

Identifying Communication Profiles from Physician–Patient Conversations Using Large Language Models

Reyhaneh Hosseinpourkhoshkbari * Yash Bipin Jain Richard M. Golden

School of Behavioral and Brain Sciences, The University of Texas at Dallas

Abstract

Large language models (LLMs) are increasingly used to score clinical communication, but validation is often limited to individual checklist items or total scores. These measures may not capture how communication behaviors occur together within a transcript. We therefore examine *communication profiles*: recurring combinations of behaviors that characterize different patterns of clinician communication and may support more targeted formative feedback. We analyzed 213 simulated respiratory OSCE transcripts rated on 15 binary Kalamazoo-derived communication items and compared final human ratings with GPT-4o, GPT-o3, and GPT-5.6 Sol. Raw item-level agreement was relatively high overall, but varied substantially across behaviors and was lower for several judgment-intensive items used for profile modeling. Bayesian latent-class analysis identified three stable human-derived profiles. Although the LLM-derived profiles showed broadly similar item-probability patterns, the models frequently assigned individual transcripts to different profiles than the human ratings. These findings show that agreement on individual communication skills does not necessarily translate into agreement on higher-level communication profiles. If LLMs are used to provide profile-based feedback, validation should therefore include profile-level agreement in addition to item-level performance.

1 Introduction

Clinical communication is central to safe and effective care. Communication quality is associated with patient satisfaction, adherence, recovery, and other clinical outcomes, while communication failures can contribute to errors and poorer care (Chavda et al., 2020; Hosseinpourkhoshkbari and Golden, 2025; Singh et al., 2013; Ahmadipour and Mortazavi, 2015). Simulated encounters and

standardized patients provide controlled settings in which these skills can be practiced and assessed (Hosseinpourkhoshkbari and Golden, 2025; Mavis et al., 2006; Johnson et al., 2020). The Kalamazoo Essential Elements framework provides a widely used structure for evaluating communication behaviors across a transcript (Makoul, 2001; Dalus et al., 2024; Yoon and Michaelsen, 2015). Because assessment outcomes can vary across evaluators, establishing and reporting the reliability of these ratings is essential.

Communication assessments are often summarized using total or domain-level scores, although reliability and agreement can vary substantially across individual communication domains (Fan et al., 2024). Examining item-level performance can identify specific strengths and weaknesses that may be hidden by similar overall scores and can therefore provide more informative diagnostic feedback (Ark et al., 2024). Moreover, two transcripts with similar total scores may show different combinations of communication strengths and weaknesses. These joint patterns may be particularly important for formative feedback.

Communication assessment is also difficult to perform consistently at scale (Brannick et al., 2011). Human raters tend to agree more readily on directly observable behaviors than on behaviors requiring more subjective judgments, such as empathy or understanding the patient’s perspective (Roh et al., 2017). LLMs have recently shown promise for automating assessment of medical interviews and OSCE performance (Shakur et al., 2024; Geathers et al., 2025), but most validation focuses on individual items or aggregate scores.

As a result, we examine communication profiles: recurring patterns in how multiple communication behaviors co-occur within a transcript. If an automated system is ultimately used to characterize these broader patterns for formative feedback, the relevant question is not only whether individ-

*reyhaneh.hosseinpour@utdallas.edu

ual items are scored similarly to human raters, but whether the same higher-level profiles are recovered.

We therefore distinguish two levels of agreement. *Item-level agreement* asks whether human and LLM raters make the same judgment on individual communication behaviors. *Profile-level agreement* asks whether they assign the same learner to the same multidimensional communication pattern. Our central question is:

Does Item-level agreement on communication behaviors translate into Profile-level agreement?

We make four contributions. First, we compare human ratings with GPT-4o, GPT-o3, and GPT-5.6 Sol across all 15 Kalamazoo-derived items. Second, we compare the three LLMs item by item on the eight informative behaviors used for profile modeling, highlighting model-specific disagreement patterns. Third, we identify latent communication profiles using Bayesian latent-class modeling and evaluate their stability with resampling and an alternative clustering method. Fourth, we compare human-derived profiles with profiles obtained from all three LLMs.

2 Data and Annotation

2.1 Dataset

We used the respiratory-case subset of the simulated patient–physician interview dataset of [Fareez et al. \(2022\)](#). The corpus contains 213 OSCE-style interviews, each approximately 50 conversational turns, conducted by senior medical students and resident physicians in respiratory history-taking scenarios.

The human annotations were produced by three coders using the same operational definitions. A 50-transcript holdout set was independently rated by all three coders to assess inter-rater agreement. Disagreements were discussed and resolved to obtain the final item ratings. Complete final human ratings and ratings from GPT-4o, GPT-o3, and GPT-5.6 Sol were available for all 213 transcripts, which formed the paired comparison sample.

2.2 Assessment Instrument

Communication was coded using 15 binary Kalamazoo-derived behaviors (Table 1). Each rater reviewed the complete transcript and marked

Item	Operationalized behavior
Q1	Greet the patient
Q2	Use words showing care and concern
Q3	Allow the opening statement without interruption
Q4	Ask “Is there anything else?” or equivalent
Q5	Explain or negotiate an agenda
Q6	Begin with the patient’s story using open questions
Q7	Clarify details with specific questions
Q8	Summarize and invite correction/addition
Q9	Transition effectively to additional questions
Q10	Ask about life circumstances affecting health
Q11	Elicit beliefs, concerns, and expectations
Q12	Respond explicitly to feelings or values
Q13	Explain using words easy for the patient to understand
Q14	Ask whether the patient has questions or other issues
Q15	Acknowledge the patient and close the interview

Table 1: The 15 Kalamazoo-derived communication behaviors used in the analysis.

whether each behavior was present. The same operational definitions were used for human and LLM ratings; the full scoring prompt is included in Appendix A.

