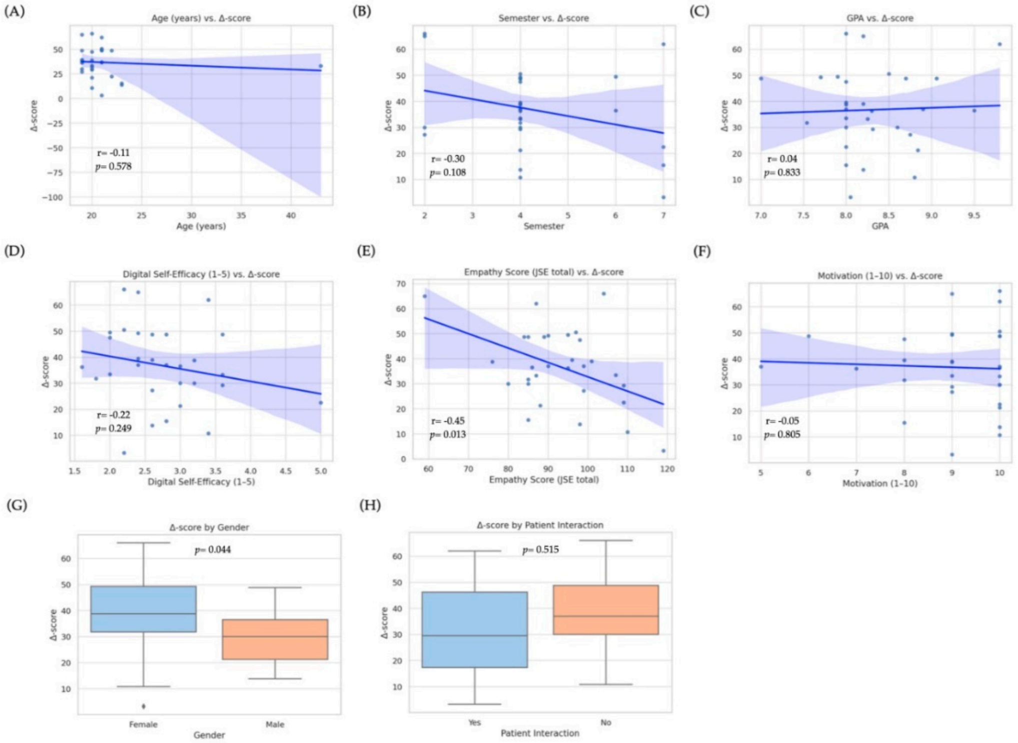

Figure 6. Associations between baseline predictors and improvement in diagnostic-communication scores (Δ-score) among medical students (n = 30). Scatterplots (top and middle rows) display bivariate Pearson correlations

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

between continuous baseline variables and the change in total rubric score (Δ-score = post − pre): (A) Age (years), (B) School term, (C) Grade Point Average (GPA), (D) Digital self-efficacy (1–5 scale), (E) Empathy score (Jefferson Scale of Empathy, JSE), and (F) Self-reported motivation (1–10 scale). Solid blue lines represent linear regression trends; shaded areas show 95 % confidence intervals. Box-and-whisker plots (bottom row) compare Δ-scores by (G) Gender (female vs. male) and (H) Prior patient interaction (yes vs. no). Each dot denotes an individual student. Only higher baseline empathy was significantly associated with lower Δ-scores in the bivariate analysis (R^2 = –0.45, *p*= 0.013); gender showed a marginal difference (*p*= 0.044). No significant bivariate correlations were observed for age, school term, GPA, digital self-efficacy, motivation, or prior patient contact. (In the multivariate model reported in Section 3.4, digital self-efficacy emerged as an additional positive predictor of improvement.).

