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.

# Quality, reliability, and readability of large language model responses to questions about immune checkpoint inhibitor-associated myocarditis: A cross-sectional study

**Lei Wu**
Cancer Center, The First Affiliated Hospital of Xi'an Jiaotong University
**Tongfei Wang**
Department of Medical Oncology, Xi 'an No.3 Hospital, The Affiliated Hospital of Northwest University
**Juan Zhang**
Cancer Center, The First Affiliated Hospital of Xi'an Jiaotong University
**Mi Zhang**
Cancer Center, The First Affiliated Hospital of Xi'an Jiaotong University
**Yan Zhou**
Cancer Center, The First Affiliated Hospital of Xi'an Jiaotong University
**Danfeng Dong**
Department of Medical Oncology, The First Affiliated Hospital of Xi'an Jiaotong University
**Jin Yang**
yangjin@mail.xjtu.edu.cn

Cancer Center, The First Affiliated Hospital of Xi'an Jiaotong University

**Article**

**Keywords:** immune checkpoint inhibitor-associated myocarditis, large language models, artificial intelligence, patient education, medical information quality, readability

**Posted Date:** November 19th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11077152/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

Immune checkpoint inhibitor-associated myocarditis (ICI-M) is a rare but often fatal immune-related adverse event. Large language models (LLMs) have provided new opportunities for patient education, yet whether they can effectively support patient education in ICI-M remains unknown. In this study, we submitted 20 standardized ICI-M-related patient questions to five LLMs (ChatGPT-5.5, Gemini-3.5 Flash, DeepSeek-V4, Qwen 3.7, and Doubao), generating 100 responses. Model performance was evaluated using the Global Quality Score (GQS), modified DISCERN (mDISCERN), Quality Analysis of Medical Artificial Intelligence (QAMAI), Ensuring Quality Information for Patients (EQIP), Patient Education Materials Assessment Tool understandability and actionability scores (PEMAT-U and PEMAT-A), and six readability indices. Significant inter-model differences were observed across most quality and readability metrics, and no single model outperformed the others across all dimensions. DeepSeek-V4 achieved the highest mDISCERN score [5.00 (IQR, 4.00–5.00)], whereas ChatGPT-5.5 achieved the highest PEMAT-A score [40.00 (IQR, 35.00–45.00)]. Readability was suboptimal, with a mean FKGL of 16.60 ± 2.64. Correlation analyses demonstrated information quality, usability, and readability represented correlated but distinct evaluation dimensions. While LLMs may serve as adjunctive tools for ICI-M patient education, substantial optimization of content accessibility and actionability is required prior to clinical implementation.


# Introduction

The application of immune checkpoint inhibitors (ICIs) has substantially transformed the therapeutic landscape of multiple malignancies by enhancing antitumor immune responses and improving patient survival outcomes[1–3]. However, while activating antitumor immunity, ICIs may also disrupt immune homeostasis and induce a spectrum of immune-related adverse events (irAEs)[4–6]. With an estimated incidence of 0.3%-0.72%, ICI-M is considered an uncommon immune-related adverse event[7]. It is characterized by insidious onset, rapid progression, and a high mortality rate ranging from 37.7% to 39.7%[8]. Therefore, ICI-M has emerged as one of the most critical cardiovascular emergencies in the field of cardio-oncology.

In light of the substantial fatality risk associated with ICI-M, current cardio-oncology guidelines emphasize the importance of timely evaluation and standardized management for patients with suspected ICI-M, including assessment of myocardial injury biomarkers, electrocardiographic evaluation, cardiac imaging, multidisciplinary collaboration, and early initiation of immunosuppressive therapy when clinically indicated[9–12]. The American Society of Clinical Oncology (ASCO) guidelines similarly recommend enhanced surveillance and management of immune-related cardiovascular adverse events during ICI treatment[13]. Effective management of ICI-M depends not only on clinicians' ability to recognize but also on patients' awareness of relevant symptoms and timely communication with healthcare providers. Clinical practice recommendations advocate providing adverse drug reaction education to all patients scheduled to receive ICIs, including information regarding potential irAEs and their clinical manifestations, in order to improve patients' self-recognition abilities and awareness of seeking medical attention[13]. However, efficiently and

accurately communicating such highly specialized medical information to patients remains a major challenge in clinical practice.

In recent years, LLMs have demonstrated considerable potential in medical information processing and patient education. By utilizing extensive medical knowledge and generating conversational responses, these models may overcome limitations associated with conventional patient education materials, such as delayed updates and poor readability. Previous studies have preliminarily explored the application of LLMs in oncology-related patient education. Rodler et al. reported that GPT-4-generated educational materials for patients with urological cancers maintained accuracy, completeness, and clarity while achieving significantly better readability compared with original guideline documents[14]. In the field of lung cancer radiotherapy, although ChatGPT-4-generated responses were rated by clinical experts as highly relevant and accurate, their readability classification remained "very difficult to read."[15] Similarly, in colorectal cancer, ChatGPT-4 responses demonstrated acceptable safety and high-quality scores, but readability assessments indicated that a relatively high level of education was required for comprehension[16]. Furthermore, several studies have identified a potential "reliability–readability paradox" in LLM-generated patient education materials, whereby models with higher information reliability often produce texts with greater information density and reduced accessibility[17].

Importantly, existing studies have primarily focused on single or limited numbers of internationally prominent models[18], with relatively narrow coverage of cancer types and irAE categories. At present, no study has systematically evaluated the performance of different LLMs in patient education regarding ICI-M. Additionally, no investigation has directly compared domestic and international mainstream models within this clinical context.

Therefore, this study aims to systematically evaluate and compare five mainstream LLMs, including ChatGPT-5.5, Gemini-3.5 Flash, DeepSeek-V4, Qwen 3.7, and Doubao, in terms of information quality, reliability, comprehensibility, actionability, and text readability when responding to patient questions related to ICI-M. The findings may provide evidence to guide clinicians in the cautious selection and appropriate application of AI tools for patient education in cardio-oncology. Moreover, they may inform future strategies for optimizing LLM-based educational approaches for patients facing complex oncological conditions and high-risk irAEs.

## Methods

# Study design

The purpose of this cross-sectional comparative investigation was to characterize the quality, reliability, comprehensibility, actionability, and readability of patient education outputs produced by five LLMs for questions related to ICI-M. A total of 20 standardized patient education questions were submitted to five LLMs, generating 100 responses (20 responses from each model). The overall study workflow is presented in Fig. 1.

Twenty patient education questions across five domains were submitted to five large language models (LLMs). Responses were independently evaluated by two senior oncology experts for quality, reliability, understandability, actionability, and readability.

## Question collection

To construct the question set, authoritative clinical guidelines for ICI-M management, including those from the European Society for Medical Oncology (ESMO), American Society of Clinical Oncology (ASCO), and Chinese Society of Clinical Oncology (CSCO), were reviewed[13,19,20]. Two senior oncologists developed 20 questions based on guideline recommendations and their clinical experience. The questions covered five domains: disease understanding, etiology and risk factors, diagnosis and examination, treatment and management, and prevention and prognosis (Table 1). Both experts had more than ten years of clinical experience and considered common concerns encountered by patients in routine clinical practice to ensure the validity and representativeness of the question set. To enhance clinical applicability, the development of this process was guided by insights from prior related research[21–23].