3 Methods

3.1 Agreement Analysis

Several binary items had highly imbalanced endorsement rates. Under this condition, Cohen’s κ can be strongly affected by prevalence because its chance-agreement correction can become large when raters frequently use the same category. We therefore used raw percent agreement and Gwet’s AC1 ([Gwet, 2008](#)). Gwet’s AC1 is a chance-corrected agreement coefficient that is less sensitive to extreme category prevalence, making it more appropriate for this rubric-based setting, where highly imbalanced response categories can produce counterintuitive chance-corrected agreement estimates despite high observed agreement. Gwet’s AC1 was interpreted as a continuous measure of chance-corrected agreement, with larger values indicating stronger agreement.

For the 50 independently rated holdout transcripts, human inter-rater agreement was summarized using mean pairwise raw agreement and multi-rater Gwet’s AC1. LLM–human agreement was calculated against the final human ratings on the full paired sample of 213 transcripts.

3.2 Selection of Informative Items

Profiles were estimated from items that showed sufficient variation across the final human ratings. Items that were endorsed by humans in fewer than 20% or more than 90% of transcripts provided little information for distinguishing one transcript from another, so they were excluded from the profile analysis. These items were Q1, Q3, Q5, Q6, Q7, Q9, and Q10. They were still included in the item-level agreement analysis. The remaining eight items—Q2, Q4, Q8, Q11, Q12, Q13, Q14, and Q15—were used to estimate the communication profiles.

3.3 Bayesian Latent Class Analysis

We used Bayesian latent-class analysis to identify recurring communication patterns across transcripts rather than reducing the ratings to a single total score (Li et al., 2018). The Bayesian approach allows uncertainty in the profile prevalences and profile-specific item probabilities to be represented through posterior distributions rather than treating these quantities as fixed estimates. Two transcripts may have the same number of endorsed items but differ in which items are present. The analysis included $N = 213$ transcripts and $J = 8$ binary communication items. Let $Y_{ij} = 1$ if item j is scored as present in transcript i , and $Y_{ij} = 0$ otherwise.

Latent profile structure. We assume that each transcript belongs to one of K latent communication profiles. These profiles are not observed directly but are inferred from recurring patterns across the binary communication items. Let $Z_i \in \{1, \dots, K\}$ denote the latent profile of transcript i . The profile proportions are

$$\boldsymbol{\pi} = (\pi_1, \dots, \pi_K), \quad \sum_{k=1}^K \pi_k = 1.$$

The latent profile indicator is modeled as

$$Z_i \mid \boldsymbol{\pi} \sim \text{Categorical}(\boldsymbol{\pi}), \quad \Pr(Z_i = k) = \pi_k,$$

where π_k represents the prevalence of profile k , that is, how common that profile is across transcripts.

Profile-specific item probabilities. Each profile is defined by the probability that each communication item is scored as present. Let

$$\theta_{kj} = \Pr(Y_{ij} = 1 \mid Z_i = k),$$

where θ_{kj} is the probability that item j is scored as present for a transcript belonging to profile k . Because the items are binary,

$$\Pr(Y_{ij} \mid Z_i = k) = \theta_{kj}^{Y_{ij}} (1 - \theta_{kj})^{1-Y_{ij}}.$$

The complete set of item probabilities defining profile k is represented by the profile-specific item-probability vector

$$\boldsymbol{\theta}_k = (\theta_{k1}, \dots, \theta_{kJ}).$$

Thus, each element of $\boldsymbol{\theta}_k$ represents the estimated probability that a particular communication item is scored as present for transcripts belonging to that profile. Throughout the paper, comparisons of human- and LLM-derived profiles are based on these profile-specific item-probability vectors.

Latent-class analysis assumes conditional independence of the items given the profile. Thus, for $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iJ})$,

$$\Pr(\mathbf{Y}_i \mid Z_i = k) = \prod_{j=1}^J \theta_{kj}^{Y_{ij}} (1 - \theta_{kj})^{1-Y_{ij}}.$$

This expression gives the probability of transcript i 's complete item pattern under profile k .

Bayesian priors. We specified two priors. First, the profile proportions were assigned a symmetric Dirichlet prior,

$$\boldsymbol{\pi} \sim \text{Dirichlet}(2, \dots, 2),$$

where

$$p(\boldsymbol{\pi}) = \frac{\Gamma(2K)}{\Gamma(2)^K} \prod_{k=1}^K \pi_k, \quad \sum_{k=1}^K \pi_k = 1.$$

This prior mildly discourages near-empty profiles while treating all profile labels symmetrically.

Second, each profile-specific item probability was assigned

$$\theta_{kj} \sim \text{Beta}(1, 1),$$

where

$$p(\theta_{kj}) = \frac{\theta_{kj}^0 (1 - \theta_{kj})^0}{B(1, 1)} = 1, \quad 0 \leq \theta_{kj} \leq 1.$$

Thus, the item-probability prior is uniform on $[0, 1]$ and does not favor high or low endorsement probabilities.

Bayesian posterior. Let $\Theta = \{\theta_{kj}\}$ denote the complete set of profile-specific item probabilities. The posterior distribution of π and Θ given the observed data $\mathbf{Y}$ is

$$p(\pi, \Theta \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \pi, \Theta) p(\pi) p(\Theta).$$

The likelihood is obtained by marginalizing over the unknown latent profile indicator for each transcript:

$$p(\mathbf{Y} \mid \pi, \Theta) = \prod_{i=1}^N \left[\sum_{k=1}^K \pi_k \prod_{j=1}^J \theta_{kj}^{Y_{ij}} (1 - \theta_{kj})^{1-Y_{ij}} \right]$$

with

$$p(\Theta) = \prod_{k=1}^K \prod_{j=1}^J p(\theta_{kj}).$$

Posterior profile probabilities and classification.

