international medical education

MDPI

Article

AI-Supported Adaptive Simulation for Diagnostic Disclosure in Medical Students: A Randomized Controlled Trial

Brenda Ofelia Jay-Jímenez [1,2], Diego Alberto Martínez-Islas [1,3,4], Axel Tonatiuh Marroquin-Aguilar [1,3,4], Fernanda Avelino-Vivas [1,3,4], Dafne Montserrat Solis-Galván [1,3,4], Alexis Arturo Laguna-González [1,3,4], Bruno Manuel García-García [1,3], Eduardo Minaya-Pérez [1,3,4], Efren Quiñones-Lara [5], Ismael Martínez-Bonilla [6], Adolfo René Méndez-Cruz [1] and Héctor Iván Saldívar-Cerón [1,3,*]

1 Carrera de Médico Cirujano, Facultad de Estudios Superiores-Iztacala, Universidad Nacional Autónoma de México, Tlalnepantla 54090, Mexico; mcbrendajay@iztacala.unam.mx (B.O.J.-J.); diegomartinezislas411@gmail.com (D.A.M.-I.); axelnrd@comunidad.unam.mx (A.T.M.-A.); f_avelinovivas@comunidad.unam.mx (F.A.-V.); dafnesg2@gmail.com (D.M.S.-G.); alexisa160502@gmail.com (A.A.L.-G.); bruno43garcia@gmail.com (B.M.G.-G.); eduardo_minaya@comunidad.unam.mx (E.M.-P.); armendez@unam.mx (A.R.M.-C.)
2 Centro Internacional de Simulación y Entrenamiento en Soporte Vital Iztacala (CISESVI), Facultad de Estudios Superiores-Iztacala, Universidad Nacional Autónoma de México, Tlalnepantla 54090, Mexico
3 Unidad de Remisión de Diabetes Mellitus (URDM), Facultad de Estudios Superiores-Iztacala, Universidad Nacional Autónoma de México, Tlalnepantla 54090, Mexico
4 Programa de Maestría en Ciencias de la Salud, Escuela Superior de Medicina, Instituto Politécnico Nacional (IPN), Ciudad de México 11350, Mexico
5 Departamento de Biomedicina Molecular, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV-IPN), Ciudad de México 07360, Mexico; efren.quinones@cinvestav.mx
6 Laboratorio Digital de Desarrollo Infantil, División de Investigación y Posgrado, Universidad Nacional Autónoma de México, Ciudad de México 04510, Mexico; ismael.m.bonilla@iztacala.unam.mx
* Correspondence: ivansaldi@iztacala.unam.mx; Tel.: +52-55-79-80-15-50

Abstract

Diagnostic disclosure is a complex communication task that requires learners to integrate interpersonal attunement, structured information delivery, and condition-specific reasoning in real time. We conducted a randomized controlled trial comparing conventional diagnostic communication training with the same training supplemented by an AI-supported adaptive virtual patient simulation designed to provide additional deliberate practice and individualized, just-in-time feedback. Eighty undergraduate medical students were randomized 1:1 and completed standardized-patient encounters involving disclosure of a new diagnosis of type 2 diabetes mellitus before and after training. Performance was assessed by blinded physician raters using an adapted Kalamazoo rubric. Among students with complete pre–post data (conventional training, $n = 25$; AI-supported training, $n = 26$), both groups showed substantial improvement. Mean gains were numerically larger in the AI-supported group, with small-to-moderate standardized effects across selected communication domains; however, baseline-adjusted group-by-time interactions did not reach conventional thresholds for statistical significance, indicating that any added mean effects beyond conventional training remain uncertain. Exploratory person-level analyses suggested greater heterogeneity of improvement in the AI-supported condition, including a higher density of large gains in higher-order communication components. These findings should therefore be interpreted as exploratory rather than confirmatory. AI-supported adaptive simulation appears feasible as an adjunct to communication training, but adequately powered studies are needed to clarify effect magnitude, mechanisms, and generalizability across training contexts.

Check for updates

Academic Editor: Branko Aleksic

Received: 10 March 2026
Revised: 25 March 2026
Accepted: 31 March 2026
Published: 1 April 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the Academic Society for International Medical Education. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: education; medical; undergraduate; generative artificial intelligence; simulation training; patient simulation; formative feedback; health communication; diabetes mellitus; type 2; physician-patient relations; randomized controlled trial

1. Introduction

Effective physician–patient communication is a foundational component of high-quality medical care and is particularly critical during the diagnostic process [1,2]. Diagnostic communication requires clinicians to integrate clinical reasoning with clear and empathic dialogue when eliciting histories, explaining diagnostic uncertainty, and disclosing new diagnoses such as type 2 diabetes mellitus (T2DM) [3–6]. Deficits in this domain can undermine patient understanding, erode trust, and contribute to dissatisfaction and complaints, even when biomedical management is technically sound [4,5]. Accordingly, communication has been established internationally as a core competency in medical education, with the expectation that it can be deliberately taught, practiced, and assessed [5,7].

Despite this consensus, substantial gaps persist between curricular intent and learner preparedness. Many trainees report limited confidence in conducting diagnostic conversations, particularly those involving uncertainty or emotionally charged information, including the disclosure of chronic disease diagnoses [2,8,9]. This discrepancy reflects a broader educational challenge: although communication is widely recognized as essential, opportunities for deliberate and repeated practice with structured feedback remain uneven and often insufficient [7,10]. As a result, diagnostic communication is frequently learned opportunistically rather than systematically, leading to variability in learners' readiness for real-world clinical encounters [11].

Simulation-based education using standardized patients has become a cornerstone for teaching communication skills because it allows learners to practice complex interactions in a safe environment and receive targeted feedback [12,13]. Evidence supports its effectiveness in improving communication behaviors, diagnostic reasoning, and clinical performance [13–15]. However, standardized patient programs are resource-intensive, requiring trained actors, faculty time, and substantial logistical coordination [12,14]. These demands limit scalability and frequency of exposure, particularly in institutions with large cohorts or constrained resources [15]. Such limitations are especially relevant in low- and middle-income contexts, where formal communication curricula and simulation infrastructure may be inconsistently implemented, further exacerbating inequities in training quality [16,17].

Recent advances in generative artificial intelligence (AI) offer a potentially valuable complement to traditional simulation-based education. Large language models can generate dynamic, context-sensitive dialogue that more closely approximates authentic patient interactions than earlier rule-based or scripted virtual patients [18–20]. Conceptually, generative AI-based simulators may function as platforms for deliberate practice by enabling repeated exposure to diagnostically challenging conversations, immediate feedback, and iterative performance refinement without the temporal and financial constraints of in-person simulation [21]. By allowing learners to rehearse communication behaviors repeatedly and reflect on feedback, these systems align with established educational frameworks emphasizing repetition, feedback, self-regulation, and individualized learning [21,22].

Early applications of generative AI in medical education suggest feasibility and acceptability, with learners reporting high engagement and perceived usefulness [23–25].

However, the current literature remains dominated by descriptive studies and self-reported outcomes. Few investigations have used rigorous experimental designs and

objective, performance-based assessments to determine whether AI-supported simulation produces measurable improvements in communication skills [20,25,26]. Moreover, little is known about how such tools perform relative to conventional training or whether their effects translate into observable behavioral improvement under blinded assessment [23]. This evidence gap is particularly pronounced outside high-income settings, where empirical data on AI-supported communication training remain scarce [27]. Cultural and linguistic contexts shape communication norms, underscoring the need to evaluate such innovations across diverse educational settings [6].

In our initial DIALOGUE (DIagnostic AI Learning through Objective Guided User Experience) single-group study, generative AI–assisted simulation was associated with improved diagnostic communication performance under blinded rating [28], but causal effects could not be established because no control group was included. Whether AI-supported simulation adds measurable value beyond conventional training therefore remains an open question.

To address this gap, we conducted a randomized controlled trial evaluating the impact of integrating generative AI-based virtual patient simulation into undergraduate medical training for diagnostic communication. The trial compared conventional instruction alone with conventional instruction supplemented by AI-supported simulation, using pre- and post-intervention standardized patient assessments scored by blinded evaluators with an adapted communication rubric [5].

Guided by a deliberate practice framework [21,29], we hypothesized that AI-supported simulation would be associated with greater improvement in diagnostic communication performance than conventional training alone. We also explored whether the intervention was associated with heterogeneous patterns of improvement across learners and communication domains. By providing controlled, performance-based evidence, this study seeks to clarify how generative AI may be integrated into medical education and whether it can strengthen the development of core communication competencies.

