Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Assessment of AETCOM-Aligned Communication Competency among Undergraduate Medical Students Using ConversationAlly-Based Simulated Consultations: A Pilot Study

Monali Kar
Siksha O Anusandhan University
Amit Pritam Swain
Siksha O Anusandhan University
Jyotiranjan Sahoo
Siksha O Anusandhan University
Venkatarao Epari
e.venkata.rao@gmail.com

Siksha O Anusandhan University
Sanjay Dash
Wellness Center
Shweta Sharma
SigmaMozak Solutions Pvt Ltd
Rajshri Mallabadi
SigmaMozak Solutions Pvt Ltd

Research Article

Keywords: AETCOM, clinical communication, simulated consultations, ConversationAlly, medical education, AI-assisted feedback, inter-rater reliability, undergraduate medical training, India

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11303061/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: Competing interest reported. MK, AS, JS, VE and SD declare no competing interests. SS and RM are inventors of ConversationAlly, which is a product of SigmaMozak Solutions. The platform was used as data-collection and transcript-review infrastructure for this study. The inventors were not involved in faculty scoring of transcripts. Faculty ratings remained the primary competency outcomes; the AI–faculty concordance analysis was post hoc and exploratory.

Abstract

Background

Effective doctor–patient communication is fundamental to patient-centered, safe and ethical clinical practice. Although embedded in India's undergraduate curriculum through the Attitude, Ethics and Communication (AETCOM) framework, scalable methods for observing, assessing and providing feedback on student communication remain limited. We assessed AETCOM-aligned communication competency among medical undergraduates using AI-enabled simulated consultations, and examined the feasibility of AI-assisted transcript capture and faculty-reviewed communication assessment.

Methods

An IEC-approved cross-sectional pilot study was conducted among Phase III Part I MBBS students at a tertiary care teaching medical college. Students voluntarily completed audio-only simulated consultations using ConversationAIly, an AI-enabled conversation analysis platform. A 20-item pilot-developed pre-participation questionnaire (Supplementary file) assessed five determinant domains. Raw platform exports were deduplicated and quality screened. Stratified transcript samples were independently scored by two faculty raters using a seven-domain rubric; inter-rater reliability was assessed using ICC (2,1). Post hoc AI rescoring of analyzable transcripts used transcript text and the rubric, with faculty ratings merged only after AI scores were frozen. AI–faculty agreement was evaluated using Lin's concordance correlation coefficient (CCC), mean difference, mean absolute error, and Bland–Altman analysis.

Results

Of 123 registered students; 89 students completed at least one consultation. Deduplication yielded 383 unique sessions; 127 transcripts passed quality screening. Faculty inter-rater reliability was excellent (mean ICC = 0.996). Mean Faculty-rated communication score was 22.55/35 (SD 7.35). Avoidance of stigmatizing language scored highest (D6: 4.59/5), while teach-back and safety-netting (D7: 2.14/5) and concern/belief elicitation (D2: 2.60/5) scored lowest. Female students had higher attitude scores than male students (4.36 vs 4.04, p = 0.003), and regular prior clinical communication exposure was associated with higher readiness (p = 0.033). Exploratory AI–faculty agreement for the total score was high (Lin's CCC = 0.960; mean difference − 1.00/35; 95% limits of agreement − 4.43 to + 2.43).

Conclusions

ConversationAIly enabled scalable capture and assessment of simulated undergraduate consultations. Baseline performance was weaker in elicitation, rapport building, and teach-back/safety-netting than in explanation and stigma-free language. The rubric demonstrated high faculty reliability. High exploratory AI–faculty concordance warrants prospective evaluation of automated assessment but should not be interpreted as formal validation of AI scoring.

Background

Communication as a clinical competency: Effective doctor-patient communication is a set of teachable, observable, and assessable behaviors with well-documented effects on clinical outcomes. Patients who feel heard are more likely to disclose symptoms accurately, adhere to treatment plans, and engage with follow-up care.[1–3] Conversely, communication failures contribute to diagnostic error, non-adherence, patient dissatisfaction, and medico-legal risk. The pathways through which communication influences health are direct and indirect: improving patient understanding, enabling shared decision-making, building therapeutic alliance, and supporting self-management.[2]

The behaviors that constitute a competent clinical consultation are well described and teachable. They include opening the encounter respectfully and establishing rapport; eliciting the patient's concern, belief, fear, or expectation before providing advice; responding empathetically to emotional cues; explaining clearly and without jargon; involving the patient in planning next steps; avoiding stigmatizing or blaming language; and confirming understanding through teach-back or safety-netting advice.[4,5] These behaviors are not peripheral. In counselling consultations involving medication adherence, chronic disease management, or culturally sensitive health beliefs, they constitute the clinical encounter.

The Indian undergraduate context and AETCOM: In India, the Attitude, Ethics and Communication (AETCOM) module, introduced by the National Medical Commission (NMC) as part of the competency-based medical education curriculum, formally recognizes communication as a distinct and assessable competency for the Indian Medical Graduate.[6] AETCOM provides a framework for teaching communication, attitude, and professionalism across all three phases of undergraduate training, and represents an important curricular advance in acknowledging these domains as learnable and teachable skills rather than innate attributes.

Despite this curricular recognition, implementation presents persistent challenges. Communication teaching in Indian medical colleges frequently relies on didactic sessions, reflective writing, or role-play exercises that, while educationally valuable, do not directly observe or score a student's actual consultation behaviour.[7,8] Large undergraduate batches, limited faculty time, variable quality of feedback, and the absence of standardized observation tools mean that assessment-driven exposure to communication theory does not reliably translate into the behavioral fluency that patient-centered practice requires.[9,10]

Simulation and AI-assisted feedback: Simulation- and skills-lab-based approaches to communication training are increasingly used in undergraduate medical education.[11,12] Simulated consultations allow students to practice realistic clinical interactions without risk to real patients, receive specific feedback on their observed communication behavior, and repeat encounters in a standardized context. The pedagogical advantage is that simulation makes communication behavior visible and reviewable.