For a given posterior draw of the model parameters, the conditional probability that transcript i belongs to profile k is

$$\tau_{ik} = \Pr(Z_i = k \mid \mathbf{Y}_i, \pi, \Theta) = \frac{s_{ik}}{\sum_{\ell=1}^K s_{i\ell}},$$

where

$$s_{ik} = \pi_k \prod_{j=1}^J \theta_{kj}^{Y_{ij}} (1 - \theta_{kj})^{1-Y_{ij}}.$$

These probabilities were computed at each posterior draw and then averaged across draws. Let

$$\bar{\tau}_{ik} = \frac{1}{S} \sum_{s=1}^S \tau_{ik}^{(s)}$$

denote the posterior-averaged probability that transcript i belongs to profile k , where S is the total number of retained posterior draws. We refer to $\bar{\tau}_{ik}$ as the *posterior profile probability*. These posterior profile probabilities reflect both how well a transcript’s item pattern fits each profile and uncertainty in the estimated model parameters. Each transcript was placed into the profile with the largest posterior profile probability,

$$\hat{Z}_i = \arg \max_k \bar{\tau}_{ik}.$$

For each transcript, the *maximum posterior profile probability*,

$$\max_k \bar{\tau}_{ik},$$

was used as a measure of classification certainty. Values closer to 1 indicate that one profile is clearly favored, whereas lower values indicate greater uncertainty across profiles.

Model estimation and profile selection. Models were estimated using Markov Chain Monte Carlo (MCMC) with the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014), implemented in PyMC (Salvatier et al., 2016). MCMC was used to generate samples from the posterior distribution of the model parameters, and NUTS provided an efficient sampling algorithm for these continuous parameters. We used four chains, 2,000 tuning iterations, 1,000 retained posterior draws per chain, and a target acceptance rate of 0.95. We fit $K = 2, 3$, and 4 solutions and compared AIC, BIC, and MCMC convergence using $\hat{R}$, while also examining whether the resulting profiles were non-negligible and interpretable. $\hat{R}$ assesses convergence across MCMC chains, with values close to 1 indicating that the chains sampled from the same posterior distribution.

Applying model to LLM ratings. After selecting K from the final human ratings, we fit the same eight-item model separately to GPT-4o, GPT-o3, and GPT-5.6 Sol. Because latent-class labels are arbitrary, each LLM-derived profile was matched to the human-derived profiles by maximizing cosine similarity between the corresponding profile-specific item-probability vectors θ_k . This matching step only relabels the profiles for comparison; it does not alter the item ratings or the transcript-level profile probabilities.

3.4 Robustness and Complementary Analyses

We assessed stability using 200 bootstrap resamples and 200 refits after randomly dropping 5% of transcripts. We also applied PAM (k-medoids) clustering with Jaccard distance to the same eight binary items as a non-mixture comparison. In contrast to latent-class analysis, PAM groups transcripts based on similarity without assuming an underlying probabilistic mixture model.

4 Results

4.1 Agreement Varies Across Items and Evaluator Types

Among the three human coders, mean pairwise raw agreement across all 15 items was 82.6%, and mean multi-rater Gwet’s AC1 was 0.712. For the

eight profile-modeling items retained based on variation in the final human ratings, mean pairwise raw agreement decreased to 70.5%, and mean multi-rater Gwet's AC1 decreased to 0.498. Thus, human agreement was lower for the eight items used to construct the communication profiles.

LLM-human agreement showed a similar pattern. Across all 15 items, mean Gwet's AC1 was 0.773 for GPT-4o, 0.718 for GPT-o3, and 0.701 for GPT-5.6 Sol. Across the eight profile-modeling items, mean Gwet's AC1 decreased to 0.600, 0.514, and 0.480, respectively. On the full 213-transcript paired sample, mean raw agreement across the eight items was 74.4% for GPT-4o, 71.2% for GPT-o3, and 71.0% for GPT-5.6 Sol (Table 2).

LLM-human agreement varied substantially across communication behaviors (Figure 1). Agreement was consistently high for Q1, Q3, Q5–Q7, Q9–Q10, Q13, and Q14, while Q2 showed consistently lower agreement across all three models. Q8, Q11, and Q12 showed the largest model-dependent differences, with particularly low agreement for some models.

Overall, agreement patterns were qualitatively similar across human-human, human-LLM, and LLM-LLM comparisons, with consistently lower agreement for Q2, Q8, Q11, and Q12, although the magnitude of disagreement varied across evaluator types. Across the eight profile-modeling items, mean Gwet's AC1 was 0.498 among the human coders and 0.600, 0.514, and 0.480 for GPT-4o, GPT-o3, and GPT-5.6 Sol versus the final human ratings, respectively; under the Landis and Koch (1977) benchmarks, these values fall within the moderate agreement range.

4.2 Bayesian Latent Class Model Selection Results

Table 3 summarizes model selection using the Bayesian Latent Class Analysis. The $K = 3$ solution had the lowest AIC and BIC. All models converged without divergent transitions; for $K = 3$, the maximum $\hat{R}$ for the profile proportions was 1.002. We therefore retained $K = 3$ as the primary solution.

The three profiles showed distinct patterns (Table 4; Figure 2). Profile 1 was characterized by high probabilities for Q4, Q13, and Q14, moderate probabilities for Q15 and Q8, and low probabilities for Q2, Q11, and Q12. Profile 2 showed the broadest pattern of high probabilities, including Q4, Q11, Q12, Q13, Q14, and Q15, with moder-

ate Q8 and low Q2. In contrast, Profile 3 showed high probabilities for Q2, Q13, Q14, and Q15, but low probabilities for Q4, Q8, Q11, and Q12. The estimated profile prevalences were 44%, 34%, and 23%, respectively.