2. Materials and Methods

2.1. Study Design and Trial Oversight

This study was a prospective, parallel-group, randomized controlled trial with a 1:1 allocation ratio designed to compare conventional diagnostic communication training with conventional training supplemented by AI-supported virtual patient simulation. The objective was to evaluate the educational impact of integrating generative AI-based simulation on diagnostic communication performance during disclosure of a new diagnosis of T2DM. Recruitment activities began in late 2025, and the intervention and outcome assessments were conducted during the intersemester period in December 2025 at the Facultad de Estudios Superiores Iztacala (FES Iztacala), Universidad Nacional Autónoma de México (UNAM), Mexico. The study protocol was finalized before participant enrollment, and no changes to eligibility criteria, interventions, outcomes, or statistical analyses were introduced after trial initiation. The study was approved by the institutional ethics committee (CE/FESI/052025/1922). All participants provided written informed consent. Because participants were undergraduate students, specific safeguards were implemented to minimize any risk of perceived coercion; participation was entirely voluntary and unrelated to academic grading or formal coursework. Patients and members of the public were not involved in the design, conduct, reporting, or dissemination of this educational trial.

2.2. Participants and Setting

Participants were undergraduate medical students enrolled in the basic sciences cycle of the MD program at FES Iztacala, UNAM. Students were in early semesters of training

and had not yet entered full clinical clerkships, a stage at which structured training in diagnostic communication may still be limited. Inclusion criteria were enrollment in the undergraduate medical program during the basic sciences cycle, availability to complete baseline and follow-up assessments, and no prior formal coursework specifically dedicated to diagnostic communication skills. Exclusion criteria included prior participation in AI-based communication training studies and incomplete baseline assessment. Participants were recruited through institutional announcements and voluntary invitations distributed by the research team. All study activities were conducted within the university's simulation facilities, with remote access used only for asynchronous components.

2.3. Randomization and Allocation Concealment

After completion of baseline assessments, participants were randomized in a 1:1 ratio to either conventional instruction alone or conventional instruction supplemented by AI-supported simulation. The random allocation sequence was generated using a computer-based procedure by a researcher not involved in participant enrollment, intervention delivery, or outcome assessment. No stratification or blocking was applied because this was an educational trial conducted in an exploratory context without sufficient prior data to justify stratification variables. Allocation concealment was maintained by releasing group assignments only after completion of baseline data collection, and the allocation sequence remained inaccessible to investigators responsible for enrollment and performance scoring.

2.4. Interventions

2.4.1. Conventional Asynchronous Instruction (Control Group)

All participants received a prerecorded asynchronous video lecture focused on diagnostic communication and disclosure of a new diagnosis of T2DM. The lecture addressed structured approaches to delivering diagnostic information and managing patient responses, integrating established communication frameworks, including SPIKES, NURSE, PEWTER, and SMART [30,31]. Content emphasized strategies for explaining the diagnosis, responding to emotional reactions, managing uncertainty, and promoting patient-centered understanding. Participants were required to complete the video module before the post-test assessment.

2.4.2. AI-Supported Simulation (Intervention Group)

In addition to the asynchronous video lecture, participants assigned to the intervention group completed an AI-supported virtual patient simulation program. Participants interacted with ChatGPT 5.2 (OpenAI, San Francisco, CA, USA), configured through standardized prompt templates developed by the research team to simulate patients receiving a new diagnosis of T2DM. The system was instructed to maintain the assigned scenario and level of emotional complexity during each interaction and to withhold formative commentary until completion of the simulated encounter. Although all participants completed the same set of simulation levels, the system adjusted the emphasis of dialogue and formative feedback according to each learner's responses, with the aim of providing more individualized practice and feedback.

The simulation consisted of 10 progressive levels of increasing emotional and psychosocial complexity, ranging from low-distress acceptance to denial, anxiety, and concern about long-term complications. Each simulated encounter required participants to communicate the diagnosis, respond to patient emotions, manage uncertainty, and organize the conversation appropriately. At the end of each level, the system generated structured formative feedback addressing clarity of explanation, empathic engagement, encounter organization, and patient-centered communication behaviors. Participants were instructed

to complete all simulation levels asynchronously during the intervention period. Prompt templates and simulation materials are available in the OSF repository.

2.5. Outcomes

The primary outcome was post-test diagnostic communication performance, measured as the total score on a diagnostic communication rubric adapted from the Kalamazoo Essential Elements Communication Checklist [5], with baseline performance included as a covariate in primary analyses. Secondary outcomes included domain-specific post-test performance across eight predefined rubric domains, also evaluated with baseline adjustment. Exploratory outcomes included descriptive characterization of individual patterns of improvement and response heterogeneity within each training modality. Baseline characteristics and learner-reported variables were collected through a structured questionnaire administered before randomization.

2.6. Assessment Procedures and Blinding

Both the pre-test and post-test consisted of a single standardized patient encounter in which participants were required to communicate a new diagnosis of T2DM to a simulated patient. Encounters were conducted in person with trained standardized patients and lasted up to 20 min. The clinical scenario was equivalent in structure across pre-test and post-test assessments to ensure consistency of task demands over time. Performance was evaluated using an eight-domain diagnostic communication rubric, with each item scored on a 5-point Likert scale. Assessments were performed by trained physician raters with experience in medical education. Outcome assessors were blinded to group allocation and were not involved in intervention delivery. Participants were aware of their assigned training condition, but outcome assessors remained blinded throughout the evaluation process.

2.7. Sample Size

Sample size planning was informed by effect estimates from our prior single-group study evaluating AI-supported diagnostic communication training using the same assessment rubric and evaluation framework. Because single-group estimates may overstate comparative effects, the present trial was conservatively planned to detect a small-to-moderate between-group effect ($f = 0.25$) in post-test performance using analysis of covariance with baseline adjustment ($\alpha = 0.05$, power > 80%). On this basis, a total sample of 72 participants (36 per group) was estimated, and a target enrollment of 80 students (40 per group) was set to account for anticipated attrition during the academic term. Of the randomized participants, 51 students (25 in the conventional training group and 26 in the AI-supported group) completed both pre-test and post-test assessments and were included in the primary performance analyses. Although this reduced statistical power for detecting small between-group differences, the final sample allowed estimation of pre–post changes, standardized effect sizes, and exploratory descriptive analyses.

2.8. Statistical Analysis

Baseline characteristics were summarized descriptively, and between-group comparability at baseline was assessed descriptively rather than through significance testing. The total post-test diagnostic communication score was prespecified as the primary confirmatory outcome. The primary analysis compared post-test total diagnostic communication scores between groups using analysis of covariance (ANCOVA), with post-test score as the dependent variable, group as the independent variable, and baseline score as a covariate. Domain-level analyses were treated as secondary outcomes and were examined using the same approach for individual rubric domains. To account for repeated measures and within-participant correlation, linear mixed-effects models with random intercepts for

participants and fixed effects for time, group, and time × group interaction were also fitted as complementary analyses. Between-group standardized effect sizes for change were estimated using Hedges' g, a small-sample corrected standardized mean difference used to quantify the magnitude and direction of between-group differences in pre–post change scores. Sensitivity analyses included complete-case analyses and multiple imputation for missing post-test data. Item-level analyses, threshold-based comparisons, and descriptive analyses of heterogeneity of response within each arm, including score distributions, heatmaps of pre–post performance, and descriptive grouping according to magnitude of change, were treated as exploratory and were not used for confirmatory hypothesis testing. Given the multiplicity of secondary and exploratory analyses, these results were interpreted cautiously. Statistical significance was defined as a two-sided p-value < 0.05. All analyses were conducted in R using RStudio (version 2025.05.1+513).

3. Results

3.1. Participant Flow and Baseline Characteristics

A total of 300 students were assessed for eligibility, of whom 80 met inclusion criteria, provided informed consent, and were randomized in a 1:1 ratio to conventional training or conventional training supplemented by AI-supported simulation (40 per group). Recruitment activities were completed in late 2025, and intervention delivery and follow-up assessments were conducted during the intersemester period in December 2025. Participant flow through enrollment, allocation, follow-up, and analysis is presented in Figure 1. Of the randomized participants, 51 completed both pre-test and post-test assessments and were included in the primary analyses (conventional training, n = 25; AI-supported training, n = 26). Attrition occurred in both groups and was primarily attributable to incomplete completion of training activities or absence from the scheduled post-test assessment. To explore potential attrition-related bias, participants who completed both assessments were compared with those who did not on baseline demographic, educational, and communication performance variables; no statistically significant differences were identified across the measured baseline variables. Nevertheless, given the level of attrition, the analyzed sample may still represent a more engaged subset of students, and selection bias cannot be excluded. Baseline demographic and educational characteristics of randomized participants are summarized in Table 1. Baseline diagnostic communication performance, assessed using the Kalamazoo-based rubric, is reported in Table 2, with item-level baseline distributions provided in Table A1 (Appendix A). Baseline performance appeared broadly comparable between groups, supporting the use of prespecified baseline-adjusted analyses in subsequent outcome comparisons.

Table 1. Baseline demographic and educational characteristics of randomized participants.