A practical constraint of simulation at scale is the faculty time required for direct observation and structured feedback. AI-assisted communication analysis platforms have emerged as one response to this constraint, enabling audio capture, automated transcription, preliminary marker extraction, and structured feedback generation at a scale that is not feasible through synchronous faculty observation alone. AI-based tools are increasingly being evaluated in health professions education, although evidence for validated automated assessment against human reference ratings remains limited.[13]

The gap this study addresses: There is limited published evidence describing the communication competency of Indian undergraduate medical students in realistic AETCOM-aligned counselling scenarios, assessed through faculty-rated transcript-level analysis of actual student behavior. Most available evidence is either self-reported, based on simulated patient ratings, or derived from structured observation that does not capture the behaviors most

relevant to AETCOM domains, such as concern elicitation, empathetic response, collaborative planning, and teach-back.[8–10] There is correspondingly limited evidence on the feasibility of AI-assisted transcript capture and faculty review as a scalable workflow in Indian undergraduate settings.

Methods

The aim and study design

We assessed AETCOM-aligned communication competency among medical undergraduates using ConversationAIly-based simulated consultations and evaluated the feasibility of AI-assisted transcript capture and faculty-reviewed communication assessment.

This was a cross-sectional analytical pilot study assessing AETCOM-aligned doctor-patient communication using scenario-based simulated consultations through ConversationAIly-assisted capture and feedback.

Setting and participants

The study was conducted in the Department of Community Medicine in a tertiary care medical college in Odisha, India, between April and July 2026. All Phase III Part I MBBS students were invited to participate following a structured orientation briefing. No formal sample-size was calculated as this was a feasibility pilot study.

Orientation, and consent

Students received an academic orientation briefing covering the purpose of the study and its relationship to AETCOM competency development; the simulated nature of consultations and absence of real patient contact; the audio-only recording procedure; confidentiality and anonymization; voluntary participation with no academic consequences; and the right to withdraw without penalty. Ethical approval was granted by the Institutional Ethics Committee prior to data collection. No real patient contact was involved at any stage; all consultations were simulated and audio-only. A secure digital registration link was shared following orientation. Students who chose to participate self-enrolled and proceeded independently with the simulated consultation tasks. Participation was voluntary. Students who did not provide digital informed consent or withdrew were not enrolled. Digital informed consent was obtained before engagement that covered audio-only recording, AI-based transcription and analysis, and use of anonymized data for research and academic improvement purposes.

Recruitment (Fig. 1)

A total of 140 registration entries were received; 17 duplicate registrations were removed, yielding 123 unique registered participants. Eighty-nine named students completed at least one simulated consultation. A group of 12 participants registered using numeric identifiers only and could not be linked at the individual level. For the remaining 22 registered participants, no linked consultation was identified in the study dataset; reasons for non-participation or non-completion were not systematically collected. At the record-screening stage, submissions consisting entirely of technical recordings, administrative discussions, or unintelligible audio were excluded from transcript analysis; these were data-quality exclusions rather than participant-level eligibility exclusions.

Pre-participation determinants questionnaire: A 20-item structured self-administered questionnaire was developed specifically for this pilot study in consultation with three subject experts. It covered five domains: attitude toward patient-centered communication (5 items), communication self-efficacy (5 items), communication anxiety and confidence (4 items), contextual awareness of language and culture (3 items), and readiness for structured and AI-

assisted feedback (3 items). All items were rated on a five-point Likert scale (1 = strongly disagree to 5 = strongly agree). Composite scores were calculated as the mean of constituent items, giving each domain a possible score range of 1–5. Demographic variables included gender, place and medium of schooling, prior formal communication training, and previous clinical communication exposure. Individual-level linkage between named registration records and session data was confirmed for all 89 named students who attempted consultations.

Simulated consultation procedure: Each participant was expected to complete standardized simulated consultation scenarios corresponding to three AETCOM-aligned scenario categories: OPD-type clinical interaction; Family Adoption Programme or community interaction; and hospital-based counselling or emotional conflict situation. Scenarios were aligned with AETCOM competency domains including empathy, patient-centered communication, cultural sensitivity, professionalism, and management of challenging conversations. Each scenario brief included patient background, the presenting concern, and the primary communication task, with embedded prompts to guide interaction while avoiding rote scripting.

AI platform, ConversationAIly

All simulated consultations were conducted through ConversationAIly, an AI-enabled conversation analysis platform developed by SigmaMozak Solutions, used for audio-only recording, automated speech-to-text transcription, NLP-based communication marker extraction, and structured feedback generation. Its role in this pilot was explicitly that of educational infrastructure for capture, transcription, and faculty review facilitation. Pre-existing platform outputs, including automated tone classification, empathy score, red-flag detection, and stigmatizing-language identification, were treated as exploratory metadata rather than primary competency outcomes, as automated outputs from conversational AI platforms require validation against faculty-rated gold standards before serving as standalone competency scores. Faculty-rated rubric scores remained the primary communication measures because automated assessment requires comparison against human reference ratings before interpretive use.[13]

Transcript quality screening

Data audit identified a technical characteristic of the raw platform export in which incremental speech-to-text fragments could be persisted as separate database records rather than only as the final complete transcript, generating multiple records per consultation. Raw platform records were therefore not treated as consultation-level observations. Records were deduplicated by session identifier and screened for repetition artefacts, short/noise transcripts, case-type mismatch, non-consultation content, mixed-language or unintelligible text, and duplicate transcripts. The analytic pipeline proceeded from 1,737 raw platform records to 383 unique sessions after deduplication and then to 127 quality-screened transcripts from participating students. These 127 transcripts constituted the sampling frame for faculty review.

