Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Large language models for automated scoring of subjective examination questions in a traditional Chinese medicine curriculum: a four-model comparative evaluation across three disciplines

Yuxin Zhao
Beijing University of Chinese Medicine
Zhuoma Xiangba
Beijing University of Chinese Medicine
Tingyu Wu
Beijing University of Chinese Medicine
Yansong Zhao
Beijing University of Chinese Medicine
Hongsheng Chang
Beijing University of Chinese Medicine
Lei Ni
Beijing University of Chinese Medicine
Xingguang Ma
himxg@126.com

Beijing University of Chinese Medicine

Research Article

Keywords: Large language model, Automated scoring, Medical education, Traditional Chinese medicine education, Subjective question assessment, Agile data management, Prompt engineering

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11316127/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background. Manual scoring of subjective questions in traditional Chinese medicine (TCM) education is labor-intensive, time-consuming, and subject to inter-rater variability. Large language models (LLMs) offer potential for automated scoring, yet systematic multi-model, multi-discipline comparative evaluations remain scarce.

Methods. Guided by agile data management methodology, we designed an intelligent scoring database and a five-dimensional evaluation framework (accuracy, reliability, stability, timeliness, and economy). Four LLMs—ChatGPT-4.0, DeepSeek-V3, ERNIE 4.0, and Tongyi Qianwen 2.5—scored 500 student responses to five questions across pharmacology, febrile diseases (TCM), and physiology using a structured scoring prompt framework. Model outputs were compared against expert raters using mean absolute error ratio, root mean square error, Spearman's rank correlation coefficient, intraclass correlation coefficient (ICC), and Wilcoxon signed-rank tests; each model performed two scoring passes plus a third adjudication pass for discrepant cases.

Results. Significant cross-model and cross-discipline differences were observed. Tongyi Qianwen 2.5 showed the most consistent performance across disciplines (ICC 0.615–0.917). ERNIE 4.0 excelled in pharmacology (MAER 0.087–0.170) but performed poorly in febrile diseases (MAER 0.257). ChatGPT-4.0 and DeepSeek-V3 performed strongly in febrile diseases and physiology. All four models struggled with a complex pharmacology mechanism question (cardiac glycoside positive inotropic action). Hallucination rates were low across all models (0.03%–0.74%). Scoring efficiency varied substantially: DeepSeek-V3 achieved the highest timeliness (mean < 7 s per response) and the lowest cost (99.8% economy score), whereas ChatGPT-4.0 incurred the highest cost (41.1% economy score).

Conclusions. LLMs show promise for automated scoring of subjective examination questions in TCM education, but model selection should be discipline-specific. Optimizing prompt frameworks and integrating human-in-the-loop validation can further enhance scoring reliability. Agile data management methodology provides a scalable approach to educational data lifecycle management.

Background

Subjective questions—including short-answer items, essay questions, and case analyses—remain essential instruments for assessing higher-order cognitive skills in medical education [1]. Unlike multiple-choice questions, these constructed-response items require students to synthesize knowledge, apply clinical reasoning, and communicate arguments coherently. However, the very features that make subjective questions educationally valuable also make them costly to score. In large-scale medical examinations, human raters must evaluate hundreds or thousands of responses against nuanced scoring rubrics, a process that is labor-intensive, time-consuming, and susceptible to inter-rater variability and rater fatigue [2].

The challenge is compounded by the expanding scale of medical education. As enrollment grows and competency-based assessment frameworks gain traction, the volume of subjective responses requiring evaluation increases disproportionately. Raters may spend several minutes per response, and consistency degrades over long scoring sessions [3]. Moreover, subjective biases—such as leniency effects, halo effects, and idiosyncratic interpretation of rubric criteria—can compromise the fairness and reliability of examination outcomes [4].

The development of automated essay scoring (AES) reflects a gradual shift from surface-level heuristics to meaning-based and data-driven approaches. Early systems such as Project Essay Grade (PEG) relied on surface features including text length and syntactic complexity [5]. Subsequent approaches incorporated latent semantic analysis (LSA) to model conceptual similarity between student responses and reference answers [6]. Machine-learning-based systems trained on human-scored corpora achieved moderate success but required substantial labelled data and struggled with open-ended responses in specialised domains [7]. These limitations motivated the exploration of general-purpose generative models for automated scoring.

The emergence of large language models (LLMs) based on the Transformer architecture has introduced a new paradigm for text understanding and generation [8]. Models such as GPT-4, ERNIE, and Tongyi Qianwen demonstrate strong performance on natural language understanding benchmarks and have been evaluated in medical contexts including the United States Medical Licensing Examination (USMLE) and Chinese medical licensing examinations [9, 10]. Their capacity for zero-shot and few-shot reasoning, combined with instruction-following capabilities, suggests potential for scoring subjective medical questions without domain-specific fine-tuning.

Nevertheless, the application of LLMs to medical subjective-question scoring remains underexplored in several respects. First, most existing studies evaluate a single model on a single discipline, leaving cross-model and cross-discipline generalization poorly understood [11]. Second, assessment criteria have typically been limited to accuracy metrics such as Pearson correlation or exact agreement, neglecting dimensions including scoring stability, hallucination-induced reliability loss, operational timeliness, and economic cost [12]. Third, the data infrastructure supporting LLM-based scoring—encompassing data collection, storage, processing, and analysis—has received insufficient attention, despite its critical role in ensuring reproducibility and scalability [13].

To address these gaps, we designed a comparative evaluation study guided by agile data management methodology. We constructed a five-dimensional evaluation framework and evaluated four commercially available LLMs—ChatGPT-4.0, DeepSeek-V3, ERNIE 4.0, and Tongyi Qianwen 2.5—on 500 student responses to subjective examination questions across three TCM-curriculum courses (pharmacology, febrile diseases, and physiology). Our research questions were: (1) How accurately do different LLMs score medical subjective questions relative to expert raters? (2) To what extent does scoring performance vary across medical disciplines? (3) How do the models compare across accuracy, reliability, stability, timeliness, and economy? (4) What are the primary sources of scoring discrepancy,

and how can they be mitigated? The Methods section describes the database design, prompt framework, and statistical analyses; the Results section reports the five-dimensional evaluation; and the Discussion section interprets the findings and their implications for medical education.

Methods

Study Design

This sample size provides high statistical power: with 100 responses per question, a two-sided Wilcoxon signed-rank test at α = 0.05 achieves more than 99% power to detect a medium effect size (Cohen's d = 0.5).

Data Collection

Student responses were collected from subjective examination questions in three core courses of the TCM curriculum at a university of Chinese medicine: pharmacology, febrile diseases (□□□, a traditional Chinese medicine discipline), and physiology. These disciplines were selected to represent diverse knowledge domains—pharmacology emphasizes drug mechanisms and clinical applications; febrile diseases requires integration of traditional Chinese medical theory with case-based reasoning; physiology focuses on mechanistic explanations of bodily functions.

The dataset comprised 500 student responses to five questions: two pharmacology short-answer questions (5 points each), one febrile diseases case-analysis question (10 points), and two physiology essay questions (10 points each). Approximately 100 responses were collected per question. Responses were originally submitted through an online examination platform in HTML format and converted to plain text for scoring.

Database Design

We designed a relational database with five core tables to support the full lifecycle of scoring data management (Fig. 1):

- - Questions table: Stored question text, discipline, question type, maximum score, and scoring rubric in both text and structured JSON formats.
- - Student Responses table: Stored de-identified student answers in both original HTML and cleaned plain-text formats, along with expert (teacher) ratings.
- - Scoring Rubric Details table: Decomposed each question's scoring rubric into individual scoring points, with point sequence number, description, and assigned score value.
- - Model Reviews table: Recorded each API call outcome, including model identifier, scoring attempt number (1, 2, or 3), processing time, total score, token usage, and JSON-formatted detailed

feedback.

