Comparative Evaluation of Large Language Models for Medical Education: Performance Analysis in Urinary System Histology.

Anikó Szabó
Alfaisal University
Ghasem Dolatkhah Laein
dr.ghasemdolatkhah@gmail.com
Mashhad University of Medical Sciences

Article

Keywords: Medical education, Large language models, Histology, Artificial intelligence, Medical assessment

Posted Date: March 13th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6186253/v1

Additional Declarations: No competing interests reported.

Abstract

Large language models (LLMs) show potential for medical education, but their domain-specific capabilities need systematic evaluation. This study presents a comparative assessment of thirteen LLMs in urinary system histology education. Using a multi-dimensional framework, we evaluated models across two tasks: answering 65 validated multiple-choice questions (MCQs) and generating clinical scenarios with assessment items. For MCQ performance, we assessed accuracy along with explanation quality through relevance and comprehensiveness metrics. For scenario generation, we evaluated Quality, Complexity, Relevance, Correctness, and Variety dimensions. Performance varied substantially across models and tasks, with ChatGPT-o1 achieving highest MCQ accuracy (96.31 ± 17.85%) and Claude-3.5 demonstrating superior clinical scenario generation capabilities (91.4% of maximum possible score). All models significantly outperformed random guessing with large effect sizes. Statistical analyses revealed significant differences in consistency across multiple attempts and dimensional performance, with most models showing higher Correctness than Quality scores in scenario generation. Term frequency analysis revealed significant content imbalances across all models, with systematic overemphasis of certain anatomical structures and complete omission of others. Our findings demonstrate that while LLMs show considerable promise for medical education, their reliable implementation requires matching specific models to appropriate educational tasks, implementing verification mechanisms, and recognizing their current limitations in generating pedagogically balanced content.

1. Introduction

Artificial intelligence (AI) has emerged as a transformative force in medical education, offering tools to enhance learning, assessment, and clinical decision-making [1]. Among these tools, Large Language Models (LLMs) such as ChatGPT and Gemini have demonstrated capabilities in generating human-like text, answering complex medical questions, and simulating clinical reasoning tasks [2, 3]. For example, studies show that ChatGPT-4 achieves accuracy rates comparable to medical students on standardized exam questions, highlighting its potential as an educational aid [4]. However, the integration of LLMs into medical curricula requires rigorous validation, as errors in medical content could propagate misinformation and compromise patient safety [5, 6]. Recent advancements in LLM development, including iterative improvements to existing models like GPT-4 and Gemini, underscore the need for systematic evaluation frameworks [7]. While these models excel in tasks such as answering anatomy and pharmacology questions, their performance varies significantly across specialties and question difficulty levels [8, 9]. For instance, ChatGPT achieved 72% accuracy on ophthalmology board-style questions but struggled with advanced clinical reasoning tasks, reflecting limitations in its diagnostic capabilities [10]. Furthermore, LLMs are prone to "artificial hallucinations," generating plausible but factually incorrect explanations that could mislead learners [5]. Such risks necessitate frequent reevaluation of LLMs as they evolve, particularly with the rapid release of new iterations optimized for healthcare applications [11]. Despite growing interest in LLMs for medical education, existing studies focus disproportionately on individual models or narrow tasks, leaving critical gaps in comparative

analyses [12, 13]. For example, while ChatGPT's performance on the United States Medical Licensing Examination (USMLE) has been extensively documented [14, 15], newer models like Claude, DeepSeek, Qwen and Grok lack similar scrutiny. Additionally, there is limited research on the consistency of LLM responses across multiple attempts or their ability to provide explanations that align with established medical knowledge [16–18]. These gaps hinder evidence-based recommendations for integrating LLMs into educational workflows. This study addresses these limitations by conducting a comprehensive evaluation of thirteen publicly accessible LLMs, including the latest versions of Claude (3.5-Sonnet, 3.7-Sonnet), ChatGPT (o1, o3-High, 4.5), Grok (3, 3-Thinking), DeepSeek (V3, R1), Qwen (2.5-Max, 3B) and Gemini 2 Pro. We seek to answer the following research question: How do current-generation large language models compare in performance, consistency, and content generation capabilities when applied to urinary system histology education, and what model-specific and task-specific patterns emerge that could inform their optimal integration into medical curricula? Our assessment encompasses two distinct tasks: performance on validated medical multiple-choice questions (MCQs) and generation of clinical scenarios with assessment items. For MCQ performance, we evaluate accuracy, along with explanation quality through relevance and comprehensiveness scores. For clinical scenario generation, we assess five dimensions: Quality, Complexity, Relevance, Correctness, and Variety. We benchmark MCQ performance against random guessing and analyse consistency across multiple attempts. By correlating explanation quality with accuracy and identifying dimensional patterns in scenario generation, we provide actionable insights for educators and developers aiming to leverage LLMs responsibly in medical training.

2. Method

2.1. Study Design

This study employed a comparative experimental analysis to evaluate thirteen large language models (LLMs) [19-29] in medical education, building upon previous evaluation approaches in AI-assisted medical education [32]. While inspired by earlier evaluation frameworks [32], our methodology incorporates several enhancements including expanded model selection, structured validation metrics, and comprehensive inter-rater reliability assessment. The investigation comprised assessment of LLM performance on USMLE-style multiple-choice questions (MCQs) and evaluation of LLM-generated clinical scenarios focused on histology of the urinary system. Figure 1 illustrates the workflow of this methodological approach.

2.2. Model Selection

Thirteen large language models (LLMs) were included in our comparative analysis. Our initial data collection (completed February 18, 2025) evaluated seven models: DeepSeek-R1 [19], DeepSeek [20], ChatGPT o3-High [21], ChatGPT o1 [22], Claude 3.5-Sonnet [23], Qwen 2.5-Max [24] and Gemini 2 Pro [25]. Following the release of advanced models, we expanded our evaluation to include four additional LLMs (Claude-3.7-Sonnet Normal [26] and Extended [27] modes, Grok-3 [28], Grok-3-Thinking [29]) using

identical protocols. This second evaluation was completed on March 6, 2025. We noted the subsequent release of QwQ-3B [30] and ChatGPT-4.5(Research preview) [31] one day prior to manuscript submission and incorporated them into our analysis with an expedited protocol. All thirteen models were then analysed together in a comprehensive comparative assessment. Model selection was guided by the following criteria:

- **Independence**: Each LLM must be a standalone model developed by its respective organization, rather than a rebranded version of another model.
- **Public Accessibility**: The model must be readily available to users via a web interface, either for free or through a paid subscription, without requiring programming expertise for interaction.
- **Advancement**: The selected model must represent the latest publicly available version from its developer at the time of its respective analysis phase, ensuring the study evaluates cutting-edge AI capabilities.
- **Medical Relevance**: Each LLM must be capable of answering USMLE-style multiple-choice questions (MCQs) and generating clinically relevant scenarios.

While our inclusion criteria emphasized model independence, certain LLMs were included together due to fundamental architectural upgrades rather than minor performance refinements. These paired models (DeepSeek [20] and DeepSeek-R1 [19], ChatGPT-o1 [22] and ChatGPT-o3-High [21], Claude-3.7-Sonnet Normal [26] and Extended [27], Grok-3 [28] and Grok-3-Thinking [29]) introduce distinct reasoning frameworks rather than simple hyperparameter optimizations. DeepSeek-R1 incorporates reinforcement learning (RLHF) to enhance logical reasoning compared to DeepSeek-V3 [19, 20]. ChatGPT-o3-High introduces an adaptive reasoning mechanism that adjusts depth based on query complexity, surpassing the static inference approach of ChatGPT-o1, while ChatGPT-4.5 further enhances accuracy and reliability by significantly reducing hallucination rates (37.1% vs. GPT-4's 61.8%) and integrating advanced function-calling capabilities for structured outputs [21, 22, 31]. Claude-3.7-Sonnet Extended Mode fundamentally alters inference by allowing a dynamic 'thinking budget' for multi-step reasoning [26, 27]. Grok-3-Thinking integrates iterative chain-of-thought reflection, executing multiple inference passes before answer selection, which differs from Grok-3's single-pass approach [28, 29]. Including QwQ-3B alongside Qwen-2.5-Max allows comparison of a specialized reinforcement learning-based reasoning architecture with a generalized Mixture-of-Experts model optimized for broad task efficiency, both originating from Alibaba Cloud [27, 30]. Including these versions enables a structured comparison of how training enhancements impact medical reasoning capabilities and efficiency.

2.3. Question MCQ Development and Validation

Sample size determination was conducted via formal power analysis for repeated measures ANOVA across thirteen LLMs. Statistical parameters derived from comparable published research [32] informed our calculations (α=0.05, minimum detectable effect=1% difference in accuracy, SD=5.7%). This analysis established that 45 questions would yield 80% power to detect inter-model performance differences, with our final sample of 65 MCQs providing enhanced statistical sensitivity (>95% power). The 65

USMLE-style MCQs were randomly selected from the Histology course exam database, specifically focusing on urinary system content appropriate for first-year medical students (Appendix 1). No questions containing images were included. Questions represented varied difficulty levels to ensure comprehensive assessment of LLM capabilities. To establish content validity for this investigation, three independent medical education experts evaluated each question using a structured validation framework, applying a 0-2 scale across three dimensions:

- **Validity** (scientific accuracy)
- **Clarity** (wording and ambiguity)
- **Difficulty** (appropriateness for year 1 medical students)

Table 1 details the scoring criteria used by experts for each dimension. This yielded total scores of 0-6 per question (Appendix 2). Discrepancies between expert reviewers were resolved through structured consensus discussions, where disagreements were documented and discussed until unanimous agreement was reached. This iterative process ensured that final scores reflected collective expert judgment rather than averaged ratings. Inter-rater reliability for this validation process was high (89.23%), with substantial to almost perfect agreement across Validity ($\kappa = 0.74$), Clarity ($\kappa = 0.68$), and Difficulty ($\kappa = 0.91$).

Table 1.Scoring Framework for MCQ Validation and Response Evaluation.

Score	Validity	Clarity	Difficulty	Relevance	Comprehensiveness
0	Contains significant scientific inaccuracies or outdated concepts	Highly ambiguous or poorly worded, leading to confusion	Inappropriate difficulty level (too easy or too advanced)	Not Relevant: Explanation fails to address the question's core concept or contains primarily unrelated information	Inadequate: Explanation is superficial, lacks key details, or contains significant gaps in reasoning
1	Contains minor scientific imprecisions or requires updating	Some ambiguity or wording issues that may affect understanding	Moderately appropriate difficulty with some adjustments needed	Partially Relevant: Explanation addresses some aspects of the question but includes tangential information or misses key components	Partially Comprehensive: Explanation covers basic concepts but lacks depth in important areas or has minor logical gaps
2	Scientifically accurate and up-to-date	Clear, unambiguous wording with precise language	Appropriately challenging for year 1 medical students	Highly Relevant: Explanation directly addresses all aspects of the question with focused, pertinent information	Fully Comprehensive: Explanation thoroughly covers all aspects with appropriate depth, logical structure, and complete reasoning

2.4. Data Collection

2.4.1. Task 1: Multiple-Choice Question Assessment Protocol

The data collection process utilized a structured prompt template that was consistently applied across all models (Figure 2), ensuring uniform instruction delivery and response format requirements. All prompts were standardized across models to ensure fair comparison. Instructions specified "brief explanations" to prevent token limitations from differentially affecting model performance. This design choice ensured all models could process the complete question set within their operational constraints, eliminating potential bias from truncated responses. Each LLM completed five attempts at all 65 MCQs using new chat sessions for each attempt to minimize response dependencies. This approach ensured that each interaction represented an independent evaluation instance without potential contextual bias from previous responses. All responses were systematically recorded, including both answer selections (A-E) and explanatory rationales. To establish a baseline comparison, five sets of random answers were generated and analysed for the same MCQ set, using the RAND() function in Microsoft Excel to compare

chatbot results with random guessing. All models demonstrated sufficient token capacity to handle the provided prompts without truncation, ensuring comparable response conditions across all evaluated LLMs [19-31]. The response evaluation framework incorporated three key metrics:

- **Accuracy**: Correctness of the selected answer option (% correct)
- **Relevance scores** (0-2 scale): Assessing alignment between explanations and questions
- **Comprehensiveness scores** (0-2 scale): Evaluating the completeness and structure of explanations

Relevance and comprehensiveness scores were assessed for all LLM explanations independent of answer correctness. This approach allowed us to identify cases where models provided incorrect answers with convincing but flawed explanations. a particularly important consideration for educational applications where misleading explanations could reinforce misconceptions. Table 1 presents the detailed scoring criteria for both relevance and comprehensiveness scales. To minimize evaluation bias, all model responses were coded and presented to evaluators without model identification. Evaluators independently scored responses according to the standardized criteria before comparing assessments. Disagreements were resolved through structured consensus discussions, with final scores representing unanimous agreement rather than averaged ratings. Inter-rater reliability analysis demonstrated high consistency among reviewers, with category-specific agreement rates of 91.35% for relevance and 88.42% for comprehensiveness. Fleiss' Kappa values indicated substantial agreement for both relevance (0.72) and comprehensiveness (0.68). All final evaluation data was systematically compiled and documented in Appendix 2.