Faculty scoring rubric

A seven-domain communication scoring rubric was developed for faculty rating, drawing on established consultation frameworks.[4,5] Each domain was scored from 1 (absent or very poor) to 5 (excellent). An overall holistic score was assigned on a separate nine-point integer scale (1–9), representing the rater's global impression of consultation quality independent of the domain scores. The holistic score was not a Likert questionnaire item; the broader 1–9 range was used to permit greater discrimination in the overall judgement while retaining the 1–5 scale for specific behavioral domains. Raters were instructed to score only behaviors demonstrably present in the transcript. Table 1 presents the seven domains and operational definitions.

Inter-rater reliability: A pragmatic stratified sample of 30 of the 127 quality-screened transcripts (23.6%) was selected for duplicate faculty rating. Stratification was performed within the 127-transcript sampling frame by case type and data-collection cohort to ensure representation of the principal consultation scenarios; it was not intended to reproduce the distribution of all 383 unique sessions proportionally. The remaining 97 quality-screened transcripts were eligible but were not selected because the inter-rater exercise was deliberately limited to a feasible duplicate-scoring subset. Two independent faculty raters scored the 30 transcripts, blinded to each other's ratings and to pre-existing AI outputs. For each transcript, raters first recorded analyzability (Yes/No/Unsure); domain and holistic scores were included only when both raters considered the transcript analyzable. Inter-rater reliability of the numerical domain ratings was quantified using ICC (2,1), a two-way random-effects, single-measure model for absolute agreement, following Koo and Li.[14]

Exploratory AI–faculty concordance analysis

The same seven-domain rubric was applied post hoc to the 29 mutually analyzable transcripts in a clean text-only AI rescoring pass. AI scores were frozen before faculty ratings were merged for comparison. Agreement between AI-generated scores and the mean of the two faculty ratings was evaluated using Lin's concordance correlation coefficient (CCC), which assesses both precision and accuracy relative to the line of identity. Mean difference (AI minus faculty) and mean absolute error (MAE) were calculated for each domain. For total seven-domain and holistic scores, Bland–Altman analysis was used to estimate mean bias and 95% limits of agreement. This was a secondary exploratory analysis and was not prespecified as a validation study.

Statistical analysis

Continuous and composite questionnaire variables were summarized using mean, standard deviation (SD), median, and interquartile range (IQR); categorical variables were expressed as frequencies and percentages. Because this was a pilot analysis, both mean (SD) and median (IQR) are reported for composite scores to show their distribution. Pre-participation composite scores were compared between gender groups, prior formal communication training groups, and regular versus occasional previous clinical communication exposure using independent-samples t-tests. Scenario-wise comparisons were descriptive. A p-value < 0.05 was used as the significance threshold. Analysis was performed using Python 3.12 with SciPy and pandas' libraries.

Table 1
Seven-domain faculty communication scoring rubric

Domain	Name	Operational definition
D1	Opening and Rapport Building	Greeting, name use, warm tone, establishing patient comfort at the outset of the consultation
D2	Concern and Belief Elicitation	Asking about patient beliefs, fears, barriers, or understanding before providing advice or explanation
D3	Empathy and Emotional Responsiveness	Acknowledging feelings, validating concern, responding non-judgmentally to patient distress or difficulty
D4	Clarity of Explanation	Plain language, minimal jargon, logical and structured presentation of information
D5	Collaborative Planning	Involving patient in decisions, offering choices, seeking agreement on next steps
D6	Avoidance of Stigmatizing Language	Person-first language, non-blaming, non-paternalistic tone throughout the consultation
D7	Teach-Back and Safety-Netting	Confirming patient understanding; advising what to do if symptoms worsen or concerns arise

Note. Scale: 1 = absent or very poor; 2 = minimal; 3 = partial; 4 = adequate; 5 = excellent. Overall holistic score rated 1–9 independently of domain scores.

Results

3.1 Participation and platform feasibility

Of the 127 quality-screened transcripts, 30 were selected for duplicate faculty review and 29 were mutually analyzable. The three most frequent scenarios among the 383 sessions were osteoarthritis/injection preference counselling (n = 137, 35.8%), iron therapy adherence during pregnancy (n = 108, 28.2%), and pregnancy nutrition/cultural food beliefs (n = 90, 23.5%).

3.2 Pre-participation determinants (n = 123)

Table 2 presents determinant composite scores and subgroup comparisons. Overall mean (SD) and median (IQR) scores were: attitude 4.24 (0.59), 4.20 (4.00–4.60); self-efficacy 3.93 (0.58), 4.00 (3.60–4.20); anxiety-confidence 3.85 (0.48), 3.75 (3.50–4.00); contextual awareness 3.91 (0.63), 4.00 (3.67–4.33); and readiness for feedback 4.14 (0.58), 4.00 (4.00–4.33). Female students had higher attitude scores than males (4.36 vs 4.04, p = 0.003). No significant differences were observed by prior formal communication training. Among students with available exposure data (n = 115), regular previous clinical communication exposure was associated with higher readiness than occasional exposure (4.37 vs 4.16, p = 0.033).

