How Model Size, Temperature, and Prompt Style Affect LLM-Human Assessment Score Alignment

Julie Jung[1], Max Lu *[1], Sina Chole Benker[2], Dogus Darici[2,3]
[1]Harvard Graduate School of Education, [2]Munster University,
[3]Institute of Anatomy and Neurobiology, University of Münster
* Joint First Authors
Correspondence: maxlu@fas.harvard.edu

Abstract

We examined how model size, temperature, and prompt style affect Large Language Models' (LLMs) alignment within itself, between models, and with human in assessing clinical reasoning skills. Model size emerged as a key factor in LLM-human score alignment. Study highlights the importance of checking alignments across multiple levels.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has introduced new possibilities for evaluating text-based and conversational assessments, offering a scalable and cost-effective alternative and complement to traditional human scoring. While earlier approaches relied on statistical models, recent studies have demonstrated that LLMs can produce accurate, consistent, and personalized evaluations, which are particularly valuable in settings with limited access to expert raters and a need for timely feedback (e.g., Pack et al., 2024; Xiao et al., 2024). In educational contexts, assessments leveraging automated scoring have enabled more frequent formative and summative evaluations by reducing teacher workload and accelerating feedback cycles (Bailey et al., 2025). These developments raise critical questions about how such technologies compare to current assessment practices and what implications they hold.

Before LLM-generated scores can be meaningfully deployed, it is essential to consider the impact of their opaque scoring processes and how different settings influence their behavior. Although LLMs may approximate human ratings at the surface level, there is often a misalignment between the cognitive processes humans rely on and the pattern-based inferences LLMs make (Baldwin et al., 2025). Unlike human raters who may draw on domain-specific reasoning, LLMs operate largely as "black boxes," making it difficult to trace how inputs lead to particular scoring outcomes (Bathaee, 2017). This lack of transparency may be problematic in domains where judgments need to be informed by specific expertise. Additionally, LLMs are not a single, homogeneous system. Different models behave differently, and even the same model can produce varying results depending on its settings. As a result, establishing a strong body of validity evidence is critical before integrating LLMs into assessment practices, particularly as a complement to human scoring.

In this study, we investigate how LLM-generated ratings, with varying model size, temperature, and prompt style, compare to current practices in a specific case study of clinical reasoning skill assessments in medical education. Clinical reasoning, the ability to recognize and interpret a patient's needs and health condition, is an essential skill for all health professionals (Tanner, 2006). It requires problem-solving abilities, experiential knowledge, and deliberate decision-making. Pattern recognition, in particular, is strongly linked to diagnostic success (Coderre et al., 2003), yet many medical students lack the clinical exposure needed to develop this skill. Thus, there is a need to provide students with opportunities to practice encounters with patients and receive structured, targeted feedback to develop their clinical reasoning skills (Haring et al., 2017). However, formative assessment of clinical reasoning is resource-intensive, typically requiring advanced medical experts to review and score student-patient dialogue. While LLMs have been used to simulate patients in such dialogues, further research is needed to determine whether they can be reliably used for scoring, and how different configurations of LLMs affect their scoring behavior and reliability (Brügge et al., 2024). This study explores how LLM-based evaluations align with or diverge from current human-based assessments and what those findings imply for the valid-

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 265–273
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

ity, fairness, and practical use of LLMs in medical education.

2 Literature

LLMs have shown great promise in evaluating complex, language-based assessments. Recent studies comparing various LLMs have demonstrated that some models, such as ChatGPT using GPT4, have high alignment with human scorers for tasks such as essay assessment, although they tend to exhibit a tendency toward inflated scores (Seßler et al., 2025). LLMs are also capable of extracting nuanced insights from text, aligning reasonably well with human ratings in tasks such as discourse coherence analysis (Naismith et al., 2023) and evaluating short-answer assessments in an undergraduate medical program (Morjaria et al., 2024) with similar performance to human expert assessors. However, most studies have focused on scoring outcomes rather than examining the processes through which LLMs generate their ratings. Given the "black box" nature of LLM reasoning (Bathaee, 2017), further research is necessary to better understand how these models arrive at their judgments.