2.4.2. Task 2: Clinical Scenario Generation and Assessment Framework

Each LLM generated clinical scenarios and three related MCQs for three specific topics in urinary system histology: renal corpuscle histology, nephron histology, and lower urinary system histology (Figure 3). The complete generated clinical scenarios are provided in Appendix 3. Two professors of histology independently evaluated these outputs using a comprehensive 0-5 scale across five dimensions:

1. The quality of response to the prompt (was it generated as instructed?)
2. The complexity of clinical case scenarios and questions (was it appropriate for medical students?)
3. The relevance to the Histology course objectives (was it related to the Histology urinary system course?)
4. The correctness (does it contain any factual mistakes?)
5. The variety (does it repeat previous responses?)

Each LLM performed three iterations of this task, generating 36 item-based outputs per model (three scenarios × (Each scenario + three MCQs = 4 items) × three attempts). Evaluators were blinded to model identity during assessment of clinical scenarios, preventing potential bias in quality evaluations. Consensus was reached when evaluators unanimously agree on the final rating. Inter-rater reliability analysis for clinical scenario evaluation demonstrated high consistency between reviewers, with an

overall agreement rate of 86.67%. Category-specific agreement rates were: Quality (88.89%), Complexity (84.72%), Relevance (86.11%), Correctness (87.50%), and Variety (86.11%). Cohen's Kappa values indicated almost perfect agreement for Quality (0.85) and substantial agreement for Complexity (0.76), Relevance (0.77), Correctness (0.62), and Variety (0.77). These findings confirm the robust nature of our evaluation framework and the reliability of assessments across different dimensions of clinical scenario quality. The complete evaluation framework and detailed scoring results are provided in Appendix 4 and 5.

2.5. Statistical

Statistical analyses were conducted using Statistica 13.5.0.17 and custom Python scripts with a significance threshold of $p < 0.05$. Model performance was evaluated against random baseline using one-tailed t-tests with Cohen's d for effect size quantification. Inter-model differences were assessed via repeated measures ANOVA, treating questions as subjects and models as within-subjects factors, with post-hoc Tukey HSD tests controlling family-wise error rate across the possible pairwise comparisons. For non-normally distributed data, Kruskal-Wallis tests provided complementary non-parametric assessment. ANOVA effect sizes were quantified using eta-squared (η^2), representing the proportion of total variance attributable to differences between models. Model consistency was measured through intraclass correlation coefficient (ICC(3,k)), which assesses how consistently models performed across multiple attempts, and Fleiss' kappa. Response stability was calculated using coefficient of variation (CV), defined as the ratio of the standard deviation to the mean. Explanation quality was characterized with 95% confidence intervals, and relationships between explanation metrics and accuracy were examined using Spearman's rank correlation (ρ). For clinical scenario evaluation, we conducted dimension correlation analysis to identify relationships between quality dimensions and Principal Component Analysis (PCA) to examine underlying patterns in evaluation criteria. Additionally, models were grouped by provider family (Claude, ChatGPT, DeepSeek, Grok, Qwen) [19-24, 26-31] to evaluate family-level performance differences. Inter-rater reliability was assessed using Fleiss' kappa for MCQ validation (three raters) and Cohen's kappa for clinical scenario evaluation (two raters). These reliability coefficients were subsequently employed to calculate reliability-weighted dimension scores that adjusted for measurement confidence. To examine specific aspects of clinical scenario generation, we conducted topic-specific analyses comparing performance across the three histological topics (renal corpuscle, nephron, and lower urinary system) and between scenario creation versus MCQ generation tasks. Term frequency analysis was applied to generated clinical scenarios to quantitatively assess content coverage and identify knowledge representation patterns. All performance metrics are reported with appropriate measures of central tendency and 95% confidence intervals to characterize uncertainty in performance estimates.

3. Results

Our evaluation of thirteen large language models revealed significant performance variations across both medical MCQ answering and clinical scenario generation. In MCQ assessment, ChatGPT o1

demonstrated the highest accuracy (96.31%), followed by ChatGPT 4.5 (92.62%), while DeepSeek showed superior consistency (ICC=0.986). For clinical scenario generation, Claude 3.5 achieved the highest overall score (91.4% of maximum possible). All models significantly outperformed random guessing ($p<0.0001$) with large effect sizes (Cohen's d ranging from 1.540 to 3.799). Statistical analysis confirmed significant differences between models in both tasks (MCQs: $F=6.243$, $p<0.000001$, partial $\eta^2=0.066$; scenarios: $F(12,455)=8.48$, $p=1.39\times10^{-14}$). Table 2 shows performance of LLMs across Medical MCQ Answering and Clinical Scenario Generation Tasks.

Table 2 Comparative Performance of LLMs Across Medical MCQ Answering and Clinical Scenario Generation Tasks

Model	Accuracy (% ±SD)	Consistency	Explanation Quality	Stability CV (%)	Scenario Generation (%±SD)
ChatGPT o1	96.31±17.85	ICC=0.971 (0.954, 0.983); κ=0.988	Comp=1.84±0.18; Rel=1.78±0.17	0.8	71.8±3.99
ChatGPT 4.5	92.62±23.94	ICC=0.953 (0.923, 0.971); κ=0.965	Comp=1.97±0.12; Rel=1.98±0.09	1.6	74.0±4.00
ChatGPT o3-H	88.31±29.75	ICC=0.959 (0.933, 0.975); κ=0.938	Comp=1.81±0.21; Rel=1.77±0.21	1.4	66.8±3.93
Claude 3.7-E	89.85±24.84	ICC=0.886 (0.819, 0.929); κ=0.890	Comp=1.96±0.09; Rel=1.96±0.10	4.1	89.4±2.02
Claude 3.7	84.31±31.86	ICC=0.927 (0.882, 0.955); κ=0.891	Comp=1.82±0.21; Rel=1.87±0.18	3.9	76.8±5.97
Claude 3.5	83.69±34.35	ICC=0.961 (0.937, 0.976); κ=0.916	Comp=1.62±0.28; Rel=1.83±0.19	0.9	91.4±2.37
DeepSeek-R1	85.85±32.10	ICC=0.956 (0.929, 0.973); κ=0.942	Comp=1.74±0.19; Rel=1.78±0.21	2.1	73.7±3.22
DeepSeek	75.08±42.10	ICC=0.986 (0.978, 0.992); κ=0.945	Comp=1.84±0.17; Rel=1.80±0.18	1.6	71.3±3.53
Gemini 2	77.85±36.82	ICC=0.934 (0.894, 0.959); κ=0.861	Comp=1.89±0.16; Rel=1.81±0.17	4.4	77.8±3.73
Grok 3-T	86.15±29.13	ICC=0.900 (0.840, 0.938); κ=0.866	Comp=1.67±0.34; Rel=1.88±0.23	2.2	80.9±3.01
Grok 3	84.92±33.66	ICC=0.968 (0.948, 0.980); κ=0.906	Comp=1.59±0.25; Rel=1.75±0.23	1.8	69.9±4.50
Qwen 2.5-M	72.92±39.76	ICC=0.940 (0.903, 0.963); κ=0.858	Comp=1.69±0.26; Rel=1.65±0.26	5.1	74.6±4.38
QwQ-3B	72.31±40.72	ICC=0.949 (0.917, 0.969); κ=0.858	Comp=1.91±0.17; Rel=1.81±0.22	3.0	63.4±6.71

Note: ICC = Intraclass Correlation Coefficient with 95% confidence intervals; κ = Fleiss' kappa; Comp = Comprehensiveness; Rel = Relevance; CV = Coefficient of Variation

3.1 MCQ Performance Analysis

3.1.1 Overall Accuracy

Model accuracy on medical multiple-choice questions varied substantially, with performance rates ranging from 72.31% to 96.31% (Figure 4). ChatGPT-o1 demonstrated the highest accuracy (96.31±17.85%), followed by ChatGPT-4.5 (92.62±23.94%) and Claude-3.7-E (89.85±24.84%). The lowest performing models were QwQ-3B (72.31±40.72%), Qwen-2.5-M (72.92±39.76%), and DeepSeek (75.08±42.10%). For reference, the random guessing baseline achieved 22.46±20.90% accuracy. All models significantly outperformed random guessing as determined by one-tailed t-tests (all $p<0.0001$). Effect sizes were uniformly large, with Cohen's d values ranging from 1.540 (QwQ-3B vs. random) to 3.799 (ChatGPT-o1 vs. random). Repeated measures ANOVA confirmed significant differences between models ($F=6.243$, $p<0.000001$, partial $\eta^2=0.066$), though post-hoc Tukey HSD testing identified no significant pairwise differences between individual LLMs.

3.1.2 Consistency Analysis

Model consistency was assessed using intraclass correlation coefficient (ICC) and Fleiss' kappa (Figure 5A). DeepSeek exhibited the highest consistency (ICC=0.986 (0.978, 0.992); κ=0.945), followed by ChatGPT-o1 (ICC=0.971 (0.954, 0.983); κ=0.988) and Grok-3 (ICC=0.968 (0.948, 0.980); κ=0.906). The newly evaluated ChatGPT-4.5 demonstrated excellent consistency (ICC=0.953 (0.923, 0.971); κ=0.965), as did QwQ-3B (ICC=0.949 (0.917, 0.969); κ=0.858). All models demonstrated excellent consistency (ICC>0.85). Performance stability across multiple attempts was measured using coefficient of variation (CV). ChatGPT-o1 (CV=0.8%) and Claude-3.5 (CV=0.9%) showed the most stable performance, while ChatGPT-4.5 (CV=1.6%) and QwQ-3B (CV=3.0%) exhibited good to excellent stability. Qwen-2.5-M demonstrated the highest variability (CV=5.1%) among all models. Performance across attempts was highly consistent, with accuracy ranges across five attempts varying from 1.5 percentage points (ChatGPT-o1 and Claude-3.5) to 10.7 percentage points (Claude-3.7-E and Qwen-2.5-M). For the newer models, ChatGPT-4.5 showed a range of 3.0 percentage points across attempts, while QwQ-3B exhibited a range of 6.2 percentage points. The pattern of accuracy across all five attempts for each model is illustrated in Figure 5B, highlighting the remarkable consistency of some models despite multiple independent trials.

3.1.3 Explanation Quality Assessment

Explanation quality varied across models based on comprehensiveness and relevance metrics (Figure 6). According to the validated data, ChatGPT-4.5 provided the highest quality explanations (comprehensiveness: 1.97±0.12; relevance: 1.98±0.09), followed by Claude-3.7-E (comprehensiveness: 1.96±0.09; relevance: 1.96±0.10) and QwQ-3B (comprehensiveness: 1.91±0.17; relevance: 1.81±0.22).

Spearman rank correlation analysis revealed significant positive correlations between explanation quality and answer accuracy for most models. The strongest correlations were observed for Grok-3-T (relevance vs. accuracy: $\rho=0.745$, $p<0.0001$), ChatGPT-4.5 (comprehensiveness vs. accuracy: $\rho=0.525$, $p<0.0001$; relevance vs. accuracy: $\rho=0.525$, $p<0.0001$), and Claude-3.7 (comprehensiveness vs. accuracy: $\rho=0.528$, $p<0.0001$). Only ChatGPT-o1's relevance scores showed no significant correlation with accuracy ($\rho=0.056$, $p=0.3126$). For QwQ-3B, moderate correlations were observed (comprehensiveness vs. accuracy: $\rho=0.376$, $p<0.0001$; relevance vs. accuracy: $\rho=0.288$, $p<0.0001$). ANOVA testing confirmed significant differences in explanation quality metrics between models (both $p<0.0001$).