- - Review Details table: Parsed each model review into per-scoring-point records, including the relevant excerpt from the student response, model-generated rationale, and point-level score.

The database was implemented using a lightweight relational database management system. Entity relationships followed standard referential integrity constraints: one question to many student responses, one question to many rubric details, one student response to many model reviews, one model review to many review details, and one rubric detail to many review details. This architecture enabled traceability from any individual scoring decision back to the original student response and rubric criterion.

Scoring Prompt Framework

We designed a six-component scoring prompt framework to standardize model behavior across all disciplines and models:

1. Role assignment: The model was instructed to adopt the persona of a senior professor in the relevant discipline (e.g., "You are a senior professor of pharmacology with extensive teaching and examination experience").
2. Task instruction: The model was directed to evaluate the student response step-by-step against the provided scoring rubric.
3. Scoring rubric: The complete rubric was presented in structured JSON format, specifying each scoring point, its description, and its assigned value.
4. Worked example: A high-scoring example response with point-by-point justification was provided to calibrate scoring standards.
5. Self-consistency instruction: The model was required to verify that the sum of point-level scores equaled the total score.
6. Output format: The model was instructed to return results in a standardized JSON structure containing per-point rationales and the total score.

Temperature was set to 0.001 to minimize randomness in scoring. The complete prompt template for each discipline is provided in Additional file 1.

LLM Selection and API Configuration

Four LLMs were selected to represent diverse training paradigms, linguistic orientations, and commercial availability:

- - ChatGPT-4.0 (OpenAI, gpt-4o): A general-purpose multilingual model with strong reasoning capabilities.

- - DeepSeek-V3 (DeepSeek, deepseek-chat): A cost-efficient Chinese-developed model with open-source weights.
- - ERNIE 4.0 (Baidu, ERNIE-4.0-8K-Latest): A knowledge-enhanced Chinese model integrating deep-learning with knowledge-graph augmentation.
- - Tongyi Qianwen 2.5 (Alibaba, qwen-max-2024-09-19): A general-purpose Chinese model with broad multilingual support.

Each model was accessed through its official API endpoint using identical prompt structures. API calls were automated using Python scripts with database connectivity. The overall study workflow is illustrated in Fig. 2.

Scoring Procedure

Each model performed two independent scoring passes on every student response using identical prompts. If the two passes produced identical total scores and identical per-point scores, the result was accepted as final. If any discrepancy was detected, a third adjudication pass was triggered. In the third pass, the model was provided with its first two scoring results and instructed to perform recursive introspection and self-consistency checking to produce a final adjudicated score. This three-pass design aimed to maximize scoring stability while maintaining computational feasibility.

Evaluation Metrics

We constructed a five-dimensional evaluation framework on the basis of information quality theory, educational measurement theory, cost–benefit analysis, and agile data management methodology, combined with a review of the evaluation literature and the content of the model scoring outputs:

Accuracy. Measured by mean absolute error ratio (MAER) and root mean square error (RMSE), both normalized by the maximum question score to enable cross-question comparison:

MAER\:=\frac{1}{n}\times\sum\limits_{i = 1}^{n}\left|\frac{y_{i}−ŷ_{i}}{S}\right|

RMSE\:=\sqrt{\frac{1}{n}\times\sum\limits_{i = 1}^{n}{\text{(}\frac{y_{i}−ŷ_{i}}{S}\text{)}}^{2}}

where yi is the expert score, ŷi is the model score, S is the maximum score, and n is the number of responses.

Overall accuracy was defined as the proportion of responses for which the model score fell within a tolerance band of ± 20% of the maximum question score relative to the expert score (i.e., ± 1 point for 5-point questions and ± 2 points for 10-point questions); it was computed as the total number of responses within the tolerance band across all questions divided by the total number of responses.

Reliability. Operationalized as the hallucination rate. A hallucination was defined as any instance in which the model cited a student response excerpt that could not be verified against the original student text. The reliability ratio was computed as 1 minus the hallucination proportion.

Stability. Defined as the proportion of responses for which the first two scoring passes produced identical results (both total and per-point scores).

Timeliness. Measured as the mean processing time per response, with a 60-second standardization threshold: standardized timeliness = (1 – mean time / 60) × 100%.

Economy. Computed from API token costs based on each provider's official pricing. Standardized economy was calculated as (1 – cost per response) × 100%, where cost per response is expressed in RMB; a benchmark cost of 0.8 RMB per response was used as the threshold for satisfactory cost-effectiveness.

Statistical Analysis

Spearman's rank correlation coefficient (r) was used to assess the monotonic relationship between model and expert scores. The Spearman correlation was computed as:

r_{s}=1−\frac{\sigma \Sigma d_{i}^{2}}{n\left(n^{2}−1\right)}

where r_s is the Spearman rank correlation coefficient, d_i is the difference between the ranks of the i^{th} observation for the two variables (model score and expert score), and n is the number of responses.

The Wilcoxon signed-rank test was used to test for systematic differences between model and expert scores (significance level α = 0.05), with effect sizes reported as:

$$r = \frac{Z}{\sqrt{n}}$$

The Wilcoxon effect size r quantifies the practical magnitude of the systematic difference between model and expert scores. Whereas the P value indicates only whether a statistically significant difference exists, the effect size r addresses the question of how large that difference is in real terms. This distinction is particularly important in the present study because, with a sample size of 100 responses per question, even trivial deviations can achieve statistical significance. Reporting effect sizes alongside P values therefore provides a more complete and clinically meaningful assessment of model scoring performance.

Intraclass correlation coefficient (ICC, two-way random, single-measure, absolute agreement) was used to quantify scoring consistency between each model and the expert rater, with 95% confidence intervals calculated using Fisher z transformation. Descriptive statistics for score differences were reported as median (Q1–Q3). All analyses were performed using SPSS 26.0.

Expert Validation

Three discipline specialists (one each in pharmacology, febrile diseases, and physiology) with extensive teaching and examination experience were invited to review a subset of model scoring outputs, particularly cases flagged as discrepant (model scores differing from expert scores by ≥ 20% of maximum score). Experts evaluated the reasonableness of model-generated rationales and adjudicated disputed cases.

Results

Overview of Scoring Data

The database contained 5,857 model review records and 29,075 review detail records. After the three-pass scoring procedure, each of the 500 student responses had a final adjudicated score from each of the four models. The mean expert scores across questions ranged from 2.81 (pharmacology Question 1, SD = 1.37) to 6.73 (febrile diseases, SD = 2.66). The distribution of model–expert score differences across all questions and models is shown in Fig. 3, and consolidated accuracy and consistency metrics are presented in Table 1.

Accuracy Analysis

Figure 3 presents box plots of the score differences (model minus expert) for each model–question combination, revealing three distinct distributional patterns. The horizontal dashed line at y = 0 represents perfect agreement. Positive values indicate model scores higher than expert scores; negative values indicate model scores lower than expert scores. On pharmacology questions, all four models produced distributions centred below zero, indicating systematic under-scoring relative to the expert. On the febrile diseases case-analysis question, all distributions shifted above zero, reflecting lenient scoring. Physiology questions showed the tightest clustering around zero, with inter-quartile ranges largely overlapping the null line, suggesting the closest agreement between models and the expert rater.

These visual patterns are quantified in Table 1. Two overarching trends emerge. First, question complexity exerted a stronger influence on scoring accuracy than model identity. Pharmacology Question 2 (cardiac glycoside positive inotropic action), the most cognitively demanding item in the set, produced a marked deterioration across all models: MAER increased by 96–179% relative to Question 1, and ICC values for DeepSeek-V3 (0.548) and ChatGPT-4.0 (0.568) fell below the 0.60 threshold conventionally regarded as the minimum for acceptable consistency in educational assessment. By contrast, the structurally simpler physiology questions yielded uniformly lower MAER (0.072–0.107) and higher ICC (0.826–0.917), with two model–question combinations (ChatGPT-4.0 on Physiology Question 2 and Tongyi Qianwen 2.5 on Physiology Question 1) showing negligible systematic bias (Wilcoxon P = 0.759 and 0.509, effect sizes r = − 0.030 and − 0.066, respectively).