Variable	Total (n = 80)	Conventional (n = 40)	AI-Supported (n = 40)
Age (years)	19.45 ± 1.83	19.15 ± 1.19	20.67 ± 2.99
Academic semester	Sem 1: 26 (32.5%) Sem 2: 48 (60%) Sem 3: 3 (3.75%) Sem 4: 3 (3.75%)	Sem 1: 10 (25%) Sem 2: 26 (65%) Sem 3: 2 (5%) Sem 4: 2 (5%)	Sem 1: 16 (40%) Sem 2: 22 (55%) Sem 3: 1 (2.5%) Sem 4: 1 (2.5%)
Sex, n (%)			
Female	48 (60.0%)	23 (57.5%)	25 (62.5%)
Male	31 (38.75%)	16 (40.0%)	15 (37.5%)
Non-binary	1 (1.25%)	1 (2.5%)	0 (0%)
Extracurricular clinical experience, n (%)	5 (6.3%)	3 (7.5%)	2 (5.0%)

Values are presented as mean ± SD or n (%).

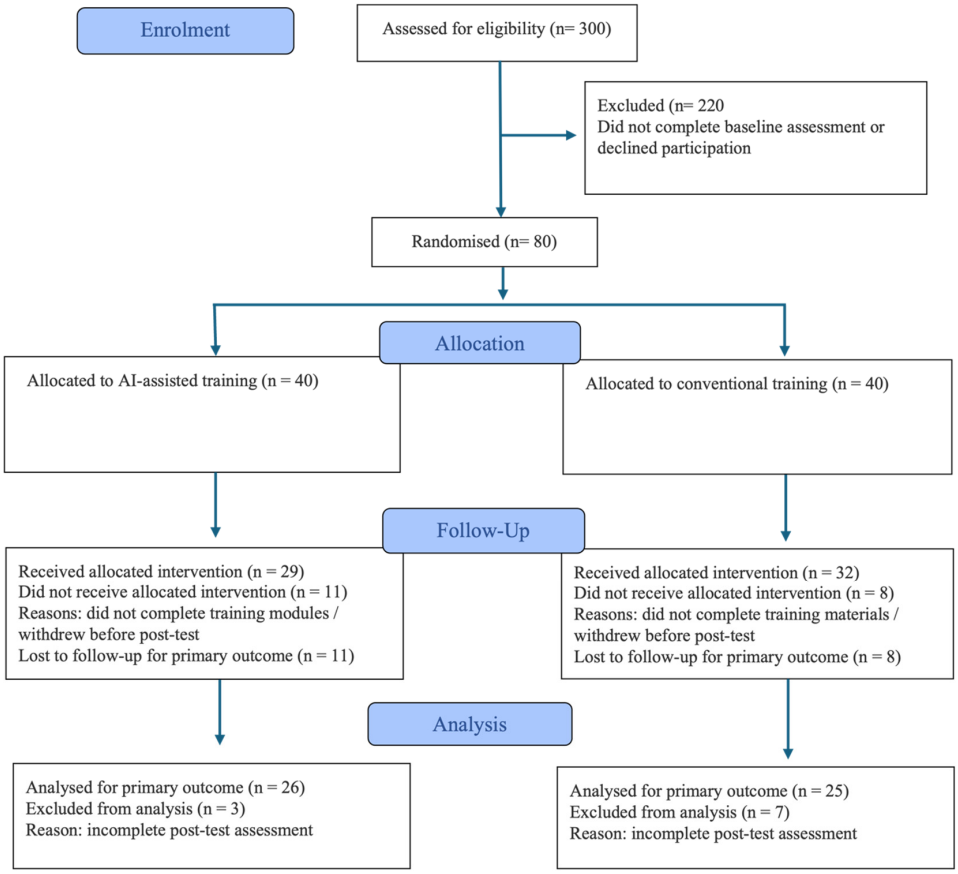


Figure 1. CONSORT flow diagram of participant enrollment, allocation, follow-up, and analysis. The figure depicts participant flow through the phases of the randomized controlled trial, including enrollment, allocation to conventional training or AI-supported training, follow-up, and inclusion in the primary outcome analysis. Of the 300 students assessed for eligibility, 80 met inclusion criteria, provided informed consent, and were randomized in a 1:1 ratio. Attrition occurred primarily because of incomplete completion of training modules or absence from the post-test assessment. The number of participants included in the primary outcome analysis for each group is indicated.

Table 2. Baseline diagnostic communication performance by Kalamazoo domains.

Item	Domain (Kalamazoo Framework)	Conventional (n = 40) (Mean ± SD)	AI-Supported (n = 40) (Mean ± SD)
1	Build a Relationship	6.57 ± 2.34	5.70 ± 1.65
2	Opening the Discussion	5.19 ± 1.56	4.74 ± 1.38
3	Gather Information	5.82 ± 1.86	5.88 ± 1.95
4	Understand the Patient's Perspective	4.77 ± 1.65	4.62 ± 1.44
5	Share Information	4.41 ± 1.44	5.31 ± 1.56
6	Reach Agreement	5.61 ± 1.98	5.73 ± 1.59
7	Provide Closure	4.98 ± 2.04	5.13 ± 1.80
8	Diabetes-specific communication	5.13 ± 1.80	5.04 ± 1.35
Total	Overall score (24–120)	42.50 ± 10.55	42.14 ± 9.67

Domain scores are reported as domain totals (range 3–15) based on three items per domain. The total score represents the sum of all 24 items (range 24–120). Values are presented as mean ± standard deviation. Baseline performance is reported descriptively to assess comparability between randomized groups. No formal statistical comparisons were conducted at baseline.

3.2. Primary Outcome: Diagnostic Communication Performance

Changes in overall diagnostic communication performance from pre-test to post-test among participants who completed both assessments (n = 51; conventional training, n = 25;

AI-supported training, $n = 26$) are summarized in Table 3. Substantial improvements in total communication scores were observed in both groups, indicating that meaningful gains occurred over time under both training conditions. Mean pre–post improvement in overall communication performance was numerically larger in the AI-supported group than in the conventional training group. At the total-score level, this difference corresponded to a moderate standardized effect size favoring AI-supported training (Hedges' g = 0.57). These effect-size estimates describe the magnitude and direction of observed differences but do not, by themselves, establish statistical superiority. In baseline-adjusted longitudinal mixed-effects models accounting for repeated measures and baseline performance (Table 4), the time × group interaction coefficients generally favored AI-supported training across most domains and for the total score; however, none reached conventional thresholds for statistical significance. At the total-score level, the estimated time × group interaction was 7.45 points (95% CI, −1.98 to 16.88; $p = 0.121$). Accordingly, although several point estimates and standardized effect sizes numerically favored the AI-supported group, the reduced final sample size limited precision, and baseline-adjusted analyses did not demonstrate statistically significant superiority over conventional instruction alone. These findings should therefore be interpreted as exploratory rather than confirmatory.

Table 3. Pre–post changes in diagnostic communication performance by Kalamazoo domains and training modality.

Item	Domain (Kalamazoo Framework)	Conventional Δ ($n = 25$) Mean ± SD	AI-Supported Δ ($n = 26$) Mean ± SD	Hedges' g (AI-Supported Minus Conventional)
1	Build a Relationship	5.88 ± 2.42	5.77 ± 2.80	−0.04
2	Opening the Discussion	5.16 ± 2.88	6.08 ± 2.23	0.35
3	Gather Information	5.40 ± 2.61	6.31 ± 2.05	0.38
4	Understand the Patient's Perspective	5.96 ± 2.91	7.19 ± 2.56	0.44
5	Share Information	6.24 ± 2.98	7.54 ± 2.18	0.49
6	Reach Agreement	4.88 ± 2.85	6.46 ± 2.85	0.55
7	Provide Closure	5.00 ± 3.54	7.23 ± 2.70	0.70
8	Diabetes-specific communication	5.04 ± 3.43	6.38 ± 2.61	0.44
Total	Overall score (24–120)	43.56 ± 19.38	52.96 ± 12.85	0.57

Values are presented as mean change (post-test − pre-test) ± standard deviation. Domain scores were derived from the Kalamazoo Essential Elements Communication Checklist. The total score ranged from 24 to 120, with higher scores indicating better diagnostic communication performance. Analyses included only participants with complete pre- and post-test assessments. Effect sizes are reported as Hedges' g.

Table 4. Longitudinal mixed-effects model of diagnostic communication performance by domain.

Item	Domain (Kalamazoo Framework)	Time × Group β	95% CI	*p*-Value	Hedges' g
1	Build a Relationship	−0.09	−0.56 to 0.38	0.707	−0.09
2	Opening the Discussion	0.41	−0.06 to 0.88	0.086	0.44
3	Gather Information	0.45	−0.03 to 0.93	0.066	0.46
4	Understand the Patient's Perspective	0.34	−0.15 to 0.83	0.173	0.35
5	Share Information	0.31	−0.12 to 0.74	0.157	0.36
6	Reach Agreement	0.24	−0.31 to 0.79	0.39	0.24
7	Provide Closure	0.49	−0.16 to 1.14	0.138	0.39
8	Diabetes-specific communication	0.38	−0.17 to 0.93	0.174	0.36
Total	Overall score	7.45	−1.98 to 16.88	0.121	0.41

Values represent the time × group interaction coefficient (β), indicating the differential pre–post change associated with AI-supported training compared with conventional training. Models included random intercepts for participants and fixed effects for time (pre vs post), group (AI-supported vs conventional), and their interaction. Positive β values indicate greater improvement over time in the AI-supported group. Effect sizes are reported as Hedges' g calculated on pre–post change scores. Analyses included participants with complete pre- and post-test assessments (conventional $n = 25$; AI-supported $n = 26$).

