

# Using LLMs to Grade Clinical Reasoning for Medical Students in Virtual Patient Dialogues

Jonathan Schiött<sup>1,2</sup>  
jonathan.schiott@ki.se

William Ivegren<sup>1</sup>  
ivegren@kth.se

Alexander Borg<sup>2</sup>  
alexander.borg@ki.se

Ioannis Parodis<sup>2</sup>  
ioannis.parodis@ki.se

Gabriel Skantze<sup>1</sup>  
skantze@kth.se

<sup>1</sup>KTH Royal Institute of Technology, Sweden

<sup>2</sup>Karolinska Institute, Sweden

## Abstract

This paper presents an evaluation of the use of large language models (LLMs) for grading clinical reasoning during rheumatology medical history virtual patient (VP) simulations. The study explores the feasibility of using state-of-the-art LLMs, including both general-purpose models, with various prompting strategies such as zero-shot, analysis-first, and chain-of-thought prompting, as well as reasoning models. The performance of these models in grading transcribed dialogues from VP simulations conducted on a Furhat robot was evaluated against human expert annotations. Human experts initially achieved a 65% inter-rater agreement, which resulted in a pooled Cohen's Kappa of 0.71 and 82.3% correctness. The best LLM, o3-mini, achieved a pooled Kappa of 0.68 and 81.5% correctness, with response times under 30 seconds, compared to approximately 6 minutes for human grading. These results indicate the possibility that automatic assessments can approach human reliability under controlled simulation conditions while delivering time and cost efficiencies.

## 1 Introduction

In recent years, the integration of new technologies into medical education has reshaped traditional pedagogical approaches by offering innovative solutions that improve learning outcomes. Among these, social robots and large language models (LLMs) have emerged as promising tools, delivering immersive, interactive experiences and enhancing virtual patients (VPs). VPs simulate clinical scenarios, enabling students to engage in diagnostic and therapeutic decision-making processes (Cook et al., 2007), while social robots use advanced technologies to facilitate intuitive interactions (Al Moubayed et al., 2012). This dual approach fosters clinical reasoning (CR) skills and meets the growing demand for complementary educational resources.

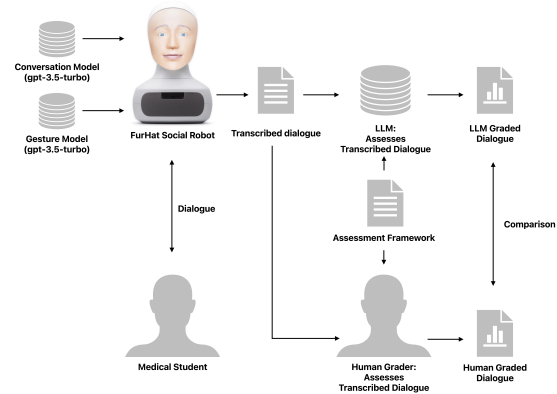


Figure 1: The Social Robot VP System Used by the Medical Students and the following Grading Schema

CR, i.e., “the application of knowledge to collect and integrate information from various sources to arrive at a diagnosis and management plan” (Cook and Triola, 2009), is an important part of medical education and widely recognized across health professions (Gordon et al., 2022). High-quality CR assessment ensures that medical schools prepare competent learners, which is vital since faulty CR is a major contributor to diagnostic errors affecting patient safety (Gordon et al., 2022). Consequently, strategies to enhance CR education and assessment have become a focus in health professions research. However, because CR relies heavily on interaction, physical and economic constraints limit hands-on training. Thus, innovative methods are needed to boost communication skills and adaptability in complex clinical settings. LLMs can transform these methods by simulating realistic patient–doctor interactions (Borg et al., 2024b; Henkel et al., 2024) and creating personalized, interactive experiences that offer more realistic VP interactions (Borg et al., 2024a). As noted in Graf et al. (2024), reviews show that VP systems improve CR and knowledge acquisition compared to no intervention, though their benefits may be mod-

est relative to conventional methods (Plackett et al., 2022; Cook et al., 2010; McGaghie et al., 2010).

In this paper, we implement and evaluate an assessment model using LLMs to support medical students in learning CR in rheumatology and to provide objective performance assessments. To validate the assessment model we benchmark its grading capabilities with practising rheumatologists as shown in Figure 1. Our research question is:

- **RQ:** How accurately can a large language model assess medical students' clinical reasoning skills in comparison to physicians that specialize in rheumatology?

## 2 Related Work

### 2.1 Virtual Patients

VPs are computer simulations that replicate clinical scenarios, enabling learners to take patient histories, perform examinations, and make diagnostic and therapeutic decisions (Cook et al., 2007). They allow students to progress through cases in controlled settings to refine diagnostic skills and CR (Kononowicz et al., 2015; Cook et al., 2010). VP designs have evolved from simple chatbots (Cameron et al., 2019) to embodied conversational agents (Campillos-Llanos et al., 2020; Pantziaras et al., 2015), as cited in (Graf et al., 2024). Recent advances incorporate social robots using facial animation and real-time emotion generation via LLMs (Mishra et al., 2023; Belpaeme et al., 2018; Al Moubayed et al., 2012) to enhance engagement and authenticity (Borg et al., 2024b). While earlier work focused on human-robot interactions, few studies have combined social robotics with LLM-based evaluation frameworks. Our work bridges this gap by integrating LLM-driven assessments into VP systems augmented by social robots (Goodrich and Schultz, 2008; Kang et al., 2024; Akhyani and Moradi, 2023; Flores, 2024). Empirical evidence supports that such VP simulations improve exam performance and learner satisfaction, though challenges in scenario complexity and authentic conversational dynamics persist (Cook and Triola, 2009; Borg et al., 2024a; Edelbring et al., 2011; Stamer et al., 2023; Holderried et al., 2024; Suárez et al., 2022).

### 2.2 Clinical Reasoning in Medical Education

CR is central to medical education, involving the interpretation of patient symptoms, diagnostic tests, and imaging to form diagnoses and treatment plans

(Richards et al., 2020; Huang et al., 2014). It includes both non-analytic (System 1) processes based on pattern recognition and intuition, and analytic (System 2) processes, characterized by deliberate evidence evaluation, and (Eva, 2005; Richards et al., 2020; Edelbring et al., 2018). This duality is evident in diagnostic reasoning, which classifies conditions from objective data (Cook et al., 2018), and management reasoning, which dynamically adjusts treatments in response to evolving patient conditions (Cook et al., 2019). VP simulations provide a safe environment for learners to iteratively apply and reflect on both strategies.