Although directly understanding the internal processes of LLM scoring remains challenging due to their nature, insights can be indirectly gained by systematically exploring how prompt style and model parameters influence scoring outcomes. Morjaria and colleagues (2024) found that the absence of a rubric in the prompt given to the LLM improved their alignments with human scoring, suggesting prompt style significantly shapes how these models evaluate responses. Yet, existing studies have largely overlooked how prompting the LLM to adopt a certain persona, such as a clinical expert, may further improve its alignment with human expert judgments by capturing nuanced reasoning processes underlying expert scoring. Additionally, model parameters, such as temperature settings, influence the randomness of and variability of generated responses and thus may affect scoring consistency and accuracy. Official guide on OpenAI suggests that lower temperature leads to more deterministic and less random outputs, while higher temperature leads to more random outputs (OpenAI, 2023). Many studies have linked the randomness to "creativity" or variability in the response (Peeperkorn et al., 2024). While the specific temperature setting for public-accessible ChatGPT is undisclosed, it is commonly believed to be around 0.7 (University of Victoria Libraries, 2024). Lastly, the choice of model type or size could also substantially influence scoring alignment due to variations in reasoning capabilities (Seßler et al., 2025). As psychometrician Andrew Ho emphasized, it is essential to ask, "Who is using what score for what purpose?" Different types of assessments, tailored to different use cases, demand varying standards for scoring, such as consistency and inter-rater reliability (Ho, 2025). Accordingly, as LLMs are increasingly employed as assessors, it becomes critical to systematically examine how variations in prompts, temperature settings, and model size affect their performance across diverse contexts. Such investigation is a necessary step toward the effective and trustworthy adoption of LLM-generated assessments.

When it comes to using LLMs to evaluate student performance for formative feedback, it is essential to understand how reliable are their ratings—how does LLM ratings compare itself to other LLMs and, importantly, to that of human—and what these comparisons imply for validity and practical use. Clinical reasoning provides a challenging test case in that it encompasses a complex cognitive process that relies on both formal and informal reasoning strategies, deeply informed by domain-specific knowledge (Simmons, 2010). It involves collecting patient information, evaluating its clinical significance, forming inferences, and generating diagnostic hypotheses. To facilitate formative assessment of these skills, the Clinical Reasoning Indicators-History Taking-Scale (CRI-HT-S) was developed, grounded in clinical reasoning indicators identified through qualitative research (Haring et al., 2017). This scale assesses dimensions such as a medical student's capacity to lead patient conversations and ask questions in a logical sequence. An initial validation has shown that the CRI-HT-S demonstrates acceptable reliability and internal consistency for assessing undergraduate medical students' clinical reasoning skills (Fürstenberg et al., 2020). Given the complexity inherent to clinical reasoning, it is particularly important to explore how LLMs evaluate this nuanced and multidimensional construct.

Despite growing interest in leveraging LLMs for formative evaluations, existing research has not systematically investigated how interactions between critical LLM parameters (model size, temperature, and prompt style) influence their scoring alignment with expert human assessors. Our study addresses this gap by comprehensively evaluating the impact of these parameters within clinical reasoning assessments, providing guidance for valid and practical implementation of LLMs in educational contexts. Specifically, we address the following research questions:

RQ1:How consistent are LLM raters when re-rating the same student in the clinical reasoning assessment?

RQ2: How do LLM design parameters (model size, temperature, and prompt style) affect the inter-rater reliability and alignment between LLM raters in the clinical reasoning assessment?

RQ3: How do LLM design parameters (model size, temperature, and prompt style) affect the inter-rater reliability and alignment with human raters in the clinical reasoning assessment?

RQ4: How do LLM design parameters (model size, temperature, and prompt style) and their interactions affect the average score levels in the clinical reasoning assessment compared to human raters?

3 Data and Sample

Our dataset consists of the transcripts of 21 third-year medical students in Germany who were each engaged in conversations with four fictional patients with differing diagnoses to assess the students' clinical reasoning skills. Two human raters with medical expertise rated each transcript with the Clinical Reasoning Indictors-History Taking-Scale, which consists of eight items measuring the quality of the medical student's clinical decision making on a Likert scale from 1 to 5 (see Appendix A). We systematically varied three LLM parameters: model size (GPT-4o as the large model vs. GPT-4o-mini as the small model), temperature (low: 0.2 vs. high: 0.7), and prompt style (regular vs. expert prompt; see Appendix A & B).These variations resulted in eight distinct LLM configurations. Each model rated every student dialogue, producing eight item-level scores per student. All transcripts were in German. One conversation (student 16 with patient 1) was excluded due to missing dialogue. Students were informed that their interactions would be assessed for clinical reasoning.