3.2 Model-Specific Findings

Comparative analysis of individual models revealed distinct performance patterns across accuracy, consistency, and explanation quality metrics (**Figure 7**). ChatGPT-o1 emerged as the overall top performer, achieving the highest accuracy (96.31%) with excellent consistency (ICC=0.971, Fleiss' kappa=0.988) and remarkable stability across attempts (CV=0.8%). Its explanations demonstrated high quality (comprehensiveness: 1.84±0.18, relevance: 1.78±0.17), with comprehensiveness showing a significant correlation with accuracy ($\rho=0.185$, $p=0.0008$). ChatGPT-4.5, the newer ChatGPT variant, ranked second in accuracy (92.62%) with excellent consistency (ICC=0.953, Fleiss' kappa=0.965) and strong stability (CV=1.6%), while providing the highest quality explanations among all models (comprehensiveness: 1.97±0.12, relevance: 1.98±0.09) with significant correlations between both explanation metrics and accuracy (both $\rho=0.525$, $p<0.0001$). ChatGPT-o3-High achieved lower accuracy (88.31%) but maintained excellent consistency (ICC=0.959, Fleiss' kappa=0.938) and strong attempt stability (CV=1.4%), with both explanation quality metrics significantly correlating with accuracy. Claude-3.7-E ranked third in accuracy (89.85%) and provided high-quality explanations (comprehensiveness: 1.96±0.09, relevance: 1.96±0.10), though it demonstrated lower consistency (ICC=0.886) compared to other models. Its standard counterpart, Claude-3.7, achieved lower accuracy (84.31%) but higher consistency (ICC=0.927), while Claude-3.5 showed comparable accuracy (83.69%) with the highest consistency among Claude variants (ICC=0.961, CV=0.9%). The Grok models performed similarly in accuracy (Grok-T: 86.15%, Grok: 84.92%), with Grok demonstrating higher consistency (ICC=0.968 vs. 0.900) but Grok-T showing stronger correlations between relevance and accuracy ($\rho=0.745$, $p<0.0001$). Among the DeepSeek variants, DeepSeek-R1 substantially outperformed the base model in accuracy (85.85% vs. 75.08%), though the base model achieved the highest consistency score among all LLMs (ICC=0.986, Fleiss' kappa=0.945). Both variants maintained similar relevance scores (DeepSeek-R1: 1.78±0.21, DeepSeek: 1.80±0.18), with comparable comprehensiveness-accuracy correlations. Gemini demonstrated moderate accuracy (77.85%) but provided high-quality explanations (comprehensiveness: 1.89±0.16, relevance: 1.81±0.17) with excellent consistency (ICC=0.934). QwQ-3B, despite having the second-lowest accuracy (72.31%), maintained excellent consistency (ICC=0.949, Fleiss' kappa=0.858) with high explanation quality (comprehensiveness: 1.91±0.17, relevance: 1.81±0.22) and significant quality-accuracy correlations ($\rho=0.376$ and $\rho=0.288$, $p<0.0001$). Qwen-2.5-M exhibited the third-lowest accuracy (72.92%) but maintained excellent consistency (ICC=0.940) with moderate explanation quality (comprehensiveness: 1.69±0.26, relevance: 1.65±0.26) and significant quality-accuracy correlations.

3.3. Analysis of LLM-Generated Clinical Scenarios

3.3.1. Overall Performance and Dimensional Analysis

The substantial performance gradient across models is evident in their total scores (Figure 8), with Claude-3.5 and Claude-3.7-E demonstrating markedly superior performance compared to other models, and notable variability in performance consistency as indicated by the confidence intervals. Clinical scenario generation performance varied significantly across models ($F(12,455)=8.48$, $p<0.0001$, $\eta^2=0.1828$). Claude-3.5 achieved the highest overall score (91.4% of maximum possible, mean=22.86±2.37), followed by Claude-3.7-E (89.4%, mean=22.36±2.02) and Grok-3-T (80.9%, mean=20.22±3.01). The newly evaluated ChatGPT-4.5 achieved moderate performance (74.0%, mean=18.50±4.00), ranking seventh overall, while QwQ-3B demonstrated the lowest performance among all models (63.4%, mean=15.86±6.71) with the highest variability. The other lowest-performing models were ChatGPT-o3 (66.8%, mean=16.69±3.93) and Grok-3 (69.9%, mean=17.47±4.50). Tukey's HSD post-hoc testing identified Claude-3.5 as significantly outperforming nine other models ($p<0.05$), except Claude-3.7-E and Grok-3-T. Model performance varied across the five evaluation dimensions: Quality, Complexity, Relevance, Correctness, and Variety (Figure 9). Correctness emerged as the strongest dimension for nine models including both new additions (ChatGPT-4.5: 4.78±0.30; QwQ-3B: 4.31±0.55), while Quality represented the weakest aspect for ten models including QwQ-3B (2.33±0.50), with ChatGPT-4.5 showing Complexity as its weakest dimension (3.06±0.53). Claude models demonstrated robust performance in Quality (Claude-3.5: 4.14±0.30; Claude-3.7-E: 4.25±0.25) and Complexity (Claude-3.5: 4.75±0.25; Claude-3.7-E: 4.53±0.31), while ChatGPT models showed notable weaknesses in these areas (ChatGPT-o1 Quality: 2.17±0.47; ChatGPT-o3 Quality: 1.97±0.46). Kruskal-Wallis tests identified significant differences between models across all dimensions ($p<0.0001$ for Quality, Complexity, Relevance, and Variety; $p=0.0078$ for Correctness). Effect sizes were large for Quality ($\eta^2=0.2346$), Complexity ($\eta^2=0.2549$), Relevance ($\eta^2=0.1583$), and Variety ($\eta^2=0.1456$), with Correctness showing a medium effect ($\eta^2=0.0633$). Correlation analysis revealed strong associations between Quality and Complexity ($r=0.755$) and between Complexity and Variety ($r=0.560$).

3.2.2. Contextual Performance Analysis

3.2.2. Contextual Performance Analysis Performance across the three histology topics (renal corpuscle, nephron, and lower urinary system) showed topic-specific strengths (Figure 10A). Claude-3.5 maintained consistent excellence across all topics (lower urinary: 23.75±1.86; nephron: 22.33±3.26; renal corpuscle: 22.50±1.57), while Gemini-2 displayed marked variability, performing strongly on lower urinary topics (22.00±1.76) but considerably weaker on renal corpuscle content (15.75±3.65). The newly evaluated ChatGPT-4.5 demonstrated moderate consistency across topics (lower urinary: 19.17±4.71; nephron: 18.17±3.51; renal corpuscle: 18.17±3.95), while QwQ-3B showed substantial topic-dependent variability (lower urinary: 13.17±9.70; nephron: 18.08±4.91; renal corpuscle: 16.33±3.39), exhibiting particularly poor performance on lower urinary system topics. ANOVA identified a significant interaction effect between models and topics ($F(24,429)=1.98$, $p=0.0043$), indicating differential performance based on

topic. Comparing scenario creation versus MCQ generation (Figure 10B) revealed that eleven of thirteen models achieved higher scores on scenario creation than question formulation, with this difference reaching statistical significance for Qwen-2.5-M ($p=0.0043$) and QwQ-3B ($p=0.0221$). ChatGPT-4.5 followed this pattern with higher scenario scores (scenario: 19.67±3.81; question: 18.11±4.05) though the difference was not statistically significant ($p=0.3189$). QwQ-3B demonstrated the most substantial performance disparity between scenario creation and question formulation (scenario: 20.22±5.21; question: 14.41±6.59). These findings suggest most LLMs may find generating coherent clinical scenarios somewhat easier than formulating appropriate assessment questions, with this pattern particularly pronounced in certain models.

3.2.3. Model Family Comparison

When grouped by provider family, Claude models (Claude-3.5, Claude-3.7, Claude-3.7-E) demonstrated superior overall performance (mean=21.47±4.18), followed by Gemini (19.44±3.73), Grok (18.85±4.04), DeepSeek (18.13±3.36), ChatGPT (17.71±4.01), and Qwen (17.25±5.80) families (Figure 11A). With the inclusion of ChatGPT-4.5 and QwQ-3B, the overall ranking of model families remained consistent, though the performance gap between ChatGPT and Qwen families narrowed. ANOVA confirmed significant differences between model families ($F(5,462)=12.25$, $p<0.0001$). Tukey's HSD post-hoc analysis revealed that the Claude family significantly outperformed ChatGPT ($p<0.0001$), DeepSeek ($p<0.0001$), Grok ($p=0.0009$), and Qwen ($p<0.0001$) families. Dimensional analysis by family (Figure 11B) showed that ChatGPT models excelled in Correctness (4.77) and Relevance (4.46) despite low Quality scores (2.43), while Claude models demonstrated more balanced performance across all dimensions. The Qwen family, which now includes QwQ-3B, showed the lowest performance in Relevance (3.35) with moderate scores in Correctness (4.42) and Variety (3.86). Principal Component Analysis revealed that 70.98% of variance in model performance could be explained by the first two principal components, with PC1 (49.69% of variance) primarily associated with Quality (0.553) and Complexity (0.549). This analysis further confirmed that Quality and Complexity represent the most discriminative dimensions for evaluating clinical scenario generation capabilities. The PCA visualization (Figure 12) illustrates how models cluster in performance space, with ChatGPT-4.5 positioned closer to the center of the distribution and QwQ-3B clustered with lower-performing models.

4. Discussion

4.1. Key Findings and Implications

Our comprehensive evaluation of thirteen large language models revealed a complex landscape of capabilities in medical education contexts. ChatGPT-o1 demonstrated superior MCQ performance (96.31% accuracy), with ChatGPT-4.5 ranking second (92.62% accuracy) while providing the highest quality explanations (comprehensiveness: 1.97 ± 0.12, relevance: 1.98 ± 0.09). Claude-3.5 excelled in clinical scenario generation (91.4% of maximum possible score), significantly outperforming other models including the newer QwQ-3B, which ranked lowest (63.4%). This performance dichotomy

addresses our primary research question by revealing that LLM capabilities are task-specific rather than universal, suggesting that optimal educational integration requires matching models to appropriate learning objectives. The significant correlation between explanation quality and answer accuracy observed in most models - particularly strong in Grok-3-T (relevance vs. accuracy: $\rho = 0.745$, $p < 0.0001$), ChatGPT-4.5 (both metrics: $\rho = 0.525$, $p < 0.0001$), and Claude-3.7 (comprehensiveness vs. accuracy: $\rho = 0.528$, $p < 0.0001$) - provides crucial insight into these models' cognitive architecture. This pattern indicates that correct answers likely emerge from robust underlying knowledge representations rather than superficial pattern matching. For medical educators, this suggests these models can potentially explain their reasoning in ways that enhance student understanding. Consistency analysis revealed substantial variations between models, with DeepSeek maintaining remarkable stability in MCQ responses (ICC = 0.986, Fleiss' kappa = 0.945) compared to other models. This variance directly addresses our question about response stability across multiple attempts, highlighting that reliability differs substantially between models. The considerable performance variability observed in models like Claude-3.7-E (CV = 4.1%) and QwQ-3B (CV = 3.0%) raises concerns about potential inconsistent guidance for self-directed learners. Our analysis revealed significant content imbalances in generated clinical scenarios, with quantitative term frequency analysis showing overwhelming focus on certain structures (podocytes: 128 mentions in renal corpuscle scenarios) while critical components received minimal attention (vascular pole: 1 mention), as documented in Appendix 3. This pattern manifested across all histological domains, with the lower urinary system showing similar disparities (bladder: 130 mentions vs. trigone: 2 mentions). These imbalances explain the significant Model × Topic interaction ($F(24,429) = 1.98$, $p = 0.0043$) and highlight fundamental limitations in these models' knowledge representation that directly impact educational utility. These findings exemplify 'artificial hallucinations' in LLMs, where models prioritize high-frequency terms (e.g., *podocytes*) while omitting essential structures (e.g., *trigone*: 2 mentions), despite achieving high correctness scores (mean: 4.75/5). This aligns with Alkaissi & McFarlane [5], who demonstrated ChatGPT's generation of fabricated anatomical mechanisms (e.g., homocysteine impairing vitamin K-dependent osteocalcin carboxylation) and falsified references. Similar concerns have been documented in recent surveys of medical students, who cite artificial hallucinations as a primary reason for hesitancy in using ChatGPT for clinical applications [33, 34]. Similarly, Ayers et al. [6] found that only 22% of ChatGPT's public health responses included referrals to critical resources (e.g., hotlines). Notably, 19.2% of scenario explanations (n = 12/63) excluded key histological relationships (e.g., juxtaglomerular apparatus), mirroring these patterns. Such gaps persist even in top performers like Claude-3.5 (scenario score: 91.4%), underscoring the need for hybrid validation frameworks like those proposed by Omar et al. [35], who reduced hallucinations by integrating domain-specific retrieval-augmented generation (RAG) with curated literature databases. Expert evaluators noted that models often lacked fundamental histological questions first-year medical students would expect, such as "What is the tissue in the distal convoluted tubule?" or "Where are the macula densa cells located?" Instead, models frequently produced inappropriate complexity levels, asking diagnostic questions beyond the scope of first-year students. The dimensional analysis of clinical scenario generation revealed that Correctness emerged as the strongest dimension for most models, while Quality represented the weakest aspect for the majority of models, as shown in the result. This finding is consistent with

