

Original Paper

Test-Retest Reliability of Large Language Models and Their Concordance With Human Experts in Rubric-Based Scoring of Clinical Summary Statements: Experimental Comparative Study

Patrick Stadler¹, MSc; Piotr Bucichowski¹, MD; Ylva Holzhausen¹, Dr phil; Inga Hege², MD, MCompSc; Harm Peters¹, MD, MHPE; Julius Josef Kaminski¹, MD, MHPE

¹Charité – University Medical Center Berlin, Dieter Scheffner Center for Medical Education and Educational Research, Dean's Office for Study Affairs, Berlin, Germany

²Faculty of Health Sciences Brandenburg, Institute of Research in Health Sciences Education, Brandenburg Medical School, Neuruppin, Germany

Corresponding Author:

Harm Peters, MD, MHPE

Charité – University Medical Center Berlin, Dieter Scheffner Center for Medical Education and Educational Research
Dean's Office for Study Affairs

Charitéplatz 1

Berlin 10117

Germany

Phone: 49 30450576207

Email: harm.peters@charite.de

Abstract

Background: Clinical summary statements convert unstructured patient information into a concise format. They are widely used in virtual patients (VPs) to train medical students in diagnostic reasoning. Most VP systems provide little individualized feedback on the quality of learners' summaries, and earlier automated approaches have required complex natural language processing, limiting their scalability. Large language models (LLMs) may offer a more flexible, rubric-guided way to score summary statements and support feedback. Still, their test-retest reliability and concordance with expert ratings in this context remain unclear.

Objective: This study aims to evaluate the test-retest reliability—specifically, whether the models produce stable ratings across repetitions—of LLM-generated medical summary statement ratings, and to measure their concordance with human expert assessments across different prompting strategies.

Methods: We conducted an experimental comparative study using 122 rubric-scored summary statements (English and German) from 6 internal medicine VP cases previously rated by 2 expert raters. Statements were evaluated along 4 ternary components (semantic qualifiers, transformation, narrowing, and global rating), and 2 binary components (factual accuracy and person). Four LLMs (Anthropic's Claude, OpenAI's GPT-4o, Mistral AI's Mistral Large, and Mistral AI's Mistral 7B) were each run 5 times under 4 prompting strategies (zero-shot, zero-shot with chain-of-thought [CoT], few-shot, and few-shot+CoT). The test-retest reliability was assessed with intraclass correlation coefficients (ICCs), Fleiss kappa (κ), and concordance with human ratings with accuracy (%) and chance-corrected interrater agreement (Gwet AC1 and AC2).

Results: Large closed-weights models (Claude, GPT-4o, and Mistral Large) showed high test-retest reliability (ICC $\geq$ 0.92; Fleiss κ =0.59-1.00). Claude was essentially deterministic (ICC and Fleiss κ $\geq$ 0.99); GPT-4o and Mistral Large reached ICC $\geq$ 0.97 and Fleiss κ $\geq$ 0.80 except under zero-shot. Mistral 7B, a small open-weights model, had high ICCs ($\geq$ 0.88) but low Fleiss κ (0.32-0.52). Few-shot+CoT prompting achieved the best concordance with human ratings for all the models (means across runs and rubric components), peaking for Mistral Large (accuracy 77%; Gwet AC1 and AC2 0.85, 95% CI 0.83-0.87), followed by Claude (71%; 0.77, 0.74-0.81), GPT-4o (67%; 0.72, 0.68-0.77), and Mistral 7B (60%; 0.64, 0.56-0.73). Concordance was higher for the binary rubric components (person and factual accuracy) than for the ternary components. Extreme deviations and invalid outputs were rare but more frequent in ternary components and the smaller local model.

Conclusions: Large, general-purpose LLMs can function as reliable, rubric-guided raters of VP-based clinical summary statements, achieving high concordance with expert ratings under clearly defined criteria in binary rubric components and when using few-shot with CoT prompting. The performance on more nuanced, ternary reasoning dimensions was more variable. LLMs can support scalable, rubric-based formative feedback on clinical reasoning within VP systems but should complement rather than replace human judgment.

JMIR Med Educ 2026;12:e94666; doi: [10.2196/94666](https://doi.org/10.2196/94666)

Keywords: medical education; large language models; automated assessment; formative feedback; clinical reasoning; virtual patients; summary statements; problem representation

Introduction

Clinical summary statements are a key element in physicians' clinical reasoning. As concise, structured narratives, they facilitate transforming unstructured patient information into an initial diagnostic interpretation and considering relevant underlying differential diagnoses [1,2]. They make observable implicit parts of the reasoning processes that underpin diagnostic processes, and over time, they can be used to evidence the development of clinical reasoning competence. For example, rather than simply noting that "Mr. Miller came to the ER with several chest pain episodes during the last 2 hours," a more advanced student would write, "Mr. Miller is a 62-year-old male with cardiovascular risk factors presenting with acute, substernal chest pressure radiating to the left arm, associated with diaphoresis, and relieved by rest." This transformation from descriptive data to diagnostic framing exemplifies a key cognitive step in diagnostic reasoning [3]. However, evaluating the quality of these summary statements reliably and at scale remains a challenge in medical education, as many learners simultaneously need individualized, timely feedback.

Clinical summary statements are also referred to as problem representations or one-liners. Depending on the context, they can take various forms and serve different functions, such as a brief patient handover at the end of one's shift or a more extensive case presentation at a clinical conference [4]. For this paper, we refer to clinical summary statements as extracts of patient information formulated by students during their learning process. Febowitz et al [4] suggested a conceptual model that explains any clinical information summary as a sequence of 5 steps: (1) aggregation and (2) organization of clinical information, (3) reduction and transformation, (4) interpretation, and (5) synthesis. In educational research on summary statements, particular attention has been given to the last 3 steps of the sequence. Well-developed and effective summary statements include both positive and negative key findings, use semantic qualifiers (SQs) (eg, acute, radiating, and substernal), and thereby contextualize the presentation within epidemiological and demographic parameters; by doing so, they offer synthesized, essential case information within an abstracted, diagnostic framework [5,6]. Accordingly, assessing summary statements extends beyond assessing the factual correctness of the information presented by the learner. Rather, the evaluation should include how effectively learners structure information and attribute meaning to it. By sharing their interpretations and abstractions of patient information in

summary statements, learners offer insight into how well they can integrate specific patient data into their illness scripts (ie, the knowledge of illnesses represented in their long-term memory [7]), the accuracy of their illness scripts, and how this process facilitates their overall clinical reasoning [1,8]. Diagnostic accuracy depends on these processes and can therefore be considered a summarizing indicator of clinical expertise [9].

Virtual patients (VPs) have become an established approach for medical students to learn clinical reasoning. These interactive, computer-based simulations allow them to train diagnostic processes in a safe, standardized, and scalable environment [10]. VPs enable repeated exposure to varied and even rare presentations and facilitate reflective, self-paced learning. Despite these advantages, providing students with reliable, individualized feedback on reasoning quality remains challenging. A recent review by Jay et al [11] revealed that among studies with learner feedback in VP systems, only a small fraction addressed feedback on the quality of clinical summary statements, and even fewer operationalized them as formative feedback for learning. This gap is relevant because feedback is a primary driver of learning in health professions education [12,13]. This points to a missed opportunity for medical students to learn in current educational approaches.

To address the gap in feedback provision, first VP studies with efforts to automate the evaluation of summary statements showed promising results: Hege et al [14] implemented a rule-based natural language processing (NLP) system grounded in a rubric by Smith et al [6], which assessed the following domains: SQ, transformation of information into higher-order concepts, diagnostic narrowing, factual accuracy, and overall quality in summary statements. Although the system achieved moderate agreement with medical expert ratings, extensive task-specific engineering, language-dependent dictionaries, and substantial training data were needed. Each new VP case demanded additional manual development, limiting scalability and adaptability. More fundamentally, rule-based systems sometimes struggled to capture the contextual nuances of clinical reasoning. Moreover, they can fail because of simple vocabulary mismatches (cf [15]), underscoring the need for approaches that interpret the meaning of summary statements rather than relying solely on pattern matching.

Recent advances in large language models (LLMs) offer a fundamentally different approach to the challenges encountered in many traditional rule-based NLP systems. LLMs

process text through learned representations of semantic and contextual relationships, enabling nuanced interpretation without handcrafted rules [16,17]. LLMs also demonstrate better handling of inferences, the transformation of information, or different wording or phrasing that often fail with plain lexical lookups [18]. Their ability to generalize across content domains suggests the potential for scalable, context-aware evaluation of learner input. Nevertheless, the performance of LLMs depends on how the models are instructed to perform a task by prompting. Prompting refers to how instructions are structured to guide LLM behavior. Prompts are typically structured into (1) task instructions, (2) contextual information such as item stems or rubrics, (3) optional examples, and (4) an output specification. Well-thought-out prompt designs—particularly regarding the inclusion and phrasing of rubrics—have been shown to affect both the accuracy and reliability of automated scoring outcomes in medical education settings such as in Objective Structured Clinical Examination (OSCE) grading [19-21]. In zero-shot prompting, the model receives only task instructions (eg, “Score this text using the rubric and the scoring instructions below”) without examples. Few-shot prompting extends this by including a few examples of the task, allowing the model to infer the scoring logic from demonstration (eg, text examples paired with rubric scores given by raters). Chain-of-Thought (CoT) prompting, which adds a reasoning step to either few-shot or zero-shot prompting, can be combined with few-shot prompting to provide the LLM with human raters’ rationales for some exemplary human ratings (Methods section).