### 2.3 Flaws with Human Evaluation

Traditional evaluation in medical education is compromised by human factors such as implicit biases, leading to inconsistent grading (Sadler, 2009; Quinn, 2020). Examiner fatigue, stress, and cognitive overload further exacerbate grading inconsistencies (Brackett et al., 2013; Klein, 2002). In addition, moral dilemmas, which balance strict rubric adherence with sensitivity to individual circumstances and varying subject expertise, undermine the consistency, objectivity, reliability, and validity of traditional evaluation systems (Tierney, 2015; Bloxham et al., 2016; Boyd and Bloxham, 2007; Brookhart, 1997).

### 2.4 LLMs in Medical Education and Evaluation

LLMs are transforming medical education by providing scalable, objective, and efficient assessment tools. Recent studies show that LLMs can evaluate short answer questions with accuracy comparable to human evaluators (Henkel et al., 2024). For instance, a study on OSCEs using the Master Interview Rating Scale reported off-by-one accuracies of 67–87% and thresholded accuracies of 75–88%, with excellent intra-rater reliability ( $\alpha = 0.98$  for gpt-4o) across clinical domains (Geathers et al., 2025). This approach reduces resource burdens (Cavalcanti et al., 2021; Kasneci et al., 2023; Wilson et al., 2021; Ruwe and Mayweg-Paus, 2024) while supporting nuanced evaluations. Beyond personalized feedback, LLMs have shown promise in assessing recorded medical conversations, potentially transforming evaluations of clinical communication skills (Butow and Hoque, 2020; Çiçek et al., 2024). Their primary advantage in this study is the standardization of assessments, reduction of grading biases, and production of consistent,

reproducible evaluations. Nevertheless, effective integration requires oversight to align with pedagogical goals and address ethical and bias concerns via transparency, audit trails, explainability, and continuous human supervision (Giannakos et al., 2024; Çiçek et al., 2024).

### 3 Methodology

Our primary objective was to evaluate the use of LLMs for objectively grading the CR skills of medical students during a rheumatology medical history simulation. This was conducted in close collaboration with Karolinska Institute in Stockholm, Sweden, as part of a clinical teaching course in rheumatology. In the course, a VP system designed to simulate realistic clinical scenarios to assess CR was used, featuring five different patient cases (Borg et al., 2024b,a).

#### 3.1 Virtual Patient Platform

The VP platform used the Furhat robot (Al Moubayed et al., 2012), integrated with the OpenAI LLM gpt-3.5-turbo (OpenAI, 2025a) designed to control interactive dialogues and mimic patient-doctor conversations. The Furhat robot features an animated face and a neck that enables realistic head movements, which have been hypothesized to enhance realism during VP encounters (Al Moubayed et al., 2012). In addition, the robot conveys subtle facial expressions and gaze behaviours to indicate the emotional state of the patient (Mishra et al., 2023). To facilitate the interaction, the Furhat robotics platform was used in conjunction with the Furhat software development kit (FurhatSDK) (Al Moubayed et al., 2012), integrated with the OpenAI gpt-3.5-turbo large language model (Brown et al., 2020) prompted to generate appropriate facial expressions at predefined anchor points in the conversation. This ensured that the robot's responses aligned with the intended emotional states (Irfan et al., 2023). These expressions were selected from a predefined set available in the FurhatSDK, determined contextually based on the VP's responses, which varied stochastically in emotional tone (Borg et al., 2024a).

To generate authentic patient responses and avoid assistant-like phrasing, the LLM was provided with a detailed patient description embedded within the prompt. This included specific instructions on generating dialogue lines, as well as the last ten dialogue turns between the student and

the VP, ensuring contextual continuity (Borg et al., 2024a). A challenge when using LLMs in turn-based dialogue is response latency, which can lead to misinterpretation of conversational flow (Irfan et al., 2023). To mitigate this, a turn-holding cue was implemented using an LED light positioned at the bottom of the robot, which signaled the turn-taking status through pre-specified colors (Skantze et al., 2015). Before engaging with the VP case on the social robotic platform, students received introductory information about the case, including relevant laboratory test results and corresponding reference values (Borg et al., 2024a).

#### 3.2 Virtual Patient Cases

The VPs used in this study are identical to those previously employed by Borg et al. (2024b) and in the course at Karolinska Institute, and were designed to simulate a range of rheumatological diseases. Each patient profile was meticulously developed in English to reflect realistic medical histories, symptoms, and family backgrounds, thus offering diverse scenarios for evaluation. The study included Angelica, a 64-year-old woman with Giant Cell Arteritis (GCA) who experiences severe headaches, jaw claudication, and visual disturbances; Hanna, a 61-year-old woman with Dermatomyositis/Antisynthetase Syndrome characterized by progressive muscle weakness, skin rash, and difficulty swallowing; Joel, a 42-year-old man with Psoriatic Arthritis manifested through intermittent joint pain, swelling, and a family history of psoriasis; Mikael, a 70-year-old man with GCA/Polymyalgia Rheumatica (PMR) marked by muscle stiffness in the shoulders and hips along with fatigue; and Peter, a 30-year-old man with Ankylosing Spondylitis suffering from chronic lower back pain and stiffness that worsens with inactivity yet improves with movement.

#### 3.3 Data Collection and Preparation

We assembled a dataset of 154 previously recorded and transcribed dialogues between medical students and the social robot from previous iterations of the rheumatology course conducted during spring and fall of 2024. The students were all from the sixth semester of the medical programme conducted at Karolinska Institute, and part of a course in internal medicine. The dialogues were conducted in English, but none of the speakers had English as their native language. All dialogues were anonymized and recordings were collected

exclusively from students who had consented to their use in future studies. An excerpt from one of the transcribed dialogues between a student and the VP Mikael can be seen in Figure 2.

```

-----
User: have you felt any other symptoms
-----
Robot: Yes, I've lost my appetite and have lost about 3 to
4 kilos in the last month. I've also been feeling sad and
depressed since my symptoms started.
-----
User: does it help when you move around
-----
Robot: No, it actually hurts more in the morning when
I get up from bed and I have a hard time bending my arms
or reaching the floor if I drop something.
-----

```

Figure 2: Example Part of Transcribed Dialogue

Each dialogue was manually categorized into one of three quality tiers: Usable (79 dialogues), Acceptable but with deviations (26 dialogues), and Unusable (49 dialogues), based on whether the students maintained their role as clinicians, avoided off-topic discussions, and whether the dialogue exhibited minimal transcription errors.

