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.

Performance of Five Mainstream Large Language Models on Common Small Cell Lung Cancer Questions: Readability, Patient-Education Suitability, and Information Quality

Ge Sun
Second Affiliated Hospital of Dalian Medical University
Jiawen Liu
Second Affiliated Hospital of Dalian Medical University
Xiaowei Liu
Dalian Municipal Central Hospital
Mo Li
limo@dmu.edu.cn

Second Affiliated Hospital of Dalian Medical University

Research Article

Keywords: large language models, artificial intelligence, small cell lung cancer, patient education, readability, C-PEMAT-P, Global Quality Score

Posted Date: September 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10667517/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

Objective

To compare five mainstream large language models (LLMs) in answering common questions about small cell lung cancer (SCLC) across patient-education suitability, overall information quality, and readability.

Methods

This cross-sectional benchmarking study analyzed 100 responses generated by DeepSeek, Doubao, GPT-5, Tongyi Qianwen, and Wenxin Yiyan to 20 standardized SCLC questions spanning five domains. Patient-education suitability and overall information quality were evaluated with the Chinese version of the Patient Education Materials Assessment Tool for Printable Materials (C-PEMAT-P) and the Global Quality Score (GQS). Readability was assessed with seven indices: the Automated Readability Index (ARI), Flesch Reading Ease Score (FRES), Gunning Fog Index (GFOG), Flesch-Kincaid Grade Level (FKGL), Coleman-Liau Index (CL), Simple Measure of Gobbledygook (SMOG), and Linsear Write (LW). Group comparisons and correlation analyses followed the statistical results provided in the study dataset.

Results

C-PEMAT-P and GQS scores differed significantly across models (both $p < 0.001$). GPT-5 had the highest mean C-PEMAT-P score (10.70 ± 1.75) and GQS score (4.70 ± 0.47). All seven readability indices also differed across models (all $p \le 0.006$). Wenxin Yiyan had the highest median FRES [25.50 (20.50, 34.50)] and relatively low GFOG, FKGL, CL, and SMOG values; DeepSeek and Doubao generally produced more linguistically complex text. C-PEMAT-P and GQS did not differ significantly across content domains ($p = 0.744$ and $p = 0.758$, respectively), whereas all seven readability indices varied by topic (all $p \le 0.027$). C-PEMAT-P correlated moderately and positively with GQS ($r = 0.57$, $p < 0.001$), while correlations between quality measures and readability indices were generally weak.

Conclusions

No model performed best across all evaluation dimensions. GPT-5 had the highest patient-education suitability and overall quality scores, whereas Wenxin Yiyan produced comparatively easier-to-read text. Model choice and question topic therefore affect different aspects of SCLC patient education. LLM-generated patient-facing content should be simplified when needed and reviewed by clinical professionals before use.

1. Introduction

Small cell lung cancer (SCLC) accounts for a minority of lung cancers but carries a high clinical burden because it grows rapidly, disseminates early, often responds well to initial treatment, and frequently relapses [1, 2]. Management depends heavily on disease stage and treatment timing; current guidelines integrate TNM staging with the conventional limited-stage/extensive-stage framework when treatment is planned [3–5]. SCLC is also molecularly heterogeneous, and biologically distinct subtypes may have different therapeutic vulnerabilities [6, 7]. Treatment now extends beyond platinum-etoposide chemotherapy and thoracic radiotherapy to immune checkpoint inhibitors, brain-metastasis surveillance and prophylactic cranial irradiation, systemic therapy for relapsed disease, and emerging targets such as delta-like ligand 3 (DLL3) [8–19]. Patients and caregivers must therefore absorb complex information on staging, treatment goals, potential benefits and risks, recurrence surveillance, and prognosis within a short period.

Patient education can help patients understand their disease and treatment options and support clinician-patient communication and shared decision-making. Its value, however, depends on whether patients can understand the information and act on it. Studies of lung-cancer and other patient-education materials have shown that publicly available health information may be poorly organized or too difficult for lay readers [20, 21]. Lower health literacy is also associated with poorer health outcomes [22]. SCLC patient-education materials should therefore be judged not only by whether the information is 'correct' or 'complete', but also by clarity, organization, actionability, and linguistic complexity relative to the intended audience.