The distribution of maximum posterior profile probabilities showed variation in classification certainty (Figure 3). Many transcripts were assigned to a profile with very high posterior probability, whereas a smaller group showed more moderate probabilities around 0.50–0.65, indicating greater uncertainty between profiles. Very few transcripts had probabilities close to the theoretical minimum of 1/3, suggesting that highly ambiguous profile classification were uncommon (Figure 3, A). In addition, most transcripts were clearly associated with one profile, while a smaller subset showed overlap across profiles, indicating greater classification uncertainty (Figure 3, B).

4.3 The Human-Derived Profiles Are Stable

The $K = 3$ solution was recovered in 99% of 200 bootstrap resamples and 100% of 200 refits after randomly dropping 5% of transcripts. PAM clustering with Jaccard distance produced a complementary partition with moderate similarity to the latent-class solution (Adjusted Rand Index [ARI] = 0.56).

4.4 LLM-Derived Profiles Show Overall Pattern Similarity but Item-Specific Differences

After matching the LLM-derived profiles to the human-derived profiles, cosine similarity between the corresponding profile-specific item-probability vectors ranged from .81 to .99 (Figure 4). These values indicate similarity in the overall direction of the eight-item profile patterns, but they do not imply close agreement for every individual item.

The degree of item-probability correspondence varied across profiles. For Profile 1, the human and LLM solutions were similar for several items, including Q4, Q11, and Q12, but differed substantially for Q13 and Q14. For Profile 3, similar probabilities were observed for Q13–Q15, whereas large differences were evident for Q2 and Q4. Profile 2 showed the strongest overall correspondence, although differences remained for Q2, Q8, and Q11. Thus, the LLM-derived profiles captured some broad features of the human-derived patterns while differing substantially on particular communication behaviors.

Item	GPT-4o		GPT-o3		GPT-5.6 Sol	
	Raw	AC1	Raw	AC1	Raw	AC1
Q2 Care	62.4%	.352	67.1%	.371	68.5%	.372
Q4 Full concerns	78.4%	.733	78.4%	.733	78.4%	.733
Q8 Summarize	73.2%	.552	69.5%	.417	54.5%	.093
Q11 Elicit	62.9%	.397	61.0%	.352	61.0%	.221
Q12 Respond	66.7%	.431	54.5%	.090	53.5%	.071
Q13 Explain	85.9%	.836	78.4%	.725	84.5%	.807
Q14 Ask questions	89.7%	.838	84.5%	.744	91.5%	.871
Q15 Close	76.1%	.660	76.5%	.679	75.6%	.675
Mean	74.4%	.600	71.2%	.514	71.0%	.480

Table 2: LLM–human agreement for the eight profile-modeling items on the full paired sample ($N = 213$). Raw agreement and Gwet’s AC1 are calculated against the final human ratings. The independent 50-transcript holdout is used separately for human inter-rater agreement.

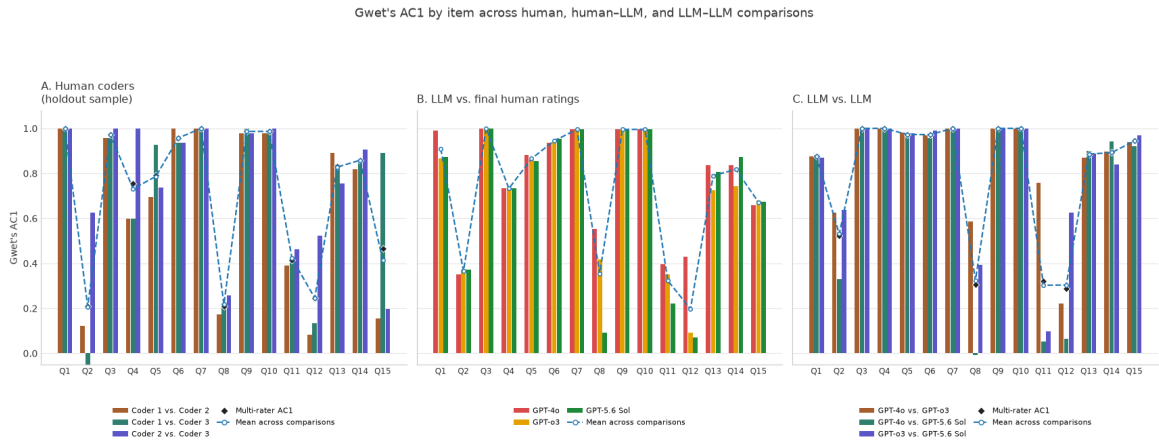


Figure 1: Item-level Gwet’s AC1 across evaluator types. Panel A shows pairwise and multi-rater agreement among the three human coders on the 50 independently rated holdout transcripts. Panel B shows each LLM against the final human ratings on all 213 transcripts. Panel C shows pairwise and multi-rater agreement among the three LLMs on all 213 transcripts. Overall, agreement patterns were qualitatively similar across human–human, human–LLM, and LLM–LLM comparisons, with consistently lower agreement for Q2, Q8, Q11, and Q12, although the magnitude of disagreement varied across evaluator types.

Profiles	AIC	BIC
$K = 2$	1897.9	1955.0
$K = 3$	1861.6	1949.0
$K = 4$	1863.2	1980.8

Table 3: Model-selection statistics for the human-derived eight-item latent-class models. Lower values indicate better fit after penalizing model complexity.

Profile prevalence also differed across the human and LLM solutions (Table 5). The GPT-4o solution estimated a substantially higher prevalence for Profile 3, whereas GPT-o3 and GPT-5.6 Sol estimated higher prevalences for Profile 2. Profile 1 had a lower estimated prevalence in all three LLM solutions (6–10%) than in the human solution (44%). The prevalences reported in Table 5 represent the model-estimated profile proportions, π_k , and there-

fore need not equal the proportions obtained by classifying each transcript according to its largest posterior profile probability.