3.5. Learner Profiles and Cluster Patterns

To examine differential response patterns, we performed an unsupervised k-means clustering (k = 3) based on participants' domain-level rubric scores before and after the intervention. The optimal k was confirmed with the elbow method and inspection of within-cluster sum-of-squares. Three discrete learner profiles emerged. Cluster A (n = 10) achieved the greatest mean improvement (Δ = 58.7 ± 10.2), followed by Cluster B (n = 12; Δ = 33.4 ± 11.5) and Cluster C (n = 8; Δ = 16.2 ± 9.3). Between-cluster differences in total Δ-score and every domain-specific Δ-score were significant (p < 0.001; Table 4), indicating meaningful heterogeneity in learning gains. Figure 7 shows domain-specific Cohen's d values, all in the moderate-to-large range. The largest effects occurred in Domain 8 (diabetes-specific explanation, d = 2.95), Domain 2 (opening the encounter, d = 2.68), and Domain 5 (information sharing, d = 2.17). The heat-map and dendrogram in Figure 8 depict these patterns: Cluster A displays uniformly high gains across all eight domains, whereas Cluster C shows marginal change, especially in "patient perspective" and "shared decision-making". Baseline trait inspection revealed that Cluster A learners started with lower empathy but higher digital self-efficacy, consistent with predictors identified in Section 3.4; Cluster C showed the opposite pattern (higher empathy, lower self-efficacy). A scatterplot of baseline total rubric score versus Δ-score (Figure 9) demonstrated a weak inverse relationship (r = –0.22, p = 0.24), supporting a compensatory effect in which students with lower starting performance realised larger gains. Collectively, these findings suggest that GenAI-mediated practice may help narrow performance gaps by especially benefiting learners who begin with weaker communication skills yet possess the technological confidence to exploit AI feedback.

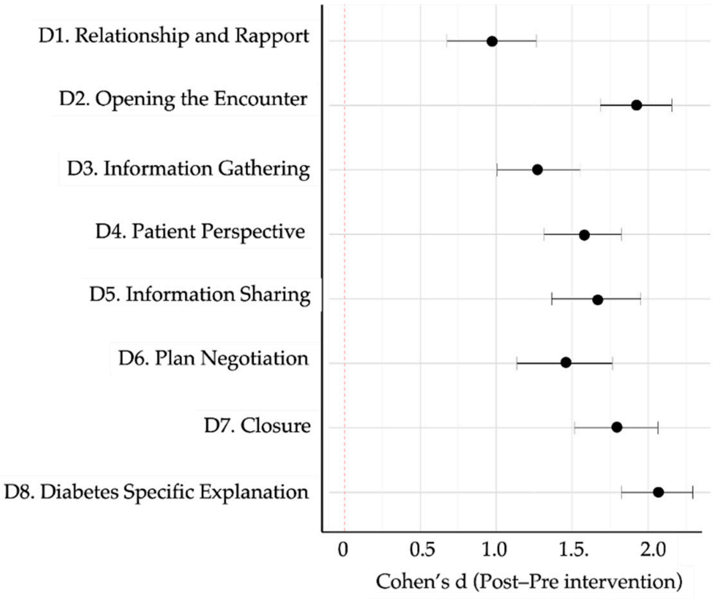

Figure 7. Forest plot showing effect sizes (Cohen's d) across the eight diagnostic communication domains. All values reflect the magnitude of improvement from pre- to post-intervention. Dots represent the mean effect size

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

per domain; horizontal lines indicate 95% confidence intervals. All domains demonstrated moderate to large effects, with the highest observed in D8 (Diabetes-Specific Explanation), D2 (Opening), and D5 (Information Sharing).

Table 4. Multiple linear regression predicting Δ-score among medical students (n = 30).

	Cluster A	Cluster B	Cluster C	
Outcome Variable	**(n = 10)**	**(n = 12)**	**(n = 8)**	***p***
Δ-score (total rubric)	58.7 ± 10.2	33.4 ± 11.5	16.2 ± 9.3	<0.001
Relationship (Domain 1)	1.23 ± 0.40	0.68 ± 0.28	0.31 ± 0.17	<0.001
Opening (Domain 2)	1.79 ± 0.52	0.96 ± 0.41	0.37 ± 0.25	<0.001
Information Gathering (D3)	1.42 ± 0.61	0.89 ± 0.36	0.34 ± 0.30	<0.001
Patient Perspective (D4)	1.51 ± 0.77	0.79 ± 0.45	0.21 ± 0.33	<0.001
Information Sharing (D5)	1.76 ± 0.91	1.03 ± 0.59	0.45 ± 0.38	<0.001
Plan Negotiation (Domain 6)	1.55 ± 0.87	0.92 ± 0.44	0.33 ± 0.28	<0.001
Closure (Domain 7)	1.62 ± 0.89	0.87 ± 0.53	0.41 ± 0.26	<0.001
DM-Specific (Domain 8)	2.04 ± 0.80	1.07 ± 0.62	0.54 ± 0.41	<0.001