Large language models (LLMs) are changing how patients obtain medical information. Through natural-language interaction, these systems can synthesize medical knowledge and generate coherent, structured answers tailored to users' questions [23–25]. Fluent output,

however, does not guarantee medical reliability: previous studies have identified factual errors, omissions, and inaccurate citations in LLM-generated responses [26, 27]. Performance may also vary with model design, version updates, post-training strategies, and deployment conditions [28, 29]. At the same time, results on standardized medical-knowledge tasks suggest applications in medical education [30]. In SCLC, where treatment is intensive and recommendations change quickly, patients who rely on unreviewed LLM output may struggle to judge which information is complete, which recommendations are actionable, and which statements need clinician confirmation.

Studies have begun to evaluate LLM-generated lung-cancer patient education with standardized questions and validated instruments. They report differences among models in content quality, patient-centered presentation, and readability, and higher information quality does not necessarily mean a lower reading burden [31–33]. Direct comparisons in SCLC remain limited, especially studies that test several mainstream LLMs with the same patient questions while assessing patient-education suitability, overall information quality, and multiple readability metrics. We therefore benchmarked responses from DeepSeek, Doubao, GPT-5, Tongyi Qianwen, and Wenxin Yiyan to 20 standardized SCLC questions. We assessed patient-education suitability with the Chinese version of the Patient Education Materials Assessment Tool for Printable Materials (C-PEMAT-P), overall information quality with the Global Quality Score (GQS), and linguistic complexity with seven readability indices. We also compared performance across question topics and examined the relationship between quality measures and readability to define the strengths and limitations of these models for SCLC patient education.

2. Materials and Methods

2.1 Study design and question development

This cross-sectional benchmarking study evaluated LLM-generated responses to common SCLC questions. On July 15, 2026, two clinical experts with 15 and 20 years of clinical experience, respectively, developed 20 frequently asked questions. One patient-education expert also reviewed the question set to improve face validity and patient relevance. The questions covered five domains: high-risk pathogenic factors, clinical symptoms and signs, staging and diagnostic methods, standardized diagnosis and treatment, and follow-up and prognosis management. Each domain contained four questions (Table 1).

Table 1
Standardized SCLC questions used for LLM benchmarking

No.	Domain	Question
1	High-Risk Pathogenic Factors	Is heavy long-term smoking the primary high-risk factor for SCLC?
2	High-Risk Pathogenic Factors	Does exposure to occupational dust and heavy metals increase SCLC risk?
3	High-Risk Pathogenic Factors	Does a family history of lung cancer increase the probability of SCLC?
4	High-Risk Pathogenic Factors	Does chronic obstructive pulmonary disease contribute to SCLC development?
5	Clinical Symptoms and Signs	Is persistent irritative cough the most common initial manifestation of SCLC?
6	Clinical Symptoms and Signs	Does unexplained short-term weight loss indicate rapid tumor proliferation?
7	Clinical Symptoms and Signs	Is refractory hyponatremia a typical paraneoplastic syndrome of SCLC?
8	Clinical Symptoms and Signs	Does superior vena cava syndrome mostly result from enlarged mediastinal lymph nodes?
9	Staging and Diagnostic Methods	Is bronchoscopic biopsy pathology the gold standard for SCLC diagnosis?
10	Staging and Diagnostic Methods	Is contrast-enhanced cranial MRI superior to CT for detecting occult brain metastases?
11	Staging and Diagnostic Methods	Can PET-CT accurately distinguish limited-stage and extensive-stage SCLC?
12	Staging and Diagnostic Methods	Does combined detection of NSE and ProGRP assist in evaluating tumor burden?
13	Standardized Diagnosis and Treatment	Is concurrent chemoradiotherapy with etoposide and platinum the standard regimen for limited-stage SCLC?
14	Standardized Diagnosis and Treatment	Is immune combined platinum-etoposide induction and maintenance recommended for untreated extensive-stage SCLC?
15	Standardized Diagnosis and Treatment	Is prophylactic cranial irradiation recommended for limited-stage SCLC patients with complete remission?
16	Standardized Diagnosis and Treatment	Can topotecan or DLL3-targeted drugs be used for platinum-resistant recurrent SCLC?
17	Follow-up and Prognosis Management	Is standardized re-examination required every 2 months within 2 years after CR of SCLC?
18	Follow-up and Prognosis Management	Can regular standardized imaging re-examination detect recurrent and metastatic lesions early?
19	Follow-up and Prognosis Management	Does continuous smoking cessation intervention significantly prolong the overall survival of SCLC patients?
20	Follow-up and Prognosis Management	Is regular thyroid and myocardial function monitoring required during immunotherapy?