Transcript-level profile classification agreement with the human ratings was 29.1% for GPT-4o, 39.0% for GPT-o3, and 38.5% for GPT-5.6 Sol. In contrast, GPT-o3 and GPT-5.6 Sol agreed with each other on 77.9% of transcript classifications (Table 6).

Figure 5 shows substantial differences in transcript-level profile classifications between the human- and LLM-derived solutions. Among the human-derived classifications, 93 transcripts were classified as Profile 1. In contrast, only 3 transcripts were classified as Profile 1 by GPT-4o, 15 by GPT-o3, and 10 by GPT-5.6 Sol. GPT-4o instead classified most transcripts as Profile 3 ($n = 156$), whereas GPT-o3 and GPT-5.6 Sol classified most

Profile	High ($\geq .70$)	Moderate (.30–.69)	Low ($< .30$)	Prev.
Profile 1	Q4 (.93), Q13 (.89), Q14 (.70)	Q15 (.59), Q8 (.37)	Q2 (.25), Q12 (.14), Q11 (.11)	44%
Profile 2	Q4 (.96), Q14 (.86), Q12 (.82), Q13 (.79), Q15 (.74), Q11 (.73)	Q8 (.40)	Q2 (.29)	34%
Profile 3	Q15 (.97), Q13 (.93), Q2 (.93), Q14 (.76)	—	Q11 (.29), Q8 (.26), Q4 (.22), Q12 (.04)	23%

Table 4: Profile-specific item-probability patterns for the three human-derived communication profiles. High, moderate, and low refer to estimated item probabilities of $\geq .70$, $.30$ – $.69$, and $< .30$, respectively. Prevalence represents the model-estimated proportion of transcripts in each profile.

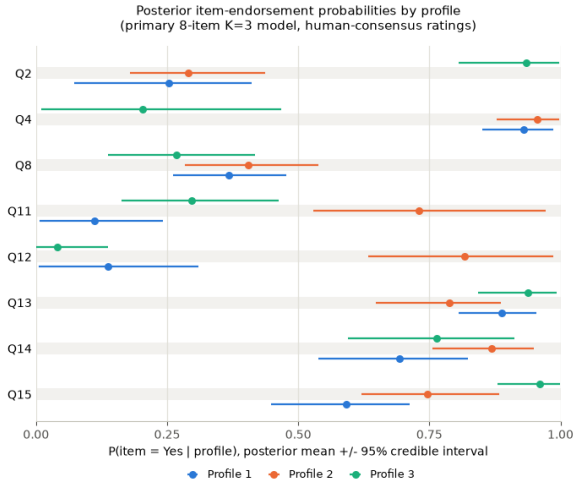


Figure 2: Posterior item-endorsement probabilities for the human-derived $K = 3$ model. Points show posterior means and intervals show 95% credible intervals.

Source	Profile 1	Profile 2	Profile 3
Human	44%	34%	23%
GPT-4o	9%	24%	67%
GPT-o3	10%	55%	35%
GPT-5.6 Sol	6%	65%	29%

Table 5: Model-estimated prevalence of the matched profiles ($N = 213$).

transcripts as Profile 2 ($n = 134$ and $n = 139$, respectively).

5 Discussion

This study examined LLM-based clinical communication assessment at both the item and profile levels. Overall item-level agreement between the LLMs and the final human ratings was generally comparable to agreement among human coders, but disagreement was concentrated in particular behaviors, especially Q2, Q8, Q11, and Q12. Bayesian latent-class analysis identified three distinct and stable communication profiles from the human ratings. The LLM-derived profiles showed overall

Pair	Profile agreement
Human–GPT-4o	29.1%
Human–GPT-o3	39.0%
Human–GPT-5.6 Sol	38.5%
GPT-4o–GPT-o3	54.0%
GPT-4o–GPT-5.6 Sol	51.6%
GPT-o3–GPT-5.6 Sol	77.9%

Table 6: Raw agreement on matched transcript-level profile classifications ($N = 213$).

pattern similarity to the human-derived profiles, but substantial differences remained for specific items. Profile prevalence also differed across human and LLM solutions, and transcript-level profile classification agreement was relatively low. Thus, similarity in the overall profile pattern did not necessarily translate into similarity in individual item probabilities, profile prevalence, or transcript-level classification. Neither GPT-o3 nor GPT-5.6 Sol showed a consistent improvement over GPT-4o across these measures.

These findings have several implications for evaluating LLM-based communication assessment. Overall agreement alone can be misleading because it may hide substantial disagreement on particular communication behaviors. Item-level analyses are therefore needed to identify where human and LLM ratings agree and where they differ.

Evaluation should also match the level at which the assessment will be interpreted. If an LLM is used to score individual checklist items, item-level agreement may be the primary concern. However, if those ratings are used to identify communication profiles or provide profile-based feedback, the resulting profile patterns, prevalences, and transcript-level classifications should also be validated.