3.3. Domain-Specific Communication Outcomes

Domain-level analyses showed that the direction and magnitude of improvement were not identical across components of diagnostic communication (Table 3; Figure 2). Standardized effect-size estimates were generally numerically larger in the AI-supported group, although the magnitude of these differences varied across Kalamazoo domains. This pattern is consistent with non-uniform improvement across communication skills. The largest effect-size estimates favoring AI-supported training were observed in domains related to structuring, negotiating, and concluding the diagnostic encounter. In particular, Provide Closure showed the largest standardized effect, followed by Reach Agreement and overall diagnostic communication performance. Domains related to information exchange and patient understanding, including Share Information, Understand the Patient's Perspective, and Gather Information, also showed small-to-moderate effect sizes favoring the AI-supported arm. By contrast, foundational interpersonal behaviors such as Build a Relationship showed comparable improvement across both training modalities, with negligible between-group differences. Diabetes-specific communication also showed a moderate effect-size estimate favoring AI-supported training, although this pattern should be interpreted cautiously given the exploratory nature of the domain-level analyses.

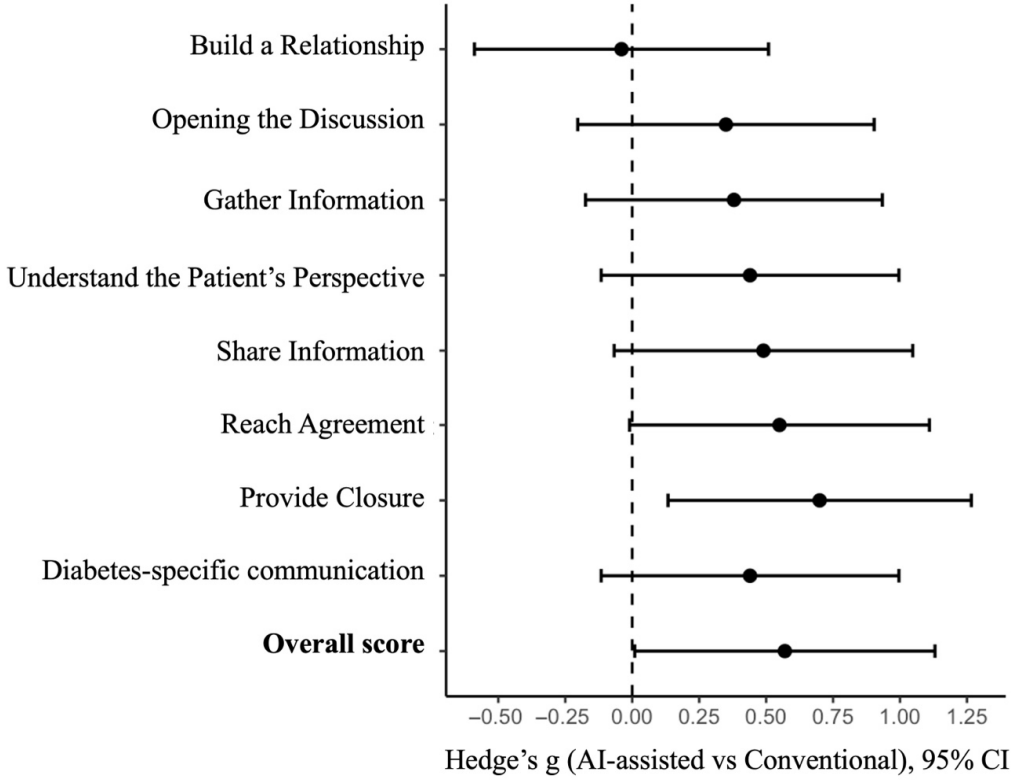


Figure 2. Domain-specific effects of AI-supported simulation on diagnostic communication. Forest plot displaying standardized mean differences (Hedges' g) and 95% confidence intervals for pre–post change scores comparing AI-supported training with conventional training. Positive values indicate numerically greater improvement in the AI-supported group. Effect-size estimates are descriptive and should be interpreted alongside baseline-adjusted analyses.

Baseline-adjusted longitudinal models (Table 4) showed a generally positive direction of time × group interaction coefficients across most domains; however, confidence intervals overlapped the null and none of the domain-specific interactions reached conventional thresholds for statistical significance. Accordingly, these domain-level findings should be interpreted as descriptive patterns rather than confirmatory evidence of differential

effectiveness. Item-level analyses were broadly consistent with these domain-level patterns and are reported in Table A2.

3.4. Individual Learning Trajectories and Heterogeneity of Response

To explore variability in learning outcomes beyond mean-based estimates, participant-level patterns of change were examined using descriptive visualizations and complementary longitudinal models. These analyses were exploratory in nature and were intended to characterize variability of response rather than support confirmatory inference. Figure 3 illustrates individual pre–post changes in overall diagnostic communication performance within each training modality, with participants ordered by magnitude of improvement. Considerable variability was observed in both groups. Although participants in both conditions showed a range of responses, the AI-supported group displayed a broader spread of improvements and a greater concentration of larger gains in overall communication performance. Item- and domain-level heatmaps (Figure A1) similarly suggested that gains were not uniformly distributed across communication domains within individuals. Participants in the AI-supported group more frequently showed concurrent improvements across multiple domains, whereas gains in the conventional group appeared more variable and often more localized across domains. These exploratory patterns should not be interpreted as evidence of consistent superiority of AI-supported training across all learners. Rather, they suggest that responses to the intervention may have been heterogeneous. This observation may be educationally relevant, but it remains descriptive and hypothesis-generating.

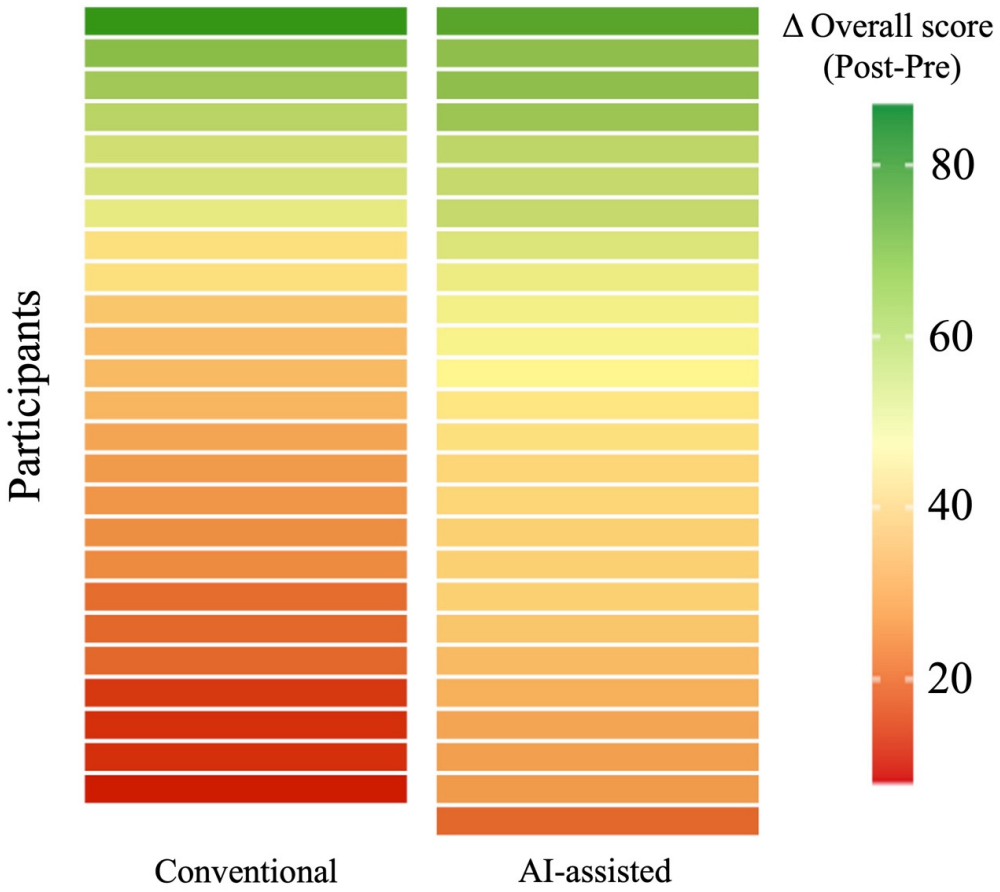


Figure 3. Individual patterns of change in diagnostic communication performance. Heatmap depicting individual pre–post changes in overall diagnostic communication scores within each training modality. Each row represents a single participant, ordered by magnitude of improvement

within group. Color intensity reflects the absolute change in total score (post-test minus pre-test), with warmer colors indicating greater improvement. The visualization illustrates the distribution of individual patterns of change across participants, with broader variability and a higher concentration of larger gains in the AI-supported training group.