Table 2
Pre-participation determinants questionnaire: composite domain scores (n = 123)

Composite domain	All Mean (SD)	Female Mean (SD)	Male Mean (SD)	p Gender	Prior training Yes Mean	Prior training No Mean	p Training	Regular vs occasional exposure Mean	p Exposure
Attitude	4.24 (0.59)	4.36 (0.40)	4.04 (0.79)	0.003	4.23	4.25	0.873	4.44 vs 4.24	0.086
Self-efficacy	3.93 (0.58)	3.96 (0.56)	3.88 (0.62)	0.446	3.98	3.89	0.432	4.12 vs 3.92	0.106
Anxiety-confidence	3.85 (0.48)	3.83 (0.49)	3.89 (0.47)	0.540	3.82	3.88	0.525	3.87 vs 3.87	0.981
Contextual awareness	3.91 (0.63)	3.89 (0.62)	3.93 (0.67)	0.736	3.95	3.88	0.555	4.03 vs 3.92	0.455
Readiness for feedback	4.14 (0.58)	4.21 (0.49)	4.01 (0.69)	0.055	4.16	4.12	0.661	4.37 vs 4.16	0.033

Note. Composite scores are means on a five-point Likert scale (1 = strongly disagree to 5 = strongly agree). Gender and prior-training comparisons use n = 123. Previous clinical communication exposure comparisons use n = 115 because exposure category data were unavailable or outside the regular/occasional comparison for eight participants. p-values are from independent-samples t-tests.

3.3 Faculty-rated baseline communication competency

Faculty-rated domain scores for the 29 mutually analyzable transcripts are presented in Table 3. The mean total score was 22.55/35 (SD 7.35; median 22.0; IQR 17.0–28.0). Domain means were D1 (2.64), D2 (2.60), D3 (3.76), D4 (3.76), D5 (3.07), D6 (4.59), and D7 (2.14). The mean holistic score was 6.14/9 (SD 2.15).

Table 3
Faculty-rated baseline communication competency by domain (n = 29 analyzable transcripts)

Domain	Mean (SD)	Median	IQR	% scoring ≤ 2 (low)
D6 Stigma-Free Language	4.59 (0.95)	5.0	5.0–5.0	3%
D3 Empathy and Emotional Resp.	3.76 (1.24)	4.0	3.0–5.0	14%
D4 Clarity of Explanation	3.76 (1.09)	4.0	3.0–5.0	17%
D5 Collaborative Planning	3.07 (1.50)	3.0	2.0–4.0	41%
D1 Opening and Rapport	2.64 (1.68)	2.0	1.0–4.0	55%
D2 Concern/Belief Elicitation	2.60 (1.46)	2.0	1.0–4.0	55%
D7 Teach-Back/Safety-Net	2.14 (1.30)	2.0	1.0–3.0	66%
Total score (sum D1–D7)	22.55 (7.35)	22.0	17.0–28.0	—
Overall Holistic Score (1–9)	6.14 (2.15)	6.0	5.0–8.0	—

Note. Scores are means of both raters on a 1–5 scale per domain (maximum total = 35). Holistic score is on a 1–9 scale. % scoring ≤ 2 indicates the proportion of transcripts where the rater mean for that domain was 2.0 or below, reflecting absent or minimal demonstration of the competency.

3.4 Inter-rater reliability

One of the 30 faculty-reviewed transcripts was marked Unsure by both raters because of residual ASR repetition artefacts and was excluded from paired scoring, yielding 29 mutually analyzable transcripts. Inter-rater reliability was excellent across domains, with a mean ICC (2,1) of 0.996. No disagreement of two or more points was observed across the 203 paired domain ratings (29 transcripts × 7 domains) as described in the Table 4.

Table 4
Inter-rater reliability using ICC (2,1) by domain (n = 29 analyzable transcripts)

Domain	R1 mean	R2 mean	% Exact agree	% Within ± 1	ICC (2,1)
D1 Opening and Rapport	2.62	2.66	97%	100%	0.994
D2 Concern/Belief Elicitation	2.59	2.62	97%	100%	0.992
D3 Empathy and Emotional Resp.	3.76	3.76	100%	100%	1.000
D4 Clarity of Explanation	3.76	3.76	100%	100%	1.000
D5 Collaborative Planning	3.03	3.10	93%	100%	0.985
D6 Stigma-Free Language	4.59	4.59	100%	100%	1.000
D7 Teach-Back/Safety-Net	2.14	2.14	100%	100%	1.000
Overall Holistic (1–9)	6.14	6.14	100%	100%	1.000

Note. ICC (2,1) = two-way random-effects, single-measure model assessing absolute agreement between faculty raters. Mean ICC = 0.996. Exact agreement and within ± 1 is presented descriptively.

3.5 Scenario-wise communication patterns

Table 5 presents domain means by scenario for the three case types with sufficient numbers for descriptive reporting. The pattern of stronger explanation and stigma-free language with weaker elicitation and teach-back was consistent across all three major case types. The osteoarthritis scenario showed the highest D5 (Collaborative Planning) mean (3.44/5), reflecting some negotiation of the patient's injection demand. The iron tablets scenario showed the lowest D1 (Opening and Rapport) mean (1.94/5), suggesting more abrupt consultation openings in this case type. Teach-back and safety-netting (D7) was weak across all scenarios (range 2.00–2.12/5). Single-transcript case types (diabetes denial, child diarrhea) are presented for completeness only and are not interpreted given insufficient N.

Table 5
Scenario-wise faculty-rated domain means: three principal case types

Scenario	n	D1	D2	D3	D4	D5	D6	D7
OA/injection preference	9	2.89	2.33	3.67	3.67	3.44	4.44	2.11
Pregnant/cultural food beliefs	8	2.62	2.62	3.88	3.88	3.56	4.62	2.00
Pregnant/iron tablets	8	1.94	2.31	3.62	3.62	2.19	4.75	2.12

Note. D1–D7 scores are means (both raters averaged) on a 1–5 scale. OA = osteoarthritis. Cases with n = 1 (diabetes denial, child diarrhea) are excluded from this table. D5 for Pregnant/iron tablets is notably low (2.19), reflecting limited collaborative planning in iron adherence counselling.

