

Detailed Rubric Prompts Improve LLM Evaluation of Clinical Reasoning Reports in Physical Therapy Education

Kanae Takahashi¹ · Tatsuhiro Yoshida² · Syunya Hatakeyama³ · Hirofumi Hori⁴ · Mitsunori Matsushita¹ · ²Kishiwada Tokushukai Hospital, Osaka · ³Kobe City Medical Center General Hospital, Hyogo · ⁴Konan Women's University, Hyogo

ETHICS / REGISTRATION, IEC Approval No. / Exemption, PROSPERO / CTRI / ClinicalTrials.gov No. : Not applicable

Funding : JSPS KAKENHI 25K15240, JST SPRING JPMJSP2150 (first author).

Conflict of Interest : No conflicts of interest

1 BACKGROUND

- The use of **rubrics** for report evaluation is **recommended**
- AI assessment** of Clinical Reasoning Reports is **desired**

2 AIMS AND OBJECTIVES

AIM 1 : To examine **the correlation** between **LLM** and **human evaluations**

AIM 2 : To examine whether a **detailed rubric improves LLM evaluation**

3 METHODOLOGY

DATA 20 Clinical Reasoning reports (20 physiotherapists with 1 to 5 years' experience)

Assessment method : Rubric (5 criteria × 3 points = 15 points)

Abstract rubric

Original broad rubric items

	Integration	Integration	Interpretation	Interpretation	problem
Level3					
Level2					
Level1					

Detailer rubric

Specific sub-items and scoring criteria

	Integration	Integration	Interpretation	Interpretation	problem
Level3					
Level2					
Level1					

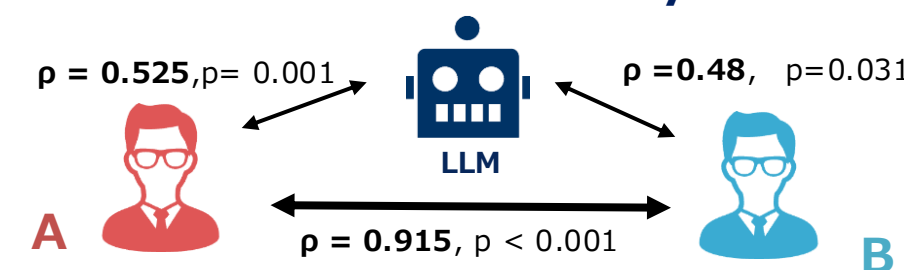
STEP1 Faculty members(A,B) and **LLM**(ChatGPT-5.5) assess the reports by referring **abstract rubric**

STEP2 **LLM only reassess** the reports by referring **detailed rubric**

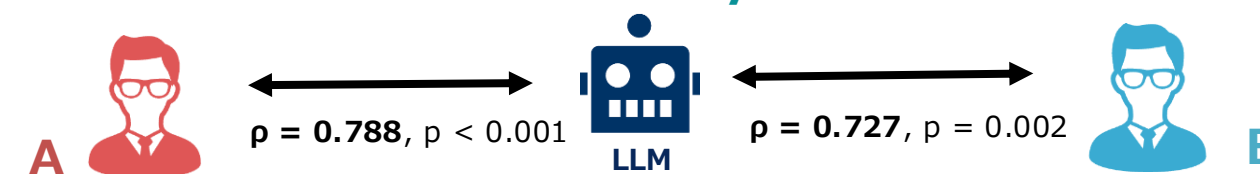
Statistics Spearman's rank correlation among evaluators

4 RESULT

Result1 Moderate positive correlation between faculty members and LLM
Strong positive correlation between faculty memberA and memberB



Result2 Providing more detailed prompts, strong positive correlations between faculty members and LLM



Point! Detailed rubric prompts strengthened correlations between LLM and faculty evaluations.

5 CONCLUSION & IMPLICATIONS

LLMs had difficulty evaluating clinical reasoning using abstract rubric.
Detailed educational rubric prompts strengthened correlations with faculty evaluations.

REFERENCES

Schiött J, Ivegren W, Borg A, Parodis I, Skantzé G. "Using LLMs to Grade Clinical Reasoning for Medical Students in Virtual Patient Dialogues." Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), 2025:750–763

6 DISCUSSION

Can LLMs support clinical reasoning assessment in practice?