Second, model performance was strongly discipline-dependent. ERNIE 4.0 dominated pharmacology (lowest MAER on both questions) but collapsed on the febrile diseases case-analysis, where its MAER (0.257) was the highest on that question among the four models and its ICC (0.533) was the lowest observed across all 20 model–question combinations. ChatGPT-4.0 and DeepSeek-V3 showed the opposite trajectory, performing strongly on febrile diseases and physiology yet struggling with pharmacology Question 2. Tongyi Qianwen 2.5 exhibited the most stable cross-disciplinary profile: its ICC never fell below 0.615—the highest minimum ICC of any model—and it achieved the highest Spearman correlation in the dataset (0.928, physiology Question 2), with a MAER range (0.072–0.257) among the narrowest observed.

A comparison of the monotonic correlation (Spearman rs) and absolute agreement (ICC) metrics offers additional insight into the nature of model–expert divergence. Across all questions, rs (range 0.560–0.928) was numerically close to ICC (range 0.533–0.913), exceeding it in 12 of the 20 model–question combinations. The largest positive gaps (rs > ICC) were observed on the febrile diseases case analysis (ERNIE 4.0: +0.088) and physiology Question 2 (DeepSeek-V3: +0.033), indicating that these models preserved the rank ordering of student responses—identifying which answers were stronger or weaker than others—more faithfully than they reproduced absolute scores. Conversely, in eight of the 20 combinations—including two of the four models on pharmacology Question 2 and three of the four models on the febrile diseases question—ICC exceeded rs (largest gap: − 0.076 for DeepSeek-V3 on febrile diseases), indicating that, once the systematic offset was accounted for, absolute agreement was comparatively strong. The direction of the systematic offset varied by discipline: negative median differences (under-scoring) on pharmacology, positive median differences (over-scoring) on febrile diseases, and near-zero differences on physiology for seven of the eight model–question combinations (ERNIE 4.0 on physiology Question 2 being the exception, median + 1.25).

Table 1
Consolidated accuracy and consistency metrics across models and questions

Question	Model	MAER	RMSE	rs	M (Q1, Q3)	P value	Effect size (r)	ICC (95% CI)
Pharm Q1	ERNIE 4.0	0.087	0.149	0.852	0.0 (-0.5, 0.0)	0.043	-0.202	0.836 (0.765–0.887)
	Tongyi Qianwen 2.5	0.093	0.165	0.841	0.0 (-0.5, 0.0)	< 0.001	-0.418	0.821 (0.745–0.876)
	ChatGPT-4.0	0.112	0.179	0.854	-0.25 (-1.0, 0.0)	< 0.001	-0.599	0.832 (0.760–0.884)
	DeepSeek-V3	0.119	0.199	0.762	0.0 (-1.0, 0.0)	< 0.001	-0.440	0.761 (0.664–0.833)
Pharm Q2	ERNIE 4.0	0.170	0.272	0.677	-0.5 (-1.5, 0.0)	< 0.001	-0.596	0.702 (0.586–0.790)
	Tongyi Qianwen 2.5	0.257	0.355	0.602	-1.0 (-2.0, 0.0)	< 0.001	-0.706	0.615 (0.476–0.724)
	DeepSeek-V3	0.297	0.412	0.560	-1.75 (-2.0, 0.0)	< 0.001	-0.698	0.548 (0.394–0.672)
	ChatGPT-4.0	0.313	0.415	0.574	-1.25 (-3.0, 0.0)	< 0.001	-0.735	0.568 (0.418–0.688)
Febrile	ChatGPT-4.0	0.111	0.140	0.822	0.5 (-0.5, 1.5)	< 0.001	-0.489	0.879 (0.825–0.917)
	DeepSeek-V3	0.125	0.159	0.787	1.0 (0.0, 2.0)	< 0.001	-0.535	0.863 (0.803–0.906)
	Tongyi Qianwen 2.5	0.163	0.212	0.728	1.25 (0.0, 2.5)	< 0.001	-0.701	0.784 (0.695–0.850)

Note. MAER, mean absolute error rate; RMSE, root-mean-square error rate (both normalized by the maximum question score); rs, Spearman rank correlation coefficient; M (Q1, Q3), median (quartiles) of score differences between model and expert ratings; r, Wilcoxon signed-rank test effect size ($r = Z/\sqrt{N}$); ICC, intraclass correlation coefficient (95% confidence intervals calculated using Fisher z transformation). P values are from Wilcoxon signed-rank tests.

Question	Model	MAER	RMSE	rs	M (Q1, Q3)	P value	Effect size (r)	ICC (95% CI)
	ERNIE 4.0	0.257	0.325	0.621	2.0 (1.0, 4.0)	< 0.001	-0.779	0.533 (0.376–0.660)
Physio Q1	Tongyi Qianwen 2.5	0.072	0.112	0.914	0.0 (0.0, 1.0)	0.509	-0.066	0.913 (0.873–0.941)
	ChatGPT-4.0	0.086	0.129	0.898	0.0 (0.0, 1.0)	0.013	-0.249	0.900 (0.855–0.932)
	DeepSeek-V3	0.101	0.148	0.879	0.0 (0.0, 1.25)	< 0.001	-0.421	0.882 (0.829–0.919)
	ERNIE 4.0	0.107	0.156	0.877	0.0 (0.0, 1.0)	< 0.001	-0.430	0.868 (0.810–0.909)
Physio Q2	ChatGPT-4.0	0.100	0.155	0.901	0.0 (-1.0, 1.0)	0.759	-0.030	0.898 (0.852–0.930)
	Tongyi Qianwen 2.5	0.100	0.138	0.928	0.0 (-1.0, 0.62)	0.698	-0.039	0.917 (0.879–0.943)
	DeepSeek-V3	0.107	0.172	0.926	0.0 (0.0, 1.0)	< 0.001	-0.414	0.893 (0.845–0.927)
	ERNIE 4.0	0.183	0.259	0.783	1.25 (0.0, 3.0)	< 0.001	-0.685	0.826 (0.752–0.880)

Note. MAER, mean absolute error rate; RMSE, root-mean-square error rate (both normalized by the maximum question score); rs, Spearman rank correlation coefficient; M (Q1, Q3), median (quartiles) of score differences between model and expert ratings; r, Wilcoxon signed-rank test effect size ($r = Z/\sqrt{N}$); ICC, intraclass correlation coefficient (95% confidence intervals calculated using Fisher z transformation). P values are from Wilcoxon signed-rank tests.

Consistency Analysis

ICC values, which quantify absolute agreement between model and expert scores, mirrored the accuracy patterns but with important nuances (Table 1). Tongyi Qianwen 2.5 demonstrated the most stable ICC profile across all five questions (range: 0.615–0.917), never falling below the 0.75 threshold for good consistency except on pharmacology Question 2 (0.615). By contrast, ERNIE 4.0 showed the widest ICC range (0.533–0.836), oscillating between excellent consistency on pharmacology Question 1 and the poorest consistency in the entire dataset on febrile diseases (0.533). This 0.303-point ICC swing within a

single model underscores that consistency is not an intrinsic model property but is tightly coupled to question characteristics.

The Bland-Altman plots in Fig. 4 visualize the limits of agreement for each model across all five questions. For each plot, the x-axis represents the mean of the model and expert scores, the y-axis represents the score difference (model minus expert), the solid horizontal line indicates the mean difference (bias), and the dashed horizontal lines indicate the 95% limits of agreement (mean difference ± 1.96 SD). Points falling within the limits of agreement indicate acceptable individual-level agreement. For DeepSeek-V3 and ChatGPT-4.0, the scatter of points was relatively symmetrical around the mean difference line, with most observations falling within ± 2 SD. ERNIE 4.0 displayed a broader vertical spread and a visible positive trend at higher score levels, indicating that its disagreement with the expert widened as student responses became more complex or higher-scoring. Tongyi Qianwen 2.5 showed the tightest clustering, consistent with its highest overall ICC values.