Profile-based methods such as Bayesian latent-class analysis can identify recurring combinations of communication behaviors that may not be visible from total scores alone. However, before such

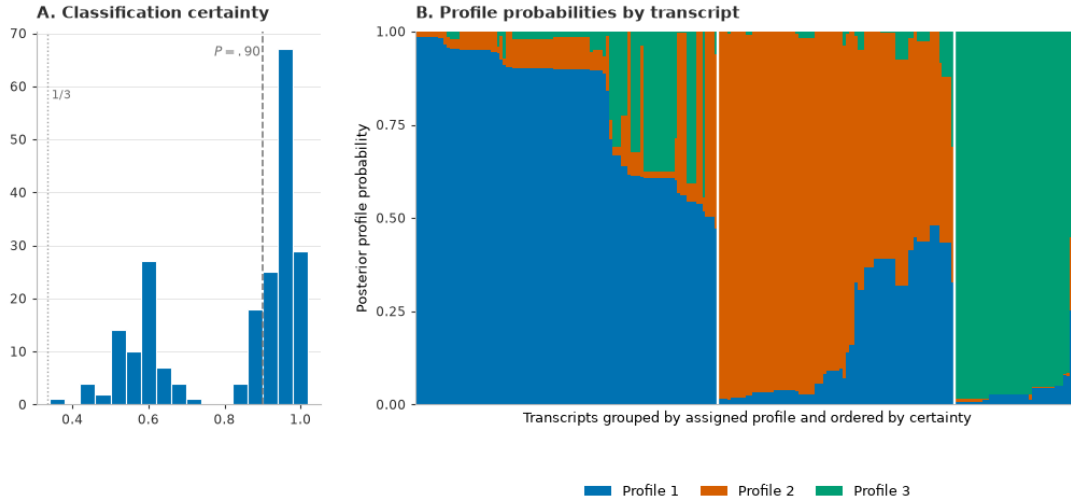


Figure 3: Posterior classification certainty for the human-derived $K = 3$ model ($N = 213$). Panel A shows each transcript’s maximum posterior profile probability. Panel B shows posterior profile probabilities, with transcripts grouped by assigned profile and sorted within each profile by certainty.

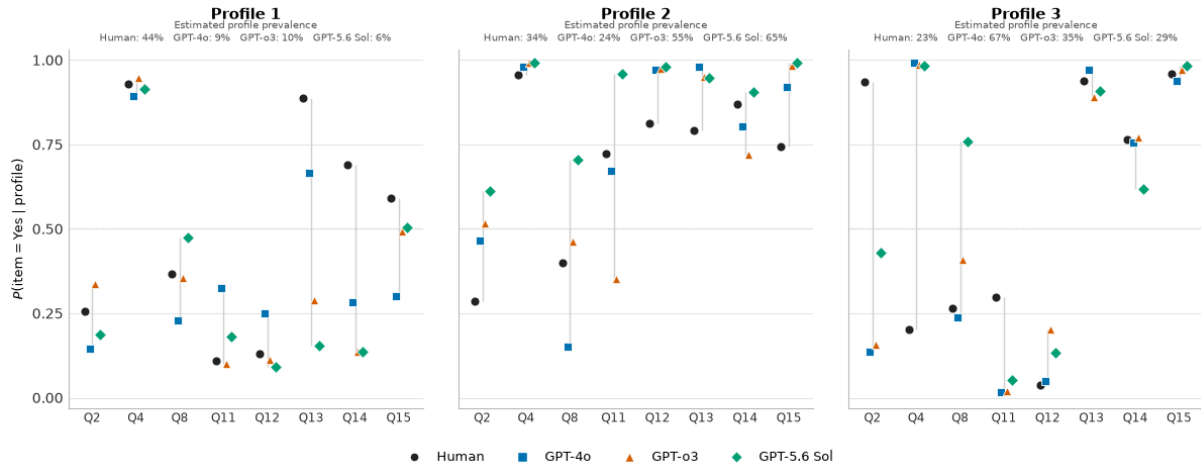


Figure 4: Profile-specific item-probability patterns for the human-derived and LLM-derived profiles after matching profile labels. Similar overall profile patterns do not necessarily imply similar transcript-level classifications.

profiles are used for assessment or feedback, future research should examine whether they are associated with meaningful differences in learner performance, patient outcomes, or the usefulness of feedback.

6 Limitations and Future Work

The study has several limitations. Each LLM was evaluated only once per transcript, so call-to-call variability was not assessed. The binary checklist also produced ceiling and floor effects that limited some items’ usefulness for profile modeling. In addition, we used a relatively simple prompting approach; future work should evaluate more advanced prompt-engineering and calibration strate-

gies. Finally, the data come from a single simulated respiratory dataset, so replication across specialties, populations, and real clinical settings is needed.

7 Conclusion

LLMs showed reasonable overall agreement with human ratings, but overall agreement alone can be misleading because substantial disagreement may remain on particular communication behaviors. Item-level analyses are therefore important for identifying which aspects of the rubric are evaluated similarly by human and LLM raters and which are not.

These differences also matter when item ratings are combined to identify communication profiles.

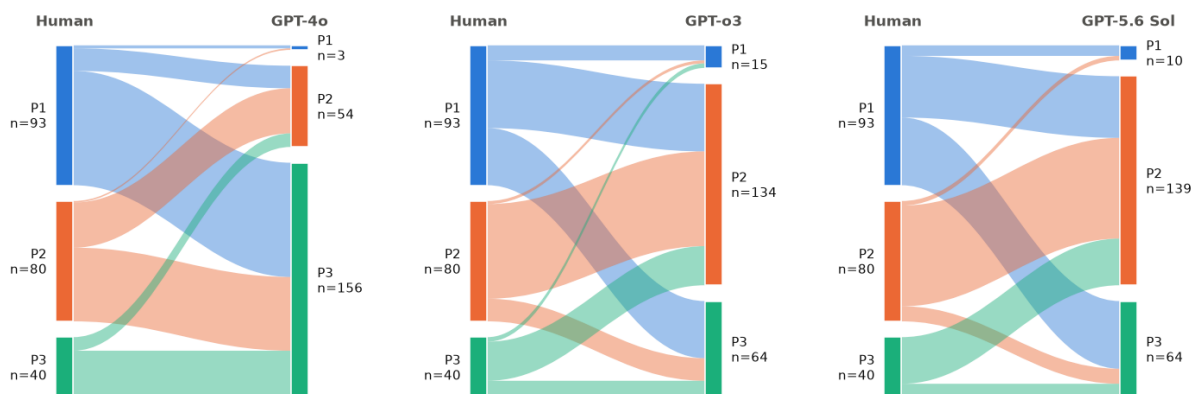


Figure 5: Human-to-LLM profile-classification transitions for GPT-4o (left), GPT-o3 (center), and GPT-5.6 Sol (right), $N = 213$. Ribbon width represents the number of transcripts.