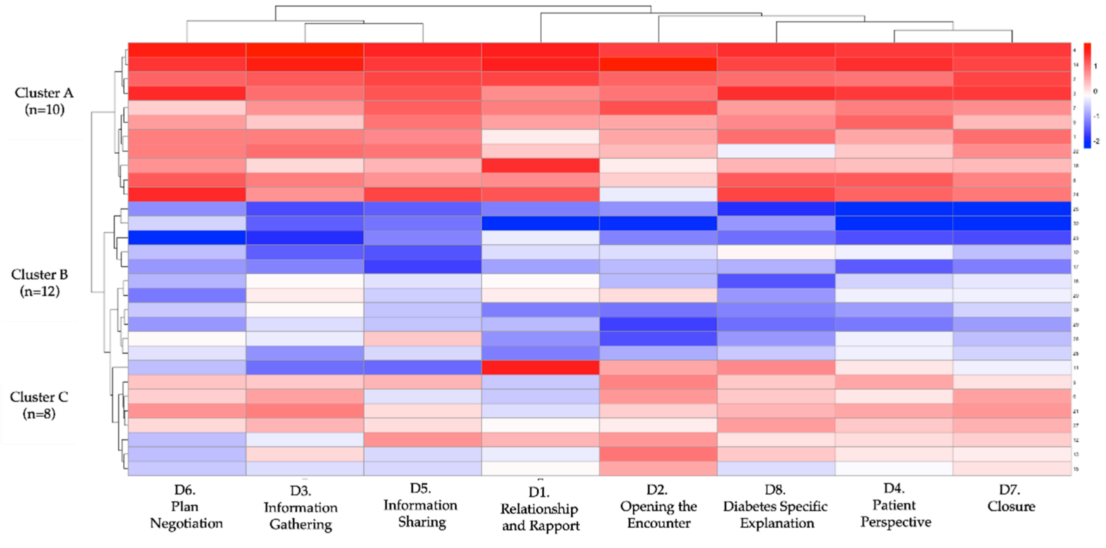

Figure 8. Cluster heatmap and hierarchical dendrogram of learner profiles based on post-intervention diagnostic communication performance. Each row represents an individual student (N = 30), and each column represents one of the eight communication domains from the adapted Kalamazoo rubric: (D1) Relationship and rapport, (D2) Opening the encounter, (D3) Information gathering, (D4) Patient perspective, (D5) Information sharing, (D6) Plan negotiation, (D7) Closure, and (D8) Diabetes-specific explanation. Scores were standardized (z-score transformation) to enable clustering and visualization. A hierarchical cluster analysis using Euclidean distance and complete linkage identified three distinct learner profiles: Cluster A (n = 10; high responders), Cluster B (n = 12; intermediate responders), and Cluster C (n = 8; low responders). Color scale indicates relative performance: red = higher scores; blue = lower scores.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

