Automated Scoring of Communication Skills

Jolie Moran

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

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

Published 2025

Significance

- Addresses a critical gap in clinical education
- Advances NLP
- Enables scalable assessment
- Promotes model interpretability
- Offers a blueprint for assessing soft skills
- Raises Ethical awareness

Background: Key Concepts

- Automated Educational Assessment with NLP
- Communication Learning Assessment Framework (CLA)
- Learning Points (LPs)
- Challenges in Scoring Communication Skills

Background: Prior Research

- Automated Short-answer grading (ASAG)
  - Haller et al., 2022; Suen et al., 2023; Clauser et al., 2024
- Essay Scoring
  - Klebanov and Madnani, 2022
- Scoring clinical patient notes written by medical students
  - Sarker et al., 2019; Harik et al., 2023; Yaneva et al., 2024

Research Questions

1. Can NLP models accurately and scalaby score physician communication skills?
2. Can learning point descriptions be expanded (manually or automatically) to improve model accuracy?
3. Can synthetic data reduce reliance on human annotation without sacrificing accuracy?

Methodology: Datasets

- 8 Clinical Scenarios
  - Each had 120 to 236 learner responses
    - Manually annotated for specific learning points (LPs)
- ~26 LPs represent communicative behaviors
  - Empathy, summarization, reassurance
- Annotation quality varied
  - Sparse and inconsistent

Methodology: Datasets

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

Methodology: Models

- Automated Communication Training Assessment (ACTA)
  - Uses DeBERTa-large transformer
    - Predicts whether a response satisfies the LP
- Each LP is treated as a separate classification task
  - Allows for fine-grained scoring

Methodology: LP Description Expansion

- ATCA-M (Manual Expansion)
  - Humans experts rewrote LP descriptions to be cleaner and more informative
- ACTA-A (Automated Expansion)
  - Qwen2.5-32B-instruct generated expanded LP descriptions using few-shots promptings

Methodology: LLM Scoring

- GPT-40 and Qwen2.5-32B-instruct
    - Used for few-shot scoring without any fine-tuning
- Prompted with LP definitions and examples to classify new reponses
- Lacked interpretability and consistency compared to ACTA

Methodology: Synthetic Data

- Generated synthetic learner responses using Qwen2.5-32Binstruct
  - 50 responses were created
  - 15 real annotated examples mixed in
- Test whether synthetic data could train models effectively with minimal human input

Findings

- ACTA – M achieved highest average binary F1 (0.939)
- LLM Scoring was less interpretable (0.906)

- Errors in Studies
  - Lack of models used
  - Limited sample size
  - Annotation inconsistencies

Findings

	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

Commentary

- Limit Sample Size
  - Only used 8 scenarios
- Annotation Inconsistencies
  - Humans vs LLMs
- Real-World Relevance
  - Physician-patient interaction

Question 1

What is an advantage of few-shot scoring with LLMs?

A. Provides transparents decision boundaries
B. Eliminates need for humans
C. Guarantees perfect accuracy
D. Allows for use without fine-tuning

Question 2

What challenges did annotation inconsistencies pose?

A. Made LP’s easier to define
B. Reduced the need for synthetic data
C. Led to mislabeling and reduced reliability
D. Decreased the amount of LP’s in the study

Question 3

What is an ethical concern discussed in the paper?

A. Data storage cost
B. How automated scoring affects learning and fairness
C. Hardware limitations
D. Lack of video examples

Question 4

What is one limitation mentioned in the paper?

A. Too much annotated data
B. Overly simple task
C. Lack of model explainability and small dataset size
D. Perfect performance across all cases

Questions?