Stability Analysis

The proportion of responses with identical first and second scoring passes was: DeepSeek-V3 (77.4%), ChatGPT-4.0 (76.8%), Tongyi Qianwen 2.5 (73.2%), and ERNIE 4.0 (55.0%). ERNIE 4.0's lower stability suggests higher sensitivity to stochastic variation despite the near-zero temperature setting.

Timeliness and Cost Analysis

Operational efficiency diverged sharply across models (Fig. 5). Panel (a) shows mean processing time per response: ChatGPT-4.0 and DeepSeek-V3 both completed scoring in approximately 7 seconds, whereas ERNIE 4.0 and Tongyi Qianwen 2.5 required roughly 40 seconds per response—nearly sixfold slower. This bifurcation reflects underlying architectural differences: the faster models likely employ more efficient inference pipelines or lower-latency API infrastructure, while the slower models may prioritize reasoning depth over speed.

Panel (b) presents the total API cost for scoring the full 500-response corpus. DeepSeek-V3 incurred only 1.15 RMB, a 256-fold cost advantage over ChatGPT-4.0 (294.64 RMB). Tongyi Qianwen 2.5 (30.69 RMB) and ERNIE 4.0 (33.49 RMB) occupied a middle tier. When extrapolated to a 10,000-response institutional examination, the implied costs are approximately 5,893 RMB for ChatGPT-4.0 versus 23 RMB for DeepSeek-V3. At this scale, the cost differential alone may outweigh the roughly three-percentage-point overall accuracy advantage of ChatGPT-4.0 over DeepSeek-V3 (80.6% versus 77.6%), particularly for formative assessment in which near-ceiling accuracy is not a prerequisite.

Hallucination Analysis

Across the 23,525 review detail records from the two initial scoring passes that were examined for hallucination (the third adjudication pass, applied only to discrepant responses, was excluded), 325 instances of hallucination were identified (1.38% overall). By model: DeepSeek-V3 (6 instances; 0.03%),

Tongyi Qianwen 2.5 (59 instances; 0.25%), ChatGPT-4.0 (86 instances; 0.36%), and ERNIE 4.0 (174 instances; 0.74%). Reliability ratios were correspondingly high: DeepSeek-V3 (99.97%), Tongyi Qianwen 2.5 (99.75%), ChatGPT-4.0 (99.64%), and ERNIE 4.0 (99.26%). Three discipline specialists with marking experience reviewed a subset of the model scoring outputs, particularly discrepant cases; they judged the model scores to be essentially consistent with expert judgments, with discrepancies observed only for isolated answers.

Cross-Model and Cross-Discipline Comparison

Figure 6 synthesizes the five-dimensional evaluation into two complementary visualizations. Five-dimensional performance comparison across four LLMs. (a) Radar chart showing normalized scores on accuracy, reliability, stability, timeliness, and economy. (b) Heatmap showing raw percentage scores for each dimension. Darker colours indicate higher scores. The radar chart reveals the shape of each model's performance profile. DeepSeek-V3 presents a broad, balanced polygon with high scores on reliability, stability, and economy, but a somewhat contracted accuracy vertex. ChatGPT-4.0 shows a more irregular shape: its accuracy and timeliness vertices extend far outward, while its economy vertex collapses toward the centre. Tongyi Qianwen 2.5 produces the most regular pentagon, indicating balanced performance without extreme strengths or weaknesses. ERNIE 4.0's polygon is the most contracted overall, with only the reliability vertex approaching the periphery.

The heatmap quantifies these patterns numerically. DeepSeek-V3 achieved the highest composite reliability score (99.97%) and the lowest hallucination rate (0.03%), yet its accuracy (77.6%) lagged behind ChatGPT-4.0 (80.6%) by three percentage points. Tongyi Qianwen 2.5 occupied the accuracy–economy sweet spot: 79.4% accuracy at 93.86% economy, compared with ChatGPT-4.0's 80.6% accuracy at only 41.07% economy. No single model occupied the outermost ring on more than two dimensions simultaneously, confirming that model selection requires explicit trade-off decisions rather than a universal "best" choice.

Discussion

Principal Findings

This study provides, to our knowledge, the first systematic comparison of four major LLMs for scoring subjective examination questions in a TCM curriculum across multiple disciplines using a multi-dimensional evaluation framework. Three principal findings emerge.

First, LLM scoring accuracy is highly discipline-dependent. ERNIE 4.0, a knowledge-enhanced Chinese model, achieved exceptional accuracy on pharmacology Question 1 (MAER = 0.087)—particularly mechanistic questions involving well-established biochemical pathways—but performed poorly on pharmacology Question 2 (MAER = 0.170) and febrile diseases case analysis (MAER = 0.257). Conversely, ChatGPT-4.0 and DeepSeek-V3, both developed with stronger English-centric training corpora, excelled on febrile diseases and physiology but lagged on pharmacology. This pattern suggests

that training data composition and knowledge-base design strongly influence scoring performance on domain-specific content.

Second, question complexity critically moderates model performance. All four models showed degraded accuracy on pharmacology Question 2, which required explanation of the cardiac glycoside positive inotropic mechanism at the cellular level (Na^+-K^+-ATPase inhibition, intracellular calcium dynamics). This question demanded integration of molecular mechanism with clinical pharmacology—a level of reasoning that appeared to challenge all evaluated models. The finding aligns with prior observations that LLMs struggle with questions requiring multi-step causal reasoning or integration of knowledge across scales [14].

Third, the five-dimensional framework reveals trade-offs that single-metric studies obscure. DeepSeek-V3 achieved near-perfect reliability and economy scores but moderate accuracy (77.6%). ChatGPT-4.0 offered strong accuracy (80.6%) but at nearly 256-fold higher cost than DeepSeek-V3. Tongyi Qianwen 2.5 provided the highest overall accuracy on physiology questions (MAER = 0.072–0.100) with competitive economy (93.86%). These trade-offs have direct practical implications for implementation at scale.

A further pattern deserving attention is the systematic over-scoring observed on the febrile diseases case-analysis question, on which all four models produced median differences above the expert ratings (Fig. 3). A tendency toward lenient grading of open-ended narrative responses is consistent with emerging reports on large language model evaluation behavior. This directional bias has practical consequences that aggregate error metrics can obscure: because MAER averages absolute deviations, systematic inflation of scores is masked when positive and negative errors cancel. Rubric-anchored, point-by-point scoring with mandatory excerpt citation, as implemented in our prompt framework, constrains but does not eliminate this tendency. Before operational use on case-analysis items, score-band calibration against historical human-rating distributions, or targeted human review of responses falling within borderline score bands, would be advisable.

The Wilcoxon effect sizes in Table 1 help distinguish statistical significance from practical importance. With N = 100, even modest deviations can yield $P < 0.001$, so effect size is essential for gauging real-world scoring bias. The largest effects appeared on the febrile diseases case-analysis question (ERNIE 4.0, $r = -0.779$; Tongyi, $r = -0.701$) and pharmacology Question 2 (all models $|r| \geq 0.596$), indicating deviations large enough to affect pass–fail decisions. Conversely, physiology questions showed small effects (ChatGPT-4.0 on physiology Question 2, $r = -0.030$; Tongyi on physiology Question 1, $r = -0.066$), with negligible practical bias. A notable nuance occurred on physiology Question 1, where DeepSeek-V3 ($r = -0.421$) and ERNIE 4.0 ($r = -0.430$) were statistically significant ($P < 0.001$) yet only moderately deviated—here significance reflects statistical power rather than large practical discrepancy, underscoring the value of reporting effect sizes alongside P values.