Despite emerging and promising evidence from automated assessments, no study has yet systematically examined the test-retest reliability of LLMs and their concordance with human expert ratings in evaluating clinical summary statements. Understanding whether LLMs can apply established rubrics is essential for determining whether using them for formative assessment in diagnostic-reasoning education is feasible. An LLM-based evaluation could help address the feedback gap identified by Jay et al [11] in current VP systems. Instead of relying on occasional, labor-intensive assessments by clinical teachers, learners could benefit from repeated VP-based practice, with LLMs providing immediate, consistent, expert-level feedback on their summary statements throughout medical training, thereby strengthening their clinical reasoning skills.

The aim of this study is to evaluate the performance of 4 transformer architecture-based, widely available LLMs with multilingual support in a rubric-guided evaluation of an existing set of clinical summary statements written by medical students during their interactions with online VPs.

Specifically, we formulate the following 2 research questions (RQs):

1. What is the test-retest reliability of LLM-generated medical summary statement ratings across 5 independent runs?

2. What is the concordance of LLM ratings with human expert ratings under different prompting configurations?

Methods

Study Design

Using an experimental comparative design, this study examined the test-retest reliability of LLMs in rubric-based ratings of medical summary statements under multiple prompting conditions and their concordance with human expert ratings.

Because LLM performance is highly dependent on the prompting techniques chosen [22], 4 prompting strategies (zero-shot and few-shot prompting, each with and without CoT) were applied. This resulted in a total of 16 experimental conditions across the 4 LLMs (4 prompting configurations × 4 LLMs). To better compare the results with those of the previous work on the automated evaluation of summary statements by Hege et al [14], the same summary statements from their study were used, as well as the evaluation rubric by Smith et al [6], which was annotated for LLM use (refer to the Rating Rubric and Prompt Development section).

Each of these combinations of models, prompt configurations (eg, zero-shot), and summary statements was run 5 times.

Models and Infrastructure

Four off-the-shelf general-purpose LLMs were tested: 3 large commercial models (Anthropic’s Claude Sonnet 4, OpenAI’s GPT-4o, and Mistral AI’s Mistral Large) and one smaller local model (Mistral AI’s Mistral 7B). The models were selected purposefully to cover the range of deployment options at the time of the study: 3 widely used high-performing commercial models from different vendors and a smaller open-weight model that can be deployed locally. Specialized domain-specific models for medical purposes, such as Google’s Med-PaLM [23] and MedAlpaca [24], or BERT (Bidirectional Encoder Representations from Transformers)-like models, such as MedBERT [25], exist. However, these models are currently either less accessible to the public and/or not mature or stable enough for reliable or automated assessment, which were 2 key research aspects of this work; thus, these specialized models were not included in our study. The local Mistral 7B model was obtained from Hugging Face, a community-driven website that shares language models and tools. We used an activation-aware weight quantization (AWQ) variant of Mistral 7B, a compressed version that is optimized for lower hardware profiles while still keeping its core features [26].

An overview of the models and specifications is provided in Table 1.

Table 1. Overview of the various LLMs^a that were used and compared in our study.

Provider	Model	Exact version
Anthropic	Claude Sonnet 4	claude-sonnet-4-20250514
OpenAI	GPT-4o	gpt-4o-2024-05-13
Mistral AI	Mistral Large	mistral-large-2411
Mistral AI	Mistral 7B	TheBloke/Mistral-7B-Instruct-v0.2-AWQ (via Hugging Face)

^aLLMs: large language models.

Data Source

The dataset from Hege et al [14] comprising 125 rubric-scored summary statements in English (n=84) and German (n=41) was used for this study. The summary statements were written by medical students in response to 6 human-written virtual patient cases in internal medicine from open-access clinical reasoning courses in Instruct's CASUS [27] VP system. They varied in length and quality, ranging from near-perfect to incomplete summaries. In a study by Hege et al [14], the statements were scored independently by 2 health care professionals with an educational background and expertise in clinical reasoning. The ratings of the statements were based on an elaborated rubric by Smith et al [6]. To calibrate their use of the rubric, the 2 experts independently rated and discussed 25 sample summaries before rating the full set of 125. Divergent scores were discussed and rating consensus was reached for all summary statements [14].

For this study, 3 of the 125 summary statements available were excluded from the current analysis because they were used as examples in few-shot prompting conditions, leaving 122 statements for evaluation.

Rating Rubric and Prompt Development

For systematic evaluation, we applied the rubric developed by Smith et al [6] as adapted by Hege et al [14], an established assessment tool with prevalidated quantitative scoring scales, consisting of binary and ternary items. This instrument was used directly as our rating standard and for developing corresponding prompt instructions. The rubric comprises 6 criteria (components): (1) SQ, (2) transformation of information, (3) narrowing of differential diagnosis, (4)

factual accuracy, (5) patient name or person (added by Hege et al [14]), and (6) a global score (a separate, subjective overall score). The rubric components no. 4 and 5 were binary (scoring: 0/1), whereas the rest were ternary (with 3 possible scorings: 0, 1, or 2).

Four prompting strategies were implemented and assessed, following established LLM practices from the prompting literature [19,22,28]:

1. Zero-shot, no CoT
2. Zero-shot with CoT
3. Few-shot, no CoT
4. Few-shot with CoT

Zero-shot prompting relied on the model's pretrained knowledge without rating examples, whereas in few-shot prompting, we included 3 manually picked summary statements of our dataset and annotated them with human expert ratings, to serve as rating examples representative for our dataset. They were chosen specifically to exemplify all possible proficiency levels (0/1 in binary and 0/1/2 in ternary rubric components) across 3 different VP cases: 2 in English and 1 in German. CoT prompting encouraged stepwise reasoning by appending the phrase "Let's think step by step" (zero-shot CoT). In few-shot prompting, CoT was additionally supported by guided step-by-step annotations on how to score provided examples [19]. On the basis of the prompting condition, different prompt segments were included, as summarized in Table 2. The exact prompt texts in Markdown format are provided in Multimedia Appendix 1. All additional data, such as the exact case vignettes, can be found in our GitHub repository (Data Availability section).

Table 2. Composition of each prompt depending on the prompting condition. The prompts contained different prompt segments.

Prompt segment	Zero-shot	Zero-shot CoT ^a	Few-shot	Few-shot CoT
Role description	✓ ^b	✓	✓	✓
Scoring rubric	✓	✓	✓	✓
VP ^c case	✓	✓	✓	✓
Three scoring examples	◦	◦ ^d	✓	✓
Case information for exemplary human case scorings	◦	◦	✓	✓
Rationales behind examples	◦	◦	◦	✓
Student response placeholder where summary statement was inserted	✓	✓	✓	✓
CoT initiator	◦	✓	◦	✓

^aCoT: chain-of-thought.

^b✓ denotes a segment included in prompt

^cVP: virtual patient.

^d◦ denotes segment not included in prompt.

The prompts were iteratively refined in the process as described by Lee et al [19]:

The rubric content and assessment rationales were formatted as a prompt and then reviewed by the study authors, including experts in educational research and AI (PB, PS, and IH). This resulted in a more detailed description of SQ based on Chang et al [1] and minor linguistic changes (eg, case vignettes → clinical cases). For the initial iterative test runs, the LLMs were prompted to provide explanatory ratings (ie, rationales) to identify cardinal ambiguities, which led to further prompt refinement (PB, PS). This optimization was performed mainly with Mistral 7B, the smallest model (ie, fewest parameters) in our study, to ensure robustness across models, under the assumption that larger models would properly process prompts modified to work with the smallest model [29]. Explanations provided by LLMs generally matched given scores, eg, when the patient name was scored with 1, the model would provide an adequate justification that the statement included the patient name. During this step, we identified that insufficient formatting led Mistral 7B and Claude Sonnet 4 to systematically confuse VP cases with the summary statement to be rated, which resulted in increased use of special characters to provide a clear structure in crucial areas of the prompt (such as the use of brackets and delimiters). Based on the ratings of summary statements excluded from the analysis, we refined the description of the scoring rubric (PB, PS, and JJK). Lastly, the final step of prompt development was to instruct all models to solely output the raw scores without further rationales.

A comparative overview of the original scoring rubric and the interpretative information and rationales that we added to minimize ambiguity for LLM interpretation are presented in [Multimedia Appendix 2](#).

Procedure

The experiments were conducted between July and September 2025 using the latest available models at that time. The 5 runs per model and prompting condition were executed consecutively, and a new run automatically started as soon as the previous run completed, resulting in a narrow time window of around 1 to 2 hours in which the data were collected per experiment condition, depending on the model and prompting condition.

The local Mistral 7B model was deployed using vLLM [30] and was run on a single NVIDIA Tesla T4 GPU via our institution's compute cloud. All other models (Claude, GPT-4o, and Mistral Large) were accessed via their official APIs using the same LangChain setup to ensure comparability between the commercial online models and the local model. No conversation memory was implemented, preventing learning effects between reruns of ratings. Each of the 122 summary statements was rated under all 16 experimental conditions across 6 rubric components, which was repeated 5 times, yielding a total of 58,560 data points collected as scores across all 6 rubric components.