4 Methods

To address the first research question, each student's responses were rated twice by the LLM raters. Each student received scores on eight items across four rounds of conversation. For each rater, we averaged a student's item scores across all rounds. We then calculated pairwise intraclass correlation coefficients (ICCs) using the icc() function from the irr package in R (Gamer et al., 2019) across all raters. A two-way model was used to appropriately assess absolute agreement across the two trials of the same rater when students are considered random effects. According to the classical test theory, a high ICC suggests that the LLM tends to agree with its own ratings, a sign of reliability but not necessarily validity. However, a low ICC would suggest poor agreement of the same LLM across trials, a sign of a lack of reliability and validity. Conventionally, ICC below 0.5, between 0.5 and 0.75, and above 0.75 suggests poor, moderate, good or excellent reliability, respectively (Koo and Li, 2016).

Rater	Model	Temperature	Prompt Style
LLM1	Small (gpt-4o-mini)	Low (0.2)	Default
LLM2	Small (gpt-4o-mini)	Low (0.2)	Expert Persona
LLM3	Small (gpt-4o-mini)	High (0.7)	Default
LLM4	Small (gpt-4o-mini)	High (0.7)	Expert Persona
LLM5	Large (gpt-4o)	Low (0.2)	Default
LLM6	Large (gpt-4o)	Low (0.2)	Expert Persona
LLM7	Large (gpt-4o)	High (0.7)	Default
LLM8	Large (gpt-4o)	High (0.7)	Expert Persona

Table 1: Color-coded settings of LLM models by Rater ID. Background colors indicate model size and temperature, with red boxes around prompt style when set to "Pretend to be expert."

To answer the second and third research questions, we evaluated interrater agreement between LLMs and across human and LLM ratings. To facilitate pairwise comparisons between LLM and human raters, we also created a composite "averaged human" rater by averaging the two human scores. We used a two-way model to compute the ICC across all pairs of raters (8 LLM raters, 2 human raters, and 1 "averaged human" rater). Comparing ICC values between rater pairs allowed us to assess the degree of alignment between them. A high ICC between LLM and human indicate that the LLM rater may be scoring similarly to humans. Meanwhile, high ICC between LLM raters alone reflects consistency between LLMs but not necessarily alignment with human judgment.

To address the fourth research question, we further aggregated the item-level scores into a single person-level mean score per student per rater. This approach minimizes the influence of within-person variability and allows us to focus on between-student differences. However, this simplification

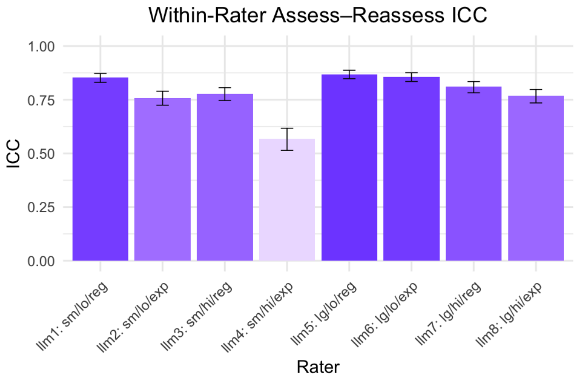


Figure 1: Bar plot of the intraclass correlations (ICCs) between two repeated assessments of the same student for each type of LLM rater. 95% confidence intervals are included.

assumes that the eight items reflect a unidimensional latent construct of performance, which is a limitation of this method.

To examine how different LLM design parameters affect LLM-generated ratings, we constructed a model including main effects and interactions among model size ($large$), temperature setting (hi_temp), and prompt style ($expert$) to reflect the eigh different LLM rater configurations. Each of these factors was coded as a binary indicator. For student i rated by rater j, we fit the following model using heteroskedasticity-consistent standard errors:

$$\begin{aligned} score_{ij} &= \beta_0 + \beta_1 large_j + \beta_2 hi\_temp_j + \beta_3 expert_j \\ &\quad + \beta_4 large_j \cdot hi\_temp_j \\ &\quad + \beta_5 large_j \cdot expert_j \\ &\quad + \beta_6 hi\_temp_j \cdot expert_j \\ &\quad + \beta_7 large_j \cdot hi\_temp_j \cdot expert_j + \varepsilon_{ij} \end{aligned} \tag{1}$$

This model estimates not only the average effects of each design parameter but also their two-way and three-way interaction effects, allowing us to explore whether the impact of one parameter depends on the levels of the others.