Table 1
List of common patient questions about immune checkpoint inhibitor-associated myocarditis (ICI-M) used as exact prompts across all LLMs.

| **Questions** |
|---|
| Understanding of the disease |
| 1. What is ICI-M? |
| 2. How common is ICI-M, and when does it typically develop after starting ICI therapy? |
| 3. What are the common manifestations and symptoms of ICI-M? |
| 4. Can ICI-M occur even if I have never had heart disease? |
| Etiology and risk factors |
| 1. Why can ICI treatment lead to ICI-M? |
| 2. Does the risk of ICI-M vary depending on the type of cancer I have? |
| 3. Which personal factors may increase my risk of ICI-M? |
| 4. Which cancer treatment regimens are associated with a higher risk of ICI-M? |
| Diagnosis and testing |
| 1. If my doctor suspects that I have ICI-M, what tests will I need? |
| 2. What abnormal findings may appear in test results if I have ICI-M? |
| 3. Why are cardiac magnetic resonance imaging (cardiac MRI) or an endomyocardial biopsy sometimes needed to diagnose ICI-M? |
| 4. How do doctors distinguish ICI-M from myocardial infarction, infection, cancer progression, or other causes of heart injury? |
| Treatment and management |
| 1. Why are high-dose glucocorticoids the main initial treatment for ICI-M, and why must they be started as soon as possible? |
| 2. Why should I not suddenly stop glucocorticoids treatment on my own while being treated for ICI-M? |
| 3. If glucocorticoids do not work well, what other immunosuppressive treatments might I receive for ICI-M? |
| 4. How will my doctors determine whether treatment for ICI-M is working? |
| Prevention and prognosis |
| 1. Are there any measures that can reduce my risk of developing ICI-M? |
| 2. What is the prognosis for ICI-M? |
| 3. After developing ICI-M, can I ever restart treatment with ICI? |
| 4. During recovery from ICI-M, can patients exercise, work, or resume daily activities, and what is the recommended schedule for follow-up visits and cardiac monitoring after discharge? |

## AI platform prompting and data collection

Five LLMs were included in this study: ChatGPT-5.5 (OpenAI, May 5, 2026), Gemini-3.5 Flash (Google, May 19, 2026), DeepSeek-V4 (DeepSeek, April 24, 2026), Qwen 3.7 (Alibaba Cloud, May 20, 2026), and Doubao (ByteDance, July 27, 2026). All models were accessed through their publicly available default versions using standard user accounts, without activation of advanced modes or enhanced features. On July 28, 2026, standardized interactions were conducted with each model through a desktop computer interface. Each question was submitted once to each LLM following an identical interaction protocol, producing 100 independent responses. The 100 original responses are provided in the Supplementary Material. To minimize the influence of contextual information, chat histories were cleared before each query and each question was asked in a new session without iterative follow-up. The AI-generated outputs were kept in their original form, and no human editing was applied. Model-identifying information was removed to anonymize the outputs. This approach aligned with common practices in LLM evaluation, where each query is submitted in a separate session to avoid contextual carryover and ensure independent, comparable outputs across platforms.

## Evaluation process and metrics

Two senior oncology specialists independently evaluated the responses generated by the five LLMs. Both evaluators had more than ten years of clinical experience. Before formal assessment, both reviewers underwent standardized training and completed pilot evaluations using predefined samples to standardize scoring criteria. During the evaluation process, model identities were concealed, and responses were reviewed under blinded conditions using coded identifiers. Any disagreement between reviewers was resolved through adjudication by a third senior oncologist, and the final adjudicated scores were used for subsequent analyses.

The GQS was used to assess the overall subjective quality, flow, and usefulness of patient education information. Scores ranged from 1 (very poor quality) to 5 (excellent quality)[24]. Reliability was evaluated using the five-item mDISCERN scale[25]. Each fulfilled criterion received one point, with total scores ranging from 0 to 5. Higher scores indicated better reliability. The QAMAI scale was applied to assess the quality of AI-generated medical information across six domains: accuracy, clarity, relevance, completeness, source attribution, and usefulness[26]. A 5-point Likert scale was applied to each item, producing an overall score ranging from 6 to 30. Higher scores denoted higher information quality. In addition, the EQIP-20 tool was used to evaluate patient education information quality[27]. EQIP-20 assesses content, source identification/attribution, and structure, while evaluating the completeness, accuracy, and appropriateness of information presentation. The EQIP-20 score was converted to a 0-100 scale in this study, with higher scores indicating better information quality.

Patient usability was assessed using the Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P), with separate percentage scores calculated for understandability (PEMAT-U) and actionability (PEMAT-A)[28]. PEMAT-P scoring was performed according to the original user guide. Applicable

items were scored using a binary approach and converted into percentage scores for each domain. Higher scores indicated greater understandability or actionability.

The readability of LLM-generated responses was analyzed using an online text readability assessment tool (https://readabilityformulas.com/). Six established English readability indices were used to evaluate linguistic complexity: Automated Readability Index (ARI)[29], Flesch Reading Ease Score (FRES)[30], Gunning Fog Index (GFI)[31], Flesch-Kincaid Grade Level (FKGL)[30], Coleman-Liau Index (CLI)[32], and Simple Measure of Gobbledygook (SMOG)[33]. Higher FRES scores reflect greater ease of reading, whereas elevated ARI, GFI, FKGL, CLI, and SMOG values denote increased reading difficulty or a higher educational level needed for comprehension. Readability metrics offered an objective assessment of linguistic complexity in LLM-generated responses, and when combined with subjective quality evaluations, they helped judge whether the content was suitable for public comprehension.

# Statistical analysis

Inter-rater agreement was assessed. For ordinal variables including (GQS, mDISCERN), quadratic weighted Cohen's kappa coefficients with 95% confidence intervals were calculated. For QAMAI, EQIP, PEMAT-U, and PEMAT-A scores, agreement was evaluated using a two-way random-effects, absolute agreement, single-measure intraclass correlation coefficient [ICC(A,1)] with 95% confidence intervals.

Because the same set of questions was answered by all five models, comparisons among LLMs were treated as repeated-measures data. GQS, mDISCERN, QAMAI, EQIP, PEMAT-U, and PEMAT-A scores were presented as medians with interquartile ranges [M (Q1, Q3)]. Overall differences among the five models were assessed using the Friedman test, followed by Dunn's post hoc test for pairwise comparisons.

Readability indices (ARI, FRES, GFI, FKGL, CLI, SMOG) scores were expressed as mean ± standard deviation (SD). Differences among models were analyzed using repeated-measures one-way analysis of variance (ANOVA). When necessary, the Geisser–Greenhouse correction was applied, and pairwise comparisons among models were adjusted using the Bonferroni correction. All statistical tests were two-sided, and $P < 0.05$ was considered statistically significant.

Because most variables were not normally distributed, associations were assessed using Spearman's rank correlation coefficients. Each cell represented the corresponding $\rho$ value, and color intensity indicated the strength and direction of correlations (red: positive correlation; blue: negative correlation). Raincloud plots were used to visualize distributions, integrating kernel density estimation, box plots, and jittered scatter points. All tests were two-sided, with $P < 0.05$ considered statistically significant.

# Ethical considerations

The study used publicly available AI-generated outputs and did not involve human participants, identifiable data, or biological samples. Therefore, institutional ethics approval was not required.

## Results

# Inter-rater reliability analysis

We submitted 20 standardized questions about ICI-M to each of the five LLMs and obtained a total of 100 responses (20 responses per model). The questions covered five domains: understanding of the disease, etiology and risk factors, diagnosis and testing, treatment and management, and prevention and prognosis. The complete set of questions is presented in Table 1, and the overall study workflow is illustrated in Fig. 1.

We assessed inter-rater reliability for all scoring instruments. Agreement between the two primary raters ranged from good to excellent (Table 2). The weighted Cohen's $\kappa$ was 0.729 (95% CI, 0.610–0.848) for the GQS and 0.718 (95% CI, 0.602–0.835) for the mDISCERN. The ICC(A,1) values for the QAMAI, EQIP, PEMAT-U, and PEMAT-A were 0.867, 0.812, 0.887, and 0.880, respectively. All reliability coefficients were statistically significant ($P < 0.001$). These findings indicate good inter-rater consistency and support the reproducibility of the evaluation process across reviewers.