Comparison with Prior Work

Our findings extend prior research on LLM-based medical assessment. Kung et al. demonstrated that ChatGPT could pass the USMLE, suggesting strong baseline medical knowledge [9]. More recently, studies have evaluated LLMs on Chinese medical licensing examinations, finding comparable performance [10]. However, these studies focused primarily on multiple-choice or single-discipline formats. Our multi-model, multi-discipline design with constructed-response scoring provides a more granular understanding of model capabilities and limitations. This focus is timely: a recent bibliometric analysis of 407 studies on ChatGPT in medical education identified educational assessment and examinations as among the most intensively pursued research fronts in the field [17], yet multi-model comparisons of constructed-response scoring with multi-dimensional evaluation frameworks remain scarce.

Recent comparative studies support the context-dependence of LLM grading behaviour that we observed. Grévisse reported that GPT-4 assigned significantly lower grades than human raters for short-answer questions across 12 undergraduate medical courses, whereas Gemini 1.0 Pro did not differ significantly from teachers [18]; Tekin et al. found that ChatGPT-4o and Gemini Flash 1.5 tended to assign higher and more consistent scores than human evaluators in objective structured clinical examination (OSCE) assessments [19]. Taken together with our observation of systematic under-scoring on pharmacology and over-scoring on the febrile diseases case analysis, these findings indicate that the direction and magnitude of LLM–human score disagreement are shaped by question format, discipline content, and rubric design, and must be characterised empirically before operational deployment in a given institution.

In the automated scoring literature, transformer-based models have shown promise for short-answer scoring [15]. Our study advances this work by evaluating four general-purpose LLMs across five questions in three medical disciplines without domain-specific fine-tuning, using a structured prompt framework and a five-dimensional evaluation framework that could be deployed with minimal technical infrastructure.

Implications for Medical Education

For high-stakes examinations, we recommend a human-in-the-loop design in which LLMs perform initial screening and flag responses requiring expert review. This approach could substantially reduce expert workload while maintaining scoring quality. For formative assessment, where perfect accuracy is less critical, single-model automated scoring with the most cost-effective model (DeepSeek-V3) may be sufficient. In expert consultation, one specialist from each of the three disciplines positively evaluated the model scoring results and rationales, considering that the models reflected students' answering levels relatively truthfully and effectively, which further supports the feasibility of LLM-based scoring in educational practice.

The discipline-specific performance differences imply that institutions should validate model selection locally. A model optimized for Western biomedicine may underperform on traditional Chinese medicine

content, and vice versa. Multi-model ensemble approaches—combining outputs from several models—may improve robustness but at increased cost and complexity.

Strengths and Limitations

Strengths of this study include the multi-model comparison design, the five-dimensional evaluation framework, the structured prompt engineering approach, and the large sample size (500 responses). The three-pass scoring procedure with self-consistency checking represents a novel methodological contribution to LLM-based assessment.

Limitations should be acknowledged. First, the study used responses from a single institution; generalizability to other medical schools and curricula requires further validation. Second, the five questions, while spanning three disciplines, represent a limited sample of the full diversity of medical subjective questions. Third, expert ratings served as the gold standard, but human raters are themselves imperfect; discrepancy between model and expert scores does not necessarily indicate model error. Fourth, API pricing and model versions change rapidly; cost and performance figures reported here reflect the December 2024–March 2025 period and may not generalize to future model iterations. Fifth, we did not evaluate fine-tuned or retrieval-augmented approaches, which may outperform zero-shot prompting. Sixth, only one expert rater scored each question, limiting the reliability of the gold-standard reference; future studies should employ multiple raters to strengthen validity.

Future Directions

Several directions merit investigation. Retrieval-augmented generation (RAG) approaches that supply models with discipline-specific textbooks or clinical guidelines may improve accuracy on complex mechanistic questions [16]. Fine-tuning on institution-specific corpora of scored responses could adapt models to local grading standards. Real-time calibration systems that adjust scoring thresholds based on ongoing quality monitoring would enhance deployment reliability. Finally, expanding evaluation to additional disciplines—including clinical skills assessment and interdisciplinary case discussions—would strengthen the evidence base for LLM-based scoring in medical education.

Conclusions

This study demonstrates that large language models can achieve scoring accuracy comparable to expert raters on subjective examination questions in a TCM curriculum, with performance varying significantly across models and disciplines. A five-dimensional evaluation framework—encompassing accuracy, reliability, stability, timeliness, and economy—provides a comprehensive basis for model selection. No single model dominates all dimensions; optimal choice depends on discipline content, accuracy requirements, and budget constraints. The structured prompt framework and agile data management approach described here offer a reproducible, scalable pathway for implementing LLM-based scoring in medical education settings. Human-in-the-loop validation remains essential for high-stakes assessments, particularly for complex mechanistic questions that challenge current models.

Abbreviations

AES
automated essay scoring
API
application programming interface
ERNIE
Enhanced Representation through kNowledge IntEgration
ICC
intraclass correlation coefficient
LLM
large language model
LSA
latent semantic analysis
MAER
mean absolute error ratio
RAG
retrieval-augmented generation
RMSE
root mean square error
USMLE
United States Medical Licensing Examination

Declarations

Ethics approval and consent to participate

This study used de-identified examination data collected as part of routine educational assessment at Beijing University of Chinese Medicine. All data were fully anonymized before analysis, and no individual student identifiers were retained. Under Article 4 and Article 73 of the Personal Information Protection Law of the People's Republic of China (2021), anonymized information does not constitute personal information. Because this research involved solely fully anonymized data, no intervention with human participants, and no collection of new personal information, it was deemed not to require ethical review under national regulations; the study was therefore not submitted for ethics committee approval, and formal ethics approval was not required. Informed consent to participate was accordingly not obtained, as the fully anonymized data do not constitute personal information for which individual consent is required under national law. The study was conducted in accordance with the institutional ethical guidelines for educational research.

Consent for publication

Not applicable. All student responses were fully anonymized before analysis, and no individual-level data that could identify any participant are reported.

Competing interests

The authors declare that they have no competing interests.

Use of AI tools in manuscript preparation

Large language models were used as the research subject of this study (for automated scoring evaluation) but were not used to generate the manuscript text.

Funding

This work was supported by the Beijing Education Science “14th Five-Year Plan” 2024 General Project (Grant No. CDGB24195) and the Beijing University of Chinese Medicine 2025 Educational and Teaching Reform Special Project (Grant No. XJYZ25011).

Author Contribution

YXZ conceptualization, methodology, scoring database and prompt framework design, data collection and analysis, writing—original draft; ZX data collection, expert validation, manuscript revision; TYW data collection, expert validation, manuscript revision; YSZ data collection, expert validation, manuscript revision; HSC data collection, expert validation, manuscript revision; LN data collection, expert validation, manuscript revision; XGM supervision, critical revision, approval of the final version. All authors read and approved the final manuscript. The authors declare no competing interests.

Acknowledgements

The authors thank the three discipline specialists who participated in expert validation. We also thank the students who contributed anonymized examination responses.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request. Scoring prompt templates are provided as Additional file 1, and the Python analysis scripts for API calling and data processing as Additional file 2.