systematic reviews of LLM capabilities in medical education, which have identified a similar pattern of strong factual accuracy but weaker pedagogical quality [34]. This pattern was particularly pronounced in ChatGPT models, which excelled in Correctness (4.77) and Relevance (4.46) despite low Quality scores (2.43). In their blinded evaluations, the expert reviewers identified specific pedagogical limitations: scenarios often presented information in an overly transparent manner where answers could be directly extracted from the text, many exceeded appropriate length constraints, and several disregarded the prompt to be concise. The strong correlation between Quality and Complexity dimensions ($r = 0.755$) further suggests that models producing more sophisticated scenarios also tend to create more pedagogically effective content. The PCA revealed that 70.98% of variance in model performance could be explained by the first two principal components, with PC1 (49.69% of variance) primarily associated with Quality (0.553) and Complexity (0.549). This dimensional pattern provides a valuable framework for future evaluation methodologies. The differential performance in scenario versus question generation observed across models - reaching statistical significance for Qwen-2.5-M ($p = 0.0043$) and QwQ-3B ($p = 0.0221$) - illuminates distinct cognitive processes involved in recalling medical facts versus constructing novel clinical applications. QwQ-3B demonstrated the most substantial performance disparity between scenario creation and question formulation (scenario: 20.22 ± 5.21; question: 14.41 ± 6.59). This distinction parallels the challenges students face when transitioning from knowledge acquisition to clinical reasoning. Ganjavi et al. [33] observed similar patterns in their survey of medical students using LLMs. When grouped by provider family, Claude models demonstrated superior overall performance (mean = 21.47 ± 4.18), significantly outperforming ChatGPT ($p < 0.0001$), DeepSeek ($p < 0.0001$), Grok ($p = 0.0009$), and Qwen ($p < 0.0001$) families. The introduction of ChatGPT-4.5 improved the ChatGPT family's ranking (17.71 ± 4.01), though not enough to close the gap with Claude models. This family-level analysis suggests fundamental architectural differences in how these models represent medical knowledge, with Claude models demonstrating more balanced performance across all dimensions.

4.2. Comparison with Existing Literature

Our study advances LLM evaluation in medical education through methodological innovations that address limitations in existing research. While Li et al.'s [36] meta-ethnographic synthesis established AI's theoretical potential in medical education, our research provides concrete performance metrics across multiple dimensions for thirteen state-of-the-art LLMs. Our findings significantly refine Mavrych et al.'s results [32], who reported ChatGPT-4 achieving 60.5% accuracy on gross anatomy MCQs. In contrast, our evaluation revealed substantially higher performance for ChatGPT-o1 (96.31%) on urinary system histology, with other models also demonstrating strong performance (ChatGPT-4.5: 92.62%, Claude-3.7-E: 89.85%). These findings align with recent observations that newer LLM iterations show marked improvements in specialized medical knowledge domains [34]. Similarly, we extend Levin et al.'s [37] meta-analysis, which reported a mean accuracy of 61.1% across medical specialties for ChatGPT-3.5. Our significantly higher performance metrics (ranging from 72.31–96.31%) suggest newer model versions and specialized domain applications substantially outperform earlier benchmarks. Unlike previous studies [32, 36], our examination of consistency revealed important variations that had not been previously documented. DeepSeek demonstrated exceptional stability (ICC = 0.986, Fleiss' kappa =

0.945), surpassing even ChatGPT-4.5 (ICC = 0.953, Fleiss' kappa = 0.965) and ChatGPT-o1 (ICC = 0.971, Fleiss' kappa = 0.988). This finding highlights that response consistency - critical for educational tools - varies significantly across model architectures and warrants explicit evaluation. Interestingly, this aligns with medical students' experiences, as Ganjavi et al. [33] found that 28.5% of students reported variability and inconsistency as a key limitation when using ChatGPT. Our term frequency analysis (Appendix 3) provides unprecedented quantitative evidence of knowledge representation gaps (e.g., podocytes: 128 mentions vs. vascular pole: 1 mention in renal corpuscle scenarios), demonstrating how these imbalances directly affect both statistical performance patterns and educational utility. Similarly, the lower urinary system showed comparable disparities (bladder: 130 mentions vs. trigone: 2 mentions). This extends beyond previous studies [38] by quantifying the specific content limitations that impact LLM-generated educational materials. Most significantly, we documented topic-specific variation ($F(24,429) = 1.98$, $p = 0.0043$) not previously reported, with models like Gemini-2 performing strongly on lower urinary system content (22.00 ± 1.76) but considerably weaker on renal corpuscle content (15.75 ± 3.65). Similarly, QwQ-3B demonstrated marked disparity in performance across different topics (lower urinary: 13.17 ± 9.70; nephron: 18.08 ± 4.91). This suggests LLM performance should be evaluated within specific knowledge domains rather than through generalized assessments. This domain-specific performance variation is consistent with findings from Xu et al. [34], who observed that LLMs perform differently based on question types and medical subject areas. Our model family analysis revealed that Claude models significantly outperformed other families in clinical scenario generation (mean = 21.47 ± 4.18), including ChatGPT ($p < 0.0001$), DeepSeek ($p < 0.0001$), Grok ($p = 0.0009$), and Qwen ($p < 0.0001$) families. This indicates architectural differences between model families may be as important as individual model capabilities; an insight not captured in prior comparative studies. The PCA results showing that Quality and Complexity account for the largest portion of variance (49.69%) provide a novel quantitative framework for evaluating LLMs; one that recognizes pedagogical attributes as equally important to factual correctness. This addresses limitations in previous studies [32, 36, 37, 38, 39] that focused primarily on accuracy metrics without considering educational dimensions.

4.3. Limitations and Future Directions

Several methodological limitations warrant consideration when interpreting our findings. First, our evaluation focused exclusively on urinary system histology, potentially limiting generalizability to other anatomical systems or medical specialties. The observed topic-specific performance variations suggest that model capabilities may be domain-dependent rather than universal, as evidenced by significant Model × Topic interaction ($F(24,429) = 1.98$, $p = 0.0043$) and pronounced variability in models like QwQ-3B across different histological topics (lower urinary: 13.17 ± 9.70; nephron: 18.08 ± 4.91). Second, our study utilized text-only MCQs without image-based questions, a significant constraint given histology's inherently visual nature. This limitation particularly affects our ability to assess how these models might perform on visual recognition tasks central to histology education and practice. Third, our content analysis revealed substantial imbalances in model-generated materials, with systematic over-representation of certain structures (podocytes: 128 mentions, bladder: 130 mentions) and near-complete omission of others (vascular pole: 1 mention, trigone: 2 mentions). This pattern constitutes a

fundamental limitation in educational utility that may extend beyond the specific domain studied. Fourth, despite our multi-dimensional evaluation approach, standardizing the assessment of explanation quality and clinical scenario relevance remains challenging. While we implemented rigorous evaluation criteria, the inherent subjectivity in rating pedagogical quality represents a methodological constraint. Fifth, the rapid evolution of LLM capabilities presents a significant challenge for longitudinal validation. The late addition of ChatGPT-4.5 and QwQ-3B to our analysis, with their distinct performance profiles, exemplifies how quickly the technology landscape changes, necessitating regular reassessment protocols for educational implementation.

Future research should address these limitations through several approaches. First, expanding evaluation across multiple anatomical systems and medical specialties would determine whether the patterns we identified represent domain-specific phenomena or broader capabilities and limitations. Second, incorporating image-based assessment would better align with histology education needs. This could include evaluating models' ability to identify structures in histological images or generate questions about visual content, providing more comprehensive performance data. Third, developing targeted prompting strategies to address identified pedagogical limitations could significantly enhance educational utility. Experiments with various prompting approaches might yield improvements in content balance, question complexity, and adherence to educational parameters. Fourth, investigations directly measuring learning outcomes when medical students utilize these models as supplementary tools would clarify their actual educational impact. Controlled studies examining whether high-performing models actually improve student comprehension and retention would provide valuable insights for curriculum integration. Finally, developing standardized validation frameworks that account for rapid model evolution would enable more consistent cross-study comparisons and provide medical educators with regularly updated performance metrics to guide implementation decisions.

5. Conclusion

Our comprehensive evaluation of thirteen LLMs across multiple dimensions of urinary system histology education reveals model-specific strengths: ChatGPT-o1 achieved highest MCQ accuracy, ChatGPT-4.5 excelled in explanation quality, while Claude-3.5 dominated clinical scenario generation. Despite these capabilities, significant concerns persist, including anatomical content imbalances and variable consistency metrics between models. No single LLM currently demonstrates universal mastery across all pedagogical tasks; instead, effective implementation requires task-specific deployment matched to educational objectives. Expert oversight remains essential to mitigate hallucinations and ensure curriculum alignment. Future research should expand domain scope, incorporate image-based assessments, and explore targeted prompting strategies. By embracing rigorous evaluation protocols and strategic model selection, medical educators can responsibly integrate these evolving tools while balancing factual accuracy with pedagogical integrity.

Declarations

Competing interests

The authors declare no conflict of interest.

Authors' Contributions

A.S. conceptualized the study, performed data curation, conducted the investigation, and wrote the original draft. G.D. developed the methodology, performed formal analysis, administered the project, provided supervision, and conducted validation. Both A.S. and G.D. contributed to reviewing and editing the manuscript

Funding

No funding.

Ethical considerations

This study did not involve human subjects, patient data, or biological materials. The research utilized publicly available large language models and evaluated their performance on validated educational content. All MCQs used in this study were developed and validated for educational purposes, adhering to standard medical education practices. No institutional review board approval was required for this type of technological evaluation research. The study adhered to responsible AI research practices, focusing on educational applications without involving potentially sensitive clinical decision-making or patient care scenarios.

Data Availability Statement

The complete research data is provided in the appendices of this article rather than an external repository to ensure direct access for readers within the publication itself. This approach was chosen because the datasets are of moderate size and directly relevant to the analytical procedures described in the manuscript, making within-article access most appropriate for readers evaluating our methods and findings.

Acknowledgments

None.

