Automated Scoring of Communication Skills in Physician-Patient Interaction: Balancing Performance and Scalability

Saed Rezayi[1], Le An Ha[2], Yiyun Zhou[1], Andrew Houriet[1], Angelo D'Addario[1], Peter Baldwin[1], Polina Harik[1], Ann King[1], and Victoria Yaneva[1]

[1]National Board of Medical Examiners, Philadelphia, USA
{srezayidemne, yyzhou, ahouriet, adaddario, pbaldwin, pharik, aking, vyaneva}@nbme.org
[2]Ho Chi Minh City University of Foreign Languages, Vietnam
anhl@huflit.edu.vn

Abstract

This paper presents an automated scoring approach for a formative assessment tool aimed at helping learner physicians enhance their communication skills through simulated patient interactions. The system evaluates transcribed learner responses by detecting key communicative behaviors, such as acknowledgment, empathy, and clarity. Built on an adapted version of the ACTA scoring framework, the model achieves a mean binary F1 score of 0.94 across 8 clinical scenarios. A central contribution of this work is the investigation of how to balance scoring accuracy with scalability. We demonstrate that synthetic training data offers a promising path toward reducing reliance on large, annotated datasets—making automated scoring more accurate and scalable.

1 Introduction

The ability to automatically evaluate free-text responses has become one of the most impactful applications of natural language processing (NLP) in education. Traditionally, research in this area has focused on automated short-answer grading (ASAG) (Haller et al., 2022; Suen et al., 2023; Clauser et al., 2024) and essay scoring (Klebanov and Madnani, 2022). Recently, the scope has expanded to scoring clinical patient notes written by medical students, which involves determining whether critical medical concepts outlined in a scoring rubric are accurately addressed (Sarker et al., 2019; Harik et al., 2023; Yaneva et al., 2024). While traditional ASAG approaches focus on evaluating factual correctness or content coverage in student responses, our work extends this paradigm to assess the quality of interpersonal communication skills—a domain where responses are more nuanced and context-dependent than typical short-answer assessments.

In this paper, we further extend NLP-based scoring in medical education by introducing a new task: automated scoring of communication skills in a learning tool for physician-patient interactions. Our work is part of the Communication Learning Assessment (CLA) framework (White et al., 2024), a structured educational program that helps medical learners practice communication skills through realistic patient interactions. In a typical CLA scenario, learners watch a brief video of a patient expressing concerns, seeking clarification about their diagnosis, or struggling with treatment adherence, among other examples. They then respond verbally to that scenario (their response can be up to one minute long), aiming to demonstrate key communication behaviors pertinent to the situation. In CLA, these expected behaviors are called learning points (LPs). For example, the LP "Praise patient's weight loss efforts" might be demonstrated by a learner saying, "I'm really proud of you for sticking with it." Evaluating these responses involves identifying specific spans of speech from the learner response that align with LPs from the scoring rubric of the scenario.

The primary contribution of this work is threefold: (1) we investigate the application of automated approaches for scoring communication skills; (2) we evaluate various techniques aimed at improving model performance; and (3) we consider the scalability of these techniques in practical deployment scenarios. While strategies such as increasing the volume of human-annotated data can enhance performance, they are inherently limited by resource constraints and thus do not support scalable solutions. To address this, we focus on approaches such as data augmentation, few-shot learning, and automated generation of training data — methods that hold promise for improving model performance while maintaining scalability.

2 Dataset and Annotation

The dataset used in this study consists of transcribed learner responses collected from simu-

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 891–897

Case ID	Total	#Positive	#Negative	#LPs
174	162	91	71	3
175	120	71	49	2
176	162	80	82	3
177	236	164	72	4
178	138	55	82	3
180	165	99	66	3
182	232	171	61	4
192	236	134	102	4

Table 1: Summary statistics per case. *#Positive* refers to the number of responses that reflect a learning point (LP). *#Negative* refers to the number of responses constructed without any such reflection (i.e., the learner did not address the LP). *#LPs* denotes the number of distinct LPs associated with each case.

lated physician-patient interactions across 8 clinical cases (see Table 1). Each case contains between 120 and 236 learner responses. The learners were 3rd and 4th year US medical students who passed the USMLE® Step 1 exam[1]. Recruitment was carried out by NBME.

Annotations were guided by a detailed rubric capturing key communication behaviors essential for effective physician-patient interactions, such as acknowledgment of patient concerns, provision of clear explanations, demonstration of empathy, and reinforcement of positive behaviors. This rubric included 26 unique Learning Points (LPs), each associated exclusively with one of the 8 clinical cases, with each case containing between 2 and 4 distinct LPs. Annotators were instructed to precisely identify reflective text spans corresponding to each LP by providing exact character-level indices within learner responses. Negative samples for each LP were systematically derived by listing all learner responses from the same clinical case that were not annotated as reflecting that specific LP, ensuring comprehensive negative examples.