2.2 LLM querying protocol

The 20 questions were submitted to five publicly accessible LLM platforms: Doubao (Doubao-Seed-2.0, ByteDance), DeepSeek (DeepSeek-V4), Wenxin Yiyan (ERNIE 5.1), Tongyi Qianwen (Qwen3.7-Max), and GPT-5 (GPT-5 Instant). Each question was entered in a new, independent conversation to minimize carryover from prior context. A zero-shot prompting strategy was used without role assignment, examples, follow-up questions, or other prompt-engineering instructions to approximate how patients or members of the public might directly seek disease-related information from LLMs. Each model-question combination was generated once, without regeneration or manual selection, yielding 100 original responses (20 questions × 5 models).

2.3 Readability evaluation

Readability was assessed with formulas implemented by ReadabilityFormulas.com, following four comparator studies [34–37]. Seven indices were recorded for each response: ARI, FRES, GFOG, FKGL, CL, SMOG, and LW. ARI, GFOG, FKGL, CL, and SMOG generally rise with linguistic complexity, whereas higher FRES values indicate easier reading. LW was retained as an additional estimate related to reading

grade level. Because the formulas capture different lexical and syntactic features, we interpreted them jointly rather than combining them into a single readability score.

Classical readability formulas measure surface-text complexity rather than comprehension, health literacy, factual accuracy, or clinical usefulness [21, 22]. Their absolute grade-level interpretation may also be less certain outside the language context in which they were developed. As in recent studies of LLM-generated patient education, we therefore used these indices mainly for relative comparisons across models and question domains [34–37].

2.4 Patient-education suitability and information quality assessment

Patient-education suitability and overall quality were evaluated with the Chinese version of the Patient Education Materials Assessment Tool for Printable Materials (C-PEMAT-P) and the Global Quality Score (GQS). The C-PEMAT-P was adapted from the original PEMAT framework and validated for Chinese printable materials [38–40]. It contains 24 items in two dimensions: understandability (16 items, including logical organization and the accessibility of professional terminology) and actionability (8 items, including specific action guidance and appropriateness for the target audience). Each item is scored 0 (criterion not met) or 1 (criterion met), giving a total score of 0–24; higher scores indicate greater accessibility for users. The GQS uses a 1–5 scale: 1 indicates poor quality, disorganized content, missing key information, and no practical value; 2 indicates generally poor quality, insufficient organization, incomplete core information, and limited practical value; 3 indicates moderate quality, with important information partially addressed and basic practical value; 4 indicates good quality, clear organization, coverage of most relevant information, and relatively high practical value; and 5 indicates excellent quality, rigorous organization, coherent presentation, and substantial practical value for users.

Two clinical experts with more than 15 years of relevant clinical experience independently evaluated all responses. Disagreements were resolved by discussion and, when necessary, adjudication by a third expert. The analysis files indicated Cohen's kappa coefficients > 0.75 for both C-PEMAT-P and GQS, meeting the prespecified criterion for excellent inter-rater agreement.

2.5 Statistical analysis

All statistical analyses were performed using IBM SPSS Statistics 25.0 (IBM Corp., Armonk, NY, USA). Continuous outcomes were summarized as mean ± standard deviation when appropriate or as median with interquartile range when non-normally distributed. Normality was assessed using the Shapiro-Wilk test, and homogeneity of variance for ANOVA using Levene's test. Outcomes meeting parametric assumptions were compared by one-way analysis of variance (ANOVA), followed by Tukey's honestly significant difference (HSD) test for post-hoc pairwise comparisons. Outcomes violating normality and/or homoscedasticity assumptions were compared by the Kruskal-Wallis H test, followed by Dunn's post-hoc pairwise comparisons. Multiple comparisons were adjusted using the Bonferroni correction. Correlations between quality metrics (GQS and C-PEMAT-P) and readability indices were assessed using Pearson's correlation coefficient (r). All tests were two-tailed, and $p < 0.05$ was considered statistically significant.