Although the LLM-derived profiles showed similarity in their overall patterns, substantial differences remained for particular item probabilities, and the models often classified transcripts differently from the human-derived solution. Thus, when profile-based assessment or feedback is the intended use, evaluation should consider not only item-level agreement but also whether human and LLM ratings produce similar profile structures and transcript-level classifications.

Acknowledgments

We thank Faria Ali, Mudra Kale, and Sohail Shaikh for their contributions as human coders for this study.

References

- Habibeh Ahmadipour and Nahid Mortazavi. 2015. Medical error and related factors during internship and residency. *Indian J Med Ethics*, 12(4):215–9.
- Tavinder Ark, Adina Kalet, Linda Tewksbury, Lisa Altshuler, Ruth Crowe, Jeffrey Wilhite, Khemraj Hardowar, Sondra Zabar, and Colleen Gillespie. 2024. [Validity evidence for the clinical communication skills assessment tool \(ccsat\) from 9 years of implementation in a high stakes medical student osce](#). *Patient Education and Counseling*, 127:108323.
- Michael T. Brannick, H. Tugba Erol-Korkmaz, and Matthew Prewett. 2011. [A systematic review of the reliability of objective structured clinical examination scores](#). *Medical Education*, 45(12):1181–1189.
- Nilesh Chavda, Priti Solanky, Jatin V. Dhanani, Aashal Shah, Nirav Patel, and Sunilkumar Bhadiyadara. 2020. [Assessment of clinical communication skills of medical students through the simulated patient approach](#). *Journal of Medical Education*, 19(3):e108661.
- A. Angel Dalus, J. Gorshia, and M. P. Brundha. 2024. [Enhancing physician-patient communication: The impact of the kalamazoo consensus statement on medical education](#). *International Journal of Medicine and Public Health*, 14(4):647–651.
- Shih-Chen Fan, Shao-Tong Tsai, Yi-Ching Wang, Meng-Lin Lee, Sheau-Ling Huang, and Ching-Lin Hsieh. 2024. [Reliability of the gap–kalamazoo communication skills assessment form in occupational therapy](#). *British Journal of Occupational Therapy*, 87(7):424–433.
- Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, Thomas Lo, and Christopher W. Smith. 2022. [A dataset of simulated patient-physician medical interviews with a focus on respiratory cases](#). *Scientific Data*, 9(1):313.
- Jadon Geathers, Yann Hicke, Colleen Chan, Niroop Rajashekar, Sarah Young, Justin Sewell, Susannah Cornes, Rene F Kizilcec, and Dennis Shung. 2025. Benchmarking generative ai for scoring medical student interviews in objective structured clinical examinations (osces). In *International Conference on Artificial Intelligence in Education*, pages 231–245. Springer.
- Kilem Li Gwet. 2008. [Computing inter-rater reliability and its variance in the presence of high agreement](#). *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.
- Matthew D. Hoffman and Andrew Gelman. 2014. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623.
- Reyhaneh Hosseinpourkhoshkbari and Richard M. Golden. 2025. [Clinical reasoning assessment methods: Current simulated environment practice and future prospects using ai and psychometrics](#). *SSRN Electronic Journal*.

- Kelly V Johnson, Allison L Scott, and Lisa Franks. 2020. Impact of standardized patients on first semester nursing students self-confidence, satisfaction, and communication in a simulated clinical case. *SAGE open nursing*, 6:2377960820930153.
- J. Richard Landis and Gary G. Koch. 1977. [The measurement of observer agreement for categorical data](#). *Biometrics*, 33(1):159–174.
- Yuelin Li, Jennifer Lord-Bessen, Mariya Shiyko, and Rebecca Loeb. 2018. [Bayesian latent class analysis tutorial](#). *Multivariate Behavioral Research*, 53(3):430–451.
- Gregory Makoul. 2001. Essential elements of communication in medical encounters: the kalamazoo consensus statement. *Academic medicine*, 76(4):390–393.
- Brian Mavis, Jane Turner, Kathryn Lovell, and Dianne Wagner. 2006. Developments: faculty, students, and actors as standardized patients: expanding opportunities for performance assessment. *Teaching and learning in medicine*, 18(2):130–136.
- HyeRin Roh, Kyung Hye Park, and Song Yi Park. 2017. [Verbal communication of students with high patient–physician interaction scores in a clinical performance examination assessed by standardized patients](#). *Korean Journal of Medical Education*, 29(4):241–251.
- John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. 2016. [Probabilistic programming in python using pymc3](#). *PeerJ Computer Science*, 2:e55.
- Ameer Hamza Shakur, Michael J Holcomb, David Hein, Shinyoung Kang, Thomas O Dalton, Krystle K Campbell, Daniel J Scott, and Andrew R Jamieson. 2024. Large language models for medical osce assessment: a novel approach to transcript analysis. *arXiv preprint arXiv:2410.12858*.
- Hardeep Singh, Traber Davis Giardina, Ashley ND Meyer, Samuel N Forjuoh, Michael D Reis, and Eric J Thomas. 2013. Types and origins of diagnostic errors in primary care settings. *JAMA internal medicine*, 173(6):418–425.
- Michelle Yoon and Veronica Michaelsen. 2015. [Critical synthesis package: The kalamazoo consensus statement assessment tools](#). *MedEdPORTAL*, 11:10098.

A Annotation Prompt

LLM Communication-Rating Prompt

Role. You are an expert medical communication rater.