3.6 Exploratory AI–faculty concordance

The clean post hoc AI rescoring of the same 29 analyzable transcripts showed high agreement with mean faculty ratings. Lin's CCC for the total D1–D7 score was 0.960, with a mean AI–faculty difference of − 1.00 point on the 35-point scale and 95% Bland–Altman limits of agreement from − 4.43 to + 2.43. For the holistic score, CCC was 0.966, mean difference was − 0.24, and 95% limits of agreement were − 1.24 to + 0.76. Domain-level CCC ranged from 0.708 for D7 (Teach-Back and Safety-Netting) to 0.960 for D6 (Avoidance of Stigmatizing Language). D7 showed the largest directional difference, with AI ratings averaging 0.52 points lower than faculty ratings. Table 6 summarizes the agreement analysis and Fig. 2 presents the Bland–Altman plot for the total score.

Table 6
Exploratory agreement between AI and faculty ratings on the seven-domain communication rubric (n = 29 analyzable transcripts)

Measure	AI mean	Faculty mean	Lin's CCC	Mean diff.	MAE	95% LoA
D1 Opening and Rapport	2.66	2.64	0.957	+ 0.02	0.22	−0.91 to + 0.94
D2 Concern/Belief Elicitation	2.24	2.60	0.883	−0.36	0.43	−1.56 to + 0.84
D3 Empathy/Emotional Responsiveness	3.55	3.76	0.851	−0.21	0.34	−1.53 to + 1.12
D4 Clarity of Explanation	3.72	3.76	0.755	−0.03	0.38	−1.47 to + 1.40
D5 Collaborative Planning	3.17	3.07	0.893	+ 0.10	0.41	−1.24 to + 1.45
D6 Stigma-Free Language	4.59	4.59	0.960	0.00	0.07	−0.52 to + 0.52
D7 Teach-Back/Safety-Netting	1.62	2.14	0.708	−0.52	0.52	−1.96 to + 0.93
Total D1–D7 (7–35)	21.55	22.55	0.960	−1.00	1.48	−4.43 to + 2.43
Holistic score (1–9)	5.90	6.14	0.966	−0.24	0.24	−1.24 to + 0.76

Note. Faculty score = mean of the two independent faculty raters. AI scores were generated in a clean text-only rescoring pass using the fixed rubric and frozen before faculty ratings were merged. CCC = Lin's concordance correlation coefficient; mean difference = AI minus faculty; MAE = mean absolute error; LoA = 95% Bland–Altman limits of agreement. The post hoc comparison is exploratory and is not presented as formal validation of automated scoring.

Discussion

4.1 Principal findings

This cross-sectional pilot demonstrates the feasibility of AI-assisted simulated consultation capture and faculty-reviewed communication assessment in an Indian undergraduate medical education setting. Faculty-rated baseline competency showed higher scores for clarity of explanation and avoidance of stigmatizing language and lower scores for concern elicitation, rapport building, and teach-back/safety-netting. Pre-participation data showed strong endorsement of patient-centered communication values alongside lower self-efficacy and anxiety-confidence

scores. A secondary post hoc analysis showed high overall AI–faculty agreement by Lin's CCC, although this exploratory comparison was not a prespecified validation study.

4.2 The elicitation gap: educational and clinical significance

The finding that concern and belief elicitation (D2) was among the lowest-scoring domains across all scenario types is the most educationally actionable result of this study. Elicitation, asking what the patient believes, fears, expects, or finds difficult before advising, is the behavior most consistently associated with patient understanding, adherence, and shared decision-making.[2] Without elicitation, even technically correct clinical advice may fail to address the patient's actual barrier to behavior change. In the osteoarthritis scenario, where patients are demanding injections and refusing exercise, the absence of elicitation means the student's counselling does not engage with the patient's reasoning for the preference. In the iron tablets scenario, failure to explore the specific side effect experience before advising continuation produces a directive rather than collaborative exchange.

This pattern is compatible with the broader undergraduate communication-training literature, which shows that communication skills require explicit teaching, repeated practice, and feedback rather than knowledge transmission alone.[7,8,12] The distinction between knowing that elicitation is important (high attitude scores) and consistently doing it in a simulated consultation (low D2 scores) is the gap that structured deliberate practice with feedback is designed to address.[12]

4.3 Teach-back as the weakest domain

D7 (Teach-Back and Safety-Netting) recorded the lowest mean faculty score (2.14/5) and the highest proportion of transcripts scored at or below 2 (66%). Teach-back, confirming that the patient has understood the key message and advising what to do if symptoms worsen, has a robust evidence base for improving patient understanding, self-management, and health outcomes across a range of clinical contexts.15 Its near-absence in baseline student consultations suggests that this behavior has not yet been integrated into students' default consultation repertoire despite its formal inclusion in AETCOM competency frameworks.6 This is consistent with evidence that teach-back, despite being broadly endorsed, is inconsistently implemented and requires explicit instruction and practice.[15]

4.4 The attitude–efficacy gap and its implications

The attitude–efficacy gap identified in the pre-participation questionnaire, students endorsing patient-centered communication values at 4.24/5 while rating their own execution at 3.93/5 and anxiety-confidence at 3.85/5, is directly mirrored in the faculty-rated consultation data. Students who believe strongly in patient-centered communication and yet default to directive, explanation-heavy consultations are not being deceitful about their values; they have not yet developed the behavioral fluency that converts attitudinal endorsement into consistent practice under the cognitive load of a real clinical interaction.

This distinction has implications for training design. If the problem were one of knowledge or attitude, AETCOM's curricular content would be sufficient. The evidence here suggests the gap is one of execution under realistic conditions. Deliberate practice with structured feedback, not additional didactic sessions, is the appropriate response.[12] Previous exposure to clinical communication was the only demographic variable with a significant association with readiness for feedback ($p = 0.033$), suggesting that experiential exposure to communication contexts may matter in developing readiness to engage with structured development.