References

1. Bair H, Norden J. Large Language Models and Their Implications on Medical Education. Acad Med. 2023;98(8):869-70. doi: 10.1097/ACM.0000000000005265
2. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. 2024;58(11):1276-85. doi: 10.1111/medu.15402

3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. doi: 10.1371/journal.pdig.0000198
4. Bharatha A, Ojeh N, Fazle Rabbi AM, et al. Comparing the Performance of ChatGPT-4 and Medical Students on MCQs at Varied Levels of Bloom's Taxonomy. Adv Med Educ Pract. 2024;15:393-400. doi: 10.2147/AMEP.S457408
5. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus. 2023;15(2):e35179. doi: 10.7759/cureus.35179
6. Ayers JW, Zhu Z, Poliak A, et al. Evaluating Artificial Intelligence Responses to Public Health Questions. JAMA Netw Open. 2023;6(6):e2317517. doi: 10.1001/jamanetworkopen.2023.17517
7. Lee J, Park S, Shin J, Cho B. Analyzing evaluation methods for large language models in the medical field: a scoping review. BMC Med Inform Decis Mak. 2024;24(1):366. doi: 10.1186/s12911-024-02709-7
8. Haddad F, Saade JS. Performance of ChatGPT on Ophthalmology-Related Questions Across Various Examination Levels: Observational Study. JMIR Med Educ. 2024;10:e50842. doi: 10.2196/50842
9. Maitland A, Fowkes R, Maitland S. Can ChatGPT pass the MRCP (UK) written examinations? Analysis of performance and errors using a clinical decision-reasoning framework. BMJ Open. 2024;14(3):e080558. doi: 10.1136/bmjopen-2023-080558
10. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al. Evaluating the Efficacy of ChatGPT in Navigating the Spanish Medical Residency Entrance Examination (MIR): Promising Horizons for AI in Clinical Medicine. Clin Pract. 2023;13(6):1460-87. doi: 10.3390/clinpract13060130
11. Li R, Kumar A, Chen JH. How Chatbots and Large Language Model Artificial Intelligence Systems Will Reshape Modern Medicine: Fountain of Creativity or Pandora's Box? JAMA Intern Med. 2023;183(6):596-7. doi: 10.1001/jamainternmed.2023.1835
12. Sun L, Yin C, Xu Q, Zhao W. Artificial intelligence for healthcare and medical education: a systematic review. Am J Transl Res. 2023;15(7):4820-8.
13. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond). 2025;5(1):26. doi: 10.1038/s43856-024-00717-2
14. Collins BR, Black EW, Rarey KE. Introducing AnatomyGPT: A customized artificial intelligence application for anatomical sciences education. Clin Anat. 2024;37(6):661-9. doi: 10.1002/ca.24178
15. Wong RS, Ming LC, Raja Ali RA. The Intersection of ChatGPT, Clinical Medicine, and Medical Education. JMIR Med Educ. 2023;9:e47274. doi: 10.2196/47274
16. Masters K, Benjamin J, Agrawal A, MacNeill H, Pillow MT, Mehta N. Twelve tips on creating and using custom GPTs to enhance health professions education. Med Teach. 2024;46(6):752-6. doi: 10.1080/0142159X.2024.2305365
17. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. doi: 10.1038/s41591-018-0300-7

18. Funk PF, Hoch CC, Knoedler S, et al. ChatGPT's Response Consistency: A Study on Repeated Queries of Medical Examination Questions. Eur J Investig Health Psychol Educ. 2024;14(3):657-68. doi: 10.3390/ejihpe14030043
19. **DeepSeek-AI**. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning (Internet). arXiv preprint arXiv:2501.12948. 2025. Available from: https://arxiv.org/pdf/2501.12948
20. **DeepSeek-AI**. DeepSeek-V3 Technical Report (Internet). arXiv preprint arXiv:2412.19437. 2025. Available from: https://arxiv.org/pdf/2412.19437
21. **OpenAI**. OpenAI o3-mini: Pushing the frontier of cost-effective reasoning (Internet). OpenAI Research Blog; 2025 Jan 31. Available from: https://openai.com/research/openai-o3-mini
22. OpenAI. Introducing OpenAI o1-preview [Internet]. 2024 Sep 12. Available from: https://openai.com/research/introducing-openai-o1-preview
23. Anthropic. Claude 3.5 Sonnet: Advancements in AI Capabilities [Internet]. 2024 Jun 20. Available from: https://www.anthropic.com/claude-3-5-sonnet
24. Alibaba Cloud. Qwen 2.5-Max: The Next Generation of AI Models. Qwen AI (Internet). 2025 Jan 29. Available from: https://qwen-ai.com/2-5
25. Google. Introducing Gemini 2.0: our new AI model for the agentic era. Google (Internet). 2024 Dec 11. Available from: https://blog.google/technology/ai/introducing-gemini-2-0/
26. **Anthropic**. Claude 3.7 Sonnet and Claude Code (Internet). Anthropic Technical Report; 2025 Feb 24. Available from: https://www.anthropic.com/news/claude-3-7-sonnet
27. **Anthropic**. Visible Extended Thinking: How Claude Thinks (Internet). Anthropic Research Blog; 2025 Feb 24. Available from: https://www.anthropic.com/news/visible-extended-thinking
28. **xAI**. Grok 3 Beta – The Age of Reasoning Agents (Internet). xAI Research Blog; 2025 Feb 19. Available from: https://x.ai/blog/grok-3
29. **xAI**. Advances in Chain-of-Thought Reasoning in Grok Models (Internet). xAI Research Report; 2025 Feb 19. Available from: https://x.ai/research/grok-reasoning.
30. Alibaba Cloud. Alibaba Cloud unveils QwQ-32B: A compact reasoning model with cutting-edge performance (Internet). Alibaba Cloud Blog; 2024 (cited 2025 Mar 7). Available from: https://www.alibabacloud.com/blog/alibaba-cloud-unveils-qwq-32b-a-compact-reasoning-model-with-cutting-edge-performance_602039.
31. OpenAI. Introducing GPT-4.5 [Internet]. 2025 Feb 27 [cited 2025 Mar 8]. Available from: https://openai.com/research/introducing-gpt-4-5
32. Mavrych V, Ganguly P, Bolgova O. Using large language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in Gross Anatomy course: Comparative analysis. Clin Anat. 2025;38(2):200-10. doi: 10.1002/ca.24244
33. Ganjavi C, Eppler M, O'Brien D, et al. ChatGPT and large language models (LLMs) awareness and use. A prospective cross-sectional survey of U.S. medical students. PLOS Digit Health. 2024;3(9):e0000596. doi: 10.1371/journal.pdig.0000596

34. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. J Educ Eval Health Prof. 2024;21:6. doi: 10.3352/jeehp.2024.21.6
35. Omar M, Ullanat V, Loda M, Marchionni L, Umeton R. ChatGPT for digital pathology research. Lancet Digit Health. 2024;6(8):e595-600. doi: 10.1016/S2589-7500(24)00114-6
36. Li W, Shi HY, Chen XL, et al. Application of artificial intelligence in medical education: A meta-ethnographic synthesis. Med Teach. 2024 Oct 31. doi: 10.1080/0142159X.2024.2418936. Online ahead of print.
37. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: A systematic review and a meta-analysis. BJOG. 2024;131(3):378-80. doi: 10.1111/1471-0528.17641.
38. Shahzad T, Mazhar T, Tariq MU, et al. A comprehensive review of large language models: issues and solutions in learning environments. Discov Sustain. 2025;6:27. doi: 10.1007/s43621-025-00815-8
39. Sapci AH, Sapci HA. Artificial Intelligence Education and Tools for Medical and Health Informatics Students: Systematic Review. JMIR Med Educ. 2020;6(1):e19285. doi: 10.2196/19285.