Table 2
Inter-rater reliability analysis for outcome measures.

| Outcome metric | Statistical test | Coefficient value | 95% Confidence Interval | *P*-value |
|---|---|---|---|---|
| GQS | Weighted Cohen's $\kappa$ (quadratic) | 0.729 | 0.610–0.848 | < 0.001 |
| mDISCERN | Weighted Cohen's $\kappa$ (quadratic) | 0.718 | 0.602–0.835 | < 0.001 |
| QAMAI | ICC(A,1), 2-way random, absolute agreement | 0.867 | 0.808–0.908 | < 0.001 |
| EQIP | ICC(A,1), 2-way random, absolute agreement | 0.812 | 0.734–0.870 | < 0.001 |
| PEMAT-U | ICC(A,1), 2-way random, absolute agreement | 0.887 | 0.836–0.922 | < 0.001 |
| PEMAT-A | ICC(A,1), 2-way random, absolute agreement | 0.880 | 0.827–0.918 | < 0.001 |

GQS, Global Quality Score, mDISCERN, modified DISCERN, QAMAI, Quality Analysis of Medical Artificial Intelligence, EQIP, Ensuring Quality Information for Patients, PEMAT-U, Patient Education Materials Assessment Tool-Understandability, PEMAT-A, Patient Education Materials Assessment Tool-Actionability.

# Quality, reliability, understandability, and actionability analysis

Overall comparisons revealed significant differences among the five LLMs in GQS, mDISCERN, QAMAI, EQIP, and PEMAT-A scores (all $P \leq 0.038$), whereas no significant between-model difference was observed for PEMAT-U ($P = 0.226$; Table 3). QAMAI demonstrated the largest between-model variation ($\chi^2 = 32.29$, $P < 0.001$), highlighting substantial differences in information quality across models. The median QAMAI scores for ChatGPT-5.5, Gemini-3.5 Flash, DeepSeek-V4, Qwen 3.7, and Doubao were 28.00, 28.00, 29.00,

27.00, and 25.50, respectively. DeepSeek-V4 had the highest median mDISCERN score [5.00 (IQR, 4.00–5.00)] and the highest median QAMAI score [29.00 (27.75–30.00)]. In contrast, Doubao had the lowest median GQS score [3.00 (3.00–4.00)] and the lowest median QAMAI score [25.50 (25.00–27.25)]. ChatGPT-5.5 had the highest median PEMAT-A score [40.00 (35.00–45.00)].

Further post hoc pairwise comparisons showed that DeepSeek-V4 scored significantly higher than Doubao on both GQS and mDISCERN (both $P < 0.01$). For QAMAI, DeepSeek-V4 scored significantly higher than Qwen 3.7 ($P < 0.05$) and Doubao ($P < 0.0001$). ChatGPT-5.5 and Gemini-3.5 Flash also outperformed Doubao in QAMAI ($P < 0.001$ and $P < 0.05$, respectively; Fig. 2). Although the overall between-model differences in EQIP and PEMAT-A were statistically significant, no pairwise comparison remained significant after post hoc adjustment.

Table 3
Comparison of response quality and reliability across different LLMs.

| Variables | Total (n = 100 ) | ChatGPT-5.5 (n = 20) | Gemini-3.5 Flash (n = 20) | DeepSeek -V4 (n = 20) | Qwen 3.7 (n = 20) | Doubao (n = 20) | Statistic | *P*-value |
|---|---|---|---|---|---|---|---|---|
| GQS, M ($Q_1$, $Q_3$) | 4.00 (3.00, 4.00) | 4.00 (4.00, 4.25) | 4.00 (3.00, 4.00) | 4.00 (4.00, 5.00) | 4.00 (3.00, 4.00) | 3.00 (3.00, 4.00) | $\chi^2$=22.84# | < 0.001 |
| mDISCERN, M ($Q_1$, $Q_3$) | 4.00 (4.00, 5.00) | 4.00 (4.00, 5.00) | 4.00 (4.00, 5.00) | 5.00 (4.00, 5.00) | 4.00 (4.00, 4.00) | 4.00 (2.00, 4.00) | $\chi^2$=18.04# | 0.001 |
| QAMAI, M ($Q_1$, $Q_3$) | 28.00 (27.00, 28.00) | 28.00 (28.00, 28.25) | 28.00 (27.00, 28.00) | 29.00 (27.75, 30.00) | 27.00 (26.75, 28.00) | 25.50 (25.00, 27.25) | $\chi^2$=32.29# | < 0.001 |
| EQIP, M ($Q_1$, $Q_3$) | 80.00 (75.00, 82.50) | 82.50 (77.50, 85.62) | 77.50 (75.00, 80.62) | 82.50 (79.38, 82.50) | 82.50 (76.88, 83.12) | 78.75 (62.50, 83.12) | $\chi^2$=10.13# | 0.038 |
| PEMAT-U, M ($Q_1$, $Q_3$) | 75.00 (68.50, 77.75) | 76.00 (69.00, 83.00) | 75.00 (67.00, 75.50) | 73.00 (69.00, 77.00) | 75.00 (73.50, 77.00) | 74.00 (58.00, 77.75) | $\chi^2$=5.66# | 0.226 |
| PEMAT-A, M ($Q_1$, $Q_3$) | 20.00 (20.00, 40.00) | 40.00 (35.00, 45.00) | 20.00 (0.00, 40.00) | 30.00 (20.00, 40.00) | 20.00 (20.00, 40.00) | 20.00 (0.00, 40.00) | $\chi^2$=12.62# | 0.013 |

#: Friedman test. M: Median, $Q_1$: 1st Quartile, $Q_3$: 3rd Quartile. GQS, Global Quality Score, mDISCERN, modified DISCERN, QAMAI, Quality Analysis of Medical Artificial Intelligence, EQIP, Ensuring Quality Information for Patients. PEMAT-U, Patient Education Materials Assessment Tool-Understandability, PEMAT-A, Patient Education Materials Assessment Tool-Actionability.

Score distributions and pairwise comparisons are shown for GQS (**a, b**), mDISCERN (**c, d**), QAMAI (**e, f**), EQIP (**g, h**), PEMAT-U (**i, j**), and PEMAT-A (**k, l**), respectively. Higher scores indicate better performance. *$P$ < 0.05, **$P$ < 0.01, ***$P$ < 0.001, ****$P$ < 0.0001. GQS, Global Quality Score, mDISCERN, modified DISCERN, QAMAI, Quality Analysis of Medical Artificial Intelligence, EQIP, Ensuring Quality Information for Patients, PEMAT-U, Patient Education Materials Assessment Tool-Understandability, PEMAT-A, Patient Education Materials Assessment Tool-Actionability.

# Readability analysis

Overall comparisons showed significant differences among the five LLMs across all six readability indices (all $P \leq 0.004$; Table 4). FRES exhibited the largest overall difference ($F = 15.880$, $P < 0.001$), indicating notable variation in text readability across models. The mean FRES scores for ChatGPT-5.5, Gemini-3.5 Flash, DeepSeek-V4, Qwen 3.7, and Doubao were 6.65, 10.00, 22.95, 11.55, and 6.50, respectively.

Table 4
Comparison of response readability across different LLMs.

| Variables | Total (n = 100 ) | ChatGPT-5.5 (n = 20) | Gemini-3.5 Flash (n = 20) | DeepSeek-V4 (n = 20) | Qwen 3.7 (n = 20) | Doubao (n = 20) | Statistic | *P*-value |
|---|---|---|---|---|---|---|---|---|
| ARI, Mean ± SD | 17.85 ± 3.09 | 16.57 ± 2.92 | 19.40 ± 3.37 | 16.27 ± 2.54 | 17.75 ± 2.84 | 19.24 ± 2.51 | *F* = 3.765 | < 0.001 |
| FRES, Mean ± SD | 11.53 ± 11.99 | 6.65 ± 8.92 | 10.00 ± 12.23 | 22.95 ± 10.56 | 11.55 ± 11.54 | 6.50 ± 9.13 | *F* = 15.880 | < 0.001 |
| GFI, Mean ± SD | 17.59 ± 2.28 | 17.21 ± 2.03 | 17.90 ± 2.05 | 16.46 ± 2.07 | 17.82 ± 2.24 | 18.54 ± 2.64 | *F* = 5.103 | 0.004 |
| FKGL, Mean ± SD | 16.60 ± 2.64 | 15.98 ± 2.49 | 17.82 ± 2.99 | 14.94 ± 2.13 | 16.65 ± 2.64 | 17.61 ± 1.92 | *F* = 7.268 | < 0.001 |
| CLI, Mean ± SD | 18.48 ± 2.54 | 19.53 ± 1.95 | 18.80 ± 2.75 | 16.32 ± 2.04 | 18.46 ± 2.23 | 19.31 ± 2.48 | *F* = 11.840 | < 0.001 |
| SMOG, Mean ± SD | 13.88 ± 2.32 | 12.18 ± 2.37 | 14.96 ± 2.22 | 13.35 ± 1.65 | 13.93 ± 2.15 | 14.98 ± 2.09 | *F* = 8.581 | < 0.001 |

*F*: RM One-way ANOVA, SD: standard deviation. ARI, Automated Readability Index, FRES, Flesch Reading Ease Score, GFI, Gunning Fog Index, FKGL, Flesch-Kincaid Grade Level, CLI, Coleman-Liau Index, SMOG, Simple Measure of Gobbledygook.