5 Results

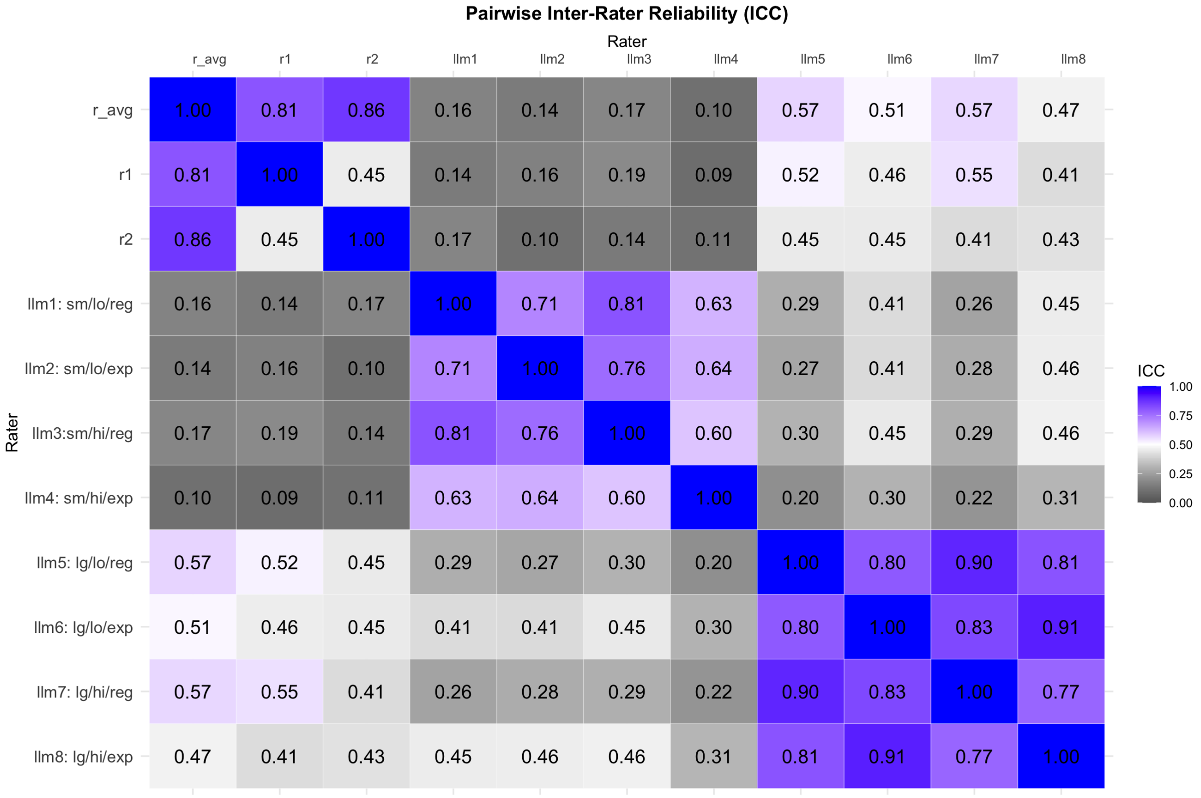
For the within model consistency (RQ1), we compared the test–retest ICCs across LLM raters (see Figure 1). Most models showed high reliability (ICC≈.76–.87). The exception was GPT-4o mini with high temperature and the expert prompt, which exhibited lower consistency. In contrast, GPT-4o with low temperature and the regular prompt achieved the highest reliability (ICC = .87).

For consistency between models and between human and model, figure 2 shows the pairwise intraclass correlation coefficients (ICCs) among all raters. As a reference, the two human raters (r1 and r2) showed moderate agreement (ICC = 0.45). To answer RQ2, which focused on consistency between models, the four small-model LLM raters (llm1–llm4) showed good to very good internal agreement (ICC range: 0.60–0.81), and the large-model LLM raters (llm5–llm8) showed even stronger internal agreement (ICC range: 0.77–0.91).

However, further inspection revealed discrepancies between human and LLM ratings (RQ3). There appeared to be poor agreement between human raters and smaller LLMs (ICC < 0.20). In contrast, there appeared to be moderate agreement between humans and larger LLMs (ICC = 0.41–0.57), comparable to or slightly exceeding human–human agreement. These results indicated that while language models, regardless of size, could be highly agreeable with each other, this does not guarantee alignment with human evaluations. Therefore, if human ratings are considered the benchmark, alignment should be assessed based on direct agreement with human raters, rather than relying solely on internal consistency among the models.

Examining the ICCs across individual items to identify discrepancies between LLM-generated and human scores revealed interesting patterns. For example, item 4, assessing whether the student's questions suggested specific causes of symptoms, showed high consistency among LLM raters but low alignment between LLM-generated and human scores. This suggested that LLMs and human raters may interpret the criterion of "suggesting specific causes" differently, emphasizing the need for improved prompts that more effectively capture human evaluative reasoning (see Appendix C).