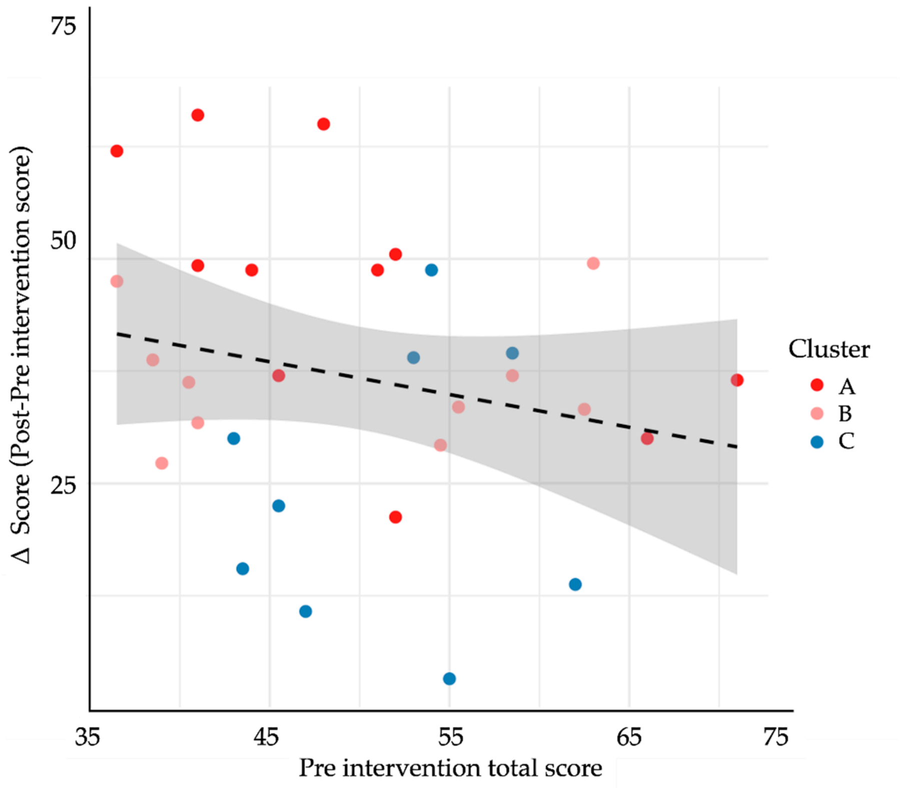

Figure 9. Association between baseline communication performance and improvement following the GenAI-based intervention. The scatterplot displays the relationship between pre-intervention rubric scores (x-axis) and Δ-scores (y-axis), calculated as post − pre total rubric score. Each point represents one student (N = 30), color-coded by cluster membership (Cluster A = red, B = pink, C = blue). The dashed regression line represents the best linear fit, and the shaded area indicates the 95% confidence interval. A negative trend was observed, suggesting that students with lower initial scores experienced greater improvements.

3.6. Post-Test Surveys

Upon completion of the post-test phase, all 30 students responded to a structured post-intervention questionnaire assessing perceived knowledge acquisition, self-efficacy in diagnostic communication, and the perceived usefulness of the GenAI training. The overall self-reported knowledge score averaged 4.2 ± 0.6 on a 5-point Likert scale, while self-efficacy in delivering a diagnosis improved from a baseline mean of 2.9 ± 0.8 to 4.3 ± 0.7. Notably, 93.3% (28/30) of participants rated the GenAI practice sessions as "very useful" or "extremely useful" for reinforcing communication skills, and 86.7% (26/30) expressed willingness to integrate AI-based simulation into future clinical training. Open-ended responses highlighted three recurring themes: (1) increased comfort navigating emotionally charged conversations, (2) appreciation for structured feedback in natural language, and (3) a desire for broader case variety and real-time faculty debriefing in future iterations. Additionally, both evaluator and SP post-revelation surveys (administered after disclosure of AI use in training) revealed no detection bias; only 1 out of 6 evaluators and 8 out of 24 SPs suspected that students had undergone prior AI-based preparation. Following disclosure, both groups acknowledged that the students' communication had exceeded their expectations.

3.7. Adverse Events and Missing Data

No adverse events or psychological distress were reported by participants during or after the intervention. Two student who completed the full AI-based remote training withdrew before the post-test simulation due to scheduling conflicts and was excluded from final analyses. No technical failures occurred during pre- or post-test evaluations. Two minor discrepancies in rubric scoring (≥2-

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

point difference between evaluators) were resolved through consensus with a third blinded reviewer. All data points from the 30 remaining participants were complete and included in the final analyses.

4. Discussion

This study provides robust evidence that a deliberately scaffolded educational intervention powered by generative artificial intelligence (GenAI) can significantly improve diagnostic communication skills among undergraduate medical students. The intervention yielded a substantial within-subject gain (Δ = +36.7 points on a 115-point rubric), with large effect sizes across all eight assessed domains. These results reflect a broad-based enhancement of both cognitive-structural and emotional-affective communication competencies rather than isolated improvements in discrete tasks. T2DM was intentionally selected as the diagnostic focus of all scenarios due to its high clinical prevalence, familiarity among medical students, and capacity to elicit both cognitive processing and empathic engagement. This choice likely facilitated student immersion and increased the ecological validity of the simulations, allowing participants to focus on communication delivery rather than clinical unfamiliarity. By anchoring the intervention in a universally relevant and emotionally resonant condition, the study maximized its potential to reveal true gains in communicative competence.

