

Finding Evidence of Communication Behaviors using LLMs

Tazin Afrin¹, Saed Rezayi¹, Janet Mee¹, Polina Harik¹, Le An Ha²

¹National Board of Medical Examiners, Philadelphia, USA
{tafrin, srezayidemne, jmee, pharik}@nbme.org

²Ho Chi Minh City University of Foreign Languages, Vietnam
anh1@hufliit.edu.vn

Abstract

This study compares two automated scoring approaches to identify communication behaviors in physician responses to patient questions: prompt-based scoring with Large Language Models (LLM) and supervised transformer-based models. Our results show that although transformer-based models provide more consistent performance, competitive LLM performance is promising in reducing the need for extensive annotated training data.

1 Background

Automated scoring of text-based responses has been used to improve the efficiency, scalability, and reliability of assessments for a variety of purposes (Shermis and Burstein, 2013). Historically, operational scoring systems have relied heavily on supervised machine learning models trained on human-scored responses. Encoder-based transformer models such as BERT, RoBERTa, and DeBERTa have demonstrated strong performance across scoring and classification tasks by leveraging large-scale pretraining followed by fine-tuning on human-annotated assessment data (Devlin et al., 2019; Liu et al., 2019; He et al., 2020; Ormerod et al., 2021; Suen et al., 2023; Rezayi et al., 2025). However, building these systems usually requires significant investment in annotation, reconciliation, and training (Bejar et al., 2016; Kaldaras and Haudek, 2022).

Generative large language models (LLMs) have introduced a new alternative to automated scoring. Models such as GPT, Claude, Gemini and Llama are trained primarily through autoregressive next-token prediction on massive text corpora rather than through task-specific supervised training for scoring or classification (Brown et al., 2020; Gilardi et al., 2023). Unlike encoder-based transformers, generative LLMs learn broad linguistic and contextual knowledge during pretraining and

can often be adapted to new tasks through prompting alone. Recent research suggests that these models can perform complex categorization and scoring tasks with little or no task-specific training data (Wei et al., 2022; Kojima et al., 2022; Gilardi et al., 2023). Their effectiveness has been further enhanced through instruction tuning, reinforcement learning from human feedback (RLHF), and prompting strategies, encouraging explicit reasoning and structured decision-making (Ouyang et al., 2022; Wei et al., 2022).

The present study compares two automated scoring paradigms for identifying evidence of patient-centered communication behaviors (PCBs) in transcribed physician responses to patient questions. For example, a span of text such as "it is quite understandable that you are frustrated by this situation" would be extracted as evidence of a PCB reflecting acknowledgment of a patient's emotions. We examine prompt-based scoring using generative large language models (LLMs), which unlike a supervised transformer-based model does not rely on substantial annotated training data and task-specific model fitting. Beyond comparing scoring performance between the two types of models, the study explores the implications of these approaches for the future development of automated scoring systems.

2 Aims

This study addresses the following research questions:

- How do prompt-based LLMs compare with fine-tuned encoder-based transformer model for PCB span extraction?
- Which modeling approach demonstrates the greatest potential for scalable automated assessment?

3 Sample(s)

The dataset consists of transcribed spoken responses produced by medical students interacting with an online communication learning environment. Students viewed video vignettes of the patient scenario and responded verbally to the prompt "What would you say next?" The student responses were then automatically transcribed.

Each student response can contain multiple examples of patient-centered communication behaviors (PCBs). PCBs are defined through functional communication categories, learning objectives, and behavioral definitions such as acknowledging emotion, explaining treatment, or clarifying instructions. Human-annotated training and evaluation examples included character-level span annotations indicating the portion of the response representing one PCB. The present analysis includes 14 cases representing a range of clinical scenarios. Each case contains between 2 and 6 PCBs and represents a distinct PCB detection scenario with varying contextual and linguistic complexity.

4 Methods

The task was formulated as a span extraction problem. Given a student response transcript and a target PCB definition, models were required to:

- Determine whether the PCB was present in the response.
- Return character-level indices corresponding to the PCB text span.

It is possible that no PCB was identified in that response.

4.1 Models

For the supervised model, a DeBERTa-based transformer model was fine-tuned for PCB span extraction using annotated examples (Rezayi et al., 2025). The model leveraged contextualized token representations to predict PCB-relevant spans within student responses.

For the LLM based approach, GPT-5 was evaluated using a structured few-shot prompting framework. The model was instructed to identify whether the PCB was fulfilled explicitly or implicitly in the student response, and return the extracted text span with character-level indices.

We also used Claude-Sonnet-4.6 using the same prompting framework and examples as GPT-5 to enable direct comparison across LLMs.

The **prompting framework** for both of the LLM models emphasized:

- Contextual grounding using clinical scenarios
- Explicit PCB definitions and examples
- Demonstrations of valid and invalid PCB matches
- Student response to evaluate
- Output formatting requirements

Two **evaluation metrics** were used:

Character-level F1 overlap: Measures span overlap F1-score between predicted and gold annotations at the character level.

Binary F1-score: Measures whether the system correctly identified the presence or absence of the PCB regardless of exact span boundaries.