Annotations were performed by NBME staff members who were trained domain experts in clinical communication. For each case, 60 responses were randomly selected for annotation development. Initially, five responses per case were independently annotated by three senior and two junior annotators, producing a total of 25 annotations. Annotators then discussed these annotations to resolve disagreements and establish consensus. Following this consensus-building step, annotators independently annotated seven additional responses each, resulting in 35 annotated responses per case. Finally, a senior annotator reviewed and finalized annotations for all 60 responses per case to ensure consistency and annotation quality. On average, each LP received approximately 45 annotations, with the exact count ranging from 20 to 80 per LP. Overall, approximately 60% of annotations explicitly reflected the targeted LPs. Table 1 summarizes detailed annotation statistics by clinical case.

[1]A high-stakes US medical licensure exam, https://www.usmle.org/

3 Model Adaptation and Training

We base our automated scoring on ACTA (Yaneva et al., 2024), which uses a DeBERTa-large architecture as a sequence-level classifier for identifying exact spans that reflect targeted Learning Points (LPs) in learner responses. Instead of predicting per-token labels, ACTA is trained to output the character-level start and end of the span corresponding to the LP, given the response and LP description as inputs. Training uses cross-entropy loss over all possible spans.

The original LP descriptions in our rubric were intentionally concise for human annotators (e.g., "Risks of MRI"), but this brevity posed challenges for ACTA's sequence classification architecture, which relies on semantic relationships between LP descriptions and response text. Terse descriptions lack the contextual cues necessary for distinguishing between superficially similar content and actual demonstrations of the targeted behavior. To address this limitation, we expanded LP descriptions to include specific behavioral indicators. For example, "Risks of MRI" became "Risks of MRI: Avoid unnecessary, costly, and risky tests," providing explicit guidance about the communication goal and enabling DeBERTa's attention mechanism to better identify relevant response segments.

We explored two approaches for generating these expanded LP descriptions:

- **ACTA-M (Manual Summaries)**: Domain experts manually created enhanced descriptions for LPs with fewer than 20 positive annotations, incorporating clinical expertise to capture nuanced communication behaviors.
- **ACTA-A (Automated Summaries)**: We used Qwen2.5-32B-instruct (4-bit) (Bai et al., 2024) to automatically generate augmented LP descriptions by synthesizing patterns from aggregated positive annotations, providing a scalable alternative to manual expansion.

	One model per case			One model for all cases			LLM scoring	
Case ID	Original LPs	ACTA-A	ACTA-M	Original LPs	ACTA-A	ACTA-M	Qwen	GPT
174	0.905	0.896	0.899	0.894	0.915	0.917	0.835	0.858
175	0.949	0.966	0.949	0.917	0.966	0.966	0.912	0.931
176	0.861	0.865	0.883	0.897	0.886	0.893	0.890	0.849
177	0.927	0.944	0.943	0.930	0.936	0.953	0.936	0.915
178	0.883	0.930	0.930	0.930	0.930	0.930	0.848	0.852
180	0.928	0.939	0.955	0.933	0.956	0.934	0.974	0.942
182	0.976	0.928	0.969	0.983	0.972	0.976	0.931	0.880
192	0.931	0.948	0.934	0.945	0.948	0.943	0.922	0.820
Average	0.920	**0.927**	**0.933**	0.929	0.938	**0.939**	0.906	0.881

Table 2: Comparison of binary F1 scores for ACTA with original and augmented learning point descriptions (ACTA-A and ACTA-M), and LLM-based scoring (i.e., a few-shot approach).

For evaluation, we employed 5-fold cross-validation at the case level, distributing the 8 clinical cases such that each fold used 6-7 cases for training and 1-2 cases for testing. This ensures the model is evaluated on entirely unseen clinical scenarios. We fine-tuned[2] DeBERTa-large on each fold's training data. For automated summaries (ACTA-A), descriptions were generated separately for each fold using only that fold's training annotations to prevent information leakage.