3.5. Secondary Outcomes

As an exploratory secondary analysis, we examined the proportion of students achieving a high level of diagnostic communication performance, defined a priori as a total post-test score $\geq$96, corresponding to a mean item score $\geq$4 across rubric items. A numerically greater proportion of students in the AI-supported training group met this threshold than in the conventional training group (46% vs. 24%). This difference did not reach statistical significance (Fisher's exact test, $p = 0.14$) and is therefore reported descriptively. No adverse events or unintended consequences related to either educational intervention were observed.

4. Discussion

In this randomized controlled trial involving undergraduate medical students, both conventional asynchronous teaching and conventional teaching supplemented with generative AI-supported virtual patient simulation were associated with substantial improvements in performance-based diagnostic disclosure skills for T2DM. Although the AI-supported arm showed numerically favorable patterns across several communication domains, particularly those related to structuring and concluding the encounter, uncertainty remained in baseline-adjusted inference, and none of the domain-level between-group differences reached conventional thresholds for statistical significance. Accordingly, these findings should be interpreted as exploratory rather than confirmatory. Rather than demonstrating clear superiority of AI-supported simulation, the study suggests that this approach is feasible and may merit further evaluation under conditions better suited to isolate its specific educational contribution [5,7,13,21].

4.1. Positioning Relative to the Literature

Most early applications of generative artificial intelligence in medical education have emphasized feasibility, acceptability, learner satisfaction, or perceived usefulness rather than controlled evaluations with objective performance outcomes [23,24]. In contrast, the present study contributes a parallel-group randomized controlled design with blinded performance assessment using standardized patients and an established communication framework [5,12,13]. In this sense, it responds to repeated calls for more rigorous evaluation of AI-enabled and technology-enhanced simulation capable of distinguishing true educational impact from novelty effects, enthusiasm bias, or self-reported benefit alone [11,13,26]. The present findings add to the emerging literature by extending evaluation beyond feasibility and learner perceptions toward controlled assessment of performance outcomes. At the same time, the results do not establish statistically significant superiority of the AI-supported approach over conventional training alone. This distinction is important. The question is no longer only whether learners find AI-based simulation engaging, but whether it produces measurable performance gains under controlled conditions. In the present trial, the answer remains provisional: although several estimates favored the AI-supported arm numerically, the study was underpowered to confirm superiority, and the findings are better viewed as exploratory signals that require replication in larger and more definitive trials.

4.2. Why AI-Supported Simulation May Work: Deliberate Practice with Feedback at Scale

A plausible explanation for the observed pattern is that the intervention aligned with principles of deliberate practice, which emphasize repeated rehearsal, timely feedback, and opportunities for iterative refinement [21]. In large undergraduate cohorts, conventional communication teaching often provides limited cycles of individualized practice because faculty time, standardized patient availability, and logistical infrastructure are constrained [12,14,15]. In contrast, the AI-supported program in this study provided repeated, level-based encounters with immediate formative feedback, potentially increasing the amount of structured rehearsal available to each learner.

This interpretation is particularly relevant for communication tasks that require organization, sequencing, and adaptive response to patient concerns. The AI-supported format may have preferentially supported these dimensions by allowing learners to rehearse them multiple times without the scheduling and resource limitations of in-person simulation. More broadly, this interpretation is consistent with educational literature identifying repetition and feedback as central drivers of improvement in complex cognitive and interpersonal skills [22,29]. At the same time, the present study cannot disentangle whether the observed pattern reflects the specific contribution of generative AI, the greater volume of structured practice and feedback received by the intervention group, or a combination of both. This design feature limits causal attribution and should temper any interpretation that the observed numerical advantages are attributable to the AI component itself.

4.3. Domain Specificity: Which Skills Improved More, and Why That Matters

The findings also reinforce the idea that diagnostic communication is not a unitary construct. Improvements were not uniform across Kalamazoo domains, and the directionally larger effects associated with the AI-supported arm were more apparent in domains related to structuring, negotiating, and concluding the encounter than in foundational rapport-building behaviors [5]. This pattern is potentially educationally meaningful because it raises the possibility that different instructional modalities may support different components of communication competence.

AI-supported simulation may be particularly promising as a tool for expanding opportunities to rehearse the procedural "architecture" of a diagnostic conversation: how to organize disclosure, explain key information, negotiate understanding, and close the encounter clearly. By contrast, faculty-guided teaching and in-person standardized-patient work may remain especially important for subtler relational and context-sensitive aspects of communication, including emotional attunement, reflective listening, and nuanced adaptation to patient cues [2,7,31]. This distinction argues against treating AI as a universal substitute for existing educational approaches and instead supports a more targeted integration aligned with specific learning objectives.

4.4. Learner-Level Variability and Patterns of Response

An additional observation from this study was the variability in learner-level response to training. Although both instructional approaches were associated with substantial average improvements, exploratory participant-level analyses suggested broader dispersion of gains in the AI-supported group, including a greater concentration of larger improvements in overall diagnostic communication performance. These patterns should not be interpreted as evidence of consistent superiority across all learners; rather, they suggest that responses to the intervention may have been heterogeneous.

This possibility is educationally relevant because the impact of instructional innovations may not be captured fully by mean differences alone. Some learners may benefit more than others from repeated asynchronous rehearsal and automated feedback, whereas

others may derive similar benefit from conventional instruction alone. However, these observations remain descriptive and hypothesis-generating. Future studies should examine this question prospectively using designs explicitly powered to assess heterogeneity of response and to identify learner characteristics associated with differential benefit.

4.5. Context and Equity: Evidence from a Mexican Public University

Empirical evidence on AI-supported communication training remains disproportionately derived from high-income settings, despite the fact that scalability constraints may be most consequential in low- and middle-income educational contexts [16]. Conducting a performance-based randomized trial within a Mexican public university therefore contributes contextually relevant data to a literature that remains concentrated in high-income and predominantly English-language settings [17,24]. Communication practices are shaped by language, culture, institutional expectations, and local educational resources, all of which may influence both the acceptability and educational impact of AI-based simulation.

The feasibility of implementing a standardized-patient evaluation framework with blinded physician raters in this setting also supports the possibility of conducting more rigorous AI-in-education research outside English-speaking and high-resource environments [6,23]. At the same time, these findings should not be generalized uncritically across institutions or regions. Context-sensitive evaluation remains essential, particularly when educational technologies may interact differently with local curricula, learner expectations, and resource constraints.

4.6. Limitations and Risks Specific to AI-Enabled Training

Several limitations warrant careful consideration. First, attrition was substantial, and complete-case analyses may have introduced bias if follow-up completion was associated with motivation, engagement, or other learner characteristics. Although baseline comparisons between completers and non-completers did not identify statistically significant differences in the measured variables, the analyzed sample may still over-represent more engaged students and may therefore not fully reflect the original randomized cohort.

Second, the study was conducted at a single institution and focused on a single diagnostic disclosure context, namely T2DM. Generalizability to other clinical topics, learner populations, and authentic practice environments is therefore limited [11,12]. Third, multiple domain-level and exploratory analyses increase the probability of chance findings; accordingly, the domain-specific patterns observed here should be interpreted as descriptive and exploratory rather than confirmatory.

Additional limitations are specific to AI-enabled education. The quality of generated feedback may vary, large language models may oversimplify or hallucinate guidance, and learners may develop inappropriate reliance on automated systems if these tools are not embedded within clear educational guardrails [23]. These considerations support the need for standardized prompts, faculty oversight, and routine auditing of AI-generated feedback in educational implementation.

A further methodological issue concerns intervention dose. Participants in the AI-supported group were exposed to a greater amount of structured practice and feedback than those in the conventional arm. Consequently, the present study cannot disentangle whether the observed pattern reflects the specific contribution of generative AI, the additional practice exposure, or a combination of both. This limits causal attribution and represents a central challenge in the evaluation of AI-enhanced educational interventions. Future studies should address this issue through designs that better match time-on-task and feedback exposure across groups.

Finally, contamination between groups cannot be excluded. Because the study was conducted in a cohort-based educational environment, participants may have discussed training content or shared experiences informally. Although such contamination would likely bias estimates toward the null, it remains an important limitation and suggests that future trials may benefit from cluster randomization, temporal separation of interventions, or both.

4.7. Implications for Implementation and Future Research

From an implementation perspective, AI-supported simulation is best understood as a scalable rehearsal layer that complements rather than replaces faculty-led instruction and standardized-patient encounters. Its most plausible near-term role may be to expand opportunities for repeated practice with immediate formative feedback, thereby helping learners prepare more efficiently for resource-intensive in-person training.

Future research should prioritize multicenter trials to improve generalizability, reduce contamination risk, and permit more explicit control of practice exposure across study arms. Larger studies will also be necessary to estimate the magnitude of any added benefit with greater precision and to determine whether the numerically favorable patterns observed here are reproducible across different curricula, learner populations, and institutional contexts. Beyond questions of efficacy, future work should address sustained engagement, minimum effective exposure, durability of gains over time, and transfer of performance to authentic clinical settings [13,14]. Comparative studies examining cost-effectiveness relative to expanded standardized-patient programs would also be valuable for curricular decision-making [12,13].