### 3.4 Dialogue Segmentation and Selection

To ensure a balanced evaluation dataset, dialogues were segmented according to their speech turn counts, providing an even distribution of conversation lengths. We then selected two balanced groups of 15 usable coded dialogues (as defined in Section 3.3). The selection process ensured diversity across the five unique virtual patient cases, while also making sure that the representation of dialogue lengths, from short to long, was balanced. The selection was intentionally limited to these objective criteria to avoid introducing subjective bias.

### 3.5 Assessment Framework

To grade the dialogues, an assessment framework was developed in collaboration with medical experts at Karolinska Institute before any of them had seen the 30 student dialogues previously selected. The experts relied solely on current CR standards and their own experience, thus keeping the framework blind to the data it would later evaluate. The goal was to establish a set of questions that medical professionals should ask when taking a medical history from patients with specific rheumatic conditions. This process resulted in five distinct sets of questions, one for each patient case in the

data set, each consisting of three parts. The framework includes Start Questions, which establish the dialogue's context; General Questions, which address broader conversational topics; and Specific Questions, which require detailed responses tailored to the patient's symptoms, therefore varying for each patient. All questions are designed to be answered with one of four options: Yes, No, Partially, and N/A, with N/A used when the patient volunteers information about a symptom without being asked, as illustrated in Figure 2. The Start Questions and General Questions are consistent across all patient cases. For instance, in the case of Mikael, the framework asks whether the student initiated the dialogue with open-ended questions, whether they inquired about unintentional weight loss, and whether they questioned the patient regarding morning stiffness. An example of a full questions set is provided in Appendix A for Mikael.

### 3.6 Human Expert Grading

Human expert grading of the selected dialogues was conducted by two independent pairs, each consisting of a junior and a senior rheumatology physician. Each group graded one set of 15 dialogues following the same instructions and a joint session where all graders could clarify any questions regarding the process. Using the assessment framework, every dialogue was evaluated based on a predefined set of questions for each patient. Initially, both experts in each group independently graded the dialogue. If discrepancies arose, the senior grader from the alternate group stepped in to grade the questions with conflicting responses. In cases where no consensus emerged, meaning that all three graders provided different answers, the senior grader from the alternate group reviewed all assessments and made the final decision. This produced a true consensus grade for each dialogue, which then served as a benchmark to evaluate the performance of both individual human graders and the LLM. The consensus process ensured that the benchmark reflected clinical judgement shaped by tacit knowledge, human subjectivity, empathy, and contextual judgement.

### 3.7 LLM Grading

Since each human graded 15 dialogues, which equals 60 total graded dialogues, each LLM was prompted to grade each of the 30 dialogues twice. This process of double grading was implemented to assess the internal consistency and reliability



of the LLM models. The average percent variation between the two independent gradings for each dialogue was then analysed to quantify this consistency. Furthermore, conducting two grading instances per dialogue enabled the calculation of performance metrics such as pooled Kappa and percent correct, facilitating a comparison of the LLMs' capabilities against the benchmark established by the human expert graders.

The assessment models were developed using various prompting techniques, including zero-shot (OS), analysis-first (AF), and Chain-of-Thought (CoT) approaches, and tested across six different OpenAI models (OpenAI, 2025c): two general-purpose models (gpt-4o and gpt-4o-mini) and four reasoning models (o3-mini, o1, o1-preview and o1-mini). We strategically included both model types to discern if inherent reasoning capabilities offered an advantage in this evaluation task. We also incorporated the reasoning model DeepSeek R1 (DeepSeek, 2025a), to broaden our assessment to include models beyond the OpenAI ecosystem. OS learning allows for the entire dataset to be used for testing, thereby enhancing the validity of the results. CoT prompting, initially introduced as a few-shot technique, improves reasoning by presenting answers as step-by-step explanations within examples. Building on this, OS-CoT employs a template-based approach for reasoning without requiring examples, relying on prompts such as "Let's think step by step" to guide the model in producing structured reasoning. This method has demonstrated significant improvements in challenging benchmarks and is most effective when paired with the AF approach, which prompts the model to generate a detailed analysis before answering (Wei et al., 2023; Kojima et al., 2022; Jia et al., 2024). Therefore, the general-purpose models were prompted using OS, AF, CoT instructions, guiding them to produce step-by-step reasoning, as seen in Figure 3. Alternative approaches such as few-shot prompting, reinforcement learning fine-tuning, and interactive feedback loops were considered but ultimately not adopted due to their potential to introduce bias, restrict generalization, or obscure the model's native reasoning process. Recent surveys on efficient prompting methods provide additional context for these choices, emphasizing that prompt engineering based on efficient design, rather than extensive fine-tuning, offers robust and scalable performance across diverse tasks (Chang et al., 2024).

The reasoning models were prompted using only

Act as a medical professor evaluating the following dialogue to assess the student's medical reasoning within rheumatology and their ability to ask relevant questions to learn about the patient's situation. Your task is to analyze the dialogue and evaluate how well the student addresses the questions: '{questions}'.

Follow these steps in your evaluation:

1. Analyze the dialogue systematically using chain of thought for the attached questions.
2. Based on your analysis, assess the student's performance for each question:
  - Did the student ask the question or a related question?
  - If there was no need to ask (e.g., the patient already discussed the topic unprompted), answer 'N/A'.
3. Provide a detailed evaluation for each question, based on the above analysis.

Dialogue:  
{str(dialogue)}

Response format:  
 - \*\*Question question\_number:\*\* Repeat the question to be answered.  
 - \*\*Analysis:\*\* Provide a structured analysis of the dialogue based on the aspects above.  
 - \*\*Answer:\*\* 'Yes', 'No', 'Partially', or 'N/A'.  
 - \*\*Reasoning:\*\* Summarize the reasoning for your assessment, highlighting key points from the dialogue.

Figure 3: General-purpose Model Prompt Used to Grade Dialogues

OS, as seen in Figure 4, since they have internal instructions that achieve AF and CoT effects with reinforcement learning, and multi-round conversations using reasoning tokens (OpenAI, 2025d,b; DeepSeek, 2025b).

Act as a medical professor evaluating a dialogue to assess the student's medical reasoning in rheumatology and their ability to ask relevant questions to learn about the patient's situation.

Evaluate the student's performance for each question: '{questions}'.

For each question:

1. Determine whether the student asked the question or a related one.
  - If the patient discussed the topic without prompting, respond 'N/A'.
2. Provide an assessment and explanation.

Dialogue:  
{str(dialogue)}

Response format:  
 - Question question\_number: repeat the question  
 - Answer: 'Yes', 'No', 'Partially', or 'N/A'  
 - Reasoning: Summarize the reasoning, referencing key points from the dialogue.