In addition to these two augmented versions of ACTA, we evaluated two other methods:

- **LLM scoring:** a few-shot scoring approach using large language models (LLMs) to detect learning points directly from learner responses.To evaluate whether few-shot classification could serve as an effective alternative or complement to ACTA without fine-tuning, we experimented with two large language models: Qwen2.5-32B-instruct (4-bit) and GPT-4o (OpenAI, 2024). Qwen was selected due to its instruction-tuning and demonstrated effectiveness in similar instructional tasks. GPT-4o was chosen based on its advanced reasoning capabilities and broad applicability to instructional scenarios (Brown et al., 2020; Wei et al., 2022). These choices align with established best practices in leveraging instruction-tuned language models for few-shot classification tasks. For each Learning Point (LP), the models were prompted with detailed task instructions alongside five positive and five negative examples, following a structured few-shot format designed to encourage consistent performance.
- **Synthetic responses**: To investigate whether synthetic data can effectively address lower scoring accuracy for Learning Points (LPs) with limited annotations, we supplemented our dataset using synthetic learner responses generated by the Qwen2.5-32B-instruct model. For each LP with sparse annotations, we prompted the LLM with task-specific instructions, definitions of the target LP, and selected positive examples from real learner data. The model then generated realistic synthetic responses explicitly demonstrating the targeted LP. This synthetic dataset augmentation enabled us to expand training data without the cost and time constraints of additional student data collection or manual annotation.

Model evaluation was performed using binary F1 score, measuring accuracy in detecting the presence or absence of communication behaviors.

4 Results

Table 2 summarizes the performance of ACTA using the original learning points compared with augmented descriptions (ACTA-A and ACTA-M) using five-fold cross-validation, as well as results from the few-shot approach using LLM scoring and the use of synthetic responses. Manual summaries (ACTA-M) achieved the highest average binary F1 scores (0.933 for the one-model-per-case setting and 0.939 for the one-model-for-all-cases setting), highlighting the moderate effectiveness of human-crafted augmentation. The automated augmentation approach (ACTA-A) also yielded moderate improvements, indicating its potential as a scalable alternative. LLM-based scoring alone did not improve performance, but it produced results

[2]epochs=10, batch_size=8, learning_rate=2e-5, max_length=256

that were nearly comparable to ACTA (e.g., 0.90 vs. 0.933). We note that this was achieved without the need for extensive data collection or human annotation, aside from the need for annotated data for evaluation purposes.

Finally, the use of synthetic responses led to substantial improvements in scalability. Table 3 shows that training ACTA solely on synthetic data (50 generated examples per case-LP pair) provided only moderate performance gains compared to a simple baseline (0.757 vs. 0.723 average binary F1). However, combining a small amount of real annotated data (15 encounters per case) with synthetic responses significantly improved results (0.878 average binary F1), clearly outperforming models trained only on limited real data (0.793 average binary F1). These results indicate that synthetic responses can effectively reduce the need for human annotation without sacrificing performance.

Performance varied across clinical cases, suggesting the benefit of tailoring augmentation strategies to specific learning point characteristics. We also note the comparability of results from using one model for all cases compared to using one model per case.

5 Error Analysis

We conducted an error analysis to understand the limitations and inform future scoring improvements. The analysis focuses on three perspectives: (1) the relationship between annotation quantity and model performance, (2) specific learning points with low model performance, and (3) cases that showed consistently low performance.

First, we observe a clear relationship between the number of positive annotations per LP and binary F1 scores (see Figure 1). LPs with more than 30 annotations generally achieve binary F1 scores above 0.87, indicating that sufficient annotation quantity is critical for model performance. This threshold is empirically derived by examining the distribution in Figure 1, where performance plateaus become apparent.

Second, to understand LPs with low binary F1 scores (< 0.85), we perform both quantitative and qualitative analyses. A systematic human review is conducted where three annotators independently examine 37 false negatives (instances where ACTA failed to identify an originally annotated LP) and 81 false positives. Among the 35 false negatives with complete reviews, only 51.4% were confirmed

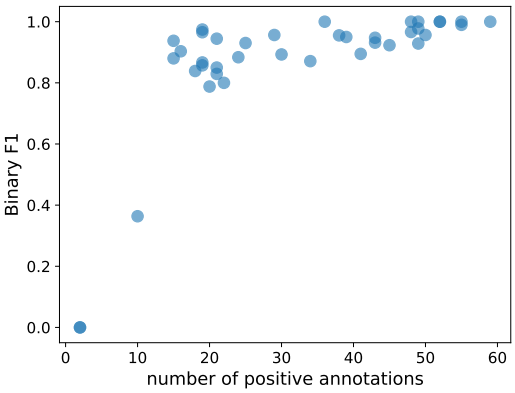

Figure 1: Relationship between the number of positive annotations per learning point and binary F1 score. LPs with more annotations generally achieve higher binary F1 scores. This indicates that sufficient annotations are important for accurate automated scoring.

as true model errors by majority vote, while 48.6% were retrospectively deemed correct predictions by ACTA. This finding reveals that nearly half of apparent model "errors" may actually reflect annotation inconsistencies rather than model failures. The analysis identified three primary factors contributing to lower performance: (1) insufficient positive training examples limiting the model's exposure to representative spans, (2) inherently ambiguous LP definitions leading to inconsistent interpretations, and (3) the intrinsic subjectivity in identifying nuanced communication behaviors despite our rigorous annotation process.