Figures

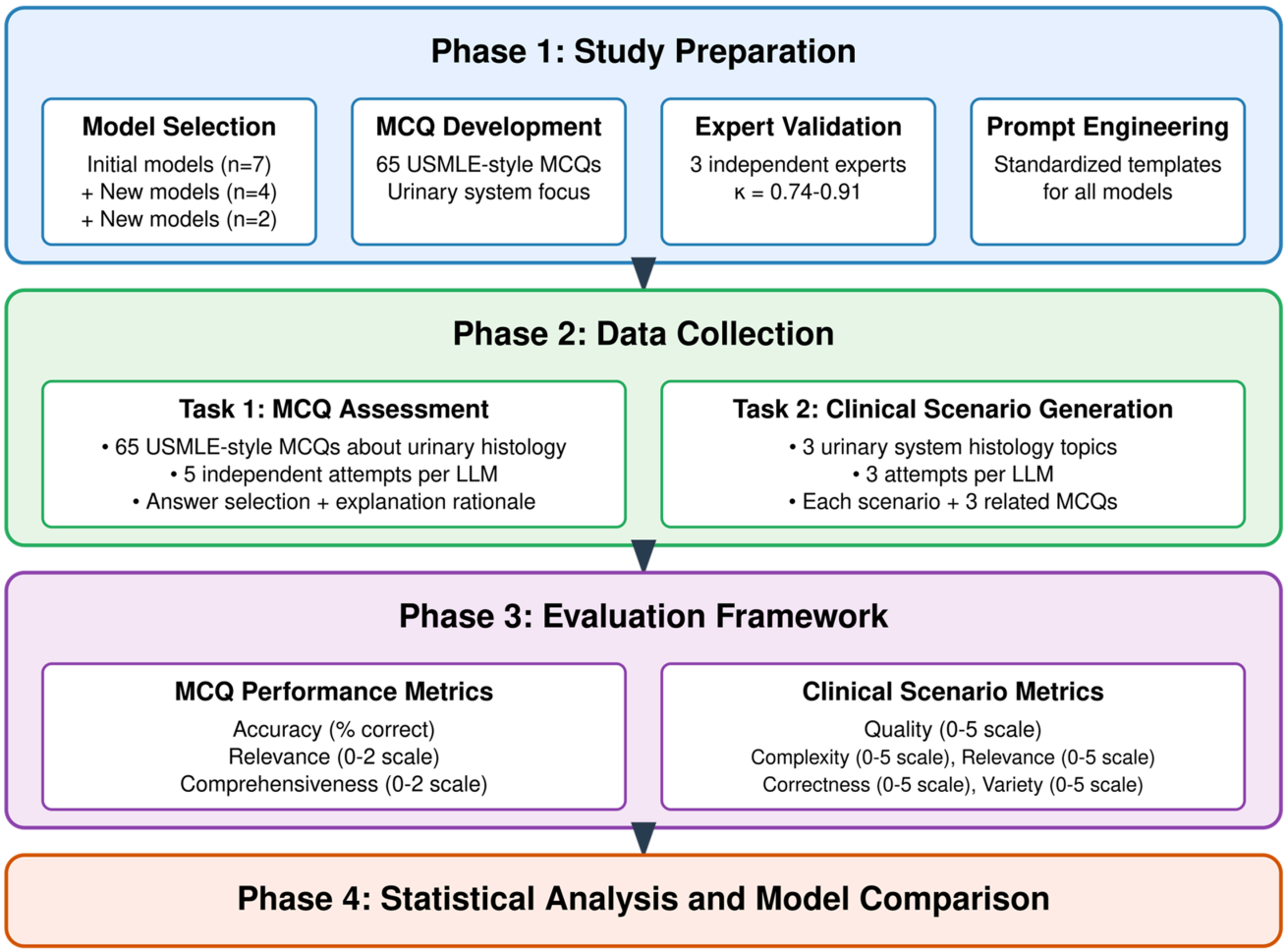


Figure 1

Flow of work

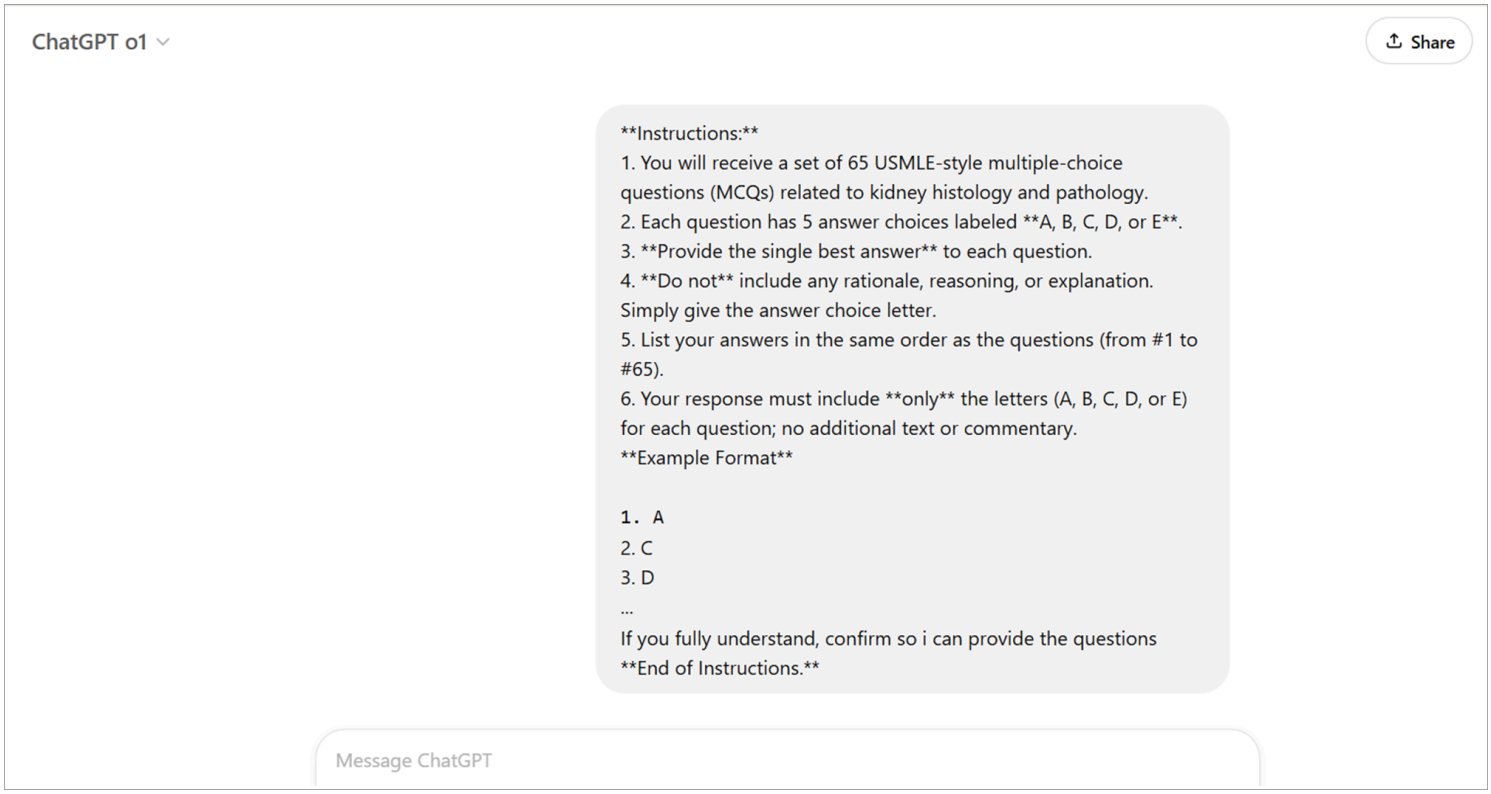


Figure 2

Example of the prompt in the ChatGPT-o1 for respond to 65 MCQs.

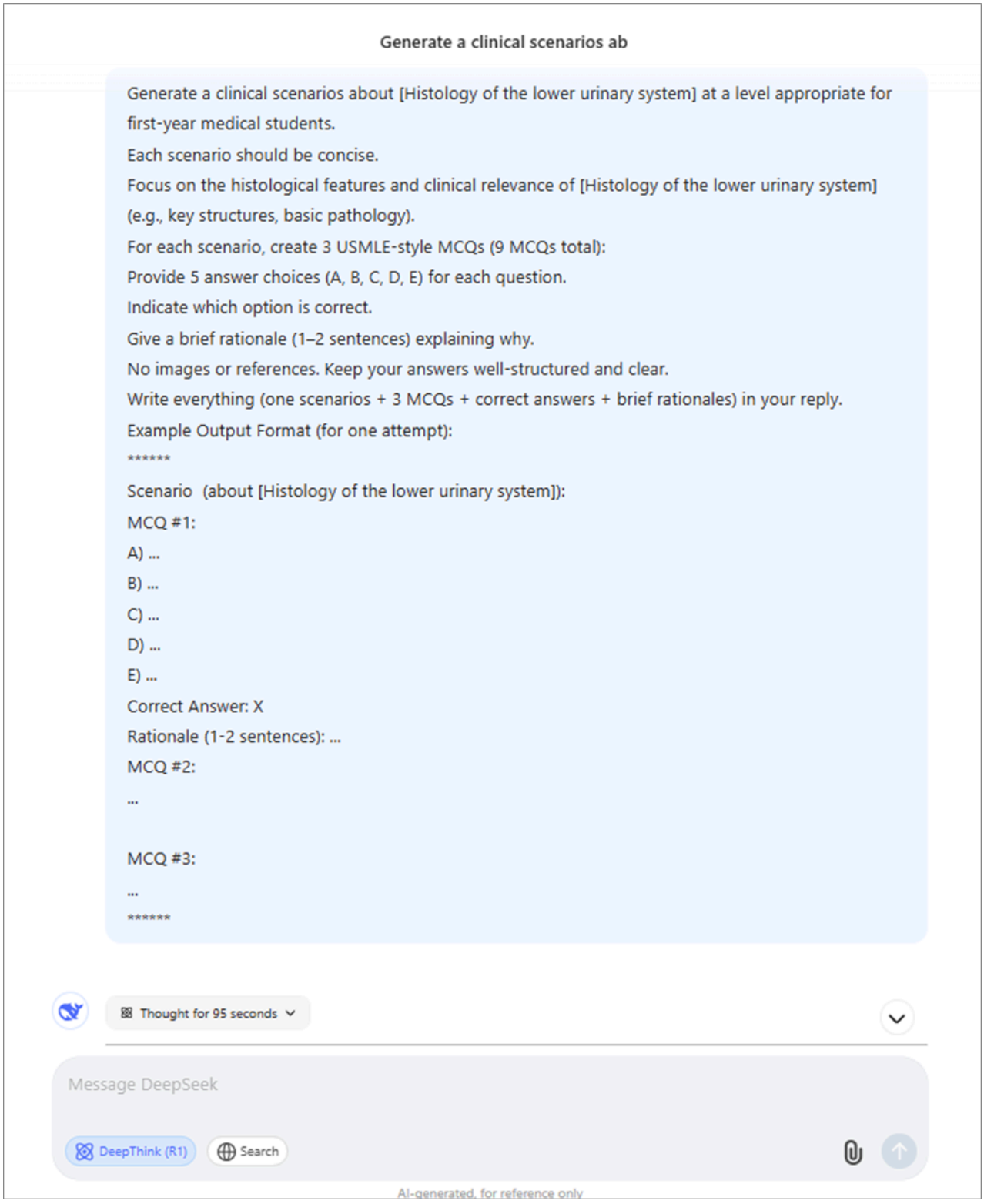


Figure 3

Example of the prompt in the DeepSeek-R1 for create clinical scenarios.

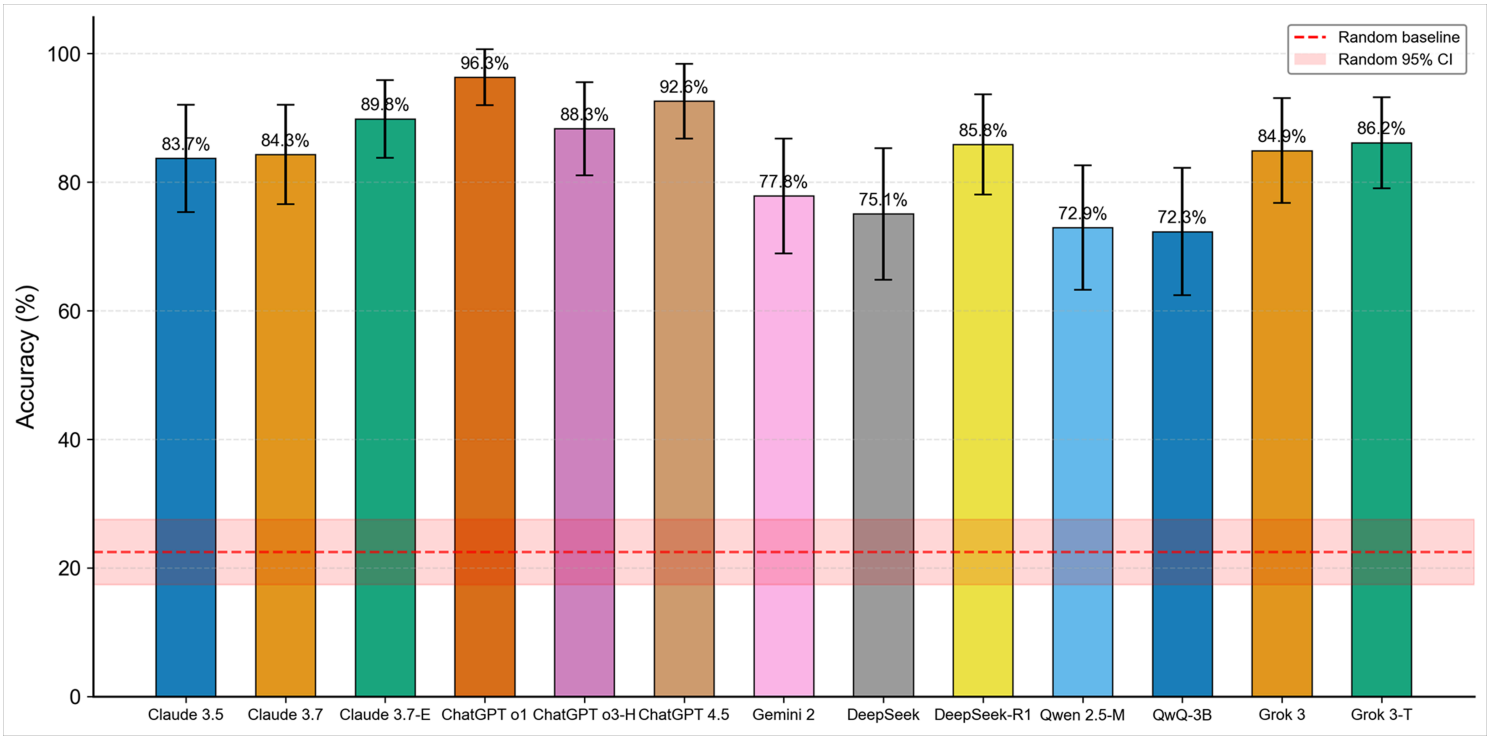
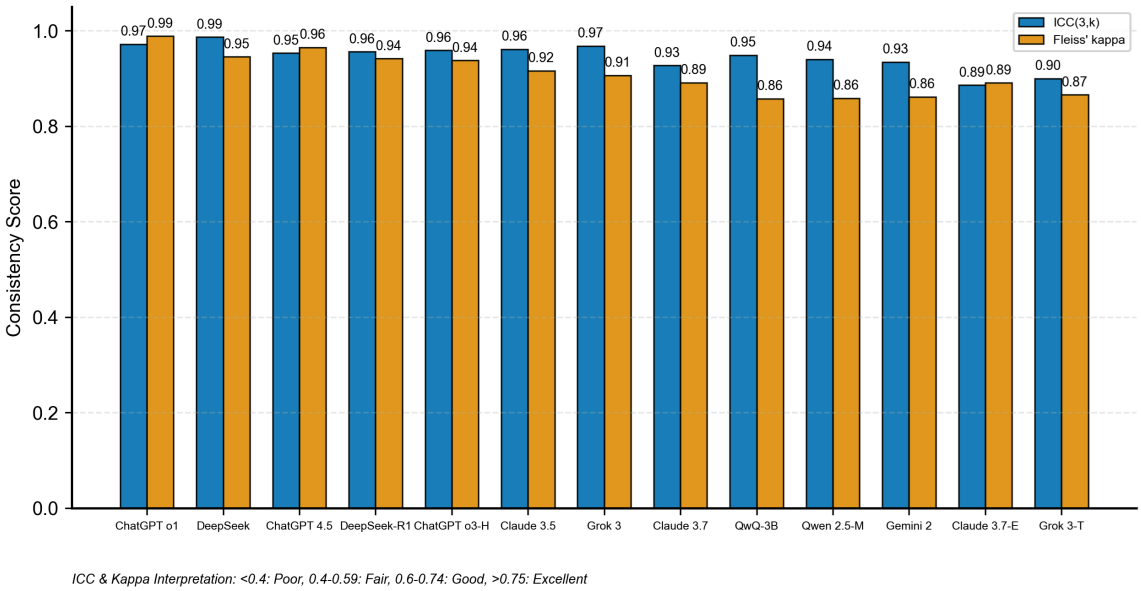


Figure 4

Model accuracy on medical MCQs with 95% confidence intervals. Bar chart displaying the percentage accuracy (mean ± 95% confidence intervals) for all 13 LLMs compared to random guessing baseline. ChatGPT-o1 achieved highest accuracy (96.31%), followed by ChatGPT-4.5 (92.62%), with all models significantly outperforming random guessing ($p<0.0001$).