DeepSeek-V4 had the lowest mean ARI (16.27 ± 2.54), GFI (16.46 ± 2.07), FKGL (14.94 ± 2.13), and CLI (16.32 ± 2.04) scores, whereas ChatGPT-5.5 had the lowest mean SMOG score (12.18 ± 2.37). Further post

hoc pairwise comparisons showed that DeepSeek-V4 had significantly higher FRES scores than ChatGPT-5.5 and Doubao (both $P < 0.0001$), as well as Gemini-3.5 Flash and Qwen 3.7 (both $P < 0.01$). DeepSeek-V4 had significantly lower CLI scores than ChatGPT-5.5 and Doubao (both $P < 0.0001$). It also had lower CLI scores than Gemini-3.5 Flash and Qwen 3.7 (both $P < 0.01$). Significant pairwise differences were observed among the five LLMs for all six readability metrics, including ARI, FRES, GFI, FKGL, CLI, and SMOG (Fig. 3).

Score distributions and pairwise comparisons are shown for ARI (**a, b**), FRES (**c, d**), GFI (**e, f**), FKGL (**g, h**), CLI (**i, j**), and SMOG (**k, l**), respectively. Higher FRES scores indicate greater readability, whereas higher scores on the other indices indicate greater reading difficulty. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$. ARI, Automated Readability Index, FRES, Flesch Reading Ease Score, GFI, Gunning Fog Index, FKGL, Flesch-Kincaid Grade Level, CLI, Coleman-Liau Index, SMOG, Simple Measure of Gobbledygook.

Despite these relative differences among models, the absolute readability of all generated responses remained poor. The overall mean FKGL and FRES scores were 16.60 ± 2.64 and 11.53 ± 11.99, respectively. These values indicate that the responses were generally difficult for the public to read and comprehend.

# Correlation analysis

Correlation analysis demonstrated two distinct correlation patterns among the evaluated metrics (Fig. 4). Quality-, reliability-, understandability-, and actionability-related measures showed consistently positive correlations. The strongest association was observed between EQIP and PEMAT-A ($\rho = 0.72$), followed by GQS and mDISCERN ($\rho = 0.48$) and mDISCERN and QAMAI ($\rho = 0.56$). GQS also demonstrated moderate positive correlations with QAMAI ($\rho = 0.41$), EQIP ($\rho = 0.36$), PEMAT-U ($\rho = 0.32$), and PEMAT-A ($\rho = 0.26$), indicating that higher perceived response quality tended to accompany better reliability, understandability, and actionability.

The heatmap shows Spearman's rank correlation coefficients ($\rho$) among GQS, mDISCERN, QAMAI, EQIP, PEMAT-U, PEMAT-A, and six readability indices. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$.

In contrast, readability indices showed stronger interrelationships. ARI was strongly correlated with FKGL ($\rho = 0.94$) and SMOG ($\rho = 0.80$), whereas FRES showed strong negative correlations with ARI ($\rho = -0.70$), GFI ($\rho = -0.79$), FKGL ($\rho = -0.77$), and CLI ($\rho = -0.90$). These findings indicate substantial overlap among readability measures and consistent evaluation of textual complexity.

Associations between readability and quality-related metrics were generally weak to moderate. Higher EQIP and PEMAT-A scores were associated with lower ARI, FKGL, and SMOG values, suggesting that responses with better usability and actionability tended to be easier to read. However, GQS and mDISCERN showed limited correlations with readability indices. These results suggest that information quality, patient usability, and readability represent related yet distinct dimensions of LLM-generated patient education.

## Discussion

This study conducted a multidimensional evaluation of five widely used LLMs for patient education related to ICI-M, focusing on information quality, reliability, understandability, actionability, and readability.

Several key findings were identified. First, the five models demonstrated different performance profiles across multiple quality and patient-oriented evaluation metrics, and no single model consistently outperformed the others across all domains. DeepSeek-V4 showed relatively better performance in mDISCERN and QAMAI, whereas ChatGPT-5.5 achieved the highest median PEMAT-A score. Second, although significant differences were observed across all six readability indices, the absolute readability of the generated responses remained suboptimal, with an overall FKGL of 16.60 ± 2.64 and FRES of 11.53 ± 11.99, indicating substantial linguistic complexity. Third, understandability scores were consistently higher than actionability scores, suggesting that current LLMs are more effective at conveying medical concepts than at translating them into clear, actionable steps for patients. Finally, correlations between readability and quality-related indicators were generally weak, underscoring that information quality, patient usability, and language accessibility represent interconnected yet distinct dimensions of LLM performance. Therefore, evaluation of LLM-generated health information should adopt a multidimensional framework rather than relying on a single quality or readability metric.

The observed differences in quality and reliability further highlight the model-specific characteristics of LLM performance. DeepSeek-V4 achieved the highest median scores in mDISCERN and QAMAI, whereas Doubao showed the lowest median scores in GQS and QAMAI. However, these differences did not persist across all evaluation tools, reflecting the heterogeneous strengths and limitations of individual models. For example, ChatGPT-5.5 achieved the highest median PEMAT-A score. Moreover, although overall statistical tests for EQIP and PEMAT-A reached significance, post hoc pairwise comparisons did not reveal significant differences between specific model pairs, suggesting that variability in these dimensions may be less pronounced. Previous comparative studies of multiple LLMs have similarly demonstrated that relative model performance is influenced by clinical topic, model architecture, prompting context, and evaluation instruments[34–36]. In different patient education scenarios, the model achieving the highest reliability or quality score may not necessarily demonstrate superior readability or consistently strong performance across other dimensions. Therefore, statistically significant differences between models should be interpreted as relative performance variations rather than direct evidence of clinical superiority. The choice of LLM platforms for developing patient-education materials should reflect an appropriate balance between content depth and understandability, aligned with the requirements of the target population.

Readability represents a major limitation shared by all five models. In this study, the overall FKGL was 16.60, indicating that the reading level of LLM-generated responses was substantially higher than commonly recommended for patient health education materials. The extremely low overall FRES score similarly reflected considerable linguistic complexity. These findings are consistent with recent studies. Previous research has shown that ChatGPT-generated prostate cancer patient education materials may be reliable but require approximately a high-school or college-level reading ability[22]. A recent Scientific Reports study on gestational diabetes mellitus also demonstrated significant differences among LLMs and reported that all models produced readability scores below the threshold considered appropriate for the general public[37]. The relatively high language complexity observed in our study may partly result from the inherent complexity of ICI-M, which involves specialized concepts including immune checkpoint inhibition, cardiac biomarkers, cardiac magnetic resonance imaging, endomyocardial biopsy, and immunosuppressive

therapy. Despite disease complexity, patient education materials should use plain language whenever possible. ICI-M often presents with nonspecific initial symptoms and may occur after patients have left the hospital. Clear and accessible patient education may therefore facilitate early symptom recognition and timely medical evaluation, ultimately improving clinical outcomes.