For RQ4, we estimated a linear model with robust standard errors to examine how model size, temperature, and prompt style, as well as their interactions, affected the average clinical reasoning scores produced by different LLM raters (see Appendix D). As shown in Figure 3, all LLM raters, regardless of configuration, produced higher average student scores than the human rater average (M = 3.19, SE = 0.054). The estimated intercept represented the expected score under the baseline



	r_avg	r1	r2	llm1	llm2	llm3	llm4	llm5	llm6	llm7	llm8
r_avg	1.00	0.81	0.86	0.16	0.14	0.17	0.10	0.57	0.51	0.57	0.47
r1	0.81	1.00	0.45	0.14	0.16	0.19	0.09	0.52	0.46	0.55	0.41
r2	0.86	0.45	1.00	0.17	0.10	0.14	0.11	0.45	0.45	0.41	0.43
llm1: sm/lo/reg	0.16	0.14	0.17	1.00	0.71	0.81	0.63	0.29	0.41	0.26	0.45
llm2: sm/lo/exp	0.14	0.16	0.10	0.71	1.00	0.76	0.64	0.27	0.41	0.28	0.46
llm3:sm/hi/reg	0.17	0.19	0.14	0.81	0.76	1.00	0.60	0.30	0.45	0.29	0.46
llm4: sm/hi/exp	0.10	0.09	0.11	0.63	0.64	0.60	1.00	0.20	0.30	0.22	0.31
llm5: lg/lo/reg	0.57	0.52	0.45	0.29	0.27	0.30	0.20	1.00	0.80	0.90	0.81
llm6: lg/lo/exp	0.51	0.46	0.45	0.41	0.41	0.45	0.30	0.80	1.00	0.83	0.91
llm7: lg/hi/reg	0.57	0.55	0.41	0.26	0.28	0.29	0.22	0.90	0.83	1.00	0.77
llm8: lg/hi/exp	0.47	0.41	0.43	0.45	0.46	0.46	0.31	0.81	0.91	0.77	1.00

Figure 2: Pairwise intraclass correlation coefficients (ICC) between human and LLM raters across rating conditions.

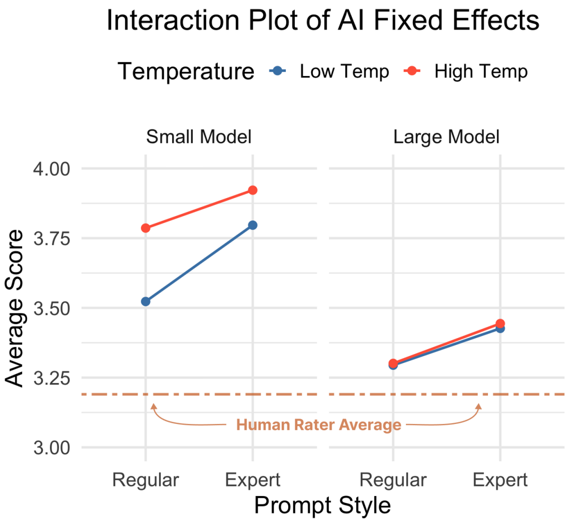


Figure 3: Interaction plot showing how model size, temperature, and prompt style influence AI-generated average scores compared to human rater average.

configuration (small model, low temperature, regular prompt) and was significantly above the human rater average (β = 3.52, p < .001), indicating that this model configuration was less stringent than the human raters.

Both high temperature and expert prompting were associated with significantly higher scores overall (β = 0.26 and β = 0.27, respectively; both p < .01). The interaction between model size and temperature was negative and statistically significant (β = –0.26, p < .05), suggesting that higher temperature inflated the score more for the small model than the large one, holding prompt style constant. On the other hand, the interaction between model size and expert prompting was nonsignificant (β = –0.14, p > .1), indicating that the effect of expert prompting on average scores did not significantly differ between the large and small models, holding temperature constant. The interaction among all three factors was also nonsignificant (β = 0.15, p > .1), suggesting that the combined effect of temperature and prompt style did not differ significantly between large and small models.

For the small model, expert prompt and high temperature each led to increases in average scores, up to almost 3.9. In contrast, the large model showed a smaller range of scores across conditions, around 3.3 to 3.4, regardless of prompt style or temperature. Notably, if the human rater average is taken as the gold standard, the configuration most closely aligned with human scoring norms was the large model with regular prompt and low temperature (M = 3.29, SE = 0.065).