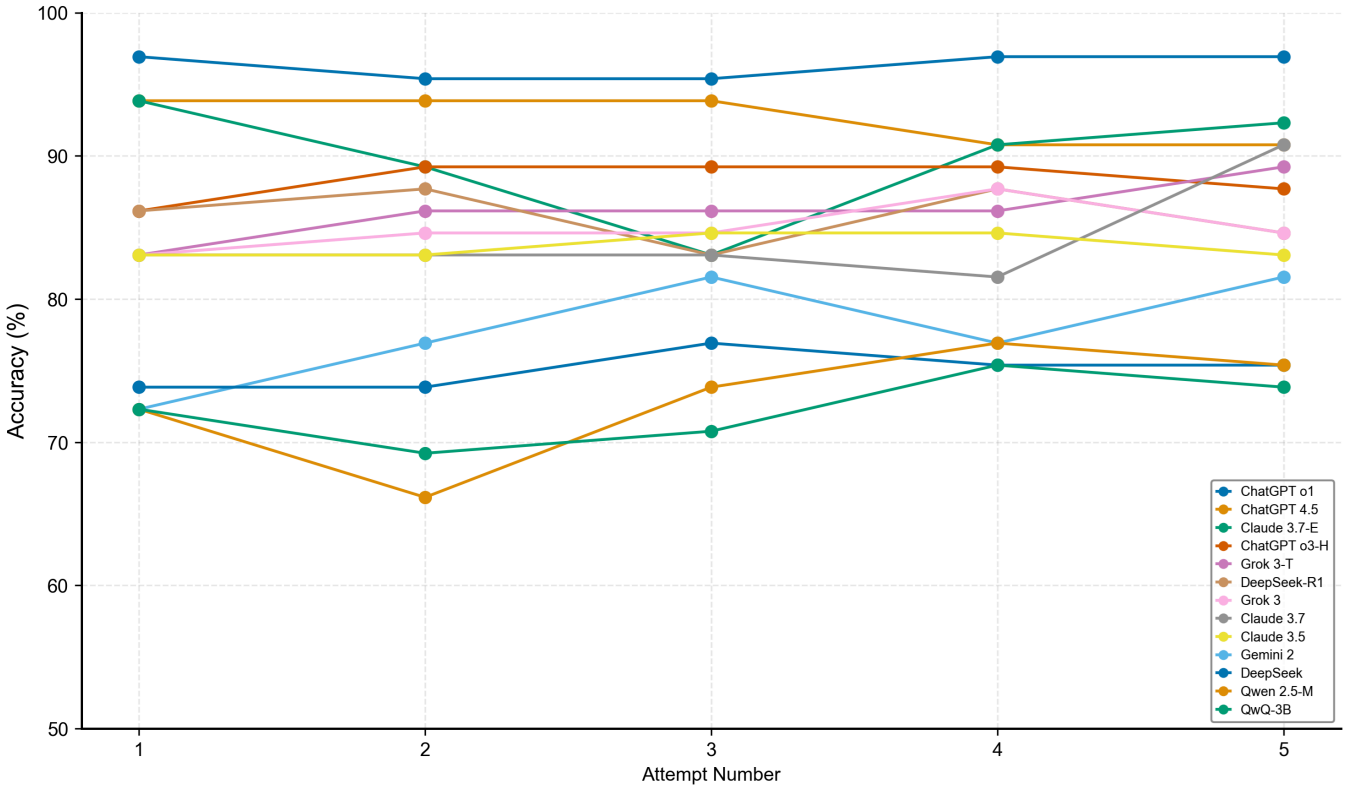


Figure 5

A. Consistency metrics across LLMs. Bar chart comparing intraclass correlation coefficients (ICC) and Fleiss' kappa values across all models, with DeepSeek demonstrating highest consistency (ICC=0.986, κ=0.945) followed by ChatGPT-o1 (ICC=0.971, κ=0.988).

B. Accuracy stability across multiple attempts for all models. Line graph showing accuracy percentages across five independent attempts for each model, highlighting remarkable stability in some models despite multiple trials, with ChatGPT-o1 and Claude-3.5 showing lowest variation (CV=0.8% and 0.9% respectively).

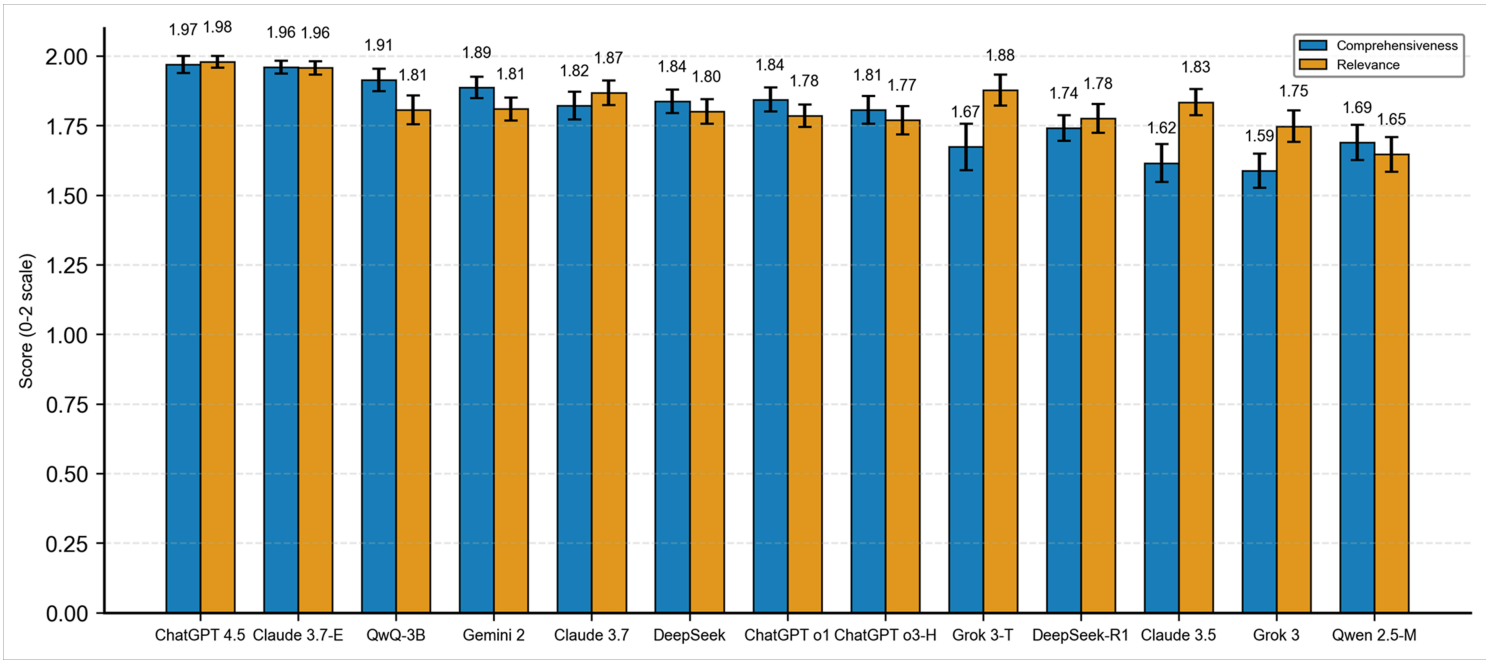


Figure 6

Explanation quality metrics across models with 95% confidence intervals. Dual-metric visualization of comprehensiveness and relevance scores (0-2 scale) with 95% confidence intervals for all LLMs. ChatGPT-4.5 provided highest quality explanations (comprehensiveness: 1.97±0.12; relevance: 1.98±0.09).

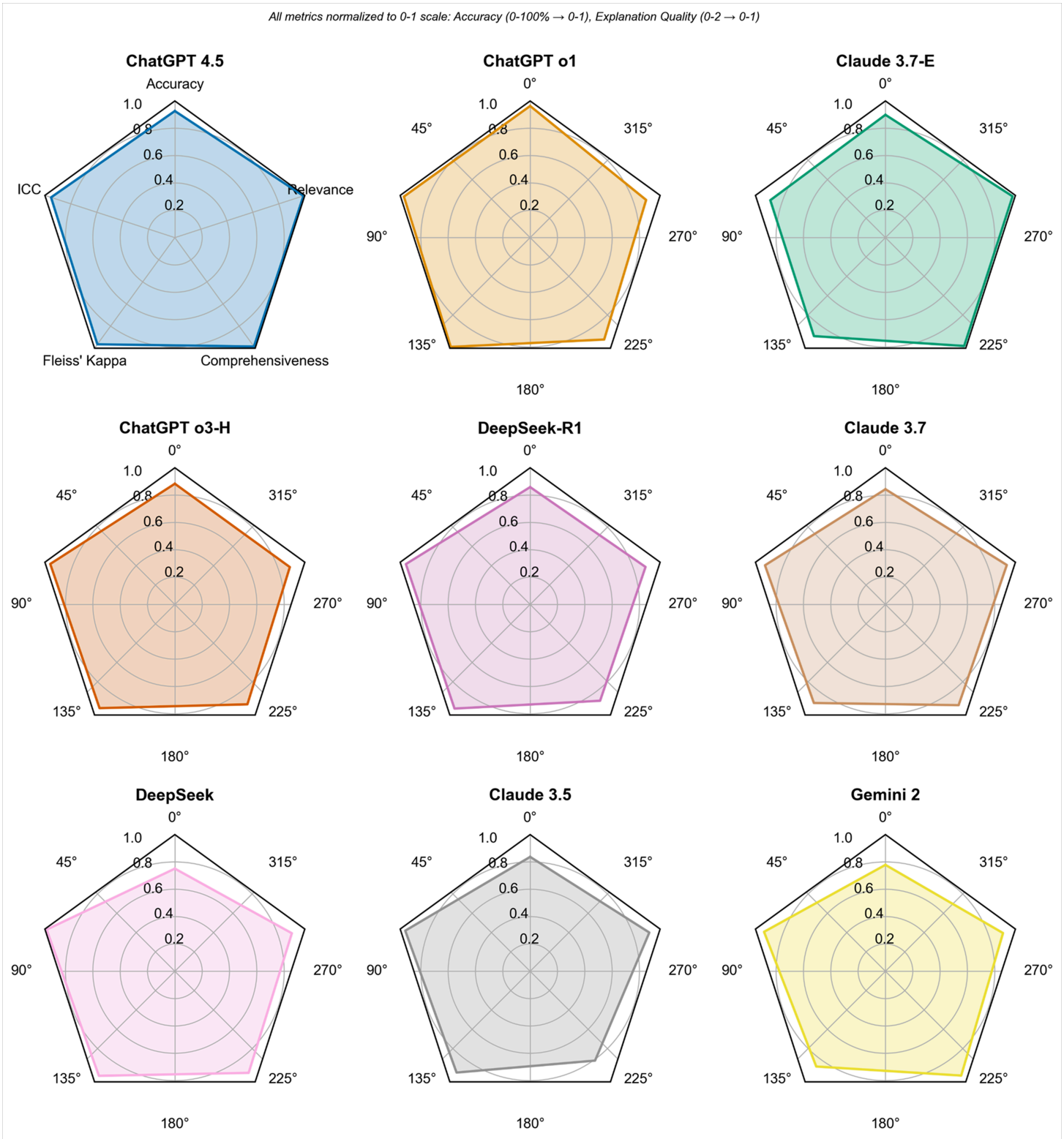


Figure 7

Normalized performance profiles across LLMs. Radar chart showing relative performance across multiple metrics (accuracy, consistency, explanation quality) for each model, enabling multi-dimensional comparison of strengths and weaknesses.