4.5 Scenario self-selection as a behavioral finding

Approximately 88% of sessions were concentrated in three of the thirteen available scenarios. Students most often selected osteoarthritis/injection preference, iron-tablet adherence, and pregnancy nutrition/cultural-belief scenarios, while several resistance-management scenarios were attempted less frequently. This distribution may reflect preference for more familiar or perceived manageable scenarios, but reasons for scenario choice were not measured. The concentration nevertheless limits the breadth of communication contexts represented and supports the use of assigned scenario sets in future studies.

4.6 ConversationAIly as feasible educational infrastructure

The platform captured simulated consultations across 13 scenario types and generated a reviewable transcript corpus suitable for asynchronous faculty assessment. The pilot also identified export and transcript-quality issues arising from persistence of incremental recognition fragments, which required deduplication and quality screening before analysis. Stable session identifiers, final-transcript flagging, suppression of interim recognition output, and reliable case-type metadata are therefore implementation priorities for future iterations. The transparent separation of faculty-rated primary outcomes from exploratory automated outputs is a methodological strength of the study. Additional strengths include IEC approval, use of a real undergraduate cohort, blinded independent faculty scoring, almost-perfect inter-rater reliability, and explicit reporting of the transcript data pipeline.

The clean post hoc AI–faculty comparison provides an exploratory assessment of transcript-level rubric scoring. Agreement was high for the total seven-domain score (Lin's CCC = 0.960) and holistic score (CCC = 0.966). The total-score mean difference was − 1.00/35, with 95% limits of agreement from − 4.43 to + 2.43. D7 showed the lowest domain-level CCC (0.708) and the largest directional difference, with AI ratings averaging 0.52 points lower than faculty ratings. Because the analysis was post hoc and not prespecified as a validation study, these findings should be interpreted as exploratory agreement rather than evidence of validated automated competency scoring.

4.7 Comparison with existing literature

The finding of a gap between communication attitudes and communication behavior is consistent with Indian undergraduate medical education literature. Varma et al. found that recently admitted medical students recognized the importance of communication-skills training while also showing substantial ambivalence and unfavorable attitudes toward communication skills as a formal subject.[9] Iqbal et al. demonstrated significant improvement in final-year medical students' communication performance after focused training and noted the likely importance of repeated training for retention.[10] D'Souza et al. demonstrated short-term empathy score improvement following a single-session communication-skills intervention.[16] The present study extends this body of evidence by providing transcript-level faculty-rated evidence of specific behavioral gaps, particularly elicitation and teach-back, rather than relying on self-report or post-intervention outcome measures alone.

Internationally, the Cochrane systematic review by Gilligan et al. found that communication-skills interventions generally show positive effects across multiple outcomes, but that effect sizes are often modest and certainty of evidence varies.[17] The present domain pattern should therefore be interpreted as a context-specific baseline finding rather than assumed to reproduce a universal sequence of communication-skill development across settings.

4.8 Limitations

Several limitations must be acknowledged. This was a single-institution pilot with a small faculty-rated sample (n = 29 analyzable transcripts), so findings are preliminary and hypothesis-generating. Consultations were simulated and audio-only, limiting ecological validity compared with real clinical encounters. Three scenario types accounted for 87.5% of sessions, restricting the breadth of communication contexts represented. The pre-participation

questionnaire was developed for this pilot and had not undergone formal psychometric validation; determinant scores should therefore be interpreted descriptively. Reasons for non-participation or non-completion among registered students without linked consultations were not systematically collected. The exploratory AI–faculty agreement analysis was post hoc and should not be interpreted as formal validation of automated scoring. No intervention was administered and no patient outcomes were measured.

Conclusions

This pilot study demonstrates the feasibility of AI-assisted simulated consultation capture and faculty-reviewed communication assessment in an Indian undergraduate medical education setting. Faculty-rated findings identify lower performance in elicitation, rapport building, and teach-back/safety-netting relative to explanation and stigma-free language across the major scenarios. The seven-domain rubric demonstrated high faculty inter-rater reliability and provides a structured basis for future transcript assessment. Exploratory AI–faculty agreement was high overall by Lin's CCC, but prospective independent validation is required before automated scores can be used as standalone competency measures. These findings provide a domain-specific training agenda and a methodological foundation for larger studies of structured communication feedback.

Declarations

Ethics approval and consent to participate

Ethical approval was granted by the Institutional Ethics Committee, IMS and SUM Hospital, Siksha O Anusandhan (Deemed to be University), Bhubaneswar (IEC Ref: IEC/IMS.SH/SOA/2026/1325 dated 24 March 2026). Digital informed consent was obtained from all participants before enrolment. Participation was voluntary with no academic consequences. The study was conducted in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable. No individual participant data, images, or identifying information are included in this manuscript.

Declarations

Competing Interests

MK, AS, JS, VE and SD declare no competing interests. SS and RM are inventors of ConversationAIly, which is a product of SigmaMozak Solutions. The platform was used as data-collection and transcript-review infrastructure for this study. The inventors were not involved in faculty scoring of transcripts. Faculty ratings remained the primary competency outcomes; the AI–faculty concordance analysis was post hoc and exploratory.

Funding

This study received no external funding. It was conducted as an internally approved pilot within the Department of Community Medicine, IMS and SUM Hospital.

Author Contribution

MK, AS, VE, SD, SS and RM conceived the study, designed the protocol, led data collection and analysis, and drafted the manuscript. All authors read and approved the final version.

Acknowledgement

The authors thank the Phase III Part I MBBS students who voluntarily participated and the faculty raters whose independent scoring made the IRR analysis possible. The authors acknowledge SigmaMozak Solutions for providing access to the ConversationAlly platform.

Data Availability

The datasets supporting the conclusions of this article are available from the corresponding author on reasonable request, subject to institutional data governance approval and de-identification of participant records.