Another clinically important finding was the substantial gap between understandability and actionability. The overall median PEMAT-U score was 75.00, whereas the median PEMAT-A score was only 20.00. This gap indicated that LLM-generated responses are considerably more effective at presenting understandable information than at providing specific, actionable guidance that patients can directly follow. Similar patterns have been reported in other LLM-based patient education studies, where understandability was generally acceptable but actionability remained limited[38]. This limitation has particular clinical relevance for ICI-M. Patients experiencing ICI-M need to recognize potential warning symptoms, including new-onset chest discomfort, dyspnea, palpitations, fatigue, or syncope, and seek urgent medical assessment. Current ESC recommendations emphasize that suspected ICI-related myocarditis requires diagnostic triage within 24 hours, followed by immediate interruption of ICI therapy and initiation of high-dose glucocorticoid after confirmation[39]. Therefore, patient-directed information in this context should not only explain the disease but also clearly specify when urgent medical attention is required.

Correlation analyses further support the multidimensional nature of LLM performance in patient education. Quality-, reliability-, understandability-, and actionability-related indicators generally showed positive correlations, whereas readability indices demonstrated stronger internal correlations within their own domain. In contrast, associations between readability and quality-related metrics were generally weak or limited. Notably, higher PEMAT-A scores were associated with lower ARI, FKGL, and SMOG values and higher FRES scores, suggesting that simpler language may facilitate actionability. However, GQS showed weak relationships with readability indices. These findings indicate that language simplification alone cannot ensure high-quality medical information, and comprehensive or reliable responses may still be difficult for patients to read. Similar separation between content quality and language accessibility has been reported in other LLM comparison studies[40]. Therefore, optimization of LLM-generated patient education materials requires simultaneous consideration of factual quality, readability, understandability, and actionability rather than improvement of a single dimension.

These limitations are particularly important in the context of ICI-M because this adverse event has a relatively low incidence but high mortality, with nonspecific clinical manifestations that may delay recognition and treatment. Current reviews and clinical guidelines emphasize its substantial clinical risk and the importance of early identification and prompt management[13]. Therefore, even in the absence of factual inaccuracies, incomplete patient education information, excessive linguistic complexity, or insufficient action guidance may have meaningful clinical consequences. LLMs should therefore be regarded as supportive tools for patient education under clinician supervision rather than independent sources of clinical guidance.

Our findings also suggest several directions for improving LLM-generated patient education content. Explicit prompting strategies may help reduce linguistic complexity and improve actionability. Such

strategies include requiring plain language, specifying target reading levels, incorporating structured step-by-step action guidance, and adding standardized safety warnings or medical attention prompts. Previous patient education studies have suggested that targeted prompting for simplification can reduce reading difficulty without substantially compromising information quality, while structured prompts may improve response consistency[14]. Future LLM outputs should clearly distinguish core information, warning signs, recommended actions, and situations requiring urgent medical attention. More importantly, these optimization strategies should be validated among real patients with different health literacy levels rather than relying solely on automated readability formulas or expert evaluations. A cross-platform system for ongoing quality control and evaluation will be essential to support the continuous monitoring and refinement of LLM performance.

This study has several strengths. First, ICI-M is a rare but potentially fatal complication, and early recognition and treatment are essential for improving outcomes. Therefore, effective patient education is clinically important. Previous studies have primarily focused on cancer treatment itself, whereas education on adverse events has received comparatively less attention. Second, we used 20 standardized patient-oriented questions to systematically compare five mainstream LLMs, covering key domains including disease knowledge, etiology and risk factors, diagnosis, treatment, and prognosis or follow-up. Third, we applied a comprehensive evaluation framework incorporating GQS, mDISCERN, QAMAI, EQIP, PEMAT-U, PEMAT-A, and six established readability indices, with expert-based evaluation metrics demonstrating generally good inter-rater reliability.

Several limitations should also be acknowledged. First, this was a cross-sectional study, and LLMs continue to evolve rapidly. Therefore, our findings represent the performance of specific model versions at a specific time point. Future studies should longitudinally monitor model performance during subsequent updates. Second, all assessments used English prompts to control baseline variables. Whether these findings generalize to other languages, cultural contexts, or populations with varying health literacy levels requires further investigation. Third, standardized single-turn prompts cannot fully simulate real-world interactions between patients and LLMs, as follow-up questions may substantially modify final responses. Fourth, evaluations were performed by medical experts and automated readability tools, which cannot directly determine whether patients truly understand or correctly implement the provided information. Fifth, although we comprehensively evaluated overall response quality, we did not quantitatively analyze clinical information omissions, guideline concordance, or clinical safety, which should be addressed in future studies.

## Conclusion

In conclusion, this study demonstrates that current LLMs provide patient education content for ICI-M with distinct strengths and limitations, and no single model consistently outperforms others across all dimensions. Although these models can provide clinically valuable medical information, their poor absolute readability and limited actionability remain major barriers for patients. The weak correlations among different evaluation domains further indicate that high information quality does not necessarily translate into understandability or practical usability. Therefore, LLMs have potential as supportive tools for patient

education under clinician supervision. However, their patient-facing outputs require substantial optimization to ensure completeness, accessibility, and clear actionable guidance before they can be safely integrated into clinical practice. Future research should broaden the scope of assessed LLMs and clinical scenarios, integrate patient-reported experiences and real-world outcome data, and establish continuous monitoring frameworks to strengthen the patient-education performance of these models over time.

# Data Availability Statement

The datasets analyzed for this study are available in online repositories. The data is provided in the article and Supplementary Material.

## Declarations

# Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Declarations
**of Generative AI and AI-Assisted Technologies**

# Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

# Author Contribution

LW and JY designed the study. TW conducted the standardized interactions with all five LLMs and collected 100 independent responses. LW performed the data analysis and drafted the initial manuscript. JZ, MZ, and YZ generated the figures and reviewed the analytical results. DD and JY revised the manuscript for important intellectual content. All authors contributed to the work and approved the final version.

# Acknowledgement

We thank Professor Biaofei Ye and Professor Shuangyan Zhang for their valuable support and assistance in this study.

# Data Availability

The datasets analyzed for this study are available in online repositories. The data is provided in the article and Supplementary Material.

## References


1. Sharma, P. et al. The Next Decade of Immune Checkpoint Therapy. *Cancer Discov.* **11**, 838–857. 10.1158/2159-8290.Cd-20-1680 (2021).
2. Björkström, K. et al. Perioperative immune checkpoint inhibitor therapy across tumors: Insights and shared lessons from a rapidly evolving field. *J. Intern. Med.* **299**, 538–569. 10.1111/joim.70073 (2026).
3. Zdenkowski, N. et al. Timing of nivolumab with neoadjuvant carboplatin and paclitaxel for early triple-negative breast cancer (BCT1902/IBCSG 61 – 20; Neo-N): a non-comparative, open-label, randomised, phase 2 trial. *Lancet Oncol.* **26**, 367–377. 10.1016/s1470-2045(24)00757-5 (2025).
4. Dougan, M., Luoma, A. M., Dougan, S. K. & Wucherpfennig, K. W. Understanding and treating the inflammatory adverse events of cancer immunotherapy. *Cell* **184**, 1575–1588. 10.1016/j.cell.2021.02.011 (2021).
5. Johnson, D. B., Nebhan, C. A., Moslehi, J. J. & Balko, J. M. Immune-checkpoint inhibitors: long-term implications of toxicity. *Nat. Rev. Clin. Oncol.* **19**, 254–267. 10.1038/s41571-022-00600-w (2022).
6. Jing, Y., Yang, J., Johnson, D. B., Moslehi, J. J. & Han, L. Harnessing big data to characterize immune-related adverse events. *Nat. Rev. Clin. Oncol.* **19**, 269–280. 10.1038/s41571-021-00597-8 (2022).
7. Hu, J. R. et al. Cardiovascular toxicities associated with immune checkpoint inhibitors. *Cardiovascular. Res.* **115**, 854–868. 10.1093/cvr/cvz026 (2019).
8. Mahmood, S. S. et al. Myocarditis in Patients Treated With Immune Checkpoint Inhibitors. *J. Am. Coll. Cardiol.* **71**, 1755–1764. 10.1016/j.jacc.2018.02.037 (2018).
9. Palaskas, N. L., Siddiqui, B. A. & Deswal, A. Steroids in Immune Checkpoint Inhibitor Myocarditis. *JACC CardioOncology*. **6**, 800–803. 10.1016/j.jaccao.2024.07.002 (2024).
10. Lyon, A. R. et al. ESC Guidelines on cardio-oncology developed in collaboration with the European Hematology Association (EHA), the European Society for Therapeutic Radiology and Oncology (ESTRO) and the International Cardio-Oncology Society (IC-OS). *Eur Heart J* 43, 4229–4361, (2022). 10.1093/eurheartj/ehac244 (2022).
11. Itzhaki, B., Zadok, O., Levi, A., Divakaran, S. & Nohria, A. Severe vs Nonsevere Immune Checkpoint Inhibitor-Induced Myocarditis: Contemporary 1-Year Outcomes. *JACC CardioOncology*. **5**, 732–744. 10.1016/j.jaccao.2023.09.004 (2023).
12. Leong, D. P. et al. Cardiovascular Considerations During Cancer Therapy: Gaps in Evidence and JACC: CardioOncology Expert Panel Recommendations. *JACC CardioOncology*. **6**, 815–834. 10.1016/j.jaccao.2024.06.005 (2024).
13. Schneider, B. J. et al. Management of Immune-Related Adverse Events in Patients Treated With Immune Checkpoint Inhibitor Therapy: ASCO Guideline Update. *J. Clin. oncology: official J. Am. Soc. Clin. Oncol.* **39**, 4073–4126. 10.1200/jco.21.01440 (2021).