6 Discussion

The purpose of this study was to explore how variations in model settings influence LLM-generated ratings and how these scores align within themselves, with each other, and with human scores, given the intended formative use of this assessment. We examined the impact of model size, temperature, and prompt style. Overall, the results showed that certain combinations of these parameters may be more suitable for specific assessment purposes. In evaluating medical students' clinical reasoning skills, findings revealed that model size influenced both alignment with human raters and internal consistency, with GPT-4o consistently outperforming GPT-4o-mini. Temperature and prompt style had a relatively minor effect when using the larger model compared to the smaller one. Thus, for formative assessment of clinical reasoning, employing a larger model such as GPT-4o is recommended to achieve greater consistency and human alignment.

Our results revealed that LLMs showed high ICC within the same model and between models of the same size, but can systematically differ from human ratings. Across almost all models, the assess–reassess reliability was fairly high, indicating fairly high reliability in reproducing their ratings across different trials. The four smaller models generally have ICC higher than 0.6 and the four larger models have ICC higher than 0.77, suggesting high internal consistency between models with the same size. However, when it comes to agreement with human, their performance vary. The smaller model (GPT-4o mini) showed very poor agreement with human raters, while the larger model showed ICC levels that are comparable to human-human ICC. Combining these insights highlight a key feature or limitation of LLM raters: high internal consistency within or across LLM raters does not imply alignment with human judgments. Simply deploying a group of LLM raters and observing agreement among them is insufficient for validating their use in scoring. Direct comparison with human ratings is necessary in those cases. However, once we are able to configure an LLM that has a high agreement with human, the high assess–reassess consistency is encouraging sign that such an LLM can produce reliable ratings.

In this study, although human raters themselves showed only moderate inter-rater reliability, the larger model aligned more closely with human scores than the smaller models did. Additionally, human raters gave lower average scores than the LLMs, suggesting that humans may apply stricter evaluation criteria - a pattern consistent with prior findings (Morjaria et al., 2024). Therefore, incorporating at least one human rater, or ensuring that LLM raters are demonstrably aligned with human evaluations, remains essential when human judgment serves as the gold standard.

Future implementations should consider periodic human validation of LLM-generated scores to promote fairness and reliability. Importantly, assessments should avoid assigning LLM raters to some students and human raters to others, as this could introduce systematic biases. Further research with larger sample sizes should also explore the use of Generalizability Theory to analyze the impact of rater variability across combinations of human and LLM raters (Shavelson et al., 1992).

Another notable finding was related to scores generated by LLMs using persona prompts. Contrary to our expectation that persona prompting (asking LLM to pretend to be an expert) would enhance alignment with human expert ratings, these prompts instead led to higher scores compared to standard prompts and showed poorer alignment with human ratings. This suggests that persona-based prompting, at least in this case, may not effectively replicate human expert evaluation processes. Future research could explore alternative prompting strategies, such as few-shot prompting, or adding a few examples within prompts, as a potentially more effective method to improve alignment between LLM-generated ratings and human ratings, particularly for items exhibiting notable discrepancies, such as Item 4.

Ultimately, advancing the use of LLMs for assessment requires careful attention not only to consistency within models, but also to variation between models and, most critically, to their alignment with human judgments, as overlooking any of these dimensions risks undermining the validity of LLM-generated scores and even leading to systematic biases.

Limitations

A major limitation of this study was the relatively small sample size. Having only two human raters restricted our ability to establish a robust human expert benchmark. This could have influenced the reliability comparisons with LLM-generated scores. Additional human raters would reduce measurement error. Moreover, with scores from only 21 medical students, we were unable to comprehensively explore or decompose sources of measurement error contributing to discrepancies between variations of LLM-generated and human ratings. Future studies should include larger datasets, enabling the application of Generalizability Theory to better identify and quantify multiple facets of error, such as variability due to raters, tasks, or specific scoring criteria.

Acknowledgments

We thank Prof. Andrew Dean Ho, Kenji Kitamura, and Mikayla Do for their valuable guidance and advice.

References

Alison L Bailey, Alexander Johnson, Natarajan Balaji Shankar, Hariram Veeramani, Julie A Washington, and Abeer Alwan. 2025. Addressing bias in spoken language systems used in the development and implementation of automated child language-based assessment. *Journal of Educational Measurement*.

Peter Baldwin, Victoria Yaneva, Kai North, Le An Ha, Yiyun Zhou, Alex J Mechaber, and Brian E Clauser. 2025. The vulnerability of ai-based scoring systems to gaming strategies: A case study. *Journal of Educational Measurement*, 62(1):172–194.

Yavar Bathaee. 2017. The artificial intelligence black box and the failure of intent and causation. *Harv. JL & Tech.*, 31:889.

Emilia Brügge, Sarah Ricchizzi, Malin Arenbeck, Marius Niklas Keller, Lina Schur, Walter Stummer, Markus Holling, Max Hao Lu, and Dogus Darici. 2024. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Medical Education*, 24(1):1391.