Interestingly, 73% of the participants were female, which closely reflects the actual gender distribution in the MD program at FES-Iztacala, where approximately 70% of enrolled students are women. Thus, the gender composition of the sample aligns with institutional demographics rather than selection bias. While female students are often reported to exhibit higher empathy scores and greater receptiveness to communication training, gender showed only a marginal effect on improvement in our multivariate model. Nonetheless, future studies may explore whether gender-related attitudes toward AI tools or communication style mediate training responsiveness. Importantly, our findings advance the current literature by moving beyond student self-report measures toward objective, evaluator-blinded behavioral assessments. Prior studies exploring LLM applications in medical education—such as Webb et al. (2023) and Chiu et al. (2025)—have primarily reported increased learner confidence following GenAI-assisted simulations but lacked external validation through SP encounters or blinded human scoring [12,17]. In contrast, our design incorporated pre–post evaluations using human SPs, a domain-anchored rubric adapted from the Kalamazoo Essential Elements Communication Checklist, and psychometric analyses confirming high internal consistency ($\alpha = 0.91$) and inter-rater reliability (ICC = 0.89). To our knowledge, this is the first study to generate such triangulated, rubric-based evidence in Spanish-speaking medical students, within a Latin American setting [15,18].

Several pedagogical mechanisms may underlie the observed gains. First, the GenAI platform provided asynchronous, low-stakes opportunities for deliberate practice—a well-established driver of communication skill acquisition in simulation-based education. Second, immediate feedback in natural language allowed for dynamic self-correction without requiring constant faculty oversight. Thematic analysis of over 300 AI-generated responses revealed alignment with communication best practices (e.g., "organize your explanation," "validate patient concerns," "check for understanding"), echoing established heuristics described by Nestel and Tierney (2007) [19]. Third, the virtual patient's emotionally neutral tone likely created a psychologically safe space, encouraging students to experiment with empathetic strategies without fear of judgment—an effect that mirrors findings from early simulated patient training [20].

Regression modeling confirmed an inverse relationship between baseline empathy (measured via the Jefferson Scale) and gains—students with lower initial empathy achieved larger improvements. Simultaneously, students with higher digital self-efficacy benefited more from the GenAI training. These predictors jointly explained over one-third of the variance in learning gains and resonate with constructs from the Technology Acceptance Model (Davis, 1989) [21], where perceived usability and personal relevance enhance engagement with digital tools. Interestingly, prior ChatGPT use, motivation level, and GPA were not significant predictors, suggesting that

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

emotional and technological readiness may outweigh cognitive or experiential variables when learning with AI. These results carry practical implications: baseline learner profiles may serve as a foundation for tailoring feedback scaffolding, simulation complexity, or engagement thresholds in future GenAI deployments.

Our unsupervised clustering further highlighted the heterogeneity of learning trajectories. Three learner profiles were identified, with "Cluster A" students—characterized by high empathy and self-efficacy—demonstrating nearly fourfold higher gains compared to their peers. These findings suggest that GenAI is not universally transformative, but particularly beneficial for specific learner subtypes. Future adaptive systems could dynamically adjust training sequences based on real-time profiling, maximizing pedagogical efficiency while personalizing learning trajectories.

From a broader curricular standpoint, our findings suggest that GenAI tools—if properly contextualized—can be feasibly integrated into clinical communication training in Mexican medical schools. Faculty shortages, limited access to standardized patients, and rigid curricula pose barriers to consistent development of communication skills. GenAI platforms, especially those adapted to local languages and clinical realities, offer a scalable, low-cost complement. However, successful adoption will require both technical infrastructure and institutional culture change. Some educators may view AI as a threat to pedagogical authority, while others may lack the AI literacy needed to evaluate feedback quality. Addressing this will require deliberate faculty training, transparency in AI decision logic, and curricular frameworks that combine human and machine feedback in meaningful ways.

Looking forward, the evolution of LLMs—such as the expected GPT-4.5 or GPT-5—may radically expand the capabilities of virtual patients. Upcoming models may integrate real-time voice, emotion recognition, and adaptive personas, allowing for multimodal simulation of complex encounters. These advances may offer unprecedented fidelity in diagnostic communication training. However, ethical challenges remain. Overreliance on AI feedback may desensitize learners or reinforce mechanical communication patterns. There is also a risk of depersonalization if students generalize from algorithmic interactions to real patient care. To mitigate these risks, GenAI tools must be embedded within reflective, supervised curricula that cultivate critical thinking, emotional sensitivity, and context-aware communication.