5 Results

Figure 1 summarizes model performance across the 14 communication cases. The DeBERTa-based model consistently achieved the strongest character-level overlap performance across most cases, with scores generally ranging from 0.81 to 0.91 with an average of 0.86. GPT-5 achieved competitive but slightly lower span overlap performance with average 0.83, while Claude-Sonnet demonstrated substantially lower character-level overlap scores, average 0.60, despite strong binary detection performance. In contrast, binary PCB detection performance was high across all three systems. Claude-Sonnet frequently achieved the highest binary F1 scores despite weaker span localization. These findings suggest that supervised fine-tuning supports precise localization of PCB spans.

GPT-5 demonstrated competitive binary and span-level performance without task-specific fine-tuning approaching DeBERTa performance in several cases. The model appeared more variable across cases, particularly for precise span localization. These results suggest that prompt-based LLMs can approximate supervised span extraction performance under carefully designed prompting conditions.

The findings indicate a divergence between PCB identification and precise span extraction. Encoder-based supervised models appear better suited for fine-grained span localization, whereas frontier LLMs demonstrate strong semantic recognition

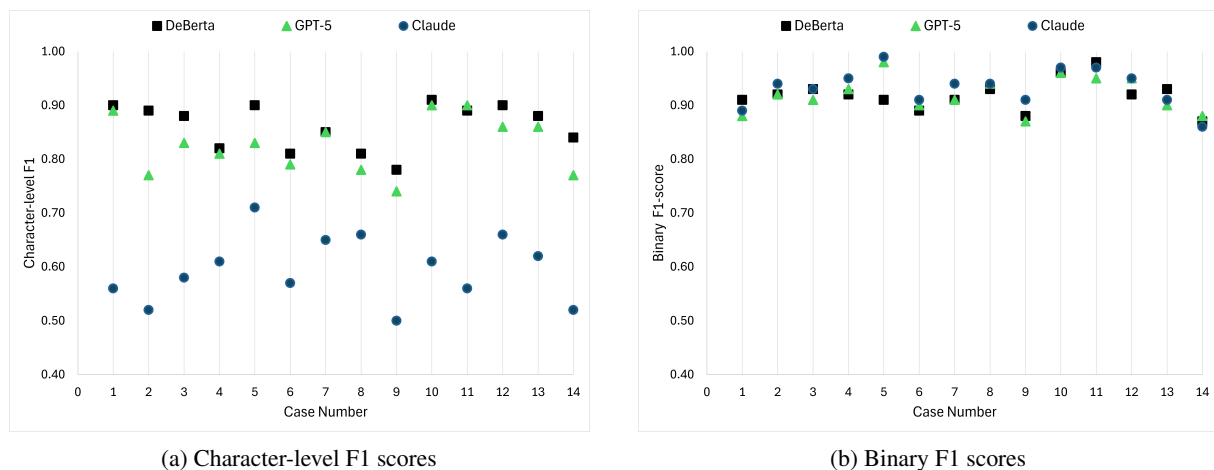


Figure 1: Performance comparison across models.

abilities without fine-tuning. The results also suggest that evaluation metric choice substantially affects conclusions about model quality.

6 Conclusions

This study demonstrates the feasibility of automated patient-centered communication assessment using both supervised transformer models and generative LLMs. The DeBERTa-based supervised model currently provides the strongest and most consistent span localization performance, supporting value of task-specific fine-tuning for precise PCB extraction. On the other hand, GPT-5 achieved competitive performance using only few-shot prompting, suggesting that modern LLMs may reduce the need for extensive supervised training data in communication assessment tasks. Claude-Sonnet’s stronger binary detection but weaker span finding highlights the importance of distinguishing between semantic recognition vs span extraction performance.

Future work will include larger-scale evaluation across individual PCB categories, error analysis of binary versus span overlap, human-model agreement analysis, and finally prompt optimization studies. Overall, this work contributes to ongoing efforts to develop scalable, automated assessment systems for physician-patient communication training in medical education.

References

Isaac I Bejar, Robert J Mislevy, and Mo Zhang. 2016. Automated scoring with validity in mind. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, pages 226–246.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [Chatgpt outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Leonora Kaldaras and Kevin C Haudek. 2022. Validation of automated scoring for learning progression-aligned next generation science standards performance assessments. In *Frontiers in education*, volume 7, page 968289. Frontiers Media SA.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher M Ormerod, Akanksha Malhotra, and Amir Jafari. 2021. Automated essay scoring using efficient

transformer-based language models. *arXiv preprint arXiv:2102.13136*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Saed Rezayi, LA Ha, Yiyun Zhou, Andrew Houriet, Angelo D’Addario, Peter Baldwin, Polina Harik, and Victoria Yaneva. 2025. Automated scoring of communication skills in physician–patient interaction: Balancing performance and scalability. In *Proc 20th BEA Workshop*.

Mark D Shermis and Jill Burstein. 2013. Handbook of automated essay evaluation. *NY: Routledge*.

King Yiu Suen, Victoria Yaneva, Janet Mee, Yiyun Zhou, Polina Harik, and 1 others. 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, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.