S Coderre, HHPH Mandin, Peter H Harasym, and Gordon H Fick. 2003. Diagnostic reasoning strategies and diagnostic success. *Medical education*, 37(8):695–703.

Sophie Fürstenberg, Tillmann Helm, Sarah Prediger, Martina Kadmon, Pascal O Berberat, and Sigrid Harendza. 2020. Assessing clinical reasoning in undergraduate medical students during history taking with an empirically derived scale for clinical reasoning indicators. *BMC Medical Education*, 20:1–7.

Matthias Gamer, Jim Lemon, Ian Fellows, and Puspendra Singh. 2019. *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1.

Catharina M Haring, Bernadette M Cools, Petra JM van Gurp, Jos WM van der Meer, and Cornelis T Postma. 2017. Observable phenomena that reveal medical students' clinical reasoning ability during expert assessment of their history taking: a qualitative study. *BMC Medical education*, 17:1–9.

Andrew D. Ho. 2025. Metaphors, mantras, and mnemonics: Communication competencies in educational measurement. In *Presidential Address, Annual Meeting of the National Council on Measurement in Education*, Denver, CO. Presidential Address.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Leo Morjaria, Levi Burns, Keyna Bracken, Anthony J Levinson, Quang N Ngo, Mark Lee, and Matthew Sibbald. 2024. Examining the efficacy of chatgpt in marking short-answer assessments in an undergraduate medical program. *International Medical Education*, 3(1):32–43.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403.

OpenAI. 2023. *Audio: Create translation*. Accessed: 2025-06-05.

Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2025. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 462–472.

Richard J Shavelson, Noreen M Webb, and Glenn L Rowley. 1992. *Generalizability theory*. American Psychological Association.

Barbara Simmons. 2010. Clinical reasoning: concept analysis. *Journal of advanced nursing*, 66(5):1151–1158.

Christine A Tanner. 2006. Thinking like a nurse: A research-based model of clinical judgment in nursing. *Journal of nursing education*, 45(6):204–211.

University of Victoria Libraries. 2024. Prompt design: Temperature parameter. https://libguides.uvic.ca/promptdesign/temp. Accessed: 2025-06-05.

Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From automation to augmentation: Large language models elevating essay scoring landscape. *arXiv e-prints*, pages arXiv–2401.

A Regular prompt

The regular prompt we used is shown below:

Regular Prompt

At this point, you will assess the quality of the third-semester medical student conducting a history-taking conversation.

Your assessment should include the following eight criteria

1. Assess whether the user has taken control of the conversation to obtain the necessary information.

2. Assess whether the user recognizes all relevant information.

3. Assess whether the user formulates targeted questions so that he can capture and specify the symptoms in detail.

4. Assess whether the questions of the user suggest that specific causes or circumstances lead to certain symptoms.

5. Assess whether the user asks questions in a logical sequence.

6. Assess whether the user reassures the patient that he has received the correct information from the patient.

7. Assess whether the user has summarized his collected information before ending the conversation.

8. Assess whether the user has collected sufficient information of high quality at an appropriate speed.

Assign each of the eight criteria a score according to the following scheme:

1 - Does not meet the criterion 2 - Rather does not meet the criterion 3 - Partially meets the criterion 4 - Rather meets the criterion 5 - Fully meets the criterion

Explain the evaluation with two sentences.

B Expert persona prompt

The expert persona prompt we used is shown below:

Expert Persona Prompt

Clinical decision-making (CDM) is a central process in healthcare where physicians gather, evaluate, and interpret relevant information about a patient's health status to make informed decisions regarding diagnosis and treatment. At this point, act as a rater with medical expertise who is evaluating medical students for their CDM mastery. You will assess the quality of the third-semester medical student conducting a history-taking conversation. Provide outputs that a rater with medical expertise would create.

Your assessment should include the following eight criteria 1. Assess whether the user has taken control of the conversation to obtain the necessary information.