We implemented custom software to automate the process using the Python programming language (Python Software Foundation) with the LangChain v0.3.26 package, a software abstraction that handles LLM inference. For each model vendor, we used the respective LangChain submodule to access the APIs. To ensure data integrity, we formulated a strict data format and passed each LLM call to a structured output function that automatically ensures that the LLMs' responses follow that format. Our custom software procedure included (1) resolving all placeholders in the prompt with the data for each individual run (the summary statement to be rated and the virtual patient case vignette; 3 examples for few-shot conditions; annotations for the examples, and a CoT reasoning initiator for CoT conditions), (2) sending the complete prompt to the LLM API via LangChain, (3) obtaining the raw results from the LLM API, (4) transforming the data into a structured format, and (5) adding the data to a CSV file for later statistical analysis. Temperature, a model-specific hyperparameter, was set to zero for all models for more deterministic results. It should be noted, though, that even a temperature of zero does not guarantee fully deterministic results when working with closed APIs [31]. This is also captured further in the Results section (see RQ1 for test-retest reliability), where we report on the test-retest reliability between consecutive runs.

Our GitHub repository includes the exact code, prompts, scoring materials, and original human reference ratings to replicate the study (see Data Availability section).

Data Analysis

Statistical analysis was performed using R v4.5.0 [32]. The complete R and Python codes used for this study can be found in our GitHub repository (see the Data Availability section).

We analyzed the resulting dataset from 2 complementary perspectives.

To address RQ1, we assessed test-retest reliability by examining whether the models produced consistent and stable ratings across the 5 run repetitions, measured as (1) consistency using intraclass correlation coefficients (ICCs) and (2) intrarater agreement using Fleiss κ (effectively treating each run as an independent measurement by the same LLM rater). The ICC was calculated using a one-way random-effects model based on the average of 5 raters according to [33]. By examining the consistency of the 5 aggregated runs, this approach was chosen to examine whether random model variances affect a model's scoring behavior, even when the API temperature was set to zero. The ICC is used to capture consistency in numeric scoring, whereas Fleiss κ is designed for exact categorical agreement. Both complement each other well in our setup: the meaning of score distances is not entirely uniform across rubric components (eg, a change from 0 to 1 does not necessarily equal a change from 1 to 2); thus, both numeric consistency and categorical agreement perspectives provide information on the degree of systematic bias.

To address RQ2, we assessed concordance with human expert ratings using 2 metrics: (1) accuracy (percentage of exact LLM-human match agreement) and (2) interrater agreement measured using Gwet AC1 and AC2 coefficients (chance-corrected indices) for comparison with prior work. Following the recommendations for robust agreement assessment, we prioritized Gwet AC1 and AC2 as our primary statistic, as suggested by Xu and Lorber [34]. AC1 is a suitable metric for binary (categorical) rubric components, whereas AC2 provides a measure for ordinal components, in our case, ternary rubric components [35,36]. Conventional indices such as Cohen kappa and Krippendorff α can underestimate agreement in cases where the distribution of ratings is imbalanced (eg, when most statements are factually correct, as in our dataset). This prevalence problem is well documented and can lead to artificially low reliability scores despite strong observed agreement. Gwet AC1 and AC2 correct for this bias [35,37] and are therefore considered more robust for skewed datasets [38,39]. For each combination of model, prompting configuration, and rubric component, Gwet AC1 and AC2 and weighted Cohen kappa coefficients were calculated separately for each of the 5 runs. For results per rubric component, we report the mean, SD, and SE across these 5 runs. For overall results per model and prompting configuration, we report the mean, SD, and SE across all 30 coefficients (6 rubric components $\times$ 5 runs). The 95% CIs were calculated using a Wald-type normal approximation (mean $\pm$ 1.96 $\times$ SE).

In accordance with prior research, we applied the Landis and Koch scale [40] to interpret the raw Gwet AC1 and AC2 values. However, this interpretation needs to be treated with some care, as Gwet AC1 and AC2 tend to produce higher

values than Cohen kappa does [37]. To allow a full comparison between all the models, prompting conditions, and rubric components, we included the full values of Gwet AC1 and AC2 and weighted Cohen kappa—for a more complete comparison with related work—over all 5 runs in [Multimedia Appendices 3 and 4](#).

Ethical Considerations

The secondary data analysis conducted in this study was approved by the Ethics Committee of the Ludwig-Maximilians-Universität München, Germany (reference number 260–15), with the primary data having been generated by Hege et al [14]. An additional informed consent was not required for the anonymous analysis of the data. The virtual patients provided to the learners were fictitious and not based on real individuals.

Results

RQ1: Test-Retest Reliability

The ICCs indicate high consistency across 5 repeated runs for all LLMs (ICC $\geq$ 0.88), with varying levels of intrarater agreement depending on the model and prompting condition (Fleiss κ = 0.32–1.00), as shown in [Table 3](#).

[Figure 1](#) shows the deviations between model and human ratings, as well as the number of invalid outputs (eg, numerical ratings outside the rubric scale), for the first of the 5 runs. The maximum possible deviation was ± 2 points for ternary components (0–2 scale) and ± 1 point for binary components (0–1 scale). Enlarged versions of the panels of [Figure 1](#) can be found in the [Multimedia Appendices 5–8](#).

Table 3. ICC^a and Fleiss κ for all the model $\times$ prompt configurations. Claude was fully stable across all runs, whereas Mistral 7B had weaker Fleiss κ values despite a high ICC.

Configuration	ICC	Fleiss κ
Mistral Large zero-shot	0.92	0.66
Mistral Large few-shot	0.97	0.80
Mistral Large zero-shot CoT ^b	0.98	0.91
Mistral Large few-shot CoT	0.97	0.80
Claude zero-shot	1.00	1.00
Claude few-shot	0.99	0.99
Claude zero-shot CoT	0.99	0.99
Claude few-shot CoT	0.99	0.99
GPT-4o zero-shot	0.92	0.59
GPT-4o few-shot	0.99	0.94
GPT-4o zero-shot CoT	0.99	0.94
GPT-4o few-shot CoT	0.99	0.92
Mistral 7B zero-shot	0.88	0.32
Mistral 7B few-shot	0.91	0.43
Mistral 7B zero-shot CoT	0.89	0.39
Mistral 7B few-shot CoT	0.92	0.52

^aICC: intraclass correlation coefficient.

^bCoT: chain-of-thought.

Figure 1. Predicted large language model ratings per model, prompting configuration, and rubric component. The stacked bars show the absolute number of equal (same) ratings, as well as ratings that were lower by 1 or 2 or higher by 1 or 2 compared with the human reference. The results of the first of the 5 runs per model and configuration were used for this figure. We have included enlarged versions of the individual panels from this figure in [Multimedia Appendices 5-8](#) to make them easier to read. CoT: chain-of-thought; LLM: large language model.



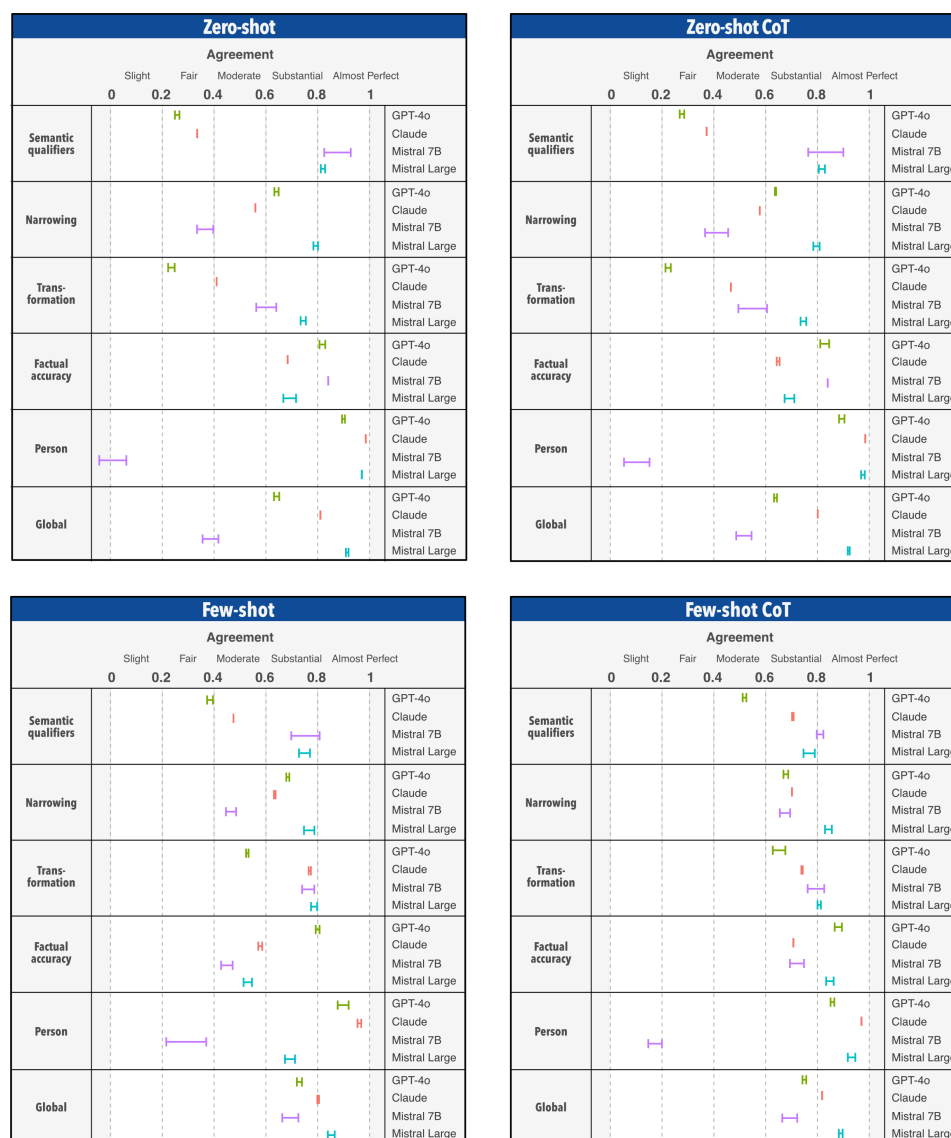
Claude was the most reliable: zero-shot runs were essentially deterministic ($ICC=1.00$, Fleiss $\kappa=1.00$), whereas all other conditions were near-deterministic ($ICC \geq 0.99$, Fleiss $\kappa \geq 0.99$). For the other models, zero-shot runs yielded the lowest ICC and Fleiss κ values. GPT-4o (excluding zero-shot) achieved very high test-retest reliability ($ICC \geq 0.99$ and Fleiss $\kappa \geq 0.92$). Mistral Large (also excluding zero-shot) showed lower values but nevertheless high consistency and intrarater agreement ($ICC \geq 0.97$, Fleiss $\kappa \geq 0.80$).