References

1. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830–7. 10.1046/j.1365-2923.2003.01594.x.
2. Norcini JJ, Anderson B, Bollela V et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):206 – 14.
3. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119(2):e1667–16.
4. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ. 2014;19(2):233–50. 10.1007/s10459-013-9458-4.
5. Attali Y, Burstein J. Automated essay scoring with e-rater(R) v.2. J Technol Learn Assess. 2006;4(3).
6. Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse Processes. 1998;25(2–3):259–84.
7. Shermis MD, Burstein JC, editors. Handbook of automated essay evaluation: Current applications and new directions. New York: Routledge; 2013.
8. Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30.
9. 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.
10. Wang X, Gong Z, Wang G, et al. ChatGPT performs on the Chinese National Medical Licensing Examination. J Med Syst. 2023;47:86. 10.1007/s10916-023-01961-0.
11. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the Written German Medical Licensing Examination: observational study. JMIR Med Educ. 2024;10:e50965.
12. Ramesh D, Sanampudi SK. An automated essay scoring systems: A systematic literature review. Artif Intell Rev. 2022;55(3):2495–527. 10.1007/s10462-021-10068-2.
13. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. 10.1038/sdata.2016.18.
14. Chi H, Li H, Yang W et al. Unveiling causal reasoning in large language models: Reality or mirage? Adv Neural Inf Process Syst. 2024;37.
15. Sung C, Ma T, Dhamecha TI et al. Pre-training BERT on domain resources for short answer grading. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019:6071-5.
16. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst. 2020;33:9459–74.
17. Zhang Y, Xie X, Xu Q. ChatGPT in Medical Education: Bibliometric and Visual Analysis. JMIR Med Educ. 2025;11:e72356. 10.2196/72356.

18. Grévisse C. LLM-based automatic short answer grading in undergraduate medical education. BMC Med Educ. 2024;24:1060. 10.1186/s12909-024-06026-5.
19. Tekin M, Yurdal MO, Toraman Ç, Korkmaz G, Uysal İ. Is AI the future of evaluation in medical education? AI vs. human evaluation in objective structured clinical examination. BMC Med Educ. 2025;25:641. 10.1186/s12909-025-07241-4.

Figures

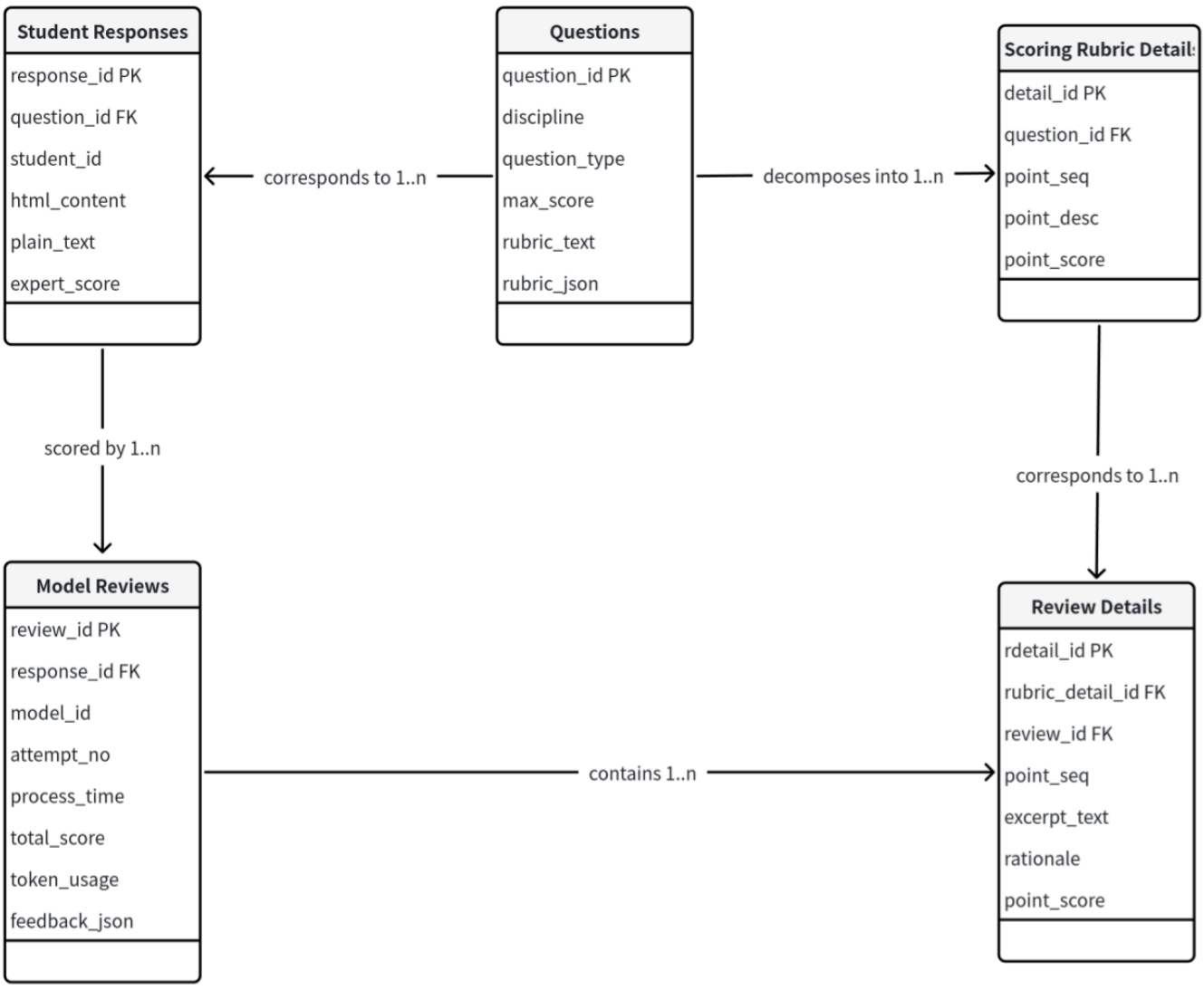


Figure 1

Database entity-relationship diagram showing five core tables and their relationships.