2.6 Ethical considerations

All data analyzed in this study were generated by large language models. No human or animal experiments were performed, and the dataset contained no personal information or biological samples from real individuals. Under the applicable academic standards, ethical review was not required for this type of study.

3. Results

3.1 Overall model comparison

C-PEMAT-P and GQS scores differed significantly across the five models ($F = 48.53$, $p < 0.001$ and $F = 52.19$, $p < 0.001$, respectively). The final dataset comprised 100 responses, with 20 generated by each LLM. GPT-5 had the highest mean C-PEMAT-P score (10.70 ± 1.75) and GQS score (4.70 ± 0.47). Tongyi Qianwen ranked second for C-PEMAT-P (8.15 ± 0.37), while Doubao ranked second for GQS (3.55 ± 0.60). Wenxin Yiyan had the lowest GQS score (2.35 ± 0.67).

All seven readability indices differed significantly across models: ARI ($\chi^2=14.99$, $p = 0.005$), FRES ($\chi^2=19.81$, $p < 0.001$), GFOG ($\chi^2=28.97$, $p < 0.001$), FKGL ($\chi^2=14.45$, $p = 0.006$), CL ($\chi^2=20.57$, $p < 0.001$), SMOG ($\chi^2=16.37$, $p = 0.003$), and LW ($\chi^2=16.25$, $p = 0.003$). Inter-platform differences therefore extended from expert-perceived quality to the lexical and syntactic features captured by readability formulas.

Table 2. Performance of the five LLMs across quality and readability measures

Variable	Total	DeepSeek	Doubao	GPT-5	Tongyi Qianwen	Wenxin Yiyan	Statistic	p
C-PEMAT-P, mean ± SD	7.96 ± 1.80	6.70 ± 1.08	7.45 ± 0.83	10.70 ± 1.75	8.15 ± 0.37	6.80 ± 0.70	F=48.53	<0.001
GQS, mean ± SD	3.28 ± 1.04	3.35 ± 0.59	3.55 ± 0.60	4.70 ± 0.47	2.45 ± 0.60	2.35 ± 0.67	F=52.19	<0.001
ARI, M (Q1, Q3)	19.24 (17.46, 20.35)	20.01 (18.52, 21.10)	20.23 (19.89, 21.27)	18.45 (17.34, 19.85)	18.82 (17.10, 19.77)	18.87 (17.35, 19.37)	χ^2=14.99	0.005
FRES, M (Q1, Q3)	14.00 (2.00, 23.25)	5.50 (0.75, 17.50)	5.50 (0.00, 19.00)	14.00 (4.25, 18.75)	12.00 (0.75, 24.25)	25.50 (20.50, 34.50)	χ^2=19.81	<0.001
GFOG, M (Q1, Q3)	17.20 (15.78, 18.50)	17.80 (16.95, 18.80)	18.80 (17.70, 19.62)	16.80 (15.30, 17.52)	18.20 (15.28, 19.50)	15.65 (14.07, 16.42)	χ^2=28.97	<0.001
FKGL, M (Q1, Q3)	16.89 (15.50, 18.50)	17.58 (16.13, 18.84)	18.43 (16.77, 20.17)	16.25 (15.29, 17.93)	16.98 (14.89, 18.03)	15.82 (15.36, 16.94)	χ^2=14.45	0.006
CL, M (Q1, Q3)	18.66 (16.66, 20.36)	19.59 (18.01, 20.48)	19.50 (17.59, 21.31)	18.93 (17.26, 20.41)	18.81 (16.45, 20.15)	15.96 (14.48, 17.59)	χ^2=20.57	<0.001
SMOG, M (Q1, Q3)	14.68 (13.81, 15.75)	15.01 (14.06, 15.70)	16.14 (15.09, 16.94)	14.28 (13.65, 14.92)	15.23 (13.61, 15.70)	14.37 (13.67, 14.75)	χ^2=16.37	0.003
LW, M (Q1, Q3)	51.00 (47.00, 54.00)	50.00 (44.75, 52.25)	48.50 (45.25, 50.00)	53.00 (49.00, 56.25)	51.50 (49.75, 54.00)	53.00 (50.00, 56.00)	χ^2=16.25	0.003

Abbreviations: SD, standard deviation; M, median; Q1, first quartile; Q3, third quartile. Higher ARI, GFOG, FKGL, CL, and SMOG values generally indicate greater reading difficulty; higher FRES values indicate easier reading.