In zero-shot settings, both GPT-4o and Mistral Large showed high consistency ($ICC=0.92$) and moderate to substantial intrarater agreement (Fleiss $\kappa=0.59$ and $=0.66$, respectively). The lowest values were reported for Mistral 7B, with high ICC (≥ 0.88) but only fair-to-moderate intrarater agreement (Fleiss $\kappa=0.32-0.52$).

RQ2: Concordance of LLM Ratings With Human Experts—Interrater Agreement (Gwet AC1 and AC2)

In our study, no single LLM-prompt configuration achieved superior agreement across all the rubric criteria. [Figure 2](#) visualizes Gwet AC1 and AC2 by model, rubric, and prompting techniques. While simple zero-shot prompting achieved adequate agreement on specific rubric components, it failed to perform consistently across the entire rubric. Supplementing models with few-shot examples and/or CoT prompting was required to achieve consistent, substantial to almost-perfect agreement across most rubric components.

Figure 2. Agreement values according to Gwet AC1 and AC2 by rubric component, prompting configuration, and model. Binary components (person, factual accuracy) generally reached higher reliability than abstract ternary components (semantic qualifiers, transformation, narrowing, and global). Agreement values are interpreted using the Landis and Koch scale [40], which ranges from slight to almost perfect agreement. CoT: chain-of-thought.



To comprehensively evaluate how LLM selection, prompting configuration, and the rubric component impact performance, we conducted a linear mixed-effects model followed by a Type III ANOVA using R's *lme4* and *lmerTest* packages. To control for the nonindependence of repeated runs, we included the unique identifier for each individual batch (combining model, prompt configuration, and run index) as a random intercept. Satterthwaite method was subsequently applied to approximate the denominator degrees of freedom for the fixed effects. The analysis evaluated run-level Gwet AC1 and AC2 coefficients across all 5 runs per experimental condition. The analysis revealed significant main effects for model ($F_{3, 64} = 2178.01, P < .001$), prompting configuration ($F_{3, 64} = 538.21, P < .001$), and rubric component ($F_{5, 320} = 714.33, P < .001$). Crucially, all 2-way

and 3-way interaction effects were also significant ($P < .001$), most notably the model $\times$ rubric component interaction ($F_{15, 320} = 1717.72, P < .001$). This indicates that while model choice and prompt configuration dictate overall performance, their efficacy is deeply tied to the specific nature of the rubric component being evaluated.

Differences by Model and Prompting Configuration

The concordance with human experts varied greatly depending on the selected model and even more for the prompt configuration. The mean Gwet AC1 and AC2 coefficients per model and prompting configuration are reported in Table 4.

Table 4. Gwet AC1 and AC2 interrater agreement coefficients for the 4 LLMs^a across 4 prompting configurations (zero-shot and few-shot, with or without CoT^b). For each model and prompting configuration, Gwet AC1 (binary components) or AC2 (ternary components) was calculated separately for each of the 6 rubric components in each of the 5 runs. The table shows the mean, SD, SE, and 95% CI of these 30 coefficients. The SD, SE, and 95% CI mainly reflect differences in agreement between rubric components.

Model and configuration	Mean (SD) ^c	SE ^c	95% CI ^c
Claude			
Few-shot	0.70 (0.16)	0.03	0.64-0.76
Few-shot CoT	0.77 (0.10)	0.02	0.74-0.81
Zero-shot	0.63 (0.23)	0.04	0.54-0.72
Zero-shot CoT	0.64 (0.21)	0.04	0.56-0.72
GPT-4o			
Few-shot	0.67 (0.17)	0.03	0.61-0.74
Few-shot CoT	0.72 (0.13)	0.02	0.68-0.77
Zero-shot	0.58 (0.26)	0.05	0.49-0.68
Zero-shot CoT	0.58 (0.26)	0.05	0.49-0.68
Mistral Large			
Few-shot	0.73 (0.10)	0.02	0.69-0.77
Few-shot CoT	0.85 (0.06)	0.01	0.83-0.87
Zero-shot	0.82 (0.10)	0.02	0.79-0.86
Zero-shot CoT	0.82 (0.10)	0.02	0.79-0.86
Mistral 7B			
Few-shot	0.57 (0.19)	0.03	0.5-0.64
Few-shot CoT	0.64 (0.22)	0.04	0.56-0.73
Zero-shot	0.52 (0.30)	0.05	0.41-0.63
Zero-shot CoT	0.54 (0.26)	0.05	0.44-0.64

^aLLM: large language model.

^bCoT: chain-of-thought.

^cGwet AC1 and AC2.

Mistral Large showed the best overall performance. Under optimal prompting, it reached almost perfect agreement on Person (AC1=0.976, 95% CI 0.969-0.983 in zero-shot CoT), Global (AC2=0.921, 95% CI 0.917-0.925 in zero-shot CoT), Narrowing (AC2=0.842, 95% CI 0.829-0.855 in few-shot CoT), SQ (AC2=0.820, 95% CI 0.811-0.828 in zero-shot), and transformation (AC2=0.807, 95% CI 0.800-0.813 in few-shot CoT). It was less effective for the binary component factual accuracy, with the lowest agreement in few-shot (AC1=0.530, 95% CI 0.514-0.546), but under few-shot CoT, it increased to 0.848 (95% CI 0.833-0.862). Interestingly, compared with zero-shot prompting, few-shot prompting reduced agreement, but this effect was mitigated when few-shot prompting was supplemented with CoT.

GPT-4o excelled in the binary rubric components, with factual accuracy peaking at AC1=0.880 (95% CI 0.866-0.894) in few-shot CoT and consistently high agreement on Person (lowest AC1=0.858, 95% CI 0.850-0.865 in few-shot CoT, highest 0.899, 95% CI 0.894-0.904 in zero-shot). However, it struggled with 2 of the 4 ternary categories in zero-shot: SQ (AC2=0.257, 95% CI 0.248-0.266) and transformation (AC2=0.235, 95% CI 0.222-0.247), corresponding to fair agreement. Only the combination of few-shot and CoT markedly improved these components (AC2=0.518, 95% CI 0.511-0.525, and AC2=0.652, 95% CI 0.627-0.676, respectively). Notably, under zero-shot, CoT had little effect across the whole rubric.

Claude reached the highest agreement in person across all prompting techniques (lowest AC1=0.961, 95% CI 0.954-0.968 in few-shot CoT; highest AC1=0.985, 95% CI 0.985-0.985 in both zero-shot and zero-shot CoT). It also achieved almost perfect agreement globally across all prompting conditions (lowest AC2=0.801, 95% CI 0.801-0.801 in zero-shot CoT; highest AC2=0.818, 95% CI 0.818-0.818 in few-shot CoT). In the other ternary categories, Claude reached fair to moderate agreement in zero-shot; however, a combination of few-shot and CoT yielded substantial gains in agreement: SQ increased from AC2=0.334 (95% CI 0.334-0.334) to 0.705 (95% CI 0.702-0.708), Transformation from 0.409 (95% CI 0.409-0.409) to 0.740 (95% CI 0.738-0.743) and narrowing from 0.559 (95% CI 0.559-0.559) to 0.701 (95% CI 0.701-0.701). CoT in zero-shot, as in GPT-4o, seemed ineffective except for minimal gains in the rubric component person.

Mistral 7B Local showed the greatest variance among all the tested models. In zero-shot, it performed well in SQ, achieving almost perfect agreement (AC2=0.876; 95% CI 0.825-0.927); however, at the same time, person collapsed to near chance (AC1=0.037; 95% CI -0.016 to 0.089). Prompting did not improve Mistral 7B's performance in person. Factual accuracy was highly dependent on the configuration: weak in few-shot (AC1=0.449, 95% CI 0.427-0.472) but strong in zero-shot and zero-shot CoT (in both AC1=0.840, 95% CI 0.840-0.840). Narrowing and

transformation benefitted from the combination of few-shot and CoT, reaching AC2=0.675 (95% CI 0.655-0.694) and 0.794 (95% CI 0.762-0.826), respectively. In Global, few-shot improved agreement independently of the presence of CoT.