References

1. Stewart MA. Effective physician-patient communication and health outcomes: a review. CMAJ. 1995;152(9):1423–33. PMID: 7728691.
2. Street RL Jr, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. Patient Educ Couns. 2009;74(3):295–301. 10.1016/j.pec.2008.11.015.
3. Stewart M, Brown JB, Donner A, McWhinney IR, Oates J, Weston WW, et al. The impact of patient-centered care on outcomes. J Fam Pract. 2000;49(9):796–804. PMID: 11032203.
4. Silverman J, Kurtz S, Draper J. Skills for Communicating with Patients. 2nd ed. Oxon: Radcliffe Publishing; 2005.
5. Kurtz SM, Silverman JD. The Calgary-Cambridge Referenced Observation Guides: an aid to defining the curriculum and organising teaching in communication training programmes. Med Educ. 1996;30(2):83–9. 10.1111/j.1365-2923.1996.tb00724.x.
6. National Medical Commission. Attitude, Ethics and Communication (AETCOM) Competencies for the Indian Medical Graduate. New Delhi: NMC. 2018. Available from: https://www.nmc.org.in/wp-content/uploads/2020/01/AETCOM_book.pdf
7. Modi JN, Anshu, Chhatwal J, Gupta P, Singh T. Teaching and assessing communication skills in medical undergraduate training. Indian Pediatr. 2016;53(6):497–504. 10.1007/s13312-016-0879-z.
8. Udgiri R, Ganganahalli P. Perceptions of Attitude, Ethics, and Communication (AETCOM) Modules Among Indian Medical Graduates in Their First Professional Year. Cureus. 2024;16(7):e64611. 10.7759/cureus.64611.
9. Varma J, Prabhakaran A, Singh S. Perceived need and attitudes towards communication skill training in recently admitted undergraduate medical students. Indian J Med Ethics. 2018;3(3):196–200. 10.20529/IJME.2018.049.
10. Iqbal N, Mookkappan S, Basheer A, Kandasamy R. Impact of focused training on communication skills of final-year medical students in a medical school in India. Australas Med J. 2015;8(10):325–32. 10.4066/AMJ.2015.2509.
11. Tanwani R, Chandki R, Joshi A, Arora VK, Nyati P, Sutay S. Perception and attitude of medical students towards Communication Skills Lab and teaching module. J Clin Diagn Res. 2017;11(6):JC12–4. 10.7860/JCDR/2017/24858.10120.

12. Issenberg SB, McGaghie WC, Petrusa ER, Gordon DL, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach. 2005;27(1):10–28. 10.1080/01421590500046924.
13. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. BMC Med Educ. 2025;25(1):129. 10.1186/s12909-025-06719-5.
14. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63. 10.1016/j.jcm.2016.02.012.
15. Talevski J, Wong Shee A, Rasmussen B, Kemp G, Beauchamp A. Teach-back: a systematic review of implementation and impacts. PLoS ONE. 2020;15(4):e0231350. 10.1371/journal.pone.0231350.
16. D'Souza PC, Rasquinha SL, D'Souza TL, Jain A, Kulkarni V, Pai K. Effect of a single-session communication skills training on empathy in medical students. Acad Psychiatry. 2020;44(3):289–94. 10.1007/s40596-019-01158-z.
17. Gilligan C, Powell M, Lynagh MC, Ward BM, Lonsdale C, Harvey P, et al. Interventions for improving medical students' interpersonal communication in medical consultations. Cochrane Database Syst Rev. 2021;2(2):CD012418. 10.1002/14651858.CD012418.pub2.