Figure 4: Reasoning Model Grading Prompt Used to Grade Dialogues

The models assessed each dialogue by applying the same set of assessment questions used by the human experts, with both dialogue and questions provided in the prompt. Examples of the assessment output of one of the general-purpose models and one of the reasoning models can be seen in Figure 5 and Figure 6.

### 3.8 Evaluation of Grading Performance

The performance of each LLM and human was evaluated by comparing the grades for each dialogue against the consensus human grades (Agreement with Correct Grades). In addition, we employed Cohen's Kappa to quantify chance-corrected agree-

ment, defined as

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where  $P_o$  is the observed proportion of agreement with the correct grades and  $P_e$  is the expected agreement by chance (McHugh, 2012; Landis and Koch, 1977). This measure provides a more robust and conservative estimate of the genuine agreement on the CR skills being assessed. Finally, efficiency was assessed by recording the time taken by the LLM to grade dialogues and benchmarking it against the grading duration of human experts, thereby emphasizing potential time and cost savings.

## 4 Results

This section presents the quantitative results of both human evaluations and automated dialogue assessment LLMs. The grading was performed on 30 dialogues comprising a total of 701 questions. In Tables 2-6 the best performances of each metric are highlighted in bold.

- \*\*question 11:\*\* did the student ask if the patient has experienced unintentional weight loss?  
- \*\*analysis:\*\* the student asked about additional symptoms, and the patient reported weight loss.  
- \*\*answer:\*\* yes  
- \*\*reasoning:\*\* this emerged during a broader inquiry into symptoms.

Figure 5: General-purpose Model gpt-4o Incorrect Grading of Specific Question

11. question 11 - "did the student ask if the patient has experienced unintentional weight loss?"  
• answer: n/a  
• reasoning: although the patient mentioned losing 3-4 kilos, the student did not explicitly ask about weight loss; this topic was spontaneously provided.

Figure 6: Reasoning Model o3-mini Correct Grading of Specific Question

### 4.1 Human Inter-Grader Agreement and Consensus Building

In the initial grading round, the graders achieved an average agreement of 65%. In the second round, where a senior grader reassessed questions with conflicting responses, consensus was reached for 92.6% of the questions (with only 52 questions remaining in disagreement). Finally, after a thorough review in which the senior graders considered the responses of the other graders and cast a decisive vote, all questions ultimately reached consensus. The final consensus has the following distribution

Grader	Number of Dialogues	Avg (%) Variation	Pooled Kappa	Percent Correct
Grader 1-(J)	15	31.3	0.80	88.0
Grader 2-(S)	15	31.3	0.59	74.9
Grader 3-(J)	15	26.1	0.66	78.9
Grader 4-(S)	15	26.1	0.79	88.0
<b>All Graders</b>	<b>60</b>	<b>6.7</b>	<b>0.71</b>	<b>82.5</b>

Table 1: Grading Performance Metrics for Human Graders

Model	Number of Dial.	Avg (%) Variation	Pooled Kappa	Percent Correct
DeepSeek R1	60	14.4	0.62	77.3
gpt-4o-mini	60	22.7	0.49	69.0
gpt-4o	60	16.4	0.57	74.8
o1-mini	60	21.0	0.56	74.9
o1-preview	60	11.1	0.65	79.6
o1	60	10.8	0.66	80.4
o3-mini	60	<b>10.4</b>	<b>0.68</b>	<b>81.8</b>

Table 2: Grading Performance Metrics for Each LLM

of correct grades per grade type: 350 questions graded with Yes, 288 with No, 53 with Partially, and 10 with N/A.

### 4.2 Individual Human Grader Performance

Table 1 summarizes performance metrics for each human grader with junior graders marked with (J) and senior graders marked with (S). Here, *Average (%) Variation* indicates the average percent variation of disagreements per dialogue between paired graders, while the *Kappa* metric quantifies individual rater reliability when compared to the correct grades. Notably, Grader 1-(J) and Grader 4-(S) (with 88.0% correct responses) outperformed Grader 2-(S) (74.9%) and Grader 3-(J) (78.9%). The combined performance of all graders shows a pooled Kappa of 0.71 and an average correctness of 82.5%.

### 4.3 LLM Grading Performance

Table 2 presents the overall performance of each assessment model, based on two independent gradings per dialogue. Metrics include the average percent variation (i.e., the mean number of discrepancies between the two gradings), the pooled Kappa (quantifying the individual model reliability when compared to the correct grades), and the overall percentage correct. The o3-mini model had the lowest average variation (10.4%), highest pooled Kappa (0.68), and the highest percent correct (81.8%).

Model	Shorter	Medium	Longer
DeepSeek-R1	77.7	79.6	74.8
gpt-4o-mini	77.1	65.1	64.5
gpt-4o	72.7	78.9	72.9
o1-mini	75.6	75.8	73.3
o1-preview	79.1	82.9	76.9
o1	80.4	<b>83.6</b>	77.4
o3-mini	<b>84.7</b>	82.4	78.2
All Human Graders	82.9	<b>83.6</b>	<b>80.6</b>

Table 3: Percent Correct Grades by Dialogue Length for each LLM and Aggregated Human Result

Model	Yes	No	Partially	N/A
DeepSeek R1	88.2	74.1	28.3	50.0
gpt-4o-mini	64.6	83.3	32.1	10.0
gpt-4o	86.6	71.2	26.4	20.0
o1-mini	91.3	68.6	12.3	15.0
o1-preview	<b>94.4</b>	74.5	21.7	15.0
o1	93.9	75.7	21.7	55.0
o3-mini	90.6	84.2	19.8	35.0
All Human Graders	83.7	<b>85.4</b>	<b>61.3</b>	<b>65.0</b>

Table 4: Percent Correct Grades by Grade Category for each LLM and Aggregated Human Result

#### 4.4 Dialogue Length Analysis

Table 3 shows the accuracy trends across three dialogue-length groups, each with 10 dialogues: Shorter (11-81 speech turns), Medium (81-113 speech turns), and Longer (125-200 speech turns). Human graders maintain a stable performance (around 81–84%), while model accuracies vary. For instance, the o3-mini model performs best on Shorter dialogues (approximately 85%) and declines to about 78% on Longer ones. In contrast, gpt-4o-mini exhibits a more pronounced drop in accuracy on Medium and Longer dialogues, while the other models have a similar pattern in performance to humans.