14. Rodler, S. Cei, Francesco, Ganjavi, Conner, Checcucci, Enrico & others. GPT-4 generates accurate and readable patient education materials aligned with current oncological guidelines: a randomized assessment. *PLoS ONE*. **20**, e0324175. 10.1371/journal.pone.0324175 (2025).
15. Richlitzki, C. et al. Assessing ChatGPT's educational potential in lung cancer radiotherapy from clinician and patient perspectives: content quality and readability analysis. *JMIR Cancer*. 10.2196/63089 (2025).
16. Siu, A. H. Y. et al. ChatGPT as a patient education tool in colorectal cancer—an in-depth assessment of efficacy, quality and readability. *Colorectal Dis.* **27**, e17240. 10.1111/codi.17240 (2025).
17. Anonymous. The readability paradox and illusion of quality in AI-generated spine surgery patient education: a systematic review and exploratory meta-analysis. *North. Am. Spine Soc. J.* 10.1016/j.xnsj.2026.100924 (2026).
18. Sun, M., Zang, D. & Chen, J. Benchmarking readability, reliability, and scientific quality of large language models in communicating organoid science. *Front. Bioeng. Biotechnol.* **14**, 1750225. 10.3389/fbioe.2026.1750225 (2026).
19. Haanen, J. et al. Management of toxicities from immunotherapy: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Annals oncology: official J. Eur. Soc. Med. Oncol.* **33**, 1217–1238. 10.1016/j.annonc.2022.10.001 (2022).
20. Chinese Society of Clinical Oncology (CSCO). *Guidelines Working Committee. Guidelines of Chinese Society of Clinical Oncology (CSCO): management of immune checkpoint inhibitor-related toxicity [in Chinese]* (People's Medical Publishing House, 2025).
21. Luo, Z., Lin, C., Kim, T. H., Shin, Y. S. & Ahn, S. T. Quality and Readability Analysis of Artificial Intelligence-Generated Medical Information Related to Prostate Cancer: A Cross-Sectional Study of ChatGPT and DeepSeek. *world J. men's health*. 10.5534/wjmh.250144 (2025).
22. Gibson, D. et al. Evaluating the Efficacy of ChatGPT as a Patient Education Tool in Prostate Cancer: Multimetric Assessment. *J. Med. Internet. Res.* **26**, e55939. 10.2196/55939 (2024).
23. Wang, Z. et al. Challenges of using generative artificial intelligence for diabetes patient education: a cross-platform analysis of the quality, readability, and actionability of text generated by large language models. *Front. public. health*. **14**, 1804524. 10.3389/fpubh.2026.1804524 (2026).
24. Bernard, A. et al. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Am. J. Gastroenterol.* **102**, 2070–2077. 10.1111/j.1572-0241.2007.01325.x (2007).
25. Charnock, D., Shepperd, S., Needham, G. & Gann, R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J. Epidemiol. Community Health*. **53**, 105–111. 10.1136/jech.53.2.105 (1999).
26. Vaira, L. et al. Validation of the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool: a new tool to assess the quality of health information provided by AI platforms. *Eur. Arch. Otorhinolaryngol.* **281**, 6123–6131. 10.1007/s00405-024-08710-0 (2024).
27. Moult, B., Franck, Linda, S. & Brady Helen. Ensuring Quality Information for Patients: development and preliminary validation of a new instrument to improve the quality of written health care information. *Health Expect.* **7**, 165–175. 10.1111/j.1369-7625.2004.00273.x (2004).

28. Shoemaker, S. J., Wolf, M. S. & Brach Cindy. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ. Couns.* **96**, 395–403. 10.1016/j.pec.2014.05.027 (2014).
29. Senter, R. J. & Smith, E. A. Automated Readability Index. *AMRL-TR-66-220* (1967).
30. Kincaid, J., Peter, Fishburne, R. P., Rogers, R. L. & Chissom, B. S. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. *Research Branch Report 8–75* (1975).
31. Gunning, R. *The Technique of Clear Writing* (McGraw-Hill, 1952).
32. Coleman, M. & Liau, T. L. A computer readability formula designed for machine scoring. *J. Appl. Psychol.* **60**, 283–284. 10.1037/h0076540 (1975).
33. McLaughlin, G. & Harry SMOG grading: a new readability formula. *J. Read.* **12**, 639–646 (1969).
34. Johri, S. et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat. Med.* **31**, 77–86. 10.1038/s41591-024-03328-5 (2025).
35. Cui, Z. et al. Performance of large language model in cross-specialty medical scenarios. *J. translational Med.* **24** 10.1186/s12967-025-07577-x (2025).
36. Xu, L., Zhao, W. & Huang, X. Diagnosis and Triage Performance of Contemporary Large Language Models on Short Clinical Vignettes. *J. Med. Syst.* **49** 10.1007/s10916-025-02284-y (2025).
37. Yigit et al. Multidimensional assessment of large language model responses to patient questions on gestational diabetes mellitus. *Sci. Rep.* **15**, 43758. 10.1038/s41598-025-27343-2 (2025).
38. Dihan, Q. A. et al. Leveraging large language models to improve patient education on dry eye disease. *Eye (London England)*. **39**, 1115–1122. 10.1038/s41433-024-03476-5 (2025).
39. Schulz-Menger, J. et al. ESC Guidelines for the management of myocarditis and pericarditis. *Eur Heart J* 46, 3952–4041, (2025). 10.1093/eurheartj/ehaf192 (2025).
40. Flanagan, R. F. & Concise Commentary AI-Generated GI Patient Education-High Quality, yet Readability May Limit Utility. *Dig. Dis. Sci.* 10.1007/s10620-026-10152-z (2026).