Despite its strengths, this study has several limitations. The absence of a control group limits causal inference, although the magnitude and distribution of effects make test–retest bias unlikely. The short duration precludes conclusions about long-term retention or clinical transfer. Our intervention was deployed within a single institution, potentially limiting generalizability. Although the GenAI platform functioned reliably in Spanish, subtle limitations in semantic nuance or cultural appropriateness may have affected feedback quality. Additionally, while we standardized system prompts and temperature settings, the inherent stochasticity of LLMs introduces minor variations that may affect reproducibility. Future research should explore the impact of AI randomness, develop seed-logging protocols, and benchmark AI feedback against expert commentary.

Finally, while this study provides compelling evidence of effectiveness, it does not compare GenAI-based feedback to traditional faculty-led instruction, nor does it evaluate downstream clinical outcomes such as patient satisfaction or diagnostic accuracy. These remain important avenues for future research. Nevertheless, our findings support the role of GenAI as a scalable and psychometrically sound adjunct—not a replacement—for clinical communication training. Although the present intervention centered on T2DM, the framework may be adaptable to other specialties such as pediatrics, psychiatry, or gynecology. Future iterations should investigate domain-specific adaptations and their differential impact on communication skill acquisition

5. Conclusions

This study demonstrates that generative AI can serve as a reliable, scalable tool to enhance diagnostic communication skills in medical students. A short, asynchronous GenAI-based intervention led to significant, domain-wide improvements, particularly among students with higher

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

baseline empathy and digital self-efficacy. While not a replacement for human teaching, GenAI offers a promising pedagogical adjunct, especially in resource-limited settings. Future work should focus on long-term outcomes, integration into curricula, and alignment with ethical, cultural, and clinical standards.

Author Contributions: Conceptualization, Ricardo Suárez-García and Héctor Saldívar-Cerón; Data curation, Jonathan Cortés-Rodríguez and Héctor Saldívar-Cerón; Formal analysis, Jonathan Cortés-Rodríguez and Héctor Saldívar-Cerón; Investigation, Ricardo Suárez-García, Quetzal Chavez-Castañeda, Rodrigo Orrico-Pérez, Sebastián Valencia-Marin, Ari Castañeda-Ramírez, Efrén Quiñones-Lara, Claudio Ramos-Cortés, Areli Gaytán-Gómez, Jonathan Cortés-Rodríguez, Jazel Jarquín-Ramírez, Nallely Aguilar-Marchand, Graciela Valdés-Hernández, Tomás Campos-Martínez, Alonso Vilches-Flores, Sonia León-Cabrera, Adolfo Méndez-Cruz, Brenda Jay-Jímenez and Héctor Saldívar-Cerón; Methodology, Héctor Saldívar-Cerón; Project administration, Brenda Jay-Jímenez; Resources, Héctor Saldívar-Cerón; Software, Jonathan Cortés-Rodríguez and Héctor Saldívar-Cerón; Supervision, Graciela Valdés-Hernández and Tomás Campos-Martínez; Validation, Adolfo Méndez-Cruz and Héctor Saldívar-Cerón; Visualization, Héctor Saldívar-Cerón; Writing – original draft, Héctor Saldívar-Cerón; Writing – review & editing, Graciela Valdés-Hernández, Tomás Campos-Martínez, Alonso Vilches-Flores, Sonia León-Cabrera, Brenda Jay-Jímenez and Héctor Saldívar-Cerón.

Funding: This research was supported by the Consejo Mexiquense de Ciencia y Tecnología (COMECYT) under grant FICDTEM-2023-131, and by the Universidad Nacional Autónoma de México (UNAM) through PAPIIT IA201725 and PAPIME 203825, awarded to H.I.S.C., as well as PAPIIT IN201724, awarded to S.L.C.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Ethics Committee of the Facultad de Estudios Superiores Iztacala, Universidad Nacional Autónoma de México (protocol code CE/FESI/052025/1922, approval date: 9 May 2025). The ethical approval was granted for the full duration of the study (valid from 7 May 2025 to 7 May 2026).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Written informed consent included permission to analyze anonymized performance data and survey responses, as well as to publish findings in academic journals. No identifiable personal information is presented in this publication.