#### 4.5 Correct Grade Category Analysis

Table 4 breaks down the accuracy by grade category: Yes, No, Partially, and N/A, for each model and all human graders. Models generally perform well in the Yes and No categories, with several achieving over 90% and 80% respectively. However, accuracy drops noticeably in the Partially and N/A categories, where human graders outperform all models. This suggests that while models can reliably identify clear-cut responses, they struggle with ambiguous or nuanced cases that humans handle more effectively.

Model	Start Questions	General Questions	Specific Questions
DeepSeek R1	77.5	88.8	69.2
gpt-4o-mini	77.1	79.4	58.9
gpt-4o	75.8	86.7	66.0
o1-mini	69.6	88.1	67.4
o1-preview	77.9	92.3	71.3
o1	80.8	91.5	72.4
o3-mini	<b>84.6</b>	<b>92.1</b>	73.6
All Human Graders	80.4	89.8	<b>77.7</b>

Table 5: Percent Correct by Question Category of the Rubric for Each LLM and Aggregated Human Results

Model	Average Time (s)	Std Dev (s)	Min Time (s)	Max Time (s)
DeepSeek R1	53.2	41.9	17.4	189.9
gpt-4o-mini	<b>27.9</b>	13.4	<b>6.0</b>	99.5
gpt-4o	30.4	9.7	10.5	59.1
o1-mini	30.3	12.5	11.5	83.4
o1-preview	57.3	27.5	31.3	230.1
o1	98.9	18.0	30.6	132.5
o3-mini	29.2	<b>6.6</b>	17.2	<b>47.9</b>

Table 6: Timing Metrics for Each LLM

#### 4.6 Question Category Analysis

This section examines model performance across different types of question categories (as distinguished in Section 3.5) to assess how well each model handles varying dialogue challenges. As shown in Table 5, while all models tend to achieve high accuracy on *General Questions*, there are notable discrepancies in the *Start Questions* and *Specific Questions* categories. For instance, the o3-mini model performs better than human graders in the *Start Questions* and *General Questions* categories, though there remains a gap in its handling of *Specific Questions* compared to human graders.

#### 4.7 Timing Performance

Table 6 summarizes the timing metrics for each model. The timing was recorded from the moment an API call was initiated until the reply was received. Notably, the gpt-4o-mini model achieved the fastest average response time (27.9 s), while the o1 model exhibited the longest (98.9 s).

## 5 Discussion

Below we integrate our observations on performance, efficiency, and the future directions for this assessment approach.

### 5.1 Impact of Human Grader Involvement on Ground Truth Construction

A key issue in this study is the reliance on human experts to establish the ground truth for LLM evaluations. Since these experts both defined the assessment criteria and graded the student dialogues, the benchmark reflects their subjective judgments. Despite a three-step consensus process aimed at reducing bias (Section 4.1), residual subjectivity affects inter-rater reliability. This variability, especially in ambiguous categories like Partially and N/A, means that the human performance numbers might reflect adherence to a specific expert-defined standard rather than absolute CR ability, while LLMs attempt to interpret these criteria without the inherent alignment of human evaluators.

### 5.2 LLM Versus Human Grading: Performance and Nuance

The o3-mini model's pooled Kappa of 0.68, close to both the pooled human Kappa of 0.71 and the initial 65% inter-rater agreement among human graders, indicates that LLMs can achieve comparable grading consistency. In Yes and No responses, LLMs such as o1-preview and o3-mini reached accuracies of over 80% (see Table 5). Additionally, the o3-mini outperformed human graders in Start Questions and General Questions. Nonetheless, both LLMs and humans struggled with more nuanced responses (i.e., Partially and N/A) and with Specific Questions, highlighting differences in performance across different types of processing. This indicates LLMs are effective at evaluating System 1 CR (Eva, 2005), i.e., adherence to basic communication protocols and identifying key words and patterns (Mirchandani et al., 2023; Xu et al., 2024), but find it challenging to evaluate System 2 CR (Eva, 2005), i.e., adherence of deeper clinical context, where integration of dialogue nuances diagnostic hypotheses is needed.

### 5.3 Time and Cost Efficiency

LLM-based assessments present a considerable time advantage by grading dialogues in under a minute compared to approximately 6 minutes for human experts. This reduction alleviates grading fatigue and cognitive overload (Brackett et al., 2013; Klein, 2002), while potentially lowering institutional costs. The resulting time savings can allow educators to focus on curriculum development, personalized mentoring, and further research,

with enhanced scalability and processing speed ultimately contributing to improved educational outcomes (Henkel et al., 2024; Ruwe and Mayweg-Paus, 2024; Gordon et al., 2022).

### 5.4 Dialogue Complexity

The impact of dialogue length on grading accuracy was evident (See Table 3). For example, the o3-mini model's accuracy declined from about 85% on short dialogues to approximately 78% on longer ones, highlighting the challenge that extended context poses for LLM architectures, especially for questions requiring the synthesis of dispersed information from earlier turns. In contrast, human graders maintained a relatively stable performance regardless of dialogue length, highlighting their robustness in complex interactions. Given that comprehensive CR skills depend on integrating scattered dialogue elements, the arbitrary categorization of dialogue lengths, determined by fixed dialogue turns, may inadequately capture interaction complexity and unfairly penalize LLMs, particularly when minor increases in length around these thresholds can lead to disproportionate accuracy drops.

### 5.5 Variability and Consistency in Grading

Both human graders and LLMs displayed variability in their assessments, but at different scales. While human experts showed an initial inter-rater agreement of 65%, the o3-mini model demonstrated impressive internal consistency with an average percent variation of 10.4%. This stability suggests that certain LLMs could help standardize and render grading more equitable by mitigating individual human biases, thereby enhancing the objectivity of CR skill evaluations.

### 5.6 Future Directions

Future research should examine the generalizability of LLM-based assessments across other medical specialties and dialogue-intensive domains (Stamer et al., 2023; Giannakos et al., 2024; Henkel et al., 2024), to assess whether the observed efficiency and objectivity extend beyond rheumatology. To reduce grader bias and obtain a more independent ground truth, future studies should include a separate group of rheumatology experts, uninformed in the framework development or consensus-building, to grade a subset of dialogues. Comparing their evaluations with the original consensus and LLM assessments could clarify the extent of



subjectivity in human grading and reveal whether initial performance metrics were inflated. Lastly, improving how ambiguous responses are handled, through model design or refined assessment criteria, will be critical for enhancing the reliability and validity of automated CR evaluations.