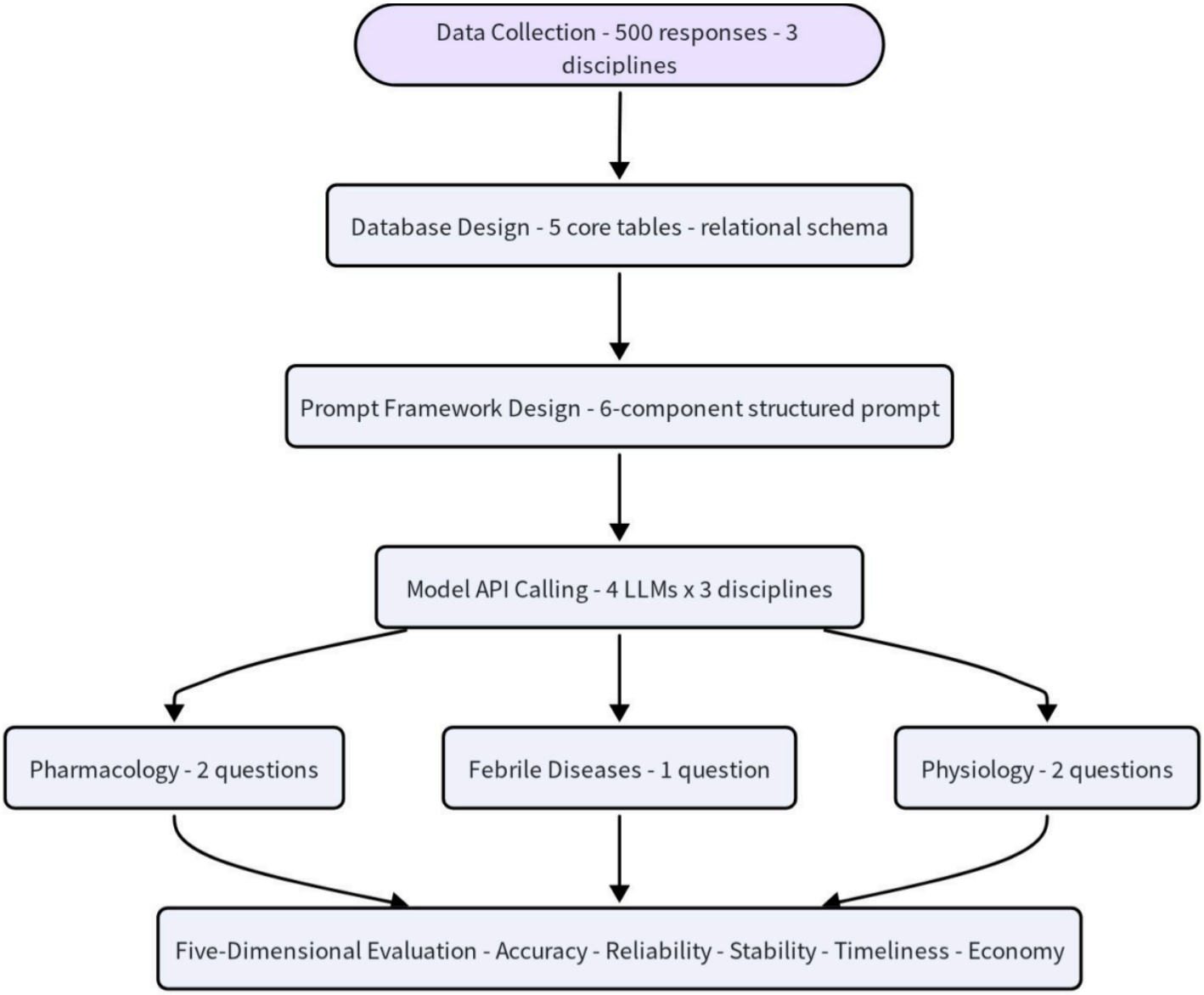


Figure 2

Study overview flowchart illustrating data collection, database design, prompt framework, model API calling, and five-dimensional evaluation.

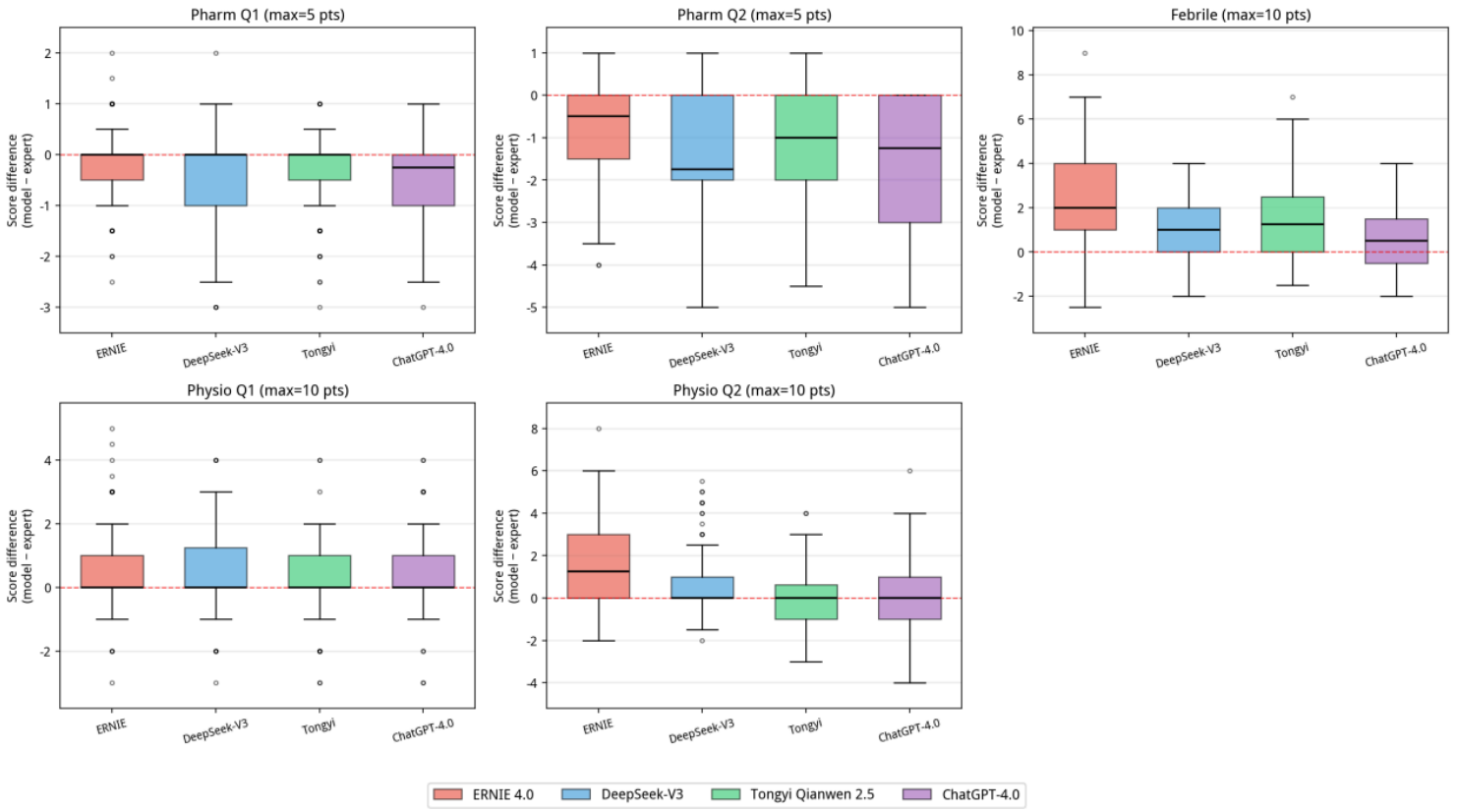


Figure 3

Box plots of score differences (model minus expert) by question and model.