Data Availability Statement: The datasets generated and analyzed during this study—including anonymized rubric scores, survey instruments, AI prompt templates, simulation scripts, and training materials—are publicly available in the Open Science Framework (OSF) repository at https://doi.org/10.17605/OSF.IO/YJF39. Audio recordings and voice files are not publicly accessible due to privacy and ethical restrictions but may be reviewed upon reasonable request and approval by the institutional ethics committee.

Acknowledgments: The authors gratefully acknowledge the standardized-patient team from the Centro Internacional de Simulación y Entrenamiento en Soporte Vital Iztacala (CISESVI), Facultad de Estudios Superiores Iztacala, Universidad Nacional Autónoma de México, whose professionalism and consistency were essential to the integrity of the post-test assessments: Rebeca Aguilar-Sánchez, Itzel Antonino-Cruz, Lizbeth Nohemí Arrazola-Live, Jorge Omar Barragán-Hermosillo, Diego Alan Calderón-Trejo, Luis Fernando Cárdenas-Gasca, Rafael Alberto Chinchilla-Herrera, María Fernanda Cruz-Carrillo, Kevin Cuandón-Quintero, Diego Díaz-Hill, Guadalupe Itzel Díaz-Roa, Jennifer Denisse Flores-Islas, Andrea Abigail Galván-Becerra, Mariana García-Morales, Samantha Karina Gómez-Islas, Yanireth Israde-Osnaya, Aida Esmeralda Juárez-Cabello, Luis Alberto Juárez-Fernández, María Elena Lara-Rodríguez, Juan Jesús López-Jiménez, Emiliano Federico Martínez-Franco, Roberto Carlos Meza-Hernández, Andrea Alejandra Murillo-Maya, Abigail Rosales-Pimentel, Luis Fernando Serrano-Trejo, Zaira Ixchel Silva-López, Evans Antonio Tierra-Lazcano, Karla Jacqueline Velázquez-Reyes, and Lourdes Aylin Villeda-Villa. We also thank the administrative staff of the Facultad de Estudios Superiores Iztacala for logistical support and the CISESVI technical team for providing audiovisual equipment and secure data storage. During the preparation of this study, the authors used ChatGPT (GPT-4o, OpenAI, May 2025 release) to generate virtual-patient dialogue, produce automated formative feedback. The authors have critically reviewed and edited all AI-generated output and take full responsibility for the final content of this publication.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AIC	Akaike Information Criterion
API	Application Programming Interface
GENAI	Generative Artificial Intelligence
GPA	Grade Point Average
GPT-4	Generative Pretrained Transformer 4
ICC	Intraclass Correlation Coefficient
ITT	Intention-To-Treat
JSE	Jefferson Scale of Empathy
LLM	Large Language Model
OSCE	Objective Structured Clinical Examination
SE	Standard Error
SP	Standardized Patient
SPIKES	Setting, Perception, Invitation, Knowledge, Emotions, Strategy (protocol for breaking bad news)
T2DM	Type 2 Diabetes Mellitus
VIF	Variance Inflation Factor
WA	Weighted Average