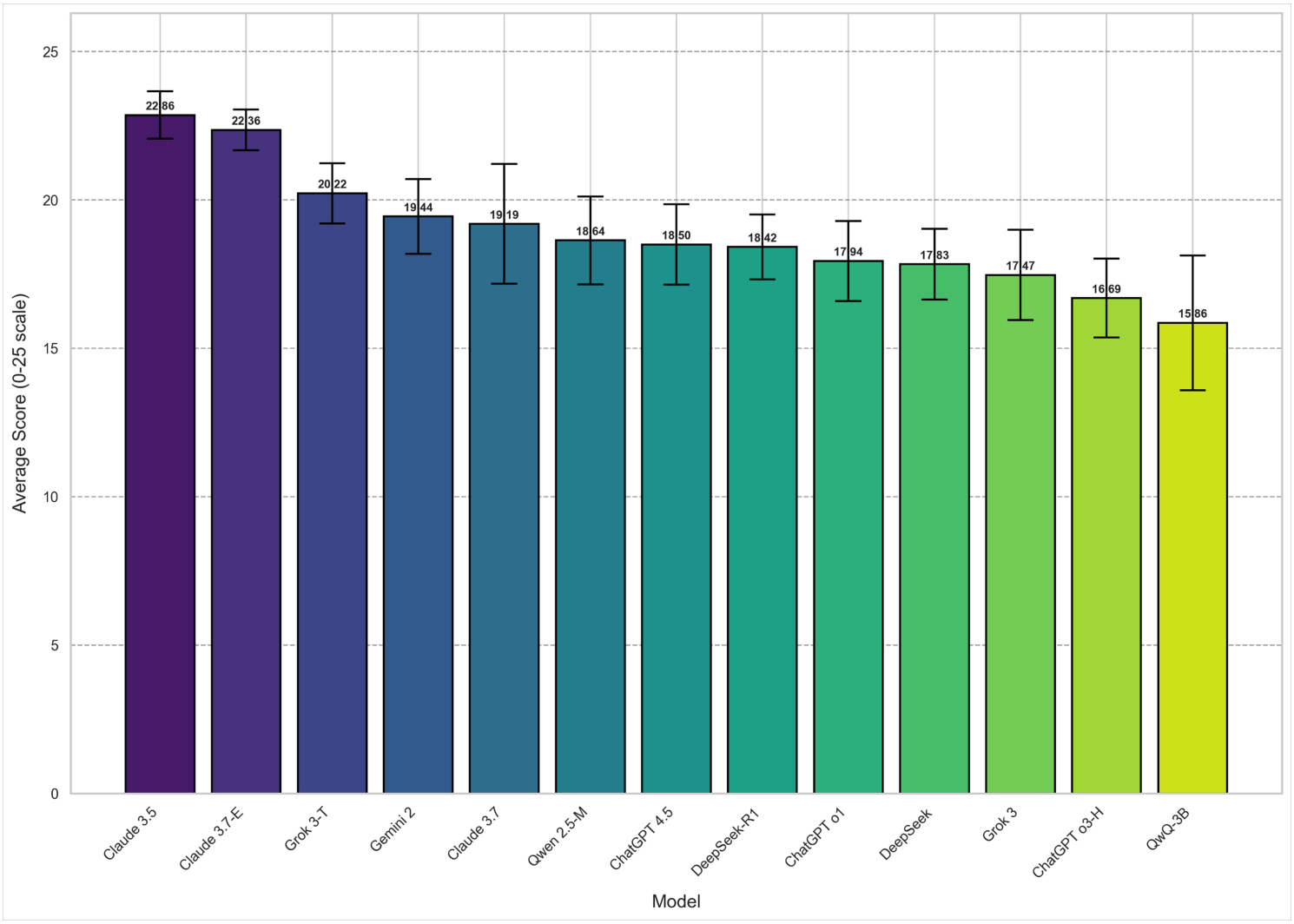
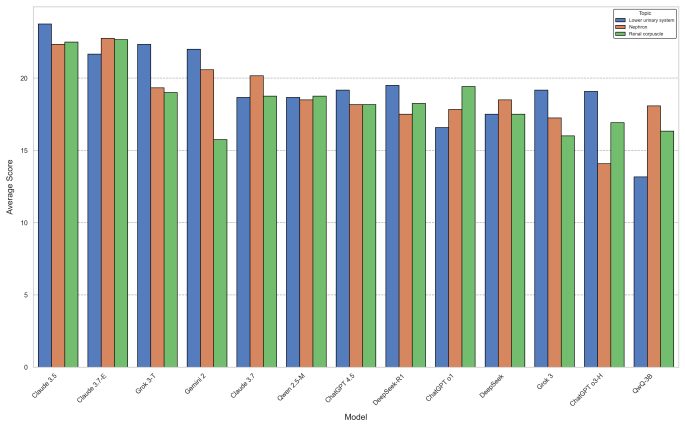


Figure 8

Overall clinical scenario generation performance across LLMs with 95% confidence intervals. Bar chart showing overall performance scores (with 95% confidence intervals) for clinical scenario generation tasks. Claude-3.5 achieved highest overall score (91.4% of maximum possible, mean=22.86±2.37), followed by Claude-3.7-E (89.4%).

Figure 9

Dimensional performance heatmap for clinical scenario generation. Color-coded heatmap visualizing model performance across five evaluation dimensions (Quality, Complexity, Relevance, Correctness, Variety), revealing Correctness as the strongest dimension for most models and Quality as the weakest.



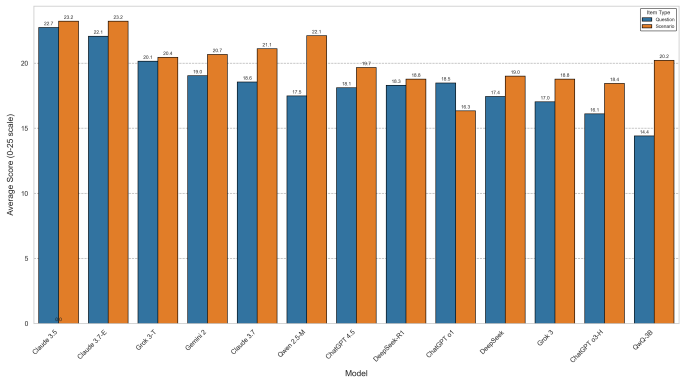
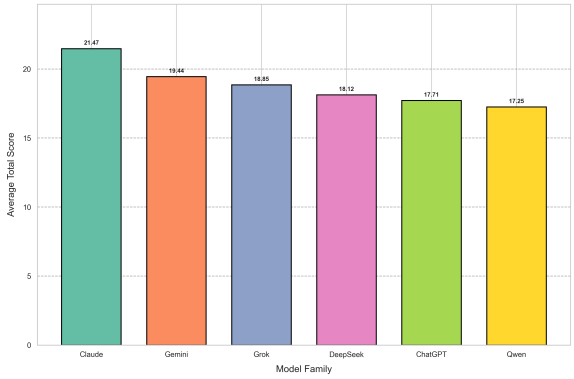


Figure 10

A. Contextual performance analysis: Performance by histology topic. Bar chart comparing model performance across three histological topics (renal corpuscle, nephron, lower urinary system), showing significant topic-dependent variability, particularly in models like Gemini-2 and QwQ-3B.

B.Contextual performance analysis: Performance by item type. Bar chart comparing performance in scenario creation versus question formulation tasks, showing 11 of 13 models achieved higher scores on scenario creation than question formulation, with QwQ-3B showing the largest disparity.



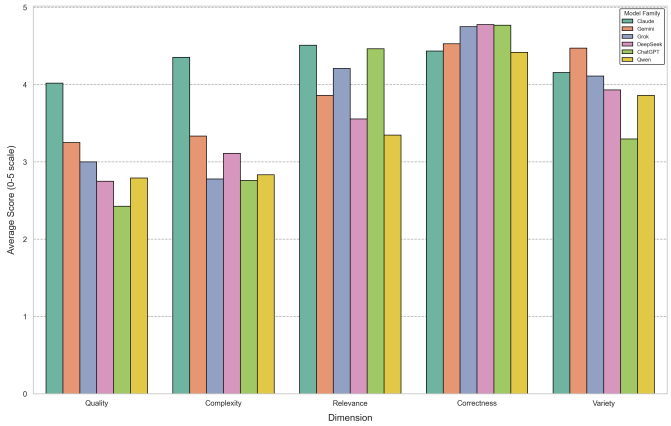


Figure 11

A. Model family analysis: Overall performance by model family**.** Bar chart showing aggregated performance of models grouped by provider family, with Claude models demonstrating superior overall performance (mean=21.47±4.18) compared to other families.

B.Model family analysis: Dimensional performance by model family. Comparison of AI model families across five evaluation dimensions (Quality, Complexity, Relevance, Correctness, and Variety) on a 0-5 scale. showing ChatGPT models excelling in Correctness (4.77) despite low Quality scores (2.43), while Claude models demonstrated more balanced performance.

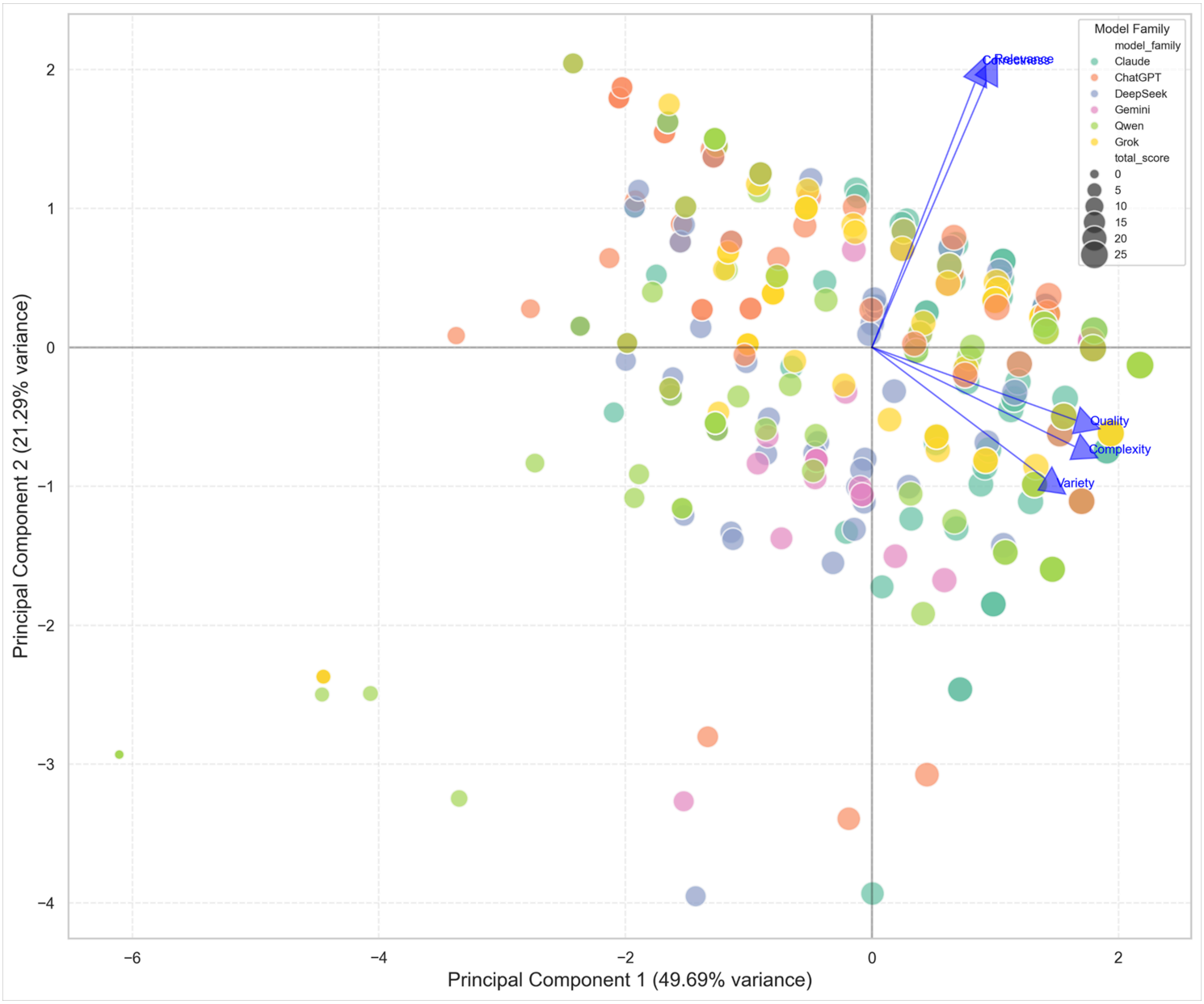


Figure 12

Principal Component Analysis of LLM performance in clinical scenario generation. Two-dimensional PCA plot visualizing clustering of models based on performance dimensions, with the first principal component (49.69% of variance) primarily associated with Quality and Complexity dimensions.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix1.MCQsusedinthestudy.pdf
- Appendix2.Validationframeworkresults.xlsx
- Appendix3.Completegeneratedclinicalscenarios.docx

- Appendix4.Evaluationframework.pdf
- Appendix5.Detailedscoringresults.xlsx