Third, cases 174 and 176 demonstrated consistently lower performance across multiple LPs. This pattern suggests these cases contain inherently more challenging communication scenarios or LPs that are particularly difficult to identify consistently. This finding emphasizes the need for targeted annotation efforts and potentially refined LP definitions for such challenging cases.

Overall, the error analysis reveals that identifying physician communication behaviors (PCBs) is highly subjective and complex. While our annotation process included consensus-building and final adjudication by senior annotators, the nuanced nature of communication skills—such as distinguishing between implicit and explicit empathy—introduces unavoidable interpretive variation. These findings underscore the importance of sufficient annotation volume and suggest that enhanced annotation guidelines with stricter quality control would be valuable for future iterations.

Case ID	Baseline	50 Synthetic	5 Real + 50 Synthetic	15 Real	15 Real + 50 Synthetic
174	0.712	0.754	0.770	0.763	**0.828**
175	0.734	0.817	0.855	0.852	**0.905**
176	0.649	0.684	0.810	0.638	**0.861**
177	0.813	0.826	0.830	**0.900**	0.894
178	0.577	0.580	0.732	**0.843**	0.842
180	0.739	0.812	0.848	0.900	**0.909**
182	0.841	0.812	0.893	0.743	**0.926**
192	0.718	0.773	0.823	0.704	**0.858**
Average	0.723	0.757	0.820	0.793	**0.878**

Table 3: Binary F1 scores across different training scenarios. **Baseline** assumes every learning point is present (i.e., full recall). **5/15 Real** uses 5 or 15 real annotated encounters per case. **Synthetic** refers to LLM-generated responses (50 per case-LP pair) created using Qwen32B-4bit. Performance is evaluated on real learner responses.

6 Discussion

This study contributes to the ongoing conversation on improving NLP-driven assessment by examining whether data augmentation, few-shot learning, and synthetic data can mitigate the scalability challenges of manual annotation.

Our experiments yielded mixed results. Neither manual (ACTA-M) nor automated (ACTA-A) data augmentation methods showed substantial improvements over the baseline model. Similarly, the few-shot learning models did not outperform the ACTA model. However, the few-shot approach performed almost comparably to the baseline model without the need for extensive data collection or human annotation, which is a significant advantage in scenarios where resources for annotation are limited. A potential explanation for these findings is the relatively small sample size of cases and annotations used, which may have limited the diversity and complexity of the learning points. Moreover, our baseline model—a DeBERTa-based classifier trained with the available annotated data—had already achieved strong F1 scores, reducing the room for significant improvement via augmentation or alternative training strategies.

A key contribution for improving scalability were the synthetic data experiments. Training ACTA exclusively on synthetic responses (generated using Qwen32B-4bit) provided moderate improvements over a naive baseline, indicating synthetic responses alone may be a viable initial training strategy in low-resource settings. However, combining a relatively small set of human-annotated responses with synthetic data significantly increased model performance (average binary F1 from 0.793 to 0.878), clearly demonstrating that synthetic responses can meaningfully reduce the need for extensive manual annotation. These results suggest that synthetic data is a practical and scalable approach to addressing annotation bottlenecks without sacrificing model accuracy.

Analysis of annotation density (Figure 1) further reaffirmed that performance improves with an increasing number of positive annotations per LP, highlighting the importance of targeted annotation efforts. Additionally, the comparable results from case-specific models and general models suggest that unified modeling strategies may be viable.

7 Limitations

Some limitations of this research stem from the small sample size of annotated responses, and the small number of vignettes. Additionally, resource constraints prevented all responses from being double-rated. While the scoring method remains interpretable by linking LPs to specific phrases in the responses, the neural models used to define phrase boundaries operate as black boxes and require careful evaluation for potential bias.

Although the few-shot LLM-based scoring approach demonstrates promising generalization without explicit fine-tuning, it shows limitations compared to ACTA models. Specifically, few-shot methods heavily depend on prompt quality and the selection of examples provided, making their performance less consistent and potentially sensitive to minor changes in prompt design. Furthermore, few-shot predictions inherently offer lower interpretability than token-level classifiers like ACTA, as LLM decisions emerge directly from prompt conditioning without explicit textual evidence or intermediate classification steps. This reduced transparency can limit their practical usefulness, especially in educational contexts where detailed feedback and justification of model predictions are often necessary. Further research is needed to investigate the

varying levels of risk that the lower explainability of few-shot learning models presents across different assessment domains. These risks can be better understood and mitigated through additional evaluation studies that provide more evidence on how to address potential concerns.