Differences by Rubric Component

The binary components usually showed the highest agreement among the rubric components. The best agreement on person was achieved in Claude zero-shot and zero-shot CoT (both ex aequo with deterministic output at AC1=0.985, 95% CI 0.985-0.985), followed closely by Mistral Large under zero-shot CoT prompting (AC1=0.976, 95% CI 0.969-0.983). Factual Accuracy peaked with GPT-4o in few-shot CoT (AC1=0.880, 95% CI 0.866-0.894). Notably, Mistral 7B Local struggled in the person category and produced outliers, failing to achieve even moderate agreement across all

prompting strategies (AC1≤0.292). Ternary rubric components led to more variable results. Global ratings were best with Mistral Large in zero-shot CoT (AC2=0.921, 95% CI 0.917-0.925); however, in terms of this component, Claude also consistently reached almost perfect agreement (the lowest AC2=0.801, 95% CI 0.801-0.801). For narrowing, Mistral Large few-shot CoT performed best (AC2=0.842, 95% CI 0.829-0.855). The agreement on the SQ was high among both Mistral models, with Mistral 7B Local in zero-shot reaching the highest AC2=0.876 (95% CI 0.825-0.927). Transformation was most reliably scored by Mistral Large in few-shot CoT, with AC2=0.807 (95% CI 0.800-0.813). Interestingly, for most models, transformation consistently benefitted from few-shot prompting but not from CoT. We summarize the best model and prompting configuration per rubric component in Table 5.

Table 5. Highest interrater agreement by rubric component. The best model × prompt configuration per rubric component was selected according to Gwet AC1 in binary and AC2 in ternary rubric components. Claude showed the same performance in zero-shot with and without CoT^a in the person rubric category.

Rubric component	Model	Configuration	Mean (SD)	SE	95% CI
Binary (AC1)					
Factual accuracy	GPT-4o	Few-shot CoT	0.880 (0.016)	0.007	0.866-0.894
Person	Claude	Zero-shot CoT	0.985 (0.000)	0.000	0.985-0.985
Person	Claude	Zero-shot	0.985 (0.000)	0.000	0.985-0.985
Ternary (AC2)					
SQs ^b	Mistral 7B	Zero-shot	0.876 (0.058)	0.026	0.825-0.927
Narrowing	Mistral Large	Few-shot CoT	0.842 (0.016)	0.007	0.829-0.855
Transformation	Mistral Large	Few-shot CoT	0.807 (0.007)	0.003	0.800-0.813
Global	Mistral Large	Zero-shot CoT	0.921 (0.004)	0.002	0.917-0.925

^aCoT: chain-of-thought.

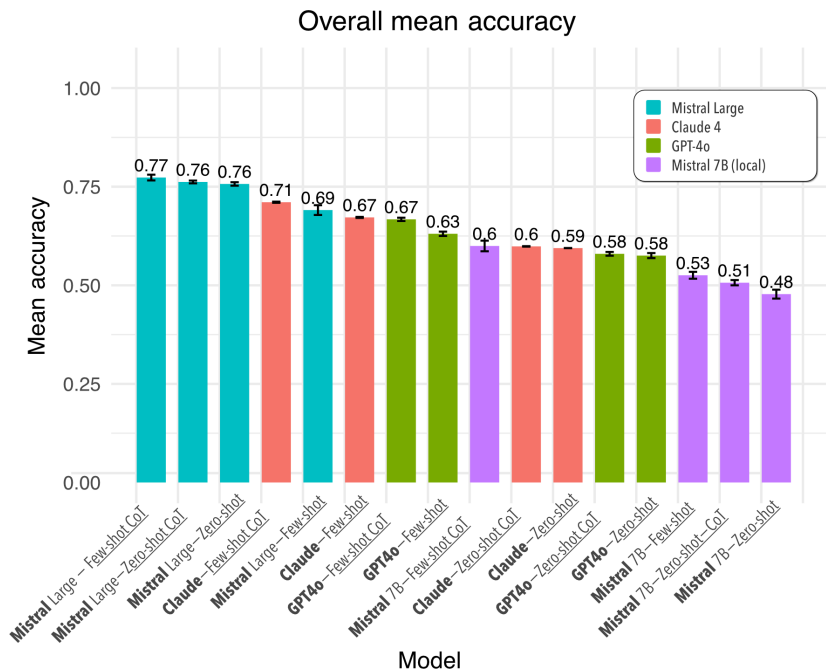
^bSQs: semantic qualifiers.

Exact-Match Agreement (Accuracy) and Absolute Deviations From Human Experts

The few-shot with CoT prompting technique achieved the highest overall accuracy across all the models (Figure 3). Accuracy was calculated as the percentage of LLM ratings

that exactly matched the human reference rating, averaged over all summary statements and rubric components. Peak values were observed for Mistral Large, where few-shot+CoT achieved 77% accuracy. The second-highest accuracy (76%) was achieved by Mistral Large zero-shot+CoT and Mistral Large zero-shot.

Figure 3. Overall mean accuracy (exact-match agreement with human expert ratings) per model and prompting configuration. Error bars show SDs across the 5 runs. CoT: chain-of-thought.



In few-shot with CoT, Claude achieved 71% accuracy, and GPT-4o reached 67%. Mistral 7B achieved 60% accuracy in few-shot CoT. It was higher than Claude in the zero-shot condition (59%) and GPT-4o in both the zero-shot and zero-shot+CoT settings (58%). Mistral 7B in zero-shot+CoT (51%) and zero-shot (48%) showed the lowest accuracy.

Deviations from human ratings were chosen as additional robustness checks. Extreme deviations of ± 2 points from the human reference ratings were rare. Mistral Large showed the fewest such deviations, whereas Claude, GPT-4o, and Mistral 7B occasionally produced ± 2 deviations, particularly in the SQs and Transformation categories (Figure 1). Rarely, invalid outputs were observed as described in the next section.

Handling Invalid Outputs

A total of 8 invalid outputs were observed (ie, scores outside the rubric scale, such as a 3 on a 0/1/2 scale, or a 2 on a binary rubric component) that only occurred with Mistral 7B as summarized in Table 6. The invalid outputs were found in all prompting configurations but few-shot with CoT. We excluded from all further analysis every summary statement entry in which an invalid data point occurred. Due to the very small number of excluded data points, their impact on reliability and agreement estimates is negligible.

Table 6. Overview of excluded invalid values from analysis. Initial sample size: 122 summary statements×5 runs (610 entries per configuration).

Configuration	Initial sample size	Number of excluded entries	Final analyzed sample size
Mistral 7B few-shot	610	1	609
Mistral 7B zero-shot	610	5	605
Mistral 7B zero-shot with CoT ^a	610	2	608

^aCoT: chain-of-thought.

Discussion

Principal Findings

The primary aim of this study was to examine whether and to what extent LLMs can rate medical students’ written summary statements from VP cases using a rubric and thereby help address the lack of individualized feedback in most VP systems. Our findings indicate that LLMs can serve as robust raters, providing a basis for scalable, automated feedback on summary statements in VP-based training. However, this requires careful consideration, as consistency

and concordance with human experts depend on several factors: model choice, prompting configuration, temperature setting, and the rubric itself. There is no single best configuration; educators need to evaluate iteratively which combination of model and prompt works for their setting.

Across 5 repeated runs per prompting configuration, Mistral Large, Claude Sonnet, and GPT-4o produced stable scores with high test-retest reliability and substantial to almost perfect agreement with human experts, whereas the smaller local model Mistral 7B showed less consistent behavior across rubric components and prompting configurations. Few-shot with CoT prompting yielded the highest

overall concordance with expert human raters, with Mistral Large performing best and Claude and GPT-4o also benefiting from illustrative examples and explicit reasoning initiators.

Divergences between high ICC values and lower Fleiss kappa values suggest a systematic scoring bias but high-light random stability, which could be due to an offset in the models' internal scoring scale that may be anchored differently from the human expert scale. This pattern was most frequently observed in GPT-4o and Mistral Large under zero-shot settings. Such systematic bias has practical implications for formative assessment, where absolute scores matter for student feedback, and argues for score calibration or normalization steps when using LLMs to score medical summary statements.

Rubric components displayed distinct performance patterns. Binary criteria (person and factual accuracy) generally achieved higher agreement than ternary criteria did. This could indicate that current LLMs capture surface-level correctness and easily operationalized features more reliably than they capture more nuanced aspects of clinical reasoning. However, this discrepancy might also stem from the frequency distribution of ratings within the human reference dataset (Multimedia Appendix 9). Future work should further investigate how rating distributions in reference datasets influence these results. To characterize agreement with the human reference standard, we prioritized Gwet AC1 and AC2, which reduced the underestimation of agreement observed with conventional coefficients such as Cohen kappa in skewed rating distributions.

Comparison With Prior Work

Recent work on automated scoring has shown that LLMs can approach or match human raters in terms of reliability and interrater agreement across a range of educational contexts. Seßler et al [41] reported high intrarater reliability (ICC up to 0.84 for closed-source models) and strong alignment with teacher ratings in multidimensional essay scoring [41], whereas Hackl et al [42] reported exceptionally high consistency for GPT-4 (ICC=0.94-0.99) in rating short written responses with minimal scoring drift under deterministic parameters. In medical education, Runyon [20] demonstrated that GPT-4o and Claude can achieve human-level accuracy when analytic rubrics are applied to post-encounter notes if the prompts and scoring instructions are carefully engineered. Complementary studies by Alshammari et al [43] and Sreedhar et al [44] showed that LLMs can reach accuracy and consistency sufficient for formative use in answer-based and critical appraisal tasks, respectively. Our finding that large LLMs yield high ICCs and AC1 and AC2 values, particularly when well prompted, aligns with these findings that LLMs can act as reliable raters for structured, rubric-based assessments.