## 6 Conclusion

Under controlled VP simulation conditions, LLM-based assessment achieved near-human reliability, suggesting that they could offer a provisional yet efficient way to assess medical students' CR skills in rheumatology history-taking. The o3-mini model achieved a Cohen's Kappa of 0.68, nearly matching the human expert consensus of 0.71, a 81.8 percent correct grades compared to 82.5 by human experts and graded dialogues in under 30 seconds compared to roughly 6 minutes for experts, substantially reducing grading time. Although challenges remain in consistently assessing ambiguous responses and accounting for human biases, these findings highlight the potential for scalable, objective assessments amid the constraints of traditional CR training. Future work should refine grading instructions and validate these tools in other clinical settings, as well as in live clinical encounters, to further enhance medical educational outcomes and patient care.

## 7 Limitations

A primary limitation of this study is that the dialogues were not originally recorded with the intent of being graded, which may have introduced additional variability in both human and LLM assessments. Moreover, the controlled simulation environment differs from real-world clinical encounters, where factors such as patient variability and dynamic interactions may further influence grading outcomes. Also, a limitation is that we only used usable-coded dialogues in our analysis. The analysis could have been broader by letting the humans and models also grade dialogues of lower quality. Finally, the grading of dialogues by models was done on different computers, with different internet connections at different times. This can affect the timing of the grading.

## 8 Acknowledgements

The authors would like to thank the medical students who participated in the study, as well as the medical staff at the Division of Rheumatology

at the Karolinska University Hospital, Stockholm, Sweden.

### 8.1 Funding

This work was supported by grants from Region Stockholm ALF Pedagogy (FoUI-977096), Karolinska Institutet Pedagogical Project Funding (FoUI-964139), the Swedish Rheumatism Association (R-1013624), King Gustaf V's 80-year Foundation (FAI-2023-1055), Swedish Society of Medicine (SLS-974449), Nyckelfonden (OLL-1023269), Professor Nanna Svartz Foundation (2021-00436), Ulla and Roland Gustafsson Foundation (2024-43), Region Stockholm (FoUI-1004114), and Karolinska Institutet.

### 8.2 Conflict of interest

IP has received research funding and/or honoraria from Amgen, AstraZeneca, Aurinia, BMS, Eli Lilly, Gilead, GSK, Janssen, Novartis, Otsuka, and Roche. GS is a co-founder and Chief Scientist at Furhat Robotics. The other authors declare that they have no conflicts of interest.

## References

- Mahta Akhyani and Hadi Moradi. 2023. [Modular Customizable ROS-Based Framework for Rapid Development of Social Robots](#). *Preprint*, arXiv:2311.15780.
- Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. [Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction](#). In *Cognitive Behavioural Systems*, pages 114–130, Berlin, Heidelberg. Springer.
- Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. [Social robots for education: A review](#). *Science Robotics*, 3(21):eaat5954.
- Sue Bloxham, Clair Hughes, and Lenore Adie. 2016. [What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices](#). *Assessment & Evaluation in Higher Education*, 41(4):638–653.
- Alexander Borg, Benjamin Jobs, Viking Huss, Cidem Gentline, Fabricio Espinosa, Mini Ruiz, Samuel Edelbring, Carina Georg, Gabriel Skantze, and Ioannis Parodis. 2024a. [Enhancing clinical reasoning skills for medical students: A qualitative comparison of LLM-powered social robotic versus computer-based virtual patients within rheumatology](#). *Rheumatology International*.
- Alexander Borg, Ioannis Parodis, and Gabriel Skantze. 2024b. [Creating Virtual Patients using Robots](#)

- and Large Language Models: A Preliminary Study with Medical Students. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 273–277, Boulder CO USA. ACM.
- Pete Boyd and Sue Bloxham. 2007. *Developing Effective Assessment in Higher Education: a practical guide*.
- Marc A. Brackett, James L. Floman, Claire Ashton-James, Lillia Cherkasskiy, and Peter Salovey. 2013. The influence of teacher emotion on grading practices: A preliminary look at the evaluation of student writing. *Teachers and Teaching*, 19(6):634–646.
- Susan M. Brookhart. 1997. A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education*, 10(2):161–180.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. *Language Models are Few-Shot Learners*. Preprint, arXiv:2005.14165.
- Phyllis Butow and Ehsan Hoque. 2020. Using artificial intelligence to analyse and teach communication in healthcare. *The Breast*, 50:49–55.
- Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O’Neill, Cherie Armour, and Michael McTear. 2019. *Assessing the Usability of a Chatbot for Mental Health Care*. In *Internet Science*, pages 121–132, Cham. Springer International Publishing.
- Leonardo Campillos-Llanos, Catherine Thomas, Éric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. 2020. Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. *Natural Language Engineering*, 26(2):183–220.
- Anderson Pinheiro Cavalcanti, Arthur Barbosa, Ruan Carvalho, Fred Freitas, Yi-Shan Tsai, Dragan Gašević, and Rafael Ferreira Mello. 2021. Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2:100027.
- Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Xiaoqian Liu, Tong Xiao, and Jingbo Zhu. 2024. Efficient Prompting Methods for Large Language Models: A Survey. Preprint, arXiv:2404.01077.
- Feray Ekin Çiçek, Müşerref Ülker, Menekşe Özer, and Yavuz Selim Kıyak. 2024. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: A randomized controlled trial. *Postgraduate Medical Journal*, page qgae170.
- David A. Cook, Steven J. Durning, Jonathan Sherbino, and Larry D. Gruppen. 2019. Management Reasoning: Implications for Health Professions Educators and a Research Agenda. *Academic Medicine*, 94(9):1310.
- David A. Cook, Patricia J. Erwin, and Marc M. Triola. 2010. Computerized Virtual Patients in Health Professions Education: A Systematic Review and Meta-Analysis. *Academic Medicine*, 85(10):1589.
- David A. Cook, Charles P. Friedman, Peter S. Greene, Tiffany Grunwald, Grace C. Huang, S. Barry Issenberg, Carol S. Kamin, Roger Kneebone, William C. McGaghie, Geoff Norman, and Eduardo Salas. 2007. *Effective Use of Educational Technology in Medical Education: Summary Report of the 2006 AAMC Colloquium on Educational Technology*. AAMC, Washington, DC.
- David A. Cook, Jonathan Sherbino, and Steven J. Durning. 2018. Management Reasoning: Beyond the Diagnosis. *JAMA*, 319(22):2267–2268.
- David A Cook and Marc M Triola. 2009. Virtual patients: A critical literature review and proposed next steps. *Medical Education*, 43(4):303–311.
- DeepSeek. 2025a. Deepseek r1 and v3. <https://www.deepseek.com/>. Accessed: 2025-03-17.
- DeepSeek. 2025b. Reasoning model (deepseek-reasoner). [https://api-docs.deepseek.com/guides/reasoning\\_model](https://api-docs.deepseek.com/guides/reasoning_model). Accessed: 2025-03-24.
- Samuel Edelbring, Maryam Dastmalchi, Håkan Hult, Ingrid E. Lundberg, and Lars Owe Dahlgren. 2011. Experiencing virtual patients in clinical learning: A phenomenological study. *Advances in Health Sciences Education: Theory and Practice*, 16(3):331–345.
- Samuel Edelbring, Ioannis Parodis, and Ingrid E. Lundberg. 2018. Increasing Reasoning Awareness: Video Analysis of Students’ Two-Party Virtual Patient Interactions. *JMIR medical education*, 4(1):e4.
- Kevin Eva. 2005. What every teacher needs to know about clinical reasoning. *Medical education*, 39:98–106.
- Inno Flores. 2024. Fudan University Unveils Revolutionary Emotional Humanoid Robot ‘Guanghua No. 1’ at AI Conference. <https://www.techtimes.com/articles/306456/20240708/fudan-university-unveils-revolutionary-emotional-humanoid-robot-guanghua-1-ai.htm>.