Likewise, the addition of synthetic data for training purposes needs to be carefully evaluated using a high-quality dataset of carefully annotated real learner responses. We note that while synthetic data can meaningfully reduce the need for data collection and human annotation, it cannot fully replace that need as such data will always be needed for a robust evaluation of any scoring system.

8 Ethical Considerations

Like many other products, automated scoring tools function as socio-technical systems, where their impact depends not only on their technical capabilities but also on how they are used and how their outputs are interpreted. Below, we outline specific aspects of the use of this system in different contexts that merit discussion of ethical implications.

In a summative assessment context, the models outlined here are designed as hybrid systems, ensuring that responses from examinees who are borderline or below the passing threshold are always reviewed by human raters. In a formative setting, it is essential to closely analyze how the system's implementation affects learning outcomes, offering critical validity evidence. This includes determining whether automated feedback aids or obstructs skill development, how examinees engage with the feedback, and whether the reliance on automated scoring impacts learning strategies over time. In the case of formative assessment, which is the primary purpose of the CLA tool, a possible negative consequence could also be "washback"—a focus on developing only the skills directly addressed by the tool. It is also crucial to evaluate whether specific learner groups benefit more than others and to identify any unintended effects, such as overdependence on the system or reinforcement of existing biases. A comprehensive exploration of these factors will help ensure that automated scoring systems function as valuable educational tools, rather than mechanical evaluation devices.

The scores provided by the automated scoring engine are currently in their raw form and have not yet been converted into feedback for students or faculty. This transformation into actionable feedback is a crucial step because raw scores alone do not provide the necessary context or guidance for improving performance. For students, feedback is essential to understand their strengths and weaknesses, guiding them on how to improve and which areas to focus on. Without clear, specific feedback, students may struggle to make meaningful improvements, as they may not fully understand the implications of their scores or how to address their performance gaps. For faculty, the feedback generated from the automated scores can provide valuable insights into student progress, helping instructors identify areas where students may be struggling, and informing instructional adjustments. This step also allows faculty to engage with the results in a more meaningful way, facilitating a deeper understanding of the learning process and ensuring that the assessment tools align with educational goals. Therefore, transforming raw scores into detailed, constructive feedback is vital to ensure that the automated scoring system contributes effectively to the learning process and supports both student development and instructional improvements.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2024. Qwen2.5: The next generation of qwen large language models. *arXiv preprint arXiv:2407.10671*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Brian E Clauser, Victoria Yaneva, Peter Baldwin, Le An Ha, and Janet Mee. 2024. Automated scoring of short-answer questions: A progress report. *Applied Measurement in Education*, 37(3):209–224.

Stefan Haller, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. 2022. Survey on automated short answer grading with deep learning: from word embeddings to transformers. *arXiv preprint arXiv:2204.03503*.

Polina Harik, Janet Mee, Christopher Runyon, and Brian E Clauser. 2023. Assessment of clinical skills: a case study in constructing an nlp-based scoring

system for patient notes. In *Advancing Natural Language Processing in Educational Assessment*, pages 58–73. Routledge.

Beata Beigman Klebanov and Nitin Madnani. 2022. *Automated essay scoring*. Springer Nature.

OpenAI. 2024. Gpt-4o system card. https://arxiv.org/abs/2410.21276. Accessed: [your access date].

Abeed Sarker, Ari Z Klein, Janet Mee, Polina Harik, and Graciela Gonzalez-Hernandez. 2019. An interpretable natural language processing system for written medical examination assessment. *Journal of biomedical informatics*, 98:103268.

King Yiu Suen, Victoria Yaneva, Janet Mee, Yiyun Zhou, Polina Harik, et al. 2023. Acta: Short-answer grading in high-stakes medical exams. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 443–447.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Andrew A White, Ann M King, Angelo E D'Addario, Karen Berg Brigham, Joel M Bradley, Thomas H Gallagher, and Kathleen M Mazor. 2024. Crowdsourced feedback to improve resident physician error disclosure skills: A randomized clinical trial. *JAMA Network Open*, 7(8):e2425923–e2425923.

Victoria Yaneva, King Yiu Suen, Janet Mee, Milton Quranda, Polina Harik, et al. 2024. Automated scoring of clinical patient notes: Findings from the kaggle competition and their translation into practice. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 87–98.