Integrating CoT reasoning with explicit rubric information improves both scoring accuracy and interpretability [19,45], but it is cautioned that simple zero-shot prompting tends to produce systematic overrating and weaker alignment with human judgments in content-rich or logic-heavy tasks [41,

46]. In medical education, Runyon [20] and Jamieson et al [20,21] reported that prompts that embed analytic rubrics and provide examples enable LLMs to score post-encounter notes with reliability comparable to that of faculty while dramatically reducing grading time. Similarly, Sreedhar et al [44] reported that rubric-guided prompting yields substantial agreement with faculty ratings. Our results replicate and extend these observations to VP-based summary statements: for all the models, few-shot with CoT prompting produced the highest agreement with human ratings, and no single LLM-prompt configuration dominated across all the rubric components. This finding supports the idea that prompt design and rubric operationalization are at least as important as model choice for achieving reliable scoring.

Our study also supports prior work emphasizing that LLMs perform best on clearly defined criteria and more variably on abstract reasoning dimensions. Across the models, the agreement was systematically greater for binary rubric components than for ternary ones. This pattern is consistent with views of clinical reasoning as a context-dependent, knowledge-intensive process that integrates multiple information sources [47]. It also resonates with findings that while LLMs handle pattern-based or shallow tasks well, they may struggle with more complex, compositional reasoning [48]. In our data, components such as transformation and narrowing benefited from few-shot with CoT prompting. Nevertheless, they did not reach the same level of robustness as binary criteria did, indicating that the more abstract the rubric dimension is, the more fragile automated scoring becomes.

The black-box character of LLMs further complicates their use as stand-alone graders. Because these models rely on learned internal representations rather than explicit symbolic rules, reconstructing exactly why a given score was assigned is difficult [48]. Runyon [20] proposed separating analytic explanation from score assignment, prompting the model first to analyze an answer and only then to derive a score to reduce complexity and improve transparency. Our findings support this line of argument: occasional extreme deviations and rare invalid outputs in our robustness checks, particularly for composite or global scores, indicate that models can still fail in ways that are hard to interpret and justify to learners.

Finally, the educational impact of LLM-based feedback must be considered in light of how learners experience it. Ali et al [49] reported that students value the immediacy of AI-generated feedback but still rate human tutor feedback higher in terms of clarity and usefulness. This aligns with the broader concept of hybrid intelligence, where human and machine capabilities are combined to compensate for each other's weaknesses [50]. Our results fit well within this framework: LLMs appear capable of providing frequent, structured ratings of VP summary statements that could meaningfully expand feedback provision.

Limitations

First, we used a single combined reference rating from 2 human expert evaluators. Although predefined criteria

and adjudicated disagreements supported consistent, context-appropriate judgments, this approach limits rater diversity and may not fully capture the variability of expert opinion. Future research should therefore involve a larger pool of raters, report interrater reliability between humans explicitly, and use these data as an additional benchmark for comparing LLM performance.

Second, the options for analyzing and comparing test-retest reliability across different LLMs (RQ1) depend statistically on the scale level of the assessment rubric. In the rubric used for this study, several components have ordinal intervals that do not represent equal distances, so the parametric ICC should be interpreted with caution. We therefore added Fleiss kappa as an exact categorical agreement metric, which served as our primary benchmark for test-retest reliability per LLM, while the ICC provides complementary information on the degree of systematic bias.

Third, the dataset comprised summary statements from selected VP cases in internal medicine only. It remains unclear whether our findings can be generalized to other disciplines, to different VP designs, or to more complex cases. Additional prompt refinement—such as further clarifying rubric descriptions or adding targeted examples—might be necessary when this approach is transferred to new contexts as LLM-generated rationales for rubric-based ratings depend on prompt wording during prompt development (for details, see Methods section). In our study, an effective prompting technique was few-shot with CoT. It included an LLM role description, the scoring rubric, the VP case context, scoring examples with corresponding cases, step-by-step rationales for those examples, an explicit CoT initiator, and a placeholder for the learner's summary statement. Future studies should systematically examine which of these elements are essential, how many examples are optimal, and whether strategies such as selecting extreme, typical, or atypical examples for few-shot prompting can further improve the scoring quality.

Fourth, our dataset contained virtual patient case vignettes and summary statements in English and German. Since the German sample size ($n=41$) was relatively small, we did not stratify our evaluation metrics by language to avoid unstable statistical estimates. Comparisons between common LLMs indicate that the performance for many high-resource languages (HRLs), such as English and German, is practically equal [51]. Future research with larger, balanced multilingual samples is needed to determine whether system performance varies significantly by language.

Fifth, regarding our model selection, new models such as GPT-5 and Claude Sonnet 4.5 became available after we completed data collection and were therefore not included. While their performance may differ, we argue that careful prompt design and robust evaluation strategies are more critical than any specific model choice for this type of task. Moreover, we deliberately restricted our analysis to baseline (ie, nonreasoning) LLMs and did not include large reasoning models (LRMs), such as OpenAI's o1 or Claude Sonnet 4.5 Thinking. This decision was made to maintain a consistent

and comparable study design. Reasoning models introduce additional training stages and inference steps that may add uncertainty and require new evaluation frameworks [52,53]. Our goal here was to provide comprehensive evidence on commonly available baseline models as a first step. Future studies should systematically investigate how LRMs handle summary statement scoring, where their behavior diverges from that of baseline models, and which evaluation methods are suitable for comparison.

Conclusions

We showed that large LLMs can achieve high consistency across repeated runs and moderate to almost perfect agreement with human raters, particularly when few-shot prompting with a CoT initiator is used. The large volume of ratings generated enabled the robust estimation of reliability indices and revealed clear patterns in model- and rubric-level performance.

Our results suggest that LLMs can be used to expand the capacity of VP systems to deliver timely, structured feedback on clinical reasoning for certain rating tasks and for well-defined rubrics. Specifically, they can provide consistent ratings on aspects such as factual accuracy and the inclusion of key patient identifiers, and they show promising, although more variable, performance on rating higher-order reasoning dimensions such as transformation and diagnostic narrowing. This finding indicates that LLM-generated scores could be used to provide immediate, rubric-based feedback after VP cases, helping learners monitor and refine their problem representations more frequently than would be feasible with exclusively human raters. However, their practical use must be further explored. The remaining variability in ternary rubric components, occasional extreme deviations, and rare invalid outputs caution against the use of LLMs as stand-alone graders, especially in summative high-stakes contexts. The opaque nature of LLM decision-making also makes it difficult to justify individual scores to learners. In line with the idea of hybrid intelligence, our findings support an architecture in which LLMs function as supportive tools: they generate rapid and scalable ratings of VP summary statements, whereas human educators can interpret borderline cases, contextualize feedback, and guide learners' metacognitive reflection on their reasoning. Further refinement of more complex rubric components, such as ternary components that require more reasoning and in-depth understanding (eg, transformation), could help to reduce human involvement in the scoring process.

Future work should examine how students interact with and perceive LLM-generated feedback within VP platforms, whether hybrid human-AI feedback models improve learning outcomes, and how domain-specific adaptations or reasoning models might increase the reliability for abstract reasoning components without compromising transparency. Overall, our findings indicate that with carefully designed prompts and annotated rubrics, current LLMs can support formative feedback on clinical reasoning in VPs, but they remain complementary to expert human judgment.

Acknowledgments

We thank Maximilian Mönch, mathematician at the Institute of Biometry and Clinical Epidemiology, Charité – Universitätsmedizin Berlin, for providing statistical counseling, particularly on the selection of appropriate interrater reliability measures. Next, we thank the authors of the iCoViP project [54] for providing the virtual patient cases, licensed as CC BY-NC-SA 4.0, as part of the Erasmus+-funded project iCoViP in CASUS [55]. We also thank the original expert raters for their expertise in rating the summary statements. A collection of 50 exemplary summary statements including the ratings, licensed as CC BY 4.0, is available at [Multimedia Appendix 10](#). Generative AI (GPT-4o and Claude Sonnet 4.5) was used for proofreading and editing, summarization, translation, and reformatting. Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was funded by the Foundation for Innovation in Higher Education (Stiftung Innovation in der Hochschullehre) within the HEDS project at Charité – Universitätsmedizin Berlin.

Data Availability

The complete R and Python codes, as well as virtual patient cases and expert reference ratings used for this study, can be found in our GitHub repository [56], licensed as GPL-3.0.

Authors' Contributions

Conceptualization: JJK (lead), PB (supporting)
Data curation: PS (lead), IH (supporting)
Formal analysis: PS
Funding acquisition: HP
Investigation: PS
Methodology: PS (lead), PB (equal)
Project administration: YH
Resources: IH (lead), HP (equal)
Supervision: JJK
Validation: PB (lead), PS (equal), JJK (equal)
Visualization: PS
Writing – original draft: PB (lead), PS (equal), JJK (equal)
Writing – review & editing: HP (lead), IH (equal), JJK (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1

Prompt design.

[\[DOC File \(Microsoft Word File\), 97 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

The Scoring Rubric.

[\[DOC File \(Microsoft Word File\), 91 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

Gwet AC1 and AC2 values.