- Jadon Geathers, Yann Hicke, Colleen Chan, Niroop Rajashekar, Justin Sewell, Susannah Cornes, Rene Kizilcec, and Dennis Shung. 2025. [Benchmarking Generative AI for Scoring Medical Student Interviews in Objective Structured Clinical Examinations \(OSCEs\)](#). *Preprint*, arXiv:2501.13957.
- Michail Giannakos, Roger Azevedo, Peter Brusilovsky, Mutlu Cukurova, Yannis Dimitriadis, Davinia Hernandez-Leo, Sanna Järvelä, Manolis Mavrikis, and Bart Rienties. 2024. [The promise and challenges of generative AI in education](#). *Behaviour & Information Technology*, 0(0):1–27.
- Michael A. Goodrich and Alan C. Schultz. 2008. [Human–Robot Interaction: A Survey](#). *Foundations and Trends® in Human–Computer Interaction*, 1(3):203–275.
- David Gordon, Joseph J. Rencic, Valerie J. Lang, Alik Thomas, Meredith Young, and Steven J. Durning. 2022. [Advancing the assessment of clinical reasoning across the health professions: Definitional and methodologic recommendations](#). *Perspectives on Medical Education*, 11(2):108–114.
- Linda Graf, Philipp Sykownik, Gertraud Gradl-Dietsch, and Maic Masuch. 2024. [Towards believable and educational conversations with virtual patients](#). *Frontiers in Virtual Reality*, 5.
- Owen Henkel, Libby Hills, Adam Boxer, Bill Roberts, and Zach Levonian. 2024. [Can large language models make the grade? an empirical study evaluating llms ability to mark short answer questions in k-12 education](#). In *Proceedings of the Eleventh ACM Conference on Learning @ Scale, L@S '24*, page 300–304, New York, NY, USA. Association for Computing Machinery.
- Friederike Holderried, Christian Stegemann–Philipps, Lea Herschbach, Julia–Astrid Moldt, Andrew Nevins, Jan Griewatz, Martin Holderried, Anne Herrmann–Werner, Teresa Festl–Wietek, and Moritz Mahling. 2024. [A Generative Pretrained Transformer \(GPT\)–Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study](#). *JMIR Medical Education*, 10(1):e53961.
- Grace C. Huang, Lori R. Newman, and Richard M. Schwartzstein. 2014. [Critical Thinking in Health Professions Education: Summary and Consensus Statements of the Millennium Conference 2011](#). *Teaching and Learning in Medicine*, 26(1):95–102.
- Bahar Irfan, Sanna–Mari Kuoppamäki, and Gabriel Skantze. 2023. [Between Reality and Delusion: Challenges of Applying Large Language Models to Companion Robots for Open-Domain Dialogues with Older Adult](#).
- Jinghan Jia, Abi Komma, Timothy Leffel, Xujun Peng, Ajay Nagesh, Tamer Soliman, Aram Galstyan, and Anoop Kumar. 2024. [Leveraging LLMs for Dialogue Quality Measurement](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 359–367, Mexico City, Mexico. Association for Computational Linguistics.
- Hangyeol Kang, Maher Ben Moussa, and Nadia Magnenat–Thalmann. 2024. [Nadine: An LLM-driven Intelligent Social Robot with Affective Capabilities and Human-like Memory](#). *Preprint*, arXiv:2405.20189.
- Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. [ChatGPT for good? On opportunities and challenges of large language models for education](#). *Learning and Individual Differences*, 103:102274.
- Joseph Klein. 2002. [The failure of a decision support system: Inconsistency in test grading by teachers](#). *Teaching and Teacher Education*, 18(8):1023–1033.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, pages 22199–22213, Red Hook, NY, USA. Curran Associates Inc.
- Andrzej A. Kononowicz, Nabil Zary, Samuel Edelbring, Janet Corral, and Inga Hege. 2015. [Virtual patients - what are we talking about? A framework to classify the meanings of the term in healthcare education](#). *BMC Medical Education*, 15(1):11.
- J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- William C McGaghie, S Barry Issenberg, Emil R Petrusa, and Ross J Scalese. 2010. [A critical review of simulation-based medical education research: 2003–2009](#). *Medical Education*, 44(1):50–63.
- Mary L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282.
- Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. 2023. [Large Language Models as General Pattern Machines](#). *Preprint*, arXiv:2307.04721.
- Chinmaya Mishra, Rinus Verdonchot, Peter Hagoort, and Gabriel Skantze. 2023. [Real-time emotion generation in human-robot dialogue using large language models](#). *Frontiers in Robotics and AI*, 10.
- OpenAI. 2025a. Gpt-3.5 turbo documentation. <https://platform.openai.com/docs/models/gpt-3.5-turbo>. Accessed: 2025-03-17.