An additional priority will be to examine whether AI-supported simulation performs similarly in more clinically and emotionally complex diagnostic contexts. Scenarios such as HIV diagnosis, communication of suspected malignancy, and conversations involving obesity and weight-related stigma introduce greater uncertainty, emotional burden, and ethical sensitivity. Evaluating AI-enabled rehearsal across such contexts will be essential to determine whether the exploratory signals observed here extend beyond relatively structured diagnostic disclosure toward more nuanced and emotionally demanding clinical communication tasks.

5. Conclusions

In undergraduate medical students learning to disclose a new diagnosis of T2DM, both conventional asynchronous teaching and an AI-supported virtual patient program were associated with substantial improvements in performance on a blinded standardized-patient assessment. Although several estimates numerically favored the AI-supported arm, the added benefit of AI-supported simulation did not consistently reach conventional thresholds for statistical significance in baseline-adjusted models. Accordingly, these findings should be interpreted as exploratory rather than confirmatory. AI-supported simulation appears feasible as an adjunct to communication training, but larger and better-controlled studies are needed to clarify its specific contribution, reproducibility, and generalizability across learners, institutions, and diagnostic contexts.

Author Contributions: Conceptualization, H.I.S.-C. and B.O.J.-J.; methodology, H.I.S.-C. and B.O.J.-J.; validation, H.I.S.-C. and B.O.J.-J.; formal analysis, H.I.S.-C. and B.O.J.-J.; investigation, B.O.J.-J., D.A.M.-I., A.T.M.-A., F.A.-V., D.M.S.-G., A.A.L.-G., B.M.G.-G., E.M.-P., E.Q.-L., I.M.-B. and A.R.M.-C.; resources, H.I.S.-C. and B.O.J.-J.; data curation, B.O.J.-J. and H.I.S.-C.; writing—original draft preparation, H.I.S.-C.; writing—review and editing, B.O.J.-J., E.Q.-L., I.M.-B., A.R.M.-C. and H.I.S.-C.; supervision, H.I.S.-C. and B.O.J.-J.; project administration, H.I.S.-C. and B.O.J.-J.; funding acquisition, H.I.S.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Ethics Committee of the Facultad de Estudios Superiores Iztacala (FES Iztacala), Universidad Nacional Autónoma de México (UNAM) (protocol code CE/FESI/052025/1922; approved in May 2025). Because participants were undergraduate medical students, specific safeguards were implemented to minimize any risk of perceived coercion. The study was conducted during the intersemester period in December 2025 as an independent research activity and did not form part of any mandatory academic course, graded exercise, or formal evaluation. Students were explicitly informed that participation, non-participation, or withdrawal would have no effect on their grades, academic standing, or relationship with faculty. Patients and members of the public were not involved in the design, conduct, reporting, or dissemination of this educational trial.

Informed Consent Statement: Written informed consent was obtained from all subjects involved in the study prior to enrollment. Participation was entirely voluntary, and students were informed that refusal to participate or withdrawal from the study at any time would not result in any penalty or academic consequence.

Data Availability Statement: The datasets generated and analyzed during the current 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/2MVW9 (accessed on 9 March 2026). Audio recordings and voice files are not publicly available due to privacy and ethical restrictions involving human participants but may be made available for review upon reasonable request, subject to approval by the institutional ethics committee.

Acknowledgments: The authors wish to thank the physicians of the Centro Internacional de Simulación y Entrenamiento en Soporte Vital Iztacala (CISESVI), Faculty of Higher Studies Iztacala, National Autonomous University of Mexico (UNAM), for their invaluable contribution as standardized patients in this study. Their professionalism, consistency, and commitment were essential to the development, implementation, and rigorous assessment of the diagnostic communication scenarios. The authors also acknowledge the administrative and technical support provided by the academic staff of FES Iztacala that facilitated the execution of this educational trial. During the preparation of this manuscript, the authors used ChatGPT 5.2 (OpenAI) for language editing and clarity improvement. The authors reviewed and edited the generated content and take full responsibility for the integrity, accuracy, and originality of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
ANCOVA	Analysis of covariance
LLM	Large language model
T2DM	Type 2 diabetes mellitus
SPIKES	Setting, Perception, Invitation, Knowledge, Emotions, Strategy, and Summary

Appendix A

Table A1 reports baseline (pre-test) diagnostic communication performance for each individual item of the Kalamazoo Essential Elements Communication Checklist among all randomized participants (n = 80). Scores are presented descriptively to characterize initial performance and variability across communication behaviors prior to the educational intervention. No between-group statistical comparisons were performed at the item level,

as the table is intended to document baseline balance and heterogeneity. Accordingly, these data should not be interpreted as evidence of baseline differences between groups.

Table A1. Baseline (pre-test) diagnostic communication performance by individual Kalamazoo items.

Item	Domain (Kalamazoo Framework)	Conventional (n = 40) (Mean ± SD)	AI-Supported (n = 40) (Mean ± SD)
1	Build a Relationship	2.19 ± 0.78	1.90 ± 0.55
1.1	Eye contact and open posture	2.30 ± 0.91	2.15 ± 0.74
1.2	Friendly body language	2.40 ± 0.96	2.02 ± 0.66
1.3	Self-introduction and role	1.88 ± 1.09	1.52 ± 0.82
2	Opening the Discussion	1.73 ± 0.52	1.58 ± 0.46
2.1	Greeting and identity verification	1.85 ± 0.62	1.85 ± 0.53
2.2	Explains purpose of the visit	1.82 ± 0.78	1.55 ± 0.68
2.3	Asks patient expectations	1.53 ± 0.60	1.35 ± 0.48
3	Gather Information	1.94 ± 0.62	1.95 ± 0.66
3.1	Uses open-ended questions	1.98 ± 0.72	1.98 ± 0.70
3.2	Listens without interrupting	2.38 ± 1.08	2.70 ± 1.20
3.3	Summarizes to confirm	1.48 ± 0.72	1.18 ± 0.45
4	Understand the Patient's Perspective	1.63 ± 0.53	1.57 ± 0.52
4.1	Explores beliefs about illness	1.50 ± 0.68	1.50 ± 0.51
4.2	Inquires about concerns/fears	1.75 ± 0.74	1.65 ± 0.62
4.3	Acknowledges impact on daily life	1.63 ± 0.70	1.57 ± 0.68
5	Share Information	1.47 ± 0.48	1.77 ± 0.52
5.1	Explains diagnosis clearly	1.55 ± 0.60	1.90 ± 0.63
5.2	Uses supports/examples	1.35 ± 0.48	1.60 ± 0.63
5.3	Checks understanding (teach-back)	1.50 ± 0.72	1.55 ± 0.68
6	Reach Agreement	1.67 ± 0.55	1.70 ± 0.61
6.1	Discusses therapeutic options	1.85 ± 0.62	1.62 ± 0.71
6.2	Involves the patient	1.50 ± 0.60	1.58 ± 0.68
6.3	Negotiates realistic goals	1.65 ± 0.66	1.65 ± 0.74
7	Provide Closure	1.76 ± 0.64	1.81 ± 0.55
7.1	Provides final summary	1.65 ± 0.70	1.40 ± 0.50
7.2	Checks final questions	1.80 ± 0.72	1.82 ± 0.59
7.3	Closes with empathy and follow-up	1.82 ± 0.68	2.20 ± 0.88
8	Diabetes-specific communication	1.72 ± 0.58	1.69 ± 0.49
8.1	Communicates results clearly (glucose/HbA1c)	1.72 ± 0.68	1.70 ± 0.56
8.2	Initial T2DM plan + anxiety reduction	1.70 ± 0.72	1.65 ± 0.57
8.3	Complications and follow-up	1.75 ± 0.74	1.72 ± 0.56
Total	Total score (24–120)	42.50 ± 10.55	42.14 ± 9.67

Table A2. Domain and Item-level pre–post changes in diagnostic communication performance by Kalamazoo items.