[\[DOC File \(Microsoft Word File\), 213 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

Weighted Cohen Kappa values.

[\[DOC File \(Microsoft Word File\), 187 KB-Multimedia Appendix 4\]](#)

Multimedia Appendix 5

Enlarged view of the Claude ratings (top left panel) from Figure 1.

[\[PNG File \(Portable Network Graphics File\), 294 KB-Multimedia Appendix 5\]](#)

Multimedia Appendix 6

Enlarged view of the GPT-4o ratings (top right panel) from Figure 1.

[\[PNG File \(Portable Network Graphics File\), 281 KB-Multimedia Appendix 6\]](#)

Multimedia Appendix 7

Enlarged view of the Mistral Large ratings (bottom left panel) from Figure 1.

[[PNG File \(Portable Network Graphics File\)](#), 301 KB-[Multimedia Appendix 7](#)]

Multimedia Appendix 8

Enlarged view of the Mistral 7B ratings (bottom right panel) from Figure 1.

[[PNG File \(Portable Network Graphics File\)](#), 311 KB-[Multimedia Appendix 8](#)]

Multimedia Appendix 9

Distribution of rating scores by the human experts.

[[DOC File \(Microsoft Word File\)](#), 101 KB-[Multimedia Appendix 9](#)]

Multimedia Appendix 10

A collection of 50 exemplary summary statements including the ratings, licensed as CC BY 4.0.

[[XLSX File \(Microsoft Excel File\)](#), 61 KB-[Multimedia Appendix 10](#)]

References

1. Chang RW, Bordage G, Connell KJ. The importance of early problem representation during case presentations. *Acad Med*. Oct 1998;73(10 Suppl):S109-S111. [doi: [10.1097/00001888-199810000-00062](#)] [Medline: [9795669](#)]
2. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. Cox M, Irby DM, editors. *N Engl J Med*. Nov 23, 2006;355(21):2217-2225. [doi: [10.1056/NEJMra054782](#)] [Medline: [17124019](#)]
3. Bowen JL, Ten Cate O. Prerequisites for learning clinical reasoning. *Principles and practice of case-based clinical reasoning education*. In: Ten Cate O, Custers EJFM, Durning SJ, editors. *Innovation and Change in Professional Education*. Vol 15. Springer; 2018:47-63. [doi: [10.1007/978-3-319-64828-6_4](#)]
4. Feblowitz JC, Wright A, Singh H, Samal L, Sittig DF. Summarization of clinical information: a conceptual model. *J Biomed Inform*. Aug 2011;44(4):688-699. [doi: [10.1016/j.jbi.2011.03.008](#)] [Medline: [21440086](#)]
5. Penner JC, Shipley LC, Minter DJ. Reasoning on rounds: a framework for teaching diagnostic reasoning in the inpatient setting. *J Gen Intern Med*. Oct 2023;38(13):3041-3046. [doi: [10.1007/s11606-023-08359-1](#)] [Medline: [37580633](#)]
6. Smith S, Kogan JR, Berman NB, Dell MS, Brock DM, Robins LS. The development and preliminary validation of a rubric to assess medical students' written summary statements in virtual patient cases. *Acad Med*. Jan 2016;91(1):94-100. [doi: [10.1097/ACM.0000000000000800](#)] [Medline: [26726864](#)]
7. Custers EJFM. Thirty years of illness scripts: theoretical origins and practical applications. *Med Teach*. May 2015;37(5):457-462. [doi: [10.3109/0142159X.2014.956052](#)] [Medline: [25180878](#)]
8. Fürstenberg S, Oubaid V, Berberat PO, Kadmon M, Harendza S. Medical knowledge and teamwork predict the quality of case summary statements as an indicator of clinical reasoning in undergraduate medical students. *GMS J Med Educ*. 2019;36(6):Doc83. [doi: [10.3205/zma001291](#)] [Medline: [31844655](#)]
9. Norman GR, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. *Acad Med*. Jan 2017;92(1):23-30. [doi: [10.1097/ACM.0000000000001421](#)] [Medline: [27782919](#)]
10. Kononowicz AA, Woodham LA, Edelbring S, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. *J Med Internet Res*. Jul 2, 2019;21(7):e14676. [doi: [10.2196/14676](#)] [Medline: [31267981](#)]
11. Jay R, Sandars J, Patel R, et al. The use of virtual patients to provide feedback on clinical reasoning: a systematic review. *Acad Med*. Feb 1, 2025;100(2):229-238. [doi: [10.1097/ACM.0000000000005908](#)] [Medline: [39485118](#)]
12. Bing-You R, Hayes V, Varaklis K, Trowbridge R, Kemp H, McKelvy D. Feedback for learners in medical education: what is known? A scoping review. *Acad Med*. Sep 2017;92(9):1346-1354. [doi: [10.1097/ACM.0000000000001578](#)] [Medline: [28177958](#)]
13. Lipnevich AA, Mattern K, Feddock C. Formative assessment and feedback in medical education: a practical guide: AMEE Guide No. 189. *Med Teach*. Jun 2026;48(6):921-940. [doi: [10.1080/0142159X.2025.2569623](#)] [Medline: [41134874](#)]
14. Hege I, Kiesewetter I, Adler M. Automatic analysis of summary statements in virtual patients - a pilot study evaluating a machine learning approach. *BMC Med Educ*. Oct 16, 2020;20(1):366. [doi: [10.1186/s12909-020-02297-w](#)] [Medline: [33066789](#)]
15. Bond WF, Zhou J, Bhat S, et al. Automated patient note grading: examining scoring reliability and feasibility. *Acad Med*. Nov 1, 2023;98(11S):S90-S97. [doi: [10.1097/ACM.0000000000005357](#)] [Medline: [37983401](#)]

16. Cho HN, Jun TJ, Kim YH, et al. Task-specific transformer-based language models in health care: scoping review. *JMIR Med Inform.* Nov 18, 2024;12:e49724. [doi: [10.2196/49724](https://doi.org/10.2196/49724)] [Medline: [39556827](https://pubmed.ncbi.nlm.nih.gov/39556827/)]
17. Nori H, Daswani M, Kelly C, et al. Sequential diagnosis with language models. *arXiv.* Preprint posted online on Jun 27, 2025. [doi: [10.48550/arXiv.2506.22405](https://doi.org/10.48550/arXiv.2506.22405)]
18. Croxford E, Gao Y, First E, et al. Evaluating clinical AI summaries with large language models as judges. *NPJ Digit Med.* Nov 5, 2025;8(1):640. [doi: [10.1038/s41746-025-02005-2](https://doi.org/10.1038/s41746-025-02005-2)] [Medline: [41193667](https://pubmed.ncbi.nlm.nih.gov/41193667/)]
19. Lee GG, Latif E, Wu X, Liu N, Zhai X. Applying large language models and chain-of-thought for automatic scoring. *Comput Educ Artif Intell.* Jun 2024;6:100213. [doi: [10.1016/j.caeai.2024.100213](https://doi.org/10.1016/j.caeai.2024.100213)]
20. Runyon C. Using large language models (LLMs) to apply analytic rubrics to score post-encounter notes. *Med Teach.* Nov 2025;47(11):1772-1780. [doi: [10.1080/0142159X.2025.2504106](https://doi.org/10.1080/0142159X.2025.2504106)] [Medline: [40380943](https://pubmed.ncbi.nlm.nih.gov/40380943/)]
21. Jamieson AR, Holcomb MJ, Dalton TO, et al. Rubrics to prompts: assessing medical student post-encounter notes with AI. *NEJM AI.* Nov 27, 2024;1(12). [doi: [10.1056/AIcs2400631](https://doi.org/10.1056/AIcs2400631)]
22. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering for large language models. *Patterns (N Y).* Jun 13, 2025;6(6):101260. [doi: [10.1016/j.patter.2025.101260](https://doi.org/10.1016/j.patter.2025.101260)] [Medline: [40575123](https://pubmed.ncbi.nlm.nih.gov/40575123/)]
23. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature.* Aug 2023;620(7972):172-180. [doi: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)] [Medline: [37438534](https://pubmed.ncbi.nlm.nih.gov/37438534/)]
24. Han T, Adams LC, Papaioannou JM, et al. MedAlpaca—an open-source collection of medical conversational AI models and training data. *arXiv.* Preprint posted online on Apr 14, 2023. [doi: [10.48550/arXiv.2304.08247](https://doi.org/10.48550/arXiv.2304.08247)]
25. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med.* May 20, 2021;4(1):86. [doi: [10.1038/s41746-021-00455-y](https://doi.org/10.1038/s41746-021-00455-y)] [Medline: [34017034](https://pubmed.ncbi.nlm.nih.gov/34017034/)]
26. Lin J, Tang J, Tang H, et al. Awq: Activation-aware weight quantization for on-device LLM compression and acceleration. *Proc Mach Learn Syst.* 2024;6:87-100. URL: https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html [Accessed 2026-09-14]
27. CASUS. CASUS online. URL: <https://crt.casus.net/pm2/app/homepage.html> [Accessed 2026-09-14]
28. Kojima T, Gu S, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. Presented at: Advances in Neural Information Processing Systems 35; Nov 28 to Dec 9, 2022:22199-22213; New Orleans, LA. 2022.[doi: [10.52202/068431-1613](https://doi.org/10.52202/068431-1613)]
29. Jiang Y, Wang Y, Zeng X, et al. FollowBench: a multi-level fine-grained constraints following benchmark for large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1. Association for Computational Linguistics; 2024:4667-4688. [doi: [10.18653/v1/2024.acl-long.257](https://doi.org/10.18653/v1/2024.acl-long.257)]
30. Kwon W, Li Z, Zhuang S, et al. Efficient memory management for large language model serving with pagedattention. In: SOSP '23: Proceedings of the 29th Symposium on Operating Systems Principles. Association for Computing Machinery; 2023:611-626. [doi: [10.1145/3600006.3613165](https://doi.org/10.1145/3600006.3613165)]
31. Atıl B, Aykent S, Chittams A. Non-determinism of “deterministic” LLM system settings in hosted environments. In: Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems. 2025:135-148. [doi: [10.18653/v1/2025.eval4nlp-1.12](https://doi.org/10.18653/v1/2025.eval4nlp-1.12)]
32. R Core Team. The R Language Definition. Vienna Austria R Found Stat Comput. 2000. URL: <https://cran.r-project.org/doc/manuals/r-devel/R-lang.html>
33. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* Jun 2016;15(2):155-163. [doi: [10.1016/j.jcm.2016.02.012](https://doi.org/10.1016/j.jcm.2016.02.012)] [Medline: [27330520](https://pubmed.ncbi.nlm.nih.gov/27330520/)]
34. Xu S, Lorber MF. Interrater agreement statistics with skewed data: evaluation of alternatives to Cohen’s kappa. *J Consult Clin Psychol.* Dec 2014;82(6):1219-1227. [doi: [10.1037/a0037489](https://doi.org/10.1037/a0037489)] [Medline: [25090041](https://pubmed.ncbi.nlm.nih.gov/25090041/)]
35. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol.* May 2008;61(Pt 1):29-48. [doi: [10.1348/000711006X126600](https://doi.org/10.1348/000711006X126600)] [Medline: [18482474](https://pubmed.ncbi.nlm.nih.gov/18482474/)]
36. Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters. Advanced Analytics, LLC; 2014. URL: https://books.google.co.in/books/about/Handbook_of_Inter_Rater_Reliability_4th.html?id=fac9BQAAQBAJ&redir_esc=y [Accessed 2026-10-07]
37. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen’s Kappa and Gwet’s AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol.* Apr 29, 2013;13(1):61. [doi: [10.1186/1471-2288-13-61](https://doi.org/10.1186/1471-2288-13-61)] [Medline: [23627889](https://pubmed.ncbi.nlm.nih.gov/23627889/)]
38. Silveira PSP, Siqueira JO. Better to be in agreement than in bad company. *Behav Res.* 2023;55(7):3326-3347. [doi: [10.3758/s13428-022-01950-0](https://doi.org/10.3758/s13428-022-01950-0)]