- OpenAI. 2025b. Learning to reason with llms. <https://openai.com/index/learning-to-reason-with-llms/>. Accessed: 2025-03-24.
- OpenAI. 2025c. Models: Explore all available models and compare their capabilities. <https://platform.openai.com/docs/models>. Accessed: 2025-03-17.
- OpenAI. 2025d. Reasoning models: Explore advanced reasoning and problem-solving models. <https://platform.openai.com/docs/guides/reasoning?api-mode=chat>. Accessed: 2025-03-17.
- Ioannis Pantziaras, Uno Fors, and Solvig Ekblad. 2015. Training With Virtual Patients in Transcultural Psychiatry: Do the Learners Actually Learn? *Journal of Medical Internet Research*, 17(2):e3497.
- Ruth Plackett, Angelos Kassianos, Sophie Mylan, Maria Kambouri, Rosalind Raine, and Jessica Sheringham. 2022. The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: A systematic review Open Access. *BMC Medical Education*, 22.
- David M. Quinn. 2020. Experimental Evidence on Teachers' Racial Bias in Student Evaluation: The Role of Grading Scales. *Educational Evaluation and Policy Analysis*, 42(3):375–392.
- Jeremy B. Richards, Margaret M. Hayes, and Richard M. Schwartzstein. 2020. Teaching Clinical Reasoning and Critical Thinking: From Cognitive Theory to Practical Application. *CHEST*, 158(4):1617–1628.
- Theresa Ruwe and Elisabeth Mayweg-Paus. 2024. Embracing LLM Feedback: The role of feedback providers and provider information for feedback effectiveness. *Frontiers in Education*, 9.
- D. Royce Sadler. 2009. Grade integrity and the representation of academic achievement. *Studies in Higher Education*, 34(7):807–826.
- Gabriel Skantze, Martin Johansson, and Jonas Beskow. 2015. Exploring Turn-taking Cues in Multi-party Human-robot Discussions about Objects. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15*, pages 67–74, New York, NY, USA. Association for Computing Machinery.
- Tjorven Stamer, Jost Steinhäuser, and Kristina Flägel. 2023. Artificial Intelligence Supporting the Training of Communication Skills in the Education of Health Care Professions: Scoping Review. *Journal of Medical Internet Research*, 25(1):e43311.
- Ana Suárez, Alberto Adanero, Víctor Díaz-Flores García, Yolanda Freire, and Juan Algar. 2022. Using a Virtual Patient via an Artificial Intelligence Chatbot to Develop Dental Students' Diagnostic Skills. *International Journal of Environmental Research and Public Health*, 19(14):8735.
- Robin Tierney. 2015. Altered grades: A grey zone in the ethics of classroom assessment. *Assessment Matters*, 8:5–30.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Preprint*, arXiv:2201.11903.
- Joshua Wilson, Cristina Ahrendt, Emily A. Fudge, Alexandria Raiche, Gaysha Beard, and Charles MacArthur. 2021. Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education*, 168:104208.
- Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.



## A Assessment Framework for VP Mikael

### Start Questions:

- Did the student ask open-ended questions initially?
- Did the student ask about the patient's own thoughts regarding their symptoms?
- Did the student ask the patient about specific concerns/worries about their symptoms?
- Did the student ask what the patient hopes to gain from the visit?

### General Questions:

- Did the student ask if the patient has any past or current disease/worries?
- Did the student ask the patient about their lifestyle factors (such as if the patient smokes, consumes alcohol, is physically active)?
- Did the student ask if the patient has any allergies?
- Did the student ask if the patient is currently taking any medications?
- Did the student ask if the patient has any hereditary diseases in their family?
- Did the student ask if the patient had a fever?
- Did the student ask if the patient has experienced unintentional weight loss?
- Did the student ask if the patient has had night sweats?

### Specific Questions:

- Did the student inquire about muscle pain?
- Did the student inquire about the character of the muscle pain?
- Did the student inquire about the location of the muscle pain?
- Did the student inquire about whether the pain radiates?
- Did the student inquire about arm claudication?
- Did the student inquire about symmetry of the muscle pain?

- Did the student inquire about variation during the day of the symptoms?
- Did the student inquire about morning stiffness?
- Did the student inquire about how long does morning stiffness last?
- Did the student inquire about medications taken for pain and their efficacy?
- Did the student inquire about new-onset headache or tenderness in the temporal artery?
- Did the student inquire about jaw claudication?
- Did the student inquire about general malaise?
- Did the student inquire about vision changes (e.g., vision loss, double vision, blurring)?

## B Example dialogue with VP Joel

The following transcript shows the automatic speech recognition output from the user, and does not necessarily reflect what the user actually said:

**Robot:** Hi, doctor. My name is Joel and I have been troubled by bodily pains for many years but no one has been able to find the cause. Can you help me?

**User:** uh of course uh tell me more

**Robot:** I have been experiencing pain in my hands, toes, and one of my achilles tendons, and recently it has spread to my right knee.

**User:** or how long have you been in pain

**Robot:** I can't remember when it started exactly, but it's been intermittent for many years and has been gradually getting worse.

**User:** so how did it start

**Robot:** It just started slowly, I can't really pinpoint a specific event or injury that could have caused it.

**User:** uh where where do the pain start

**Robot:** The pain started in my hands and toes and then gradually spread to my achilles tendon and now my right knee.

**User:** how would you describe the pain

**Robot:** The pain feels like a burning sensation and can be quite unbearable at times.

**User:** when when uh when is the pain uh

**Robot:** The pain is worst in the morning when I first wake up, but it is present throughout the day and sometimes makes it difficult for me to sleep at night.

**User:** do you wake up from the pain

**Robot:** Yes, sometimes I wake up from the pain in my knee.

**User:** do you feel like the pain gets worse or better with exercise

**Robot:** I've noticed that sometimes the pain feels a bit better after I've exercised, but it doesn't last for very long and the pain always comes back.

**User:** have you experienced any stiffness

**Robot:** Yes, I have experienced morning stiffness, and it's usually worst when I first wake up in the morning.

**User:** is the joint pain symmetrical

**Robot:** I do believe the pain is located in my joints, but it's not symmetrical.

**User:** that is

**Robot:** I have also been feeling fatigued and tired lately, and I'm worried that my mental state is being affected by the pain.

**User:** uh have you noticed any any rashes or other skin symptoms

**Robot:** No, I haven't noticed any new rashes or lesions on my skin, but a few months back I did notice that my nails changed in character.

**User:** tell me more about that

**Robot:** I wasn't sure if it was relevant to mention, but my father and grandfather both had psoriasis, and my father had plaques on his elbows.

**User:** control

**Robot:** I have actually started to wonder if I might have psoriasis too and was interested in investigating that further.

**User:** yeah we will look into that do you have any other uh illnesses or

**Robot:** I did have an irritable bowel when I was younger, but it got better after I started to diet and exercise better.

**User:** any medications

**Robot:** Yes, I currently take paracetamol for my pain, Propavane to sleep at night, and Escitalopram for my depression.

**User:** do you drink alcohol

**Robot:** No, I don't drink alcohol often, maybe once or twice a year on special occasions.

**User:** and how about smoking

**Robot:** Actually, I tried smoking a couple of times during my college years, but that was a long time ago and I haven't touched a cigarette since then.

**User:** okay uh goodbye

**Robot:** Goodbye Doctor