Your task is to read a clinical interview transcript and evaluate **only the doctor's communication behavior** using the coding rules below.

Important instructions

- Doctor utterances begin with D: .
- Patient utterances begin with P: .
- Score **only** the doctor's behavior.
- Use the patient's responses only as context.
- Do not infer behaviors that are not supported by the transcript.
- Return binary scores only:
 - 1 = YES, behavior is present
 - 0 = NO, behavior is absent
- Apply the coding rules exactly as written below.

Q1. Did the doctor greet the patient?

Score 1 if the doctor acknowledges or greets the patient before beginning health-related questions.

Examples:

- "Hello"
- "Hi"
- "How are you?"

Otherwise, score 0.

Q2. Did the doctor use words that show care and concern throughout the interview?

Score 1 if the doctor demonstrates compassion, empathy, care, or concern beyond routine medical questioning.

This may include responding compassionately when the patient discusses:

- A traumatic event
- Death of a loved one
- Emotional distress
- Fear or worry
- Other difficult personal experiences

Otherwise, score 0.

Q3. Did the doctor allow the patient to complete their opening statement without interruption?

Because the transcripts are text-based and interruptions cannot be reliably assessed, always score:

$$Q3 = 1$$

Q4. Did the doctor ask "Is there anything else?" or otherwise elicit a full set of concerns?

Score 1 if the doctor asks sufficiently broad and comprehensive background or follow-up questions to obtain a full picture of the patient's problem.

Examples include asking:

- What makes the symptom better or worse?
- Whether other symptoms occurred before or with the main symptom
- About associated symptoms
- About symptom progression
- Other relevant medical questions beyond the main symptom itself

For example, for a patient with a cough:

- “Does anything relieve the cough?”
- “Does anything make the cough worse?”
- “Did another symptom appear before the cough?”

Otherwise, score 0.

Q5. Did the doctor explain and/or negotiate an agenda for the visit?

Score 1 if, near the beginning of the interview, the doctor:

- Explains the purpose of the visit
- Explains what the patient is being seen for
- Describes what will be discussed
- Explains how the appointment may proceed

Otherwise, score 0.

Q6. Did the doctor begin with the patient’s story using open-ended questions?

Score 1 if the doctor begins by inviting the patient to describe the problem or reason for the visit in their own words.

Examples:

- “What brings you in today?”
- “Tell me about your symptoms.”
- “Can you tell me what has been going on?”

Otherwise, score 0.

Q7. Did the doctor clarify details as necessary using more specific questions?

Score 1 if, after the patient describes the main problem, the doctor asks more specific questions to clarify symptoms, history, risk factors, or related details.

Examples:

- “Is anyone around you sick?”
- “Have you been traveling?”
- “Have you lost or gained weight?”
- Questions about symptom characteristics or associated symptoms

Otherwise, score 0.

Q8. Did the doctor summarize information and give the patient an opportunity to correct or add information?

Score 1 if the doctor summarizes or restates information provided by the patient, particularly toward the end of the interview. The summary should reflect the patient’s reason for the visit, symptoms, or information previously discussed.

Otherwise, score 0.

Q9. Did the doctor transition effectively to additional questions?

Score 1 if the doctor’s questions follow a logical and organized sequence.

The doctor should generally finish exploring one symptom or topic before moving to another rather than jumping abruptly between unrelated topics.

Example: For a patient with headache, the doctor may first complete headache-related questions and then move to related vision or other neurological symptoms.

Otherwise, score 0.

Q10. Did the doctor ask about life events, circumstances, or other people that might affect the patient’s health?

Score 1 if the doctor asks about relevant contextual factors that may affect the patient’s health.

Examples:

- Travel
- Sick contacts
- Stress
- Family medical history
- Social circumstances
- Environmental exposures

- Other relevant life events

Otherwise, score 0.

Q11. Did the doctor elicit the patient's beliefs, concerns, and expectations about the illness and treatment?

Score 1 if the doctor identifies, explores, or meaningfully responds to the patient's beliefs, concerns, fears, expectations, or worries about the illness or treatment.

Example: If a patient says they are afraid to take a medication because it causes brain fog, score 1 if the doctor acknowledges and addresses that concern with explanation, discussion, or empathy.

Do not score 1 merely because the doctor provides medical information. There must be evidence that the patient's own belief, concern, expectation, or worry was addressed.

Otherwise, score 0.

Q12. Did the doctor respond explicitly to the patient's statements about ideas, feelings, or values?

Score 1 if the patient expresses an idea, belief, feeling, fear, or personal value and the doctor directly responds to it.

Examples of appropriate responses include:

- Explanation
- Reassurance
- Empathy
- Discussion of the patient's concern

Example: If the patient says they are worried their symptoms may indicate cancer, score 1 if the doctor directly responds to that concern with explanation and/or empathy.

Otherwise, score 0.

Q13. Did the doctor explain using words that are easy for the patient to understand?

Score 1 if the doctor uses clear, everyday, patient-friendly language when explaining:

- Symptoms
- Diagnoses
- Tests
- Procedures
- Treatment
- Next steps

Score 0 if the explanation relies mainly on unexplained medical jargon.

If the doctor does not provide any meaningful explanation, score 0.

Q14. Did the doctor ask whether the patient had any questions, concerns, or other issues?

Score 1 if, usually near the end of the interview, the doctor explicitly asks whether the patient:

- Has any questions
- Has any concerns
- Wants to add anything
- Has another issue to discuss
- Needs clarification

Otherwise, score 0.

Q15. Did the doctor acknowledge the patient and appropriately close the interview?

Score 1 if the doctor ends the interview appropriately rather than stopping abruptly.

An appropriate close may include:

- A conclusion
- Explanation of next steps
- Follow-up instructions
- Safety instructions
- A closing or acknowledgement of the patient

Otherwise, score 0.