39. Popplewell M, Reizes J, Zaslawski C. Appropriate statistics for determining chance-removed interpractitioner agreement. *J Altern Complement Med*. Nov 2019;25(11):1115-1120. [doi: [10.1089/acm.2017.0297](https://doi.org/10.1089/acm.2017.0297)] [Medline: [29851495](https://pubmed.ncbi.nlm.nih.gov/29851495/)]
40. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. Mar 1977;33(1):159-174. [doi: [10.2307/2529310](https://doi.org/10.2307/2529310)] [Medline: [843571](https://pubmed.ncbi.nlm.nih.gov/843571/)]
41. Seßler K, Fürstenberg M, Bühler B, Kasneci E. Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In: *LAK '25. Association for Computing Machinery*; 2025:462-472. [doi: [10.1145/3706468.3706527](https://doi.org/10.1145/3706468.3706527)]
42. Hackl V, Müller AE, Granitzer M, Sailer M. Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Front Educ*. Dec 5, 2023;8:1272229. [doi: [10.3389/educ.2023.1272229](https://doi.org/10.3389/educ.2023.1272229)]
43. Alshammari AF, Madfa AA, Anazi BA, Alenezi YE, Alkurdi KA. Comparison of accuracy and consistency of AI Language models when answering standardised dental MCQs. *BMC Med Educ*. Oct 28, 2025;25(1):1507. [doi: [10.1186/s12909-025-07624-7](https://doi.org/10.1186/s12909-025-07624-7)] [Medline: [41152856](https://pubmed.ncbi.nlm.nih.gov/41152856/)]
44. Sreedhar R, Chang L, Gangopadhyaya A, et al. Comparing scoring consistency of large language models with faculty for formative assessments in medical education. *J Gen Intern Med*. Jan 2025;40(1):127-134. [doi: [10.1007/s11606-024-09050-9](https://doi.org/10.1007/s11606-024-09050-9)] [Medline: [39402411](https://pubmed.ncbi.nlm.nih.gov/39402411/)]
45. Xiao C, Ma W, Song Q, et al. Human-AI collaborative essay scoring: a dual-process framework with llms. In: *LAK '25. Association for Computing Machinery*; 2025:293-305. [doi: [10.1145/3706468.3706507](https://doi.org/10.1145/3706468.3706507)]
46. Wu X, Saraf PP, Lee G, Latif E, Liu N, Zhai X. Unveiling scoring processes: dissecting the differences between LLMs and human graders in automatic scoring. *Tech Know Learn*. Jun 2026;31(2):669-684. [doi: [10.1007/s10758-025-09836-8](https://doi.org/10.1007/s10758-025-09836-8)]
47. Olson A, Kämmer JE, Taher A, et al. The inseparability of context and clinical reasoning. *J Eval Clin Pract*. Jun 2024;30(4):533-538. [doi: [10.1111/jep.13969](https://doi.org/10.1111/jep.13969)] [Medline: [38300231](https://pubmed.ncbi.nlm.nih.gov/38300231/)]
48. Yuan Z, Yuan H, Tan C, Wang W, Huang S. How well do large language models perform in arithmetic tasks? *arXiv*. Preprint posted online on Mar 16, 2023. [doi: [10.48550/ARXIV.2304.02015](https://doi.org/10.48550/ARXIV.2304.02015)]
49. Ali M, Harbieh I, Haider KH. Bytes versus brains: a comparative study of AI-generated feedback and human tutor feedback in medical education. *Med Teach*. Jan 2026;48(1):131-141. [doi: [10.1080/0142159X.2025.2519639](https://doi.org/10.1080/0142159X.2025.2519639)] [Medline: [40528415](https://pubmed.ncbi.nlm.nih.gov/40528415/)]
50. Dellermann D, Ebel P, Söllner M, Leimeister JM. Hybrid Intelligence. *Bus Inf Syst Eng*. Oct 2019;61(5):637-643. [doi: [10.1007/s12599-019-00595-2](https://doi.org/10.1007/s12599-019-00595-2)]
51. Lai V, Ngo N, Pouran Ben Veyseh A, et al. ChatGPT beyond English: towards a comprehensive evaluation of large language models in multilingual learning. Presented at: Findings of the Association for Computational Linguistics: EMNLP 2023; Dec 6-10, 2023:13171-13189; Singapore. [doi: [10.18653/v1/2023.findings-emnlp.878](https://doi.org/10.18653/v1/2023.findings-emnlp.878)]
52. Shojae P, Mirzadeh I, Alizadeh K, Horton M, Bengio S, Farajtabar M. The illusion of thinking: understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv*. Preprint posted online on Jun 7, 2025. [doi: [10.48550/ARXIV.2506.06941](https://doi.org/10.48550/ARXIV.2506.06941)]
53. Chen Y, Benton J, Radhakrishnan A, et al. Reasoning models don't always say what they think. *arXiv*. Preprint posted online on May 8, 2025. [doi: [10.48550/ARXIV.2505.05410](https://doi.org/10.48550/ARXIV.2505.05410)]
54. International Collection of Virtual Patients (iCoViP). URL: <https://icovip.eu> [Accessed 2026-10-08]
55. CASUS. URL: <https://icovip.casus.net> [Accessed 2026-10-08]
56. MedicalSummaryStatements_LLMs. GitHub. URL: https://github.com/DSFZ-Team/MedicalSummaryStatements_LLMs [Accessed 2026-10-08]

Abbreviations

AWQ: activation-aware weight quantization
BERT: Bidirectional Encoder Representations from Transformers
CoT: chain-of-thought
HRL: high-resource language
ICC: intraclass correlation coefficient
LLM: large language model
LRM: large reasoning model
NLP: natural language processing
OSCE: Objective Structured Clinical Examination

RQ: research question

SQ: semantic qualifier

VP: virtual patient

Edited by Blake Lesselroth; peer-reviewed by Emmanouil Karampinis, Oudong Xia; submitted 05.Mar.2026; final revised version received 12.Aug.2026; accepted 26.Aug.2026; published 21.Oct.2026

Please cite as:

Stadler P, Bucichowski P, Holzhausen Y, Hege I, Peters H, Kaminski JJ

Test-Retest Reliability of Large Language Models and Their Concordance With Human Experts in Rubric-Based Scoring of Clinical Summary Statements: Experimental Comparative Study

JMIR Med Educ 2026;12:e94666

URL: <https://mededu.jmir.org/2026/1/e94666>

doi: [10.2196/94666](https://doi.org/10.2196/94666)

© Patrick Stadler, Piotr Bucichowski, Ylva Holzhausen, Inga Hege, Harm Peters, Julius Josef Kaminski. Originally published in JMIR Medical Education (<https://mededu.jmir.org>), 21.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on <https://mededu.jmir.org/>, as well as this copyright and license information must be included.