Item	Domain (Kalamazoo Framework)	Conventional Δ Mean ± SD	AI-Supported Δ Mean ± SD	Hedges' g
1	Build a Relationship	5.88 ± 2.42	5.77 ± 2.80	−0.04
1.1	Eye contact and open posture	1.85 ± 0.97	1.84 ± 1.18	−0.01
1.2	Friendly body language	1.88 ± 0.99	1.52 ± 1.23	−0.32
1.3	Self-introduction and role	2.12 ± 1.03	2.24 ± 1.23	0.11
2	Opening the Discussion	5.16 ± 2.88	6.08 ± 2.23	0.35
2.1	Greeting and identity verification	1.27 ± 1.37	1.80 ± 1.15	0.41
2.2	Explains purpose of the visit	1.85 ± 1.16	2.00 ± 1.08	0.14
2.3	Asks patient expectations	2.23 ± 1.03	2.04 ± 0.84	−0.20
3	Gather Information	5.40 ± 2.61	6.31 ± 2.05	0.38

Table A2. *Cont.*

Item	Domain (Kalamazoo Framework)	Conventional Δ Mean ± SD	AI-Supported Δ Mean ± SD	Hedges' g
3.1	Uses open-ended questions	1.77 ± 0.91	2.00 ± 1.00	0.24
3.2	Listens without interrupting	1.54 ± 1.27	1.76 ± 1.16	0.18
3.3	Summarizes to confirm	2.12 ± 1.21	2.32 ± 1.11	0.17
4	Understand the Patient's Perspective	5.96 ± 2.91	7.19 ± 2.56	0.44
4.1	Explores beliefs about illness	2.31 ± 1.16	2.40 ± 1.19	0.08
4.2	Inquires about concerns/fears	1.92 ± 1.16	2.24 ± 1.16	0.27
4.3	Acknowledges impact on daily life	1.88 ± 1.31	2.16 ± 0.90	0.24
5	Share Information	6.24 ± 2.98	7.54 ± 2.18	0.49
5.1	Explains diagnosis clearly	2.12 ± 1.24	2.16 ± 1.03	0.04
5.2	Uses supports/examples	2.08 ± 1.35	2.32 ± 0.99	0.2
5.3	Checks understanding (teach-back)	2.23 ± 1.14	2.60 ± 1.26	0.3
6	Reach Agreement	4.88 ± 2.85	6.46 ± 2.85	0.55
6.1	Discusses therapeutic options	1.46 ± 1.27	1.60 ± 1.22	0.11
6.2	Involves the patient	1.88 ± 1.03	2.08 ± 1.32	0.16
6.3	Negotiates realistic goals	1.81 ± 1.30	1.92 ± 1.29	0.09
7	Provide Closure	5.00 ± 3.54	7.23 ± 2.70	0.7
7.1	Provides final summary	1.92 ± 1.52	2.04 ± 1.21	0.08
7.2	Checks final questions	1.50 ± 1.24	2.36 ± 1.32	0.66
7.3	Closes with empathy and follow-up	1.77 ± 1.39	2.24 ± 1.01	0.38
8	Diabetes-specific communication	5.04 ± 3.43	6.38 ± 2.61	0.44
8.1	Communicates results clearly (glucose/HbA1c)	2.00 ± 1.10	1.96 ± 0.84	−0.04
8.2	Initial T2DM plan + anxiety reduction	1.50 ± 1.24	2.12 ± 1.20	0.5
8.3	Complications and follow-up	1.69 ± 1.59	1.84 ± 1.34	0.1
Total	Overall score	43.56 ± 19.38	52.96 ± 12.85	0.57

Table A2 presents pre–post change scores (post-test minus pre-test) for each individual Kalamazoo communication item among participants with complete assessments (n = 51). Values are reported separately by training modality and are intended to illustrate item-level patterns of improvement across diagnostic communication behaviors. Effect-size estimates at the item level are descriptive and exploratory and should not be interpreted as confirmatory evidence of differential effectiveness.

Table A3. Domain and Item-level longitudinal mixed-effects model results for diagnostic communication performance.

Item	Domain (Kalamazoo Framework)	Time × Group β	95% CI	*p*-Value	Hedges' g
1	Build a Relationship	−0.09	−0.56 to 0.38	0.707	−0.09
1.1	Eye contact and open posture	−0.02	−0.29 to 0.25	0.876	−0.03
1.2	Friendly body language	−0.18	−0.52 to 0.15	0.286	−0.21
1.3	Self-introduction and role	0.13	−0.22 to 0.48	0.466	0.15
2	Opening the Discussion	0.41	−0.06 to 0.88	0.086	0.44
2.1	Greeting and identity verification	0.41	−0.05 to 0.87	0.081	0.48
2.2	Explains purpose of the visit	0.22	−0.26 to 0.69	0.369	0.25
2.3	Asks patient expectations	−0.09	−0.52 to 0.34	0.676	−0.10
3	Gather Information	0.45	−0.03 to 0.93	0.066	0.46
3.1	Uses open-ended questions	0.29	−0.18 to 0.76	0.227	0.33
3.2	Listens without interrupting	0.18	−0.33 to 0.69	0.488	0.2
3.3	Summarizes to confirm	0.21	−0.29 to 0.71	0.405	0.24
4	Understand the Patient's Perspective	0.34	−0.15 to 0.83	0.173	0.35
4.1	Explores beliefs about illness	0.12	−0.36 to 0.60	0.624	0.14
4.2	Inquires about concerns/fears	0.34	−0.12 to 0.80	0.145	0.38
4.3	Acknowledges impact on daily life	0.28	−0.21 to 0.77	0.257	0.31

Table A3. *Cont.*

Item	Domain (Kalamazoo Framework)	Time × Group β	95% CI	*p*-Value	Hedges' g
5	Share Information	0.31	−0.12 to 0.74	0.157	0.36
5.1	Explains diagnosis clearly	0.05	−0.40 to 0.50	0.822	0.06
5.2	Uses supports/examples	0.26	−0.20 to 0.72	0.265	0.3
5.3	Checks understanding (teach-back)	0.39	−0.09 to 0.87	0.111	0.45
6	Reach Agreement	0.24	−0.31 to 0.79	0.39	0.24
6.1	Discusses therapeutic options	0.14	−0.37 to 0.65	0.586	0.16
6.2	Involves the patient	0.19	−0.36 to 0.74	0.495	0.22
6.3	Negotiates realistic goals	0.11	−0.43 to 0.65	0.685	0.12
7	Provide Closure	0.49	−0.16 to 1.14	0.138	0.39
7.1	Provides final summary	0.1	−0.45 to 0.65	0.722	0.11
7.2	Checks final questions	0.61	0.04 to 1.18	0.036	0.66
7.3	Closes with empathy and follow-up	0.38	−0.19 to 0.95	0.191	0.4
8	Diabetes-specific communication	0.38	−0.17 to 0.93	0.174	0.36
8.1	Communicates results clearly (glucose/HbA1c)	−0.04	−0.52 to 0.44	0.866	−0.04
8.2	Initial T2DM plan + anxiety reduction	0.49	−0.02 to 1.00	0.061	0.55
8.3	Complications and follow-up	0.15	−0.40 to 0.70	0.594	0.17
Total	Overall score	7.45	−1.98 to 16.88	0.121	0.41

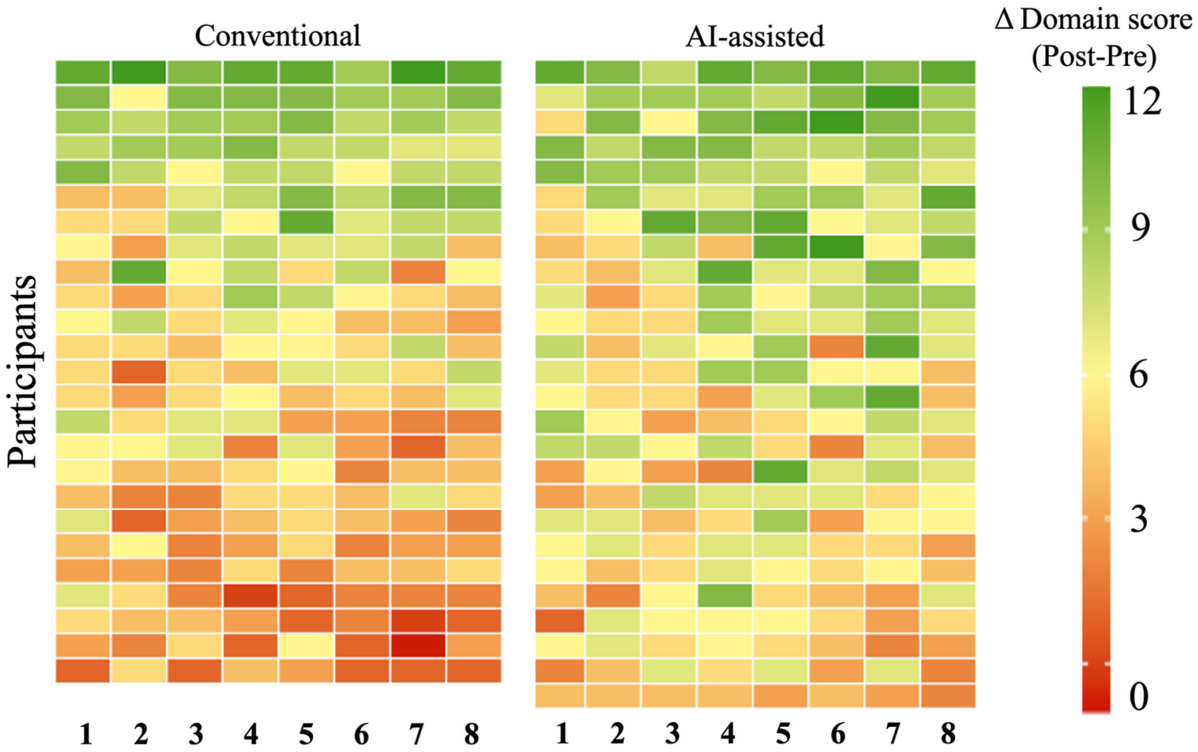


Figure A1. Item-level heatmap of individual pre–post changes in diagnostic communication performance. Figure A1 displays item-level heatmaps of pre–post change scores (post-test minus pre-test) for individual diagnostic communication behaviors based on the Kalamazoo Essential Elements Communication Checklist. Each row represents a single participant, and each column corresponds to an individual rubric item. Color intensity reflects the magnitude and direction of change, with warmer colors indicating greater improvement and cooler colors indicating smaller gains or declines. Heatmaps are presented separately for the conventional training group and the AI-supported training group among participants with complete assessments (n = 51). This visualization is intended to illustrate heterogeneity of patterns of change across communication behaviors and is presented for exploratory and descriptive purposes only, without inferential statistical testing.