2. Assess whether the user recognizes all relevant information.

3. Assess whether the user formulates targeted questions so that he can capture and specify the symptoms in detail.

4. Assess whether the questions of the user suggest that specific causes or circumstances lead to certain symptoms.

5. Assess whether the user asks questions in a logical sequence.

6. Assess whether the user reassures the patient that he has received the correct information from the patient.

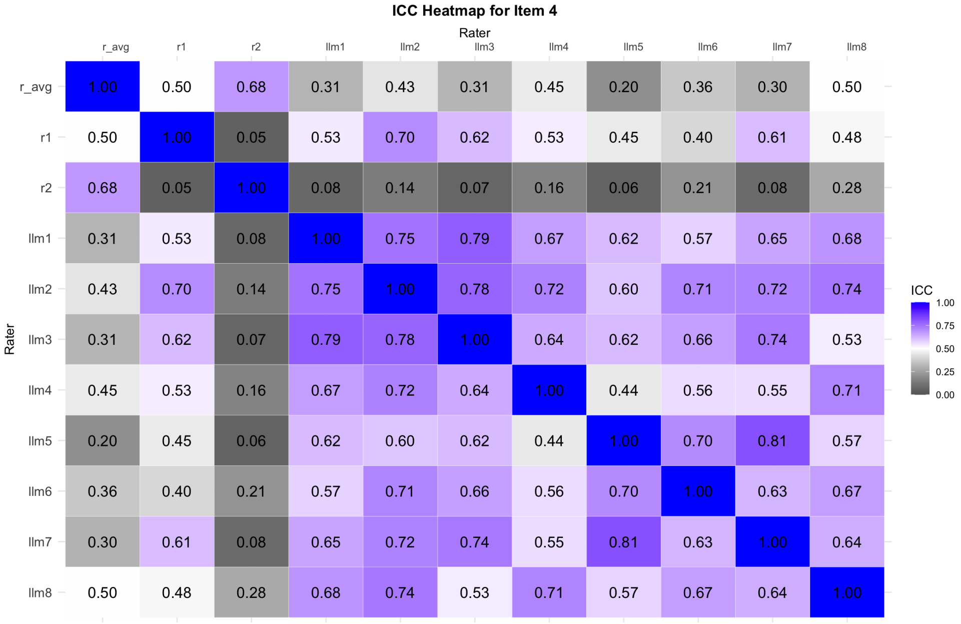
7. Assess whether the user has summarized his collected information before ending the conversation.

8. Assess whether the user has collected sufficient information of high quality at an appropriate speed.

Assign each of the eight criteria a score according to the following scheme:

1 - Does not meet the criterion 2 - Rather does not meet the criterion 3 - Partially meets the criterion 4 - Rather meets the criterion 5 - Fully meets the criterion

Explain the evaluation with two sentences.



	r_avg	r1	r2	llm1	llm2	llm3	llm4	llm5	llm6	llm7	llm8
r_avg	1.00	0.50	0.68	0.31	0.43	0.31	0.45	0.20	0.36	0.30	0.50
r1	0.50	1.00	0.05	0.53	0.70	0.62	0.53	0.45	0.40	0.61	0.48
r2	0.68	0.05	1.00	0.08	0.14	0.07	0.16	0.06	0.21	0.08	0.28
llm1	0.31	0.53	0.08	1.00	0.75	0.79	0.67	0.62	0.57	0.65	0.68
llm2	0.43	0.70	0.14	0.75	1.00	0.78	0.72	0.60	0.71	0.72	0.74
llm3	0.31	0.62	0.07	0.79	0.78	1.00	0.64	0.62	0.66	0.74	0.53
llm4	0.45	0.53	0.16	0.67	0.72	0.64	1.00	0.44	0.56	0.55	0.71
llm5	0.20	0.45	0.06	0.62	0.60	0.62	0.44	1.00	0.70	0.81	0.57
llm6	0.36	0.40	0.21	0.57	0.71	0.66	0.56	0.70	1.00	0.63	0.67
llm7	0.30	0.61	0.08	0.65	0.72	0.74	0.55	0.81	0.63	1.00	0.64
llm8	0.50	0.48	0.28	0.68	0.74	0.53	0.71	0.57	0.67	0.64	1.00

Figure 4: Pairwise Intraclass Correlation Coefficients (ICC) between human and LLM raters across rating conditions for Item 4

C Pairwise ICC for Item 4 (Figure 4)

D Main effects & interaction regression table (Table 2)

	Coefficient (SE)
(Intercept)	3.523*** (0.070)
Large Model	-0.228* (0.096)
High Temperature	0.263** (0.093)
Expert Prompt	0.274** (0.087)
Large Model:High Temperature	-0.257* (0.130)
Large Model:Expert Prompt	-0.142 (0.132)
High Temperature:Expert Prompt	-0.138 (0.121)
Large Model:High Temp.:Expert Prompt	0.149 (0.185)
Num. Obs.	189
R^2	0.293
Adj. R^2	0.265
AIC	137.3
BIC	166.5
RMSE	0.33

Table 2: Fixed effects regression on average clinical reasoning scores, with predictors for model size (Large Model), temperature (High Temperature), and prompt style (Expert Prompt), including interaction terms. Standard errors in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$.