Figures

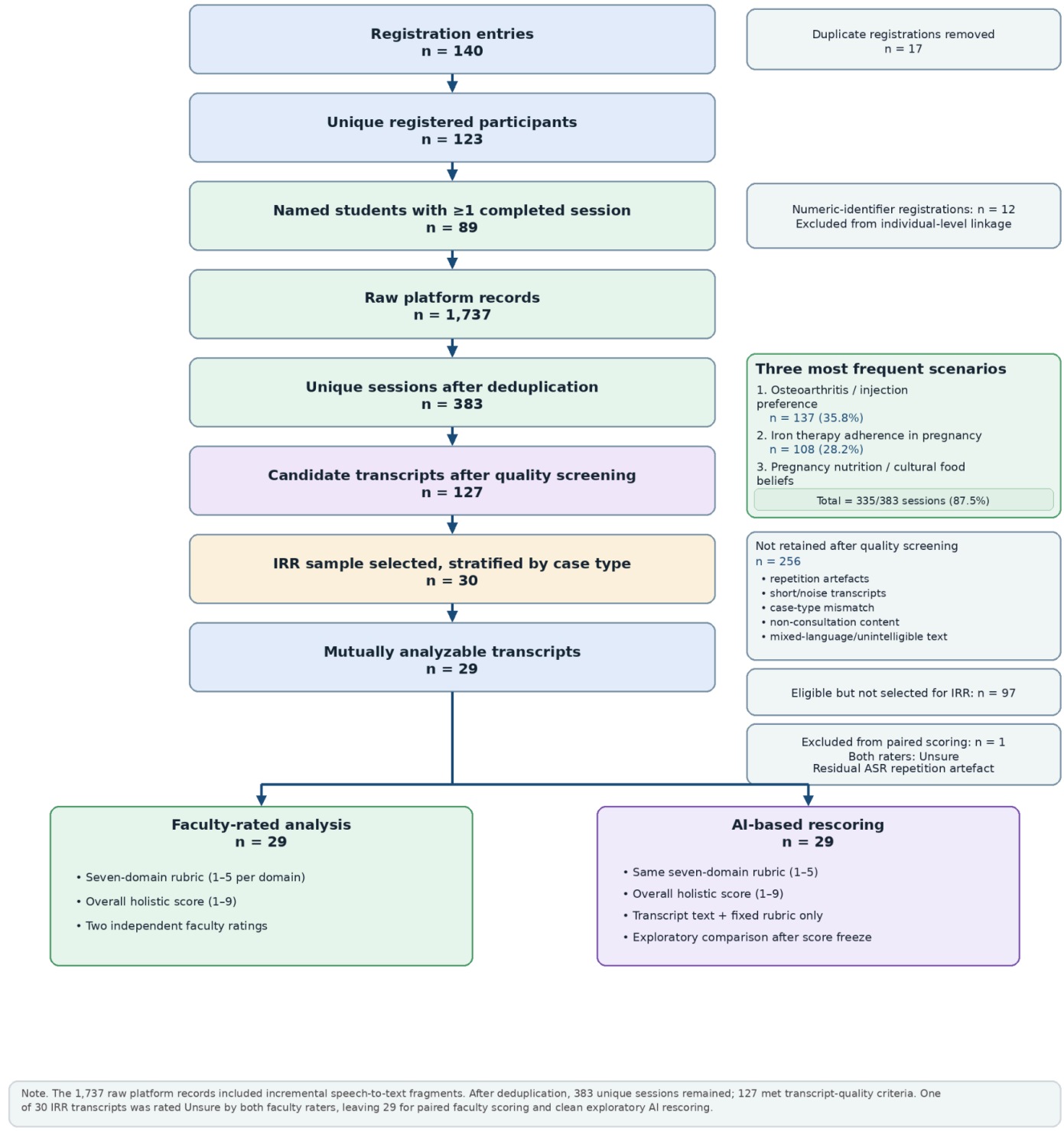


Figure 1

Study flow of registration, session capture and transcript selection. From registration to the final faculty-rated and exploratory AI-rescored analytical sample.

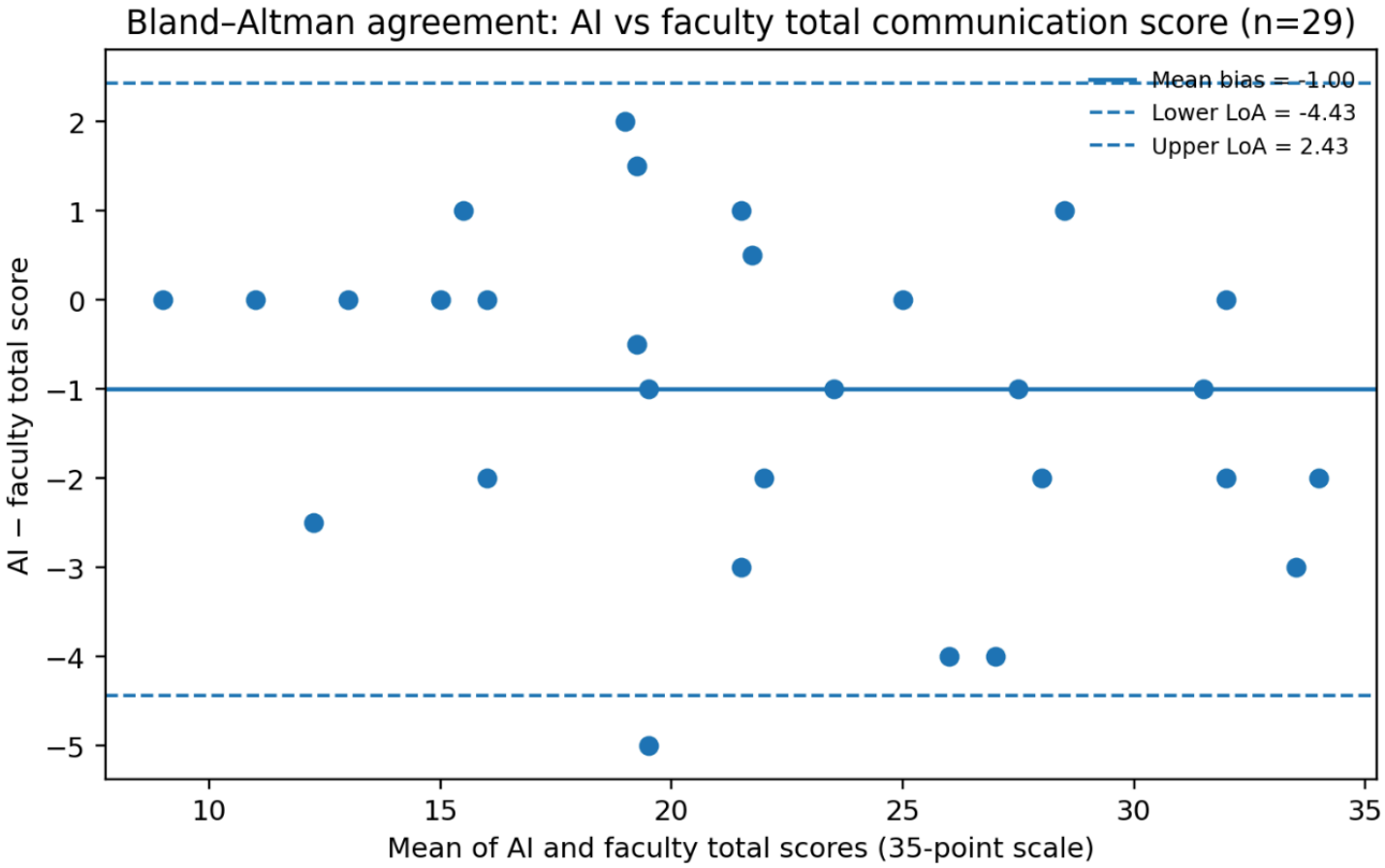


Figure 2

Bland–Altman plot of agreement between AI and mean faculty total seven-domain scores (n = 29). The solid line shows mean bias (−1.00 point); dashed lines show the 95% limits of agreement (−4.43 to +2.43).