Table A3 shows item-level longitudinal effects derived from linear mixed-effects models evaluating changes in diagnostic communication performance over time. Models included fixed effects for time (pre-test vs post-test), training group (AI-supported vs conventional), and their interaction, with random intercepts for participants. The Time × Group interaction coefficient reflects the differential change associated with AI-supported training relative to conventional instruction. Results are presented for exploratory purposes to characterize patterns of learning across individual communication behaviors. Given the number of item-level models estimated, no correction for multiple comparisons was applied; therefore, individual p-values should be interpreted cautiously and in the context of the overall pattern rather than as isolated statistically significant findings.

References

1. Epstein, R.M.; Street, R.L., Jr. The values and value of patient-centered care. *Ann. Fam. Med.* **2011**, *9*, 100–103. [CrossRef]
2. Holmes-Rovner, M. Skills for Communicating with Patients (2e). *Health Expect.* **2005**, *8*, 277–278. [CrossRef]
3. Al Shahrani, A.; Baraja, M. Patient Satisfaction and it's Relation to Diabetic Control in a Primary Care Setting. *J. Fam. Med. Prim. Care* **2014**, *3*, 5–11. [CrossRef]
4. Levinson, W.; Roter, D.L.; Mullooly, J.P.; Dull, V.T.; Frankel, R.M. Physician-patient communication. The relationship with malpractice claims among primary care physicians and surgeons. *JAMA* **1997**, *277*, 553–559. [CrossRef]
5. Makoul, G. Essential elements of communication in medical encounters: The Kalamazoo consensus statement. *Acad. Med.* **2001**, *76*, 390–393. [CrossRef]
6. Street, R.L., Jr.; Makoul, G.; Arora, N.K.; Epstein, R.M. How does communication heal? Pathways linking clinician-patient communication to health outcomes. *Patient Educ. Couns.* **2009**, *74*, 295–301. [CrossRef] [PubMed]
7. von Fragstein, M.; Silverman, J.; Cushing, A.; Quilligan, S.; Salisbury, H.; Wiskin, C. UK consensus statement on the content of communication curricula in undergraduate medical education. *Med. Educ.* **2008**, *42*, 1100–1107. [CrossRef] [PubMed]
8. Kee, J.W.; Khoo, H.S.; Lim, I.; Koh, M.Y. Communication Skills in Patient-Doctor Interactions: Learning from Patient Complaints. *Health Prof. Educ.* **2018**, *4*, 97–106. [CrossRef]
9. Yedidia, M.J.; Gillespie, C.C.; Kachur, E.; Schwartz, M.D.; Ockene, J.; Chepaitis, A.E.; Snyder, C.W.; Lazare, A.; Lipkin, M., Jr. Effect of communications training on medical student performance. *JAMA* **2003**, *290*, 1157–1165. [CrossRef]
10. Back, A.L.; Arnold, R.M.; Tulsky, J.A.; Baile, W.F.; Fryer-Edwards, K.A. Teaching communication skills to medical oncology fellows. *J. Clin. Oncol.* **2003**, *21*, 2433–2436. [CrossRef]
11. Hodges, B.D.; Albert, M.; Arweiler, D.; Akseer, S.; Bandiera, G.; Byrne, N.; Charlin, B.; Karazivan, P.; Kuper, A.; Maniate, J.; et al. The future of medical education: A Canadian environmental scan. *Med. Educ.* **2011**, *45*, 95–106. [CrossRef]
12. Barrows, H.S. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *AAMC. Acad. Med.* **1993**, *68*, 443–451; discussion 451–453. [CrossRef]
13. Cook, D.A.; Hamstra, S.J.; Brydges, R.; Zendejas, B.; Szostek, J.H.; Wang, A.T.; Erwin, P.J.; Hatala, R. Comparative effectiveness of instructional design features in simulation-based education: Systematic review and meta-analysis. *Med. Teach.* **2013**, *35*, e867–e898. [CrossRef]
14. Cleland, J. Simulation-based education: What's it all about? *Perspect. Med. Educ.* **2018**, *7*, 30–33. [CrossRef]
15. Nestel, D.; Groom, J.; Eikeland-Husebø, S.; O'Donnell, J.M. Simulation for learning and teaching procedural skills: The state of the science. *Simul. Healthc.* **2011**, *6*, S10–S13. [CrossRef]
16. Frenk, J.; Chen, L.; Bhutta, Z.A.; Cohen, J.; Crisp, N.; Evans, T.; Fineberg, H.; Garcia, P.; Ke, Y.; Kelley, P.; et al. Health professionals for a new century: Transforming education to strengthen health systems in an interdependent world. *Lancet* **2010**, *376*, 1923–1958. [CrossRef] [PubMed]
17. Ross, S.; Poth, C.N.; Donoff, M.; Humphries, P.; Steiner, I.; Schipper, S.; Janke, F.; Nichols, D. Competency-based achievement system: Using formative feedback to teach and assess family medicine residents' skills. *Can. Fam. Physician* **2011**, *57*, e323–e330. [PubMed]
18. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med. Educ.* **2023**, *9*, e45312. [CrossRef]
19. 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. [CrossRef] [PubMed]

20. Moldt, J.A.; Festl-Wietek, T.; Fuhl, W.; Zabel, S.; Claassen, M.; Wagner, S.; Nieselt, K.; Herrmann-Werner, A. Assessing AI Awareness and Identifying Essential Competencies: Insights From Key Stakeholders in Integrating AI Into Medical Education. *JMIR Med. Educ.* **2024**, *10*, e58355. [CrossRef]
21. Ericsson, K.A.; Krampe, R.T.; Tesch-Römer, C. The role of deliberate practice in the acquisition of expert performance. *Psychol. Rev.* **1993**, *100*, 363–406. [CrossRef]
22. Hattie, J.; Timperley, H. The Power of Feedback. *Rev. Educ. Res.* **2007**, *77*, 81–112. [CrossRef]
23. Lee, P.; Bubeck, S.; Petro, J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N. Engl. J. Med.* **2023**, *388*, 1233–1239. [CrossRef] [PubMed]
24. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. [CrossRef]
25. Weisman, A.; Yona, T.; Masharawi, Y. Education about pain and experience with cognitive-based interventions do not reduce healthcare professionals' chronic pain. *PeerJ* **2025**, *13*, e19448. [CrossRef]
26. Plackett, R.; Kassianos, A.P.; 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 Med. Educ.* **2022**, *22*, 365. [CrossRef]
27. Quail, N.P.A.; Boyle, J.G. Virtual Patients in Health Professions Education. In *Biomedical Visualisation: Volume 4*; Rea, P.M., Ed.; Springer International Publishing: Cham, Switzerland, 2019; pp. 25–35.
28. Suárez-García, R.X.; Chavez-Castañeda, Q.; Orrico-Pérez, R.; Valencia-Marin, S.; Castañeda-Ramírez, A.E.; Quiñones-Lara, E.; Ramos-Cortés, C.A.; Gaytán-Gómez, A.M.; Cortés-Rodríguez, J.; Jarquín-Ramírez, J.; et al. DIALOGUE: A Generative AI-Based Pre–Post Simulation Study to Enhance Diagnostic Communication in Medical Students Through Virtual Type 2 Diabetes Scenarios. *Eur. J. Investig. Health Psychol. Educ.* **2025**, *15*, 152. [CrossRef]
29. McGaghie, W.C.; Issenberg, S.B.; Cohen, E.R.; Barsuk, J.H.; Wayne, D.B. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Acad. Med.* **2011**, *86*, 706–711. [CrossRef] [PubMed]
30. Baile, W.F.; Buckman, R.; Lenzi, R.; Glober, G.; Beale, E.A.; Kudelka, A.P. SPIKES-A six-step protocol for delivering bad news: Application to the patient with cancer. *Oncologist* **2000**, *5*, 302–311. [CrossRef]
31. Childers, J.W.; Bulls, H.; Arnold, R. Beyond the NURSE Acronym: The Functions of Empathy in Serious Illness Conversations. *J. Pain Symptom Manag.* **2023**, *65*, e375–e379. [CrossRef]

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.