# Figures

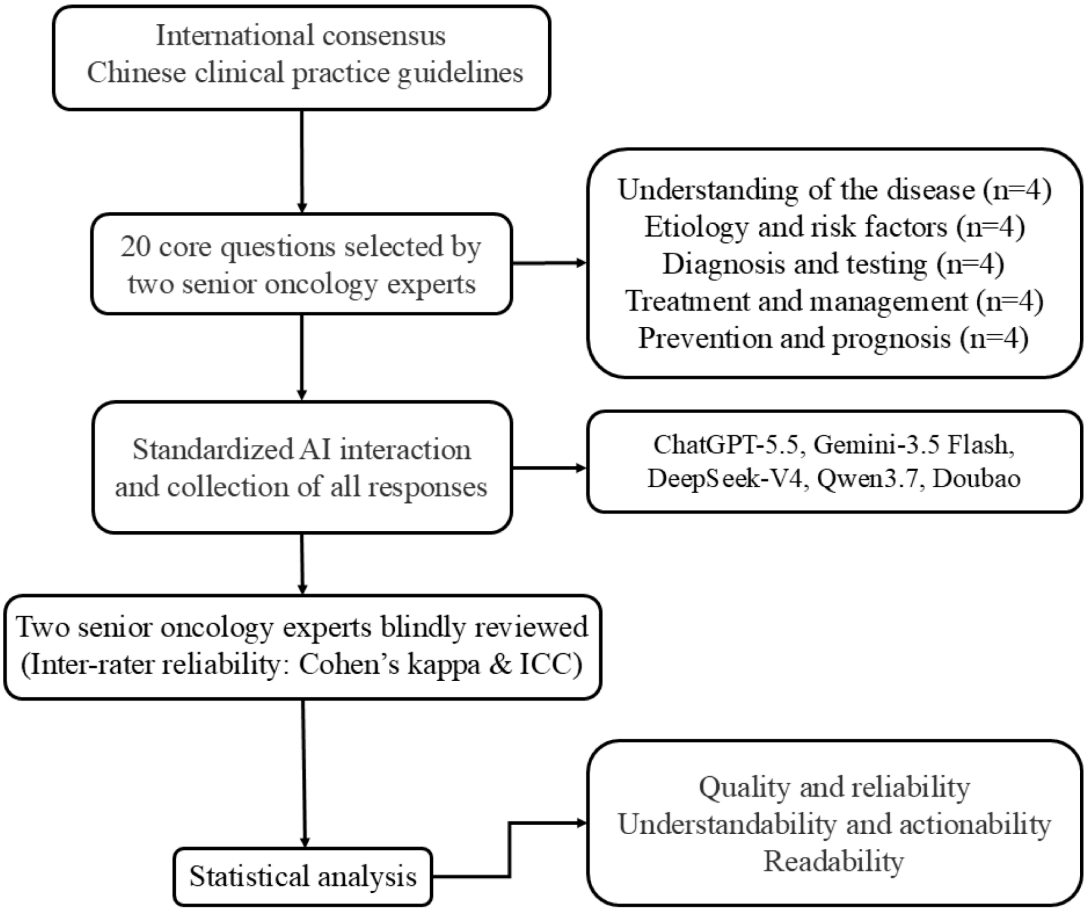
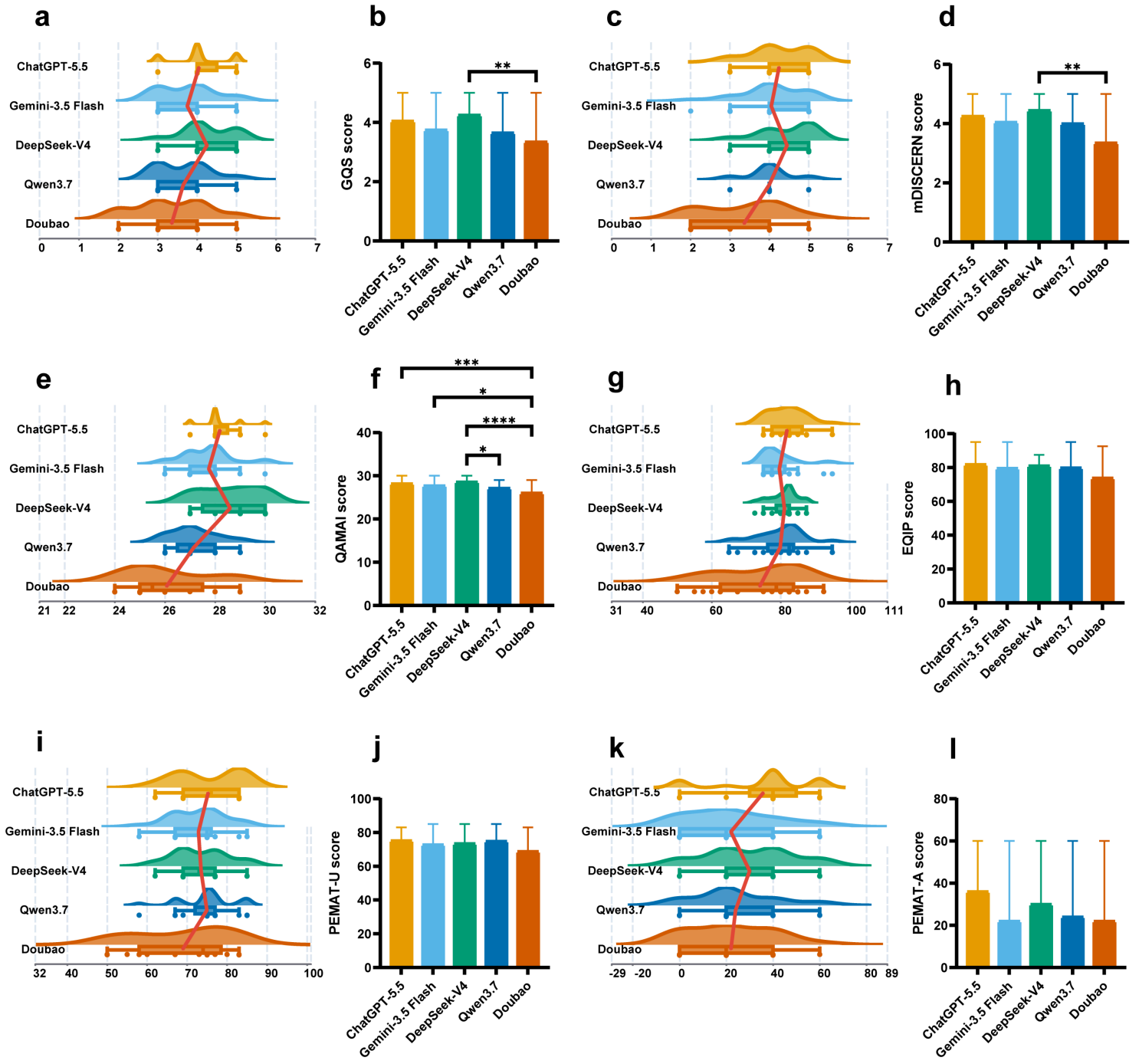