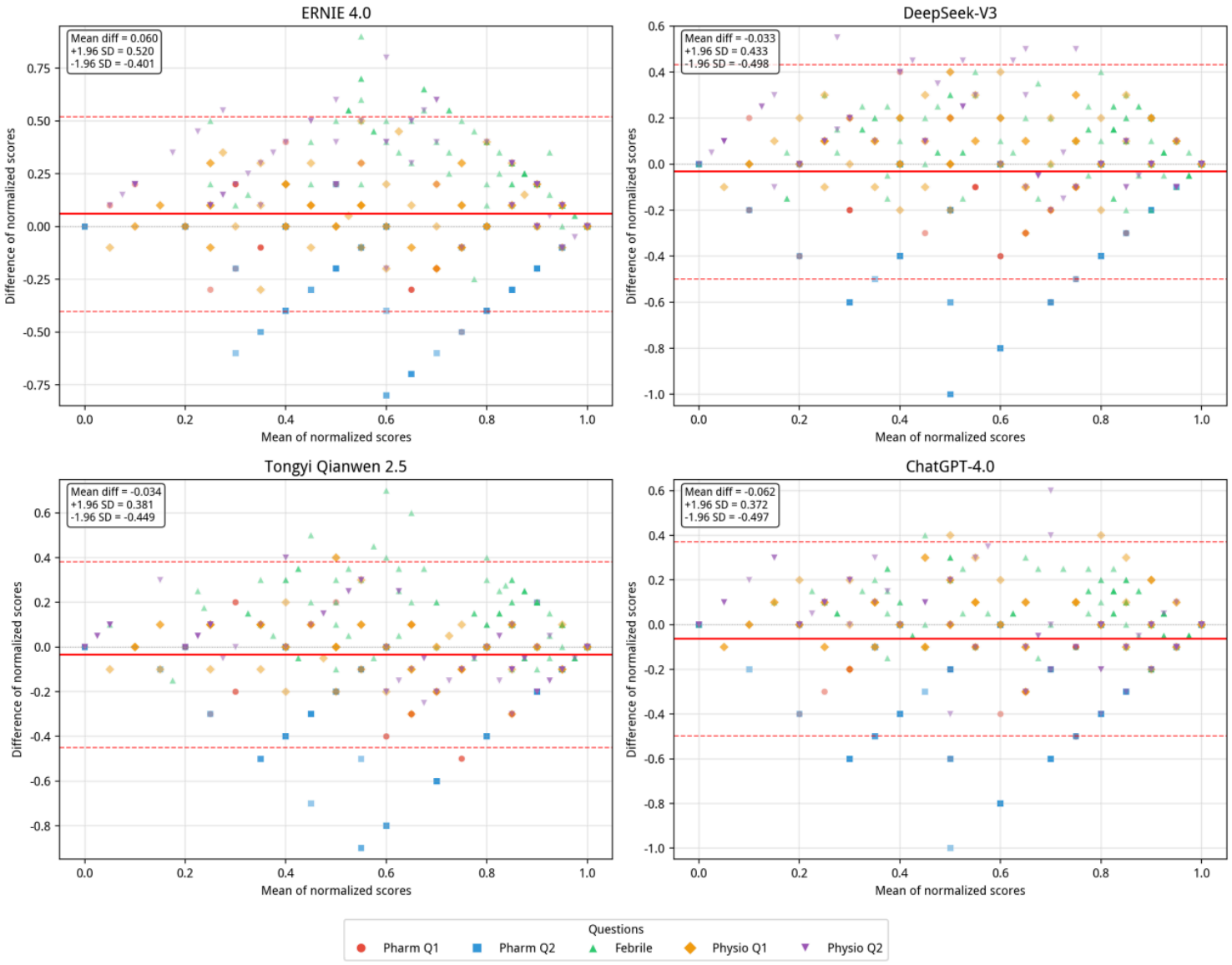


Figure 4

Bland-Altman plots showing model–expert agreement for each LLM.

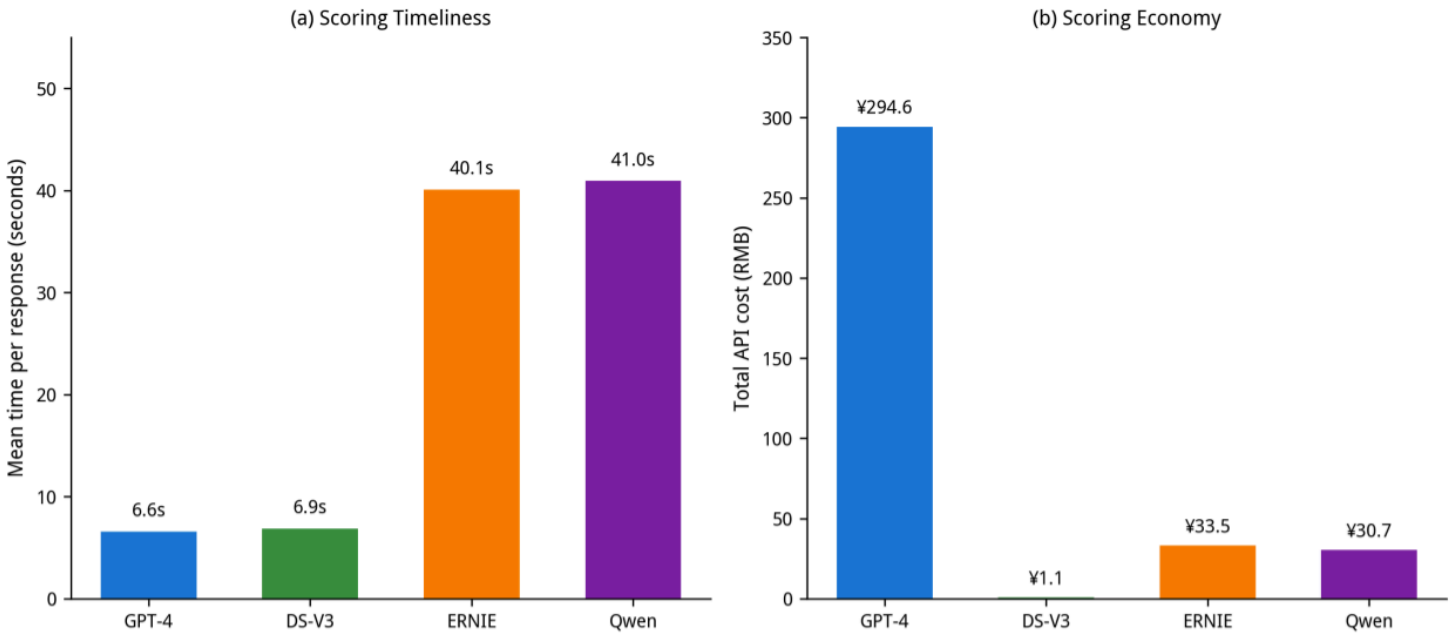


Figure 5

(a) Scoring timeliness (mean processing time per response) and (b) scoring economy (total API cost for 500 responses) across four LLMs.

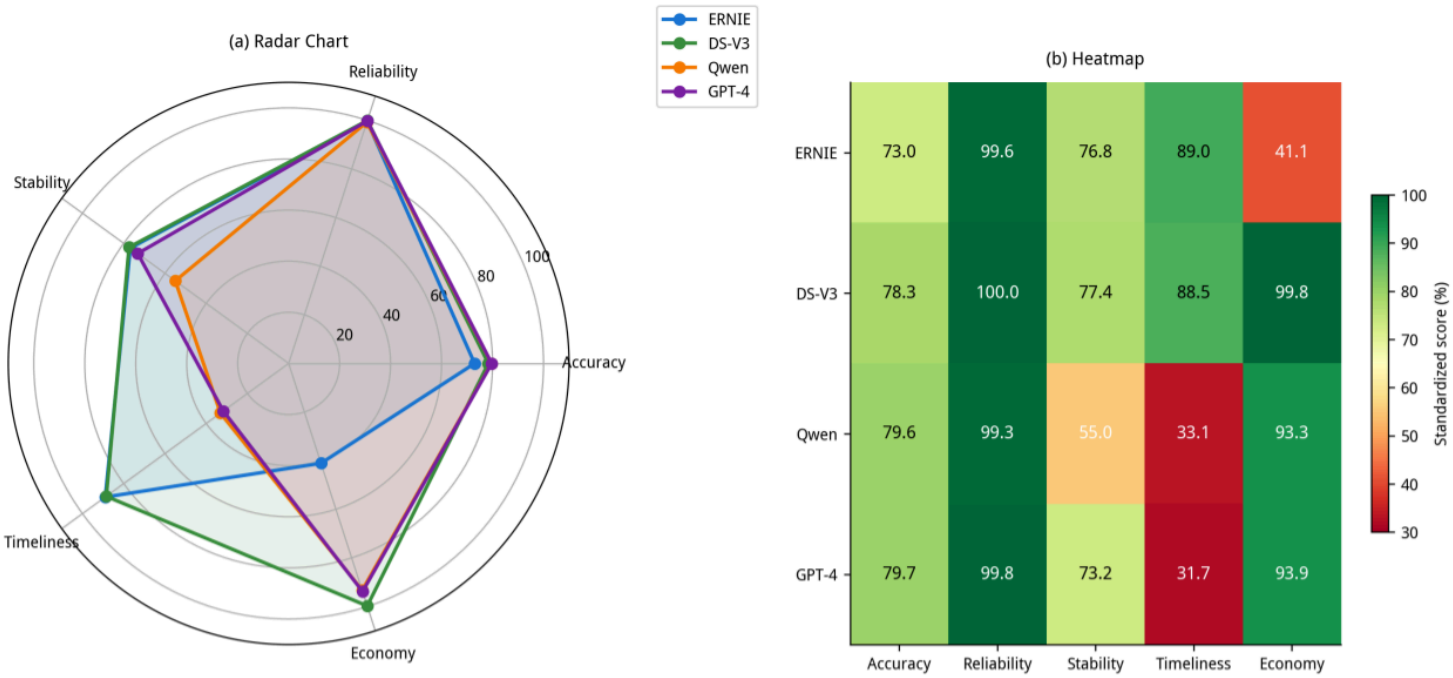


Figure 6

Five-dimensional performance comparison: (a) radar chart and (b) heatmap.