References

1. Feng, S.; Shen, Y. ChatGPT and the Future of Medical Education. *Academic Medicine* **2023**, *98*, 867-868, doi:10.1097/acm.0000000000005242.
2. Chokkakula, S.; Chong, S.; Yang, B.; Jiang, H.; Yu, J.; Han, R.; Attitalla, I.H.; Yin, C.; Zhang, S. Quantum leap in medical mentorship: exploring ChatGPT's transition from textbooks to terabytes. *Front Med (Lausanne)* **2025**, *12*, 1517981, doi:10.3389/fmed.2025.1517981.
3. Moezzi, M.; Rasekh, S.; Zare, E.; Karimi, M. Evaluating clinical communication skills of medical students, assistants, and professors. *BMC Medical Education* **2024**, *24*, 19, doi:10.1186/s12909-023-05015-4.
4. Santos, M.S.; Cunha, L.M.; Ferreira, A.J.; Drummond-Lage, A.P. From classroom to clinic: Addressing gaps in teaching and perceived preparedness for breaking bad news in medical education. *BMC Medical Education* **2025**, *25*, 449, doi:10.1186/s12909-024-06498-5.
5. Elendu, C.; Amaechi, D.C.; Okatta, A.U.; Amaechi, E.C.; Elendu, T.C.; Ezeh, C.P.; Elendu, I.D. The impact of simulation-based training in medical education: A review. *Medicine* **2024**, *103*, e38813, doi:10.1097/md.0000000000038813.
6. Skryd, A.; Lawrence, K. ChatGPT as a Tool for Medical Education and Clinical Decision-Making on the Wards: Case Study. *JMIR Form Res* **2024**, *8*, e51346, doi:10.2196/51346.
7. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Basel)* **2023**, *11*, doi:10.3390/healthcare11060887.
8. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health* **2023**, *2*, e0000198, doi:10.1371/journal.pdig.0000198.
9. Weisman, D.; Sugarman, A.; Huang, Y.M.; Gelberg, L.; Ganz, P.A.; Comulada, W.S. Development of a GPT-4-Powered Virtual Simulated Patient and Communication Training Platform for Medical Students to Practice Discussing Abnormal Mammogram Results With Patients: Multiphase Study. *JMIR Form Res* **2025**, *9*, e65670, doi:10.2196/65670.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2025 doi:10.20944/preprints202507.0877.v1

10. Öncü, S.; Torun, F.; Ülkü, H.H. AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians. *BMC Medical Education* **2025**, *25*, 278, doi:10.1186/s12909-025-06877-6.
11. Scherr, R.; Halaseh, F.F.; Spina, A.; Andalib, S.; Rivera, R. ChatGPT Interactive Medical Simulations for Early Clinical Education: Case Study. *JMIR Med Educ* **2023**, *9*, e49877, doi:10.2196/49877.
12. Chiu, J.; Castro, B.; Ballard, I.; Nelson, K.; Zarutskie, P.; Olaiya, O.K.; Song, D.; Zhao, Y. Exploration of the Role of ChatGPT in Teaching Communication Skills for Medical Students: A Pilot Study. *Medical Science Educator* **2025**, doi:10.1007/s40670-025-02394-9.
13. Poei, D.M.; Tang, M.N.; Kwong, K.M.; Sakai, D.H.t.; Choi, S.Y.t.; Chen, J.J. Increasing Medical Students' Confidence in Delivering Bad News Using Different Teaching Modalities. *Hawaii J Health Soc Welf* **2022**, *81*, 302-308.
14. Shorey, S.; Mattar, C.; Pereira, T.L.; Choolani, M. A scoping review of ChatGPT's role in healthcare education and research. *Nurse Educ Today* **2024**, *135*, 106121, doi:10.1016/j.nedt.2024.106121.
15. Makoul, G. Essential Elements of Communication in Medical Encounters: The Kalamazoo Consensus Statement. *Academic Medicine* **2001**, *76*, 390-393.
16. Hopewell, S.; Chan, A.-W.; Collins, G.S.; Hróbjartsson, A.; Moher, D.; Schulz, K.F.; Tunn, R.; Aggarwal, R.; Berkwits, M.; Berlin, J.A.; et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. *BMJ* **2025**, *389*, e081123, doi:10.1136/bmj-2024-081123.
17. Webb, J.J. Proof of Concept: Using ChatGPT to Teach Emergency Physicians How to Break Bad News. *Cureus* **2023**, *15*, e38755, doi:10.7759/cureus.38755.
18. Makoul, G. The SEGUE Framework for teaching and assessing communication skills. *Patient Educ Couns* **2001**, *45*, 23-34, doi:10.1016/s0738-3991(01)00136-7.
19. Nestel, D.; Tierney, T. Role-play for medical students learning about communication: Guidelines for maximising benefits. *BMC Medical Education* **2007**, *7*, 3, doi:10.1186/1472-6920-7-3.
20. Kalet, A.; Pugnaire, M.P.; Cole-Kelly, K.; Janicik, R.; Ferrara, E.; Schwartz, M.D.; Lipkin, M., Jr.; Lazare, A. Teaching communication in clinical clerkships: models from the macy initiative in health communications. *Acad Med* **2004**, *79*, 511-520, doi:10.1097/00001888-200406000-00005.
21. Davis, F.D. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* **1989**, *13*, 319-340, doi:10.2307/249008.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.