**Figure 1**

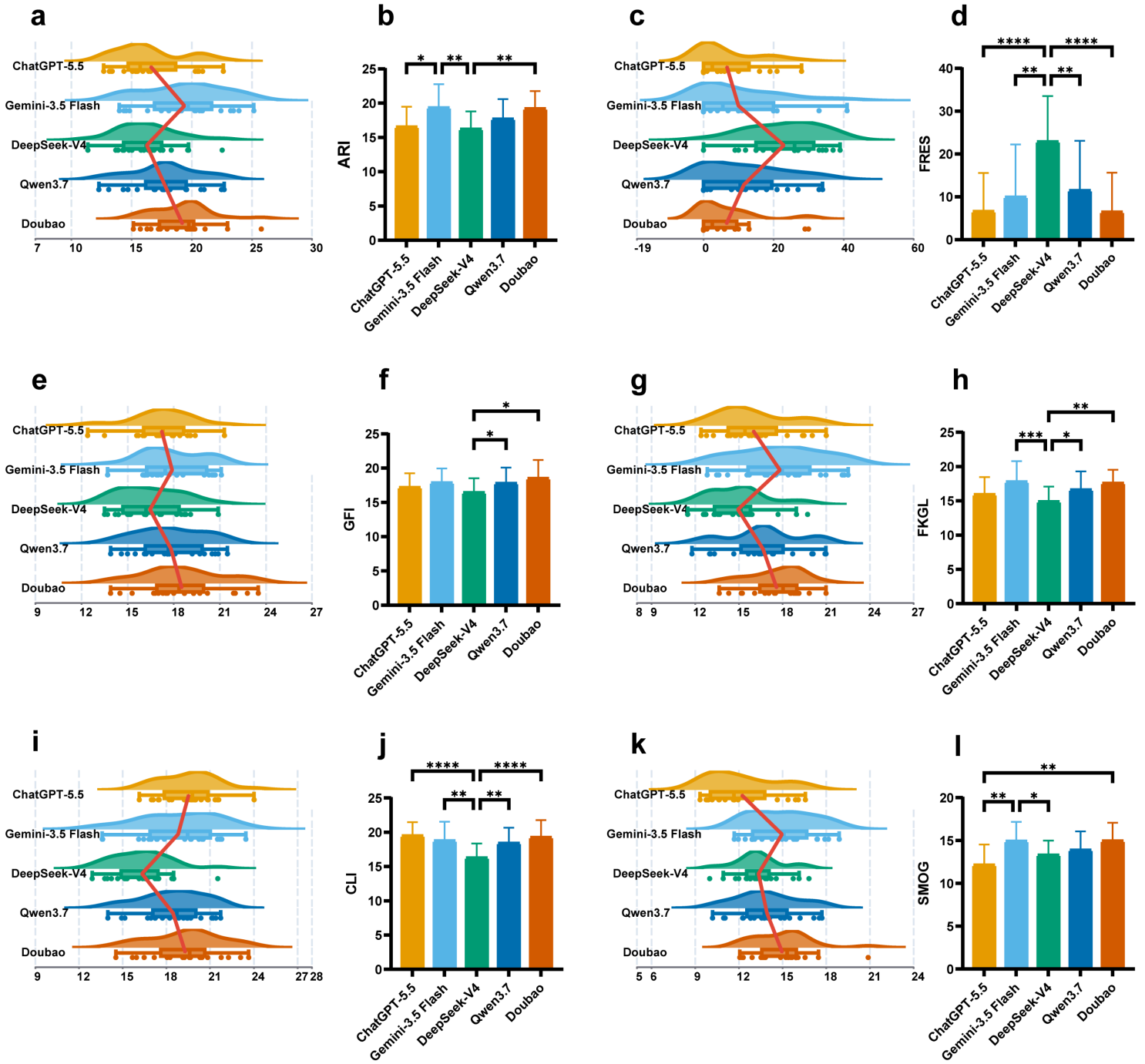
Overview of the study design and evaluation workflow.
Twenty patient education questions across five domains were submitted to five large language models (LLMs). Responses were independently evaluated by two senior oncology experts for quality, reliability, understandability, actionability, and readability.




## Figure 2

Comparison of the quality, reliability, understandability, and actionability of patient education responses generated by five large language models.

Score distributions and pairwise comparisons are shown for GQS (**a, b**), mDISCERN (**c, d**), QAMAI (**e, f**), EQIP (**g, h**), PEMAT-U (**i, j**), and PEMAT-A (**k, l**), respectively. Higher scores indicate better performance. *$P$ < 0.05, **$P$ < 0.01, ***$P$ < 0.001, ****$P$ < 0.0001. GQS, Global Quality Score, mDISCERN, modified DISCERN, QAMAI, Quality Analysis of Medical Artificial Intelligence, EQIP, Ensuring Quality Information for Patients, PEMAT-U, Patient Education Materials Assessment Tool-Understandability, PEMAT-A, Patient Education Materials Assessment Tool-Actionability.




## Figure 3

Comparison of the readability of patient education responses generated by five large language models. Score distributions and pairwise comparisons are shown for ARI (**a, b**), FRES (**c, d**), GFI (**e, f**), FKGL (**g, h**), CLI (**i, j**), and SMOG (**k, l**), respectively. Higher FRES scores indicate greater readability, whereas higher scores on the other indices indicate greater reading difficulty. *$P$< 0.05, **$P$ < 0.01, ***$P$ < 0.001, ****$P$ < 0.0001. ARI, Automated Readability Index, FRES, Flesch Reading Ease Score, GFI, Gunning Fog Index, FKGL, Flesch-Kincaid Grade Level, CLI, Coleman-Liau Index, SMOG, Simple Measure of Gobbledygook.

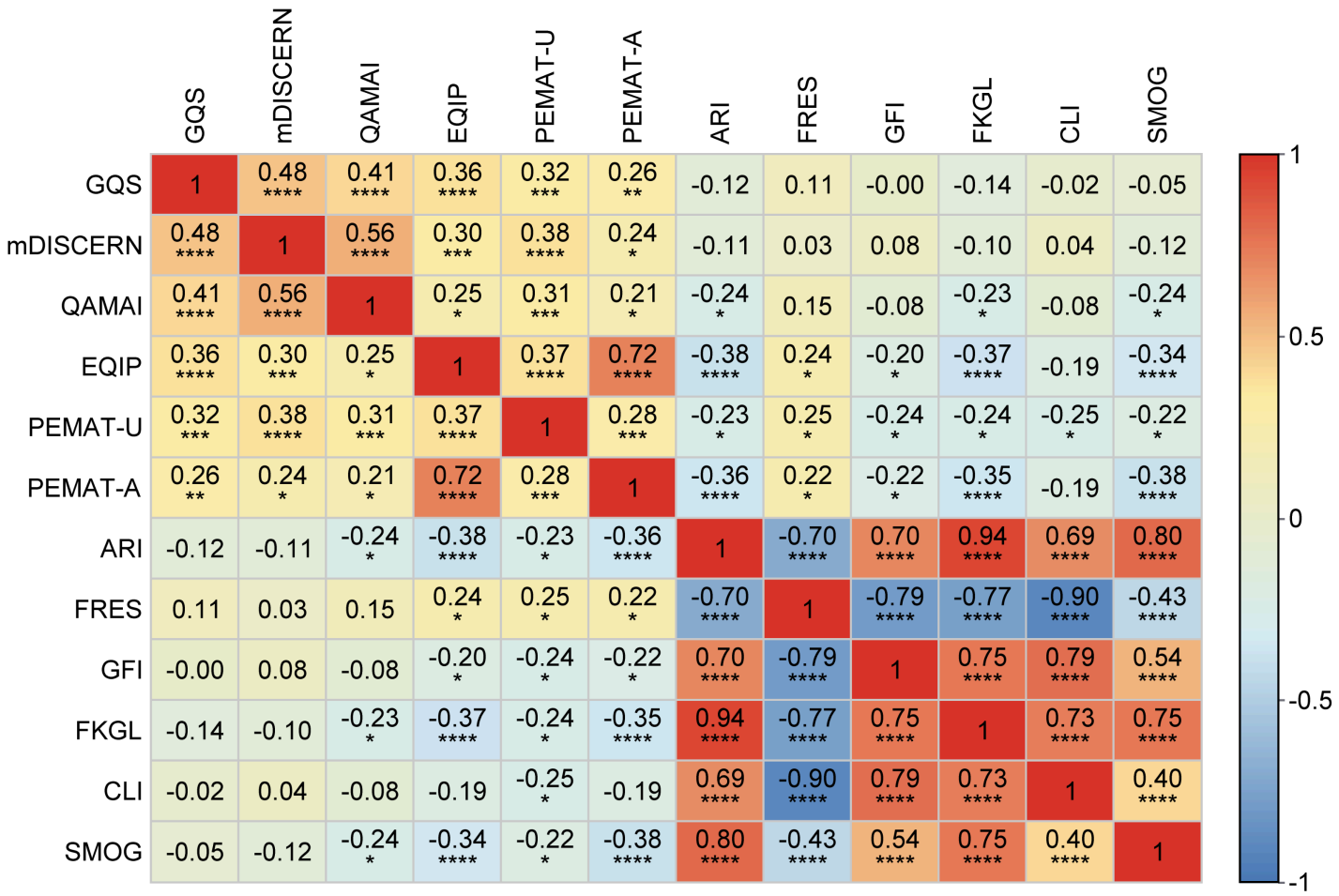



**Figure 4**

Correlations among quality, reliability, understandability, actionability, and readability measures of large language model-generated patient education responses.
The heatmap shows Spearman's rank correlation coefficients ($\rho$) among GQS, mDISCERN, QAMAI, EQIP, PEMAT-U, PEMAT-A, and six readability indices. *$P$ < 0.05, **$P$ < 0.01, ***$P$< 0.001, ****$P$ < 0.0001.

# Supplementary Files

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

- SupplementaryMaterial.rar