3.2 Patient-education suitability and overall quality

C-PEMAT-P scores shifted toward higher values for GPT-5 (Figure 1). Its median score was 10.50 (Q1-Q3, 9.75-12.00), with a range of 8-13. Tongyi Qianwen had a median score of 8.00 and a relatively narrow distribution, suggesting more stable scores. Scores for Doubao, DeepSeek, and Wenxin Yiyan were concentrated mainly between 6 and 8. These distributions were consistent with the mean comparisons in Table 2.

GQS separated the models more clearly (Figure 2). GPT-5 responses clustered at scores of 4-5, with a median of 5.00 (Q1-Q3, 4.00-5.00), and no response scored below 4. Doubao and DeepSeek responses clustered mainly at scores of 3-4. Tongyi Qianwen and Wenxin Yiyan both had median GQS values of 2.00, and Wenxin Yiyan reached a minimum score of 1. Apart from GPT-5, higher patient-education suitability did not map directly onto the same ranking in expert-perceived overall quality.

3.3 Readability by model

Model rankings differed across readability metrics. DeepSeek and Doubao had relatively high values on several difficulty-oriented indices. For Doubao, median ARI, GFOG, FKGL, CL, and SMOG values were 20.23, 18.80, 18.43, 19.50, and 16.14, respectively, while the median FRES was 5.50. DeepSeek showed a similar pattern, with a median ARI of 20.01 and median FRES of 5.50. GPT-5 had lower median ARI (18.45), GFOG (16.80), FKGL (16.25), and SMOG (14.28) values than DeepSeek or Doubao, but was not the easiest-to-read model across all indices.

Wenxin Yiyan had the highest median FRES [25.50 (20.50, 34.50)] and the lowest median GFOG [15.65 (14.07, 16.42)], FKGL [15.82 (15.36, 16.94)], and CL [15.96 (14.48, 17.59)] values, with a relatively low SMOG [14.37 (13.67, 14.75)]. This lower linguistic complexity coexisted with its lowest GQS score. No model performed best simultaneously on patient-education suitability, overall quality, and every readability index.

3.4 Differences across question topics

C-PEMAT-P and GQS scores did not differ significantly across topics (F=0.49, p=0.744 and F=0.47, p=0.758, respectively), whereas all seven readability indices varied across content domains: ARI (χ^2=24.24, p<0.001), FRES (χ^2=11.00, p=0.027), GFOG (χ^2=18.56, p<0.001), FKGL (χ^2=14.21, p=0.007), CL (χ^2=16.16, p=0.003), SMOG (χ^2=24.29, p<0.001), and LW (χ^2=23.42, p<0.001). The 100 responses were evenly distributed across five question domains, with 20 responses per domain.

The clinical symptoms and signs domain was generally easier to read, with the highest FRES [22.00 (15.25, 32.00)] and relatively low median GFOG [15.40 (14.02, 16.85)] and SMOG [13.55 (12.87, 13.97)]. Standardized diagnosis and treatment had the highest median ARI [20.14 (19.86, 21.41)] and CL [19.80 (18.84, 20.63)], whereas follow-up and prognosis management had the highest median FKGL [18.41 (15.60, 19.27)] and SMOG [15.62 (14.39, 16.43)] and the lowest FRES [3.50 (0.00, 20.00)]. Linguistic complexity may therefore be more sensitive to question topic than the two quality measures.

Table 3. Quality and readability across SCLC question domains

Variable	Total	Symptoms/signs	Follow-up/prognosis	Risk factors	Staging/diagnosis	Diagnosis/treatment	Statistic	p
C-PEMAT-P, mean ± SD	7.96 ± 1.80	8.45 ± 2.48	7.95 ± 0.60	7.80 ± 1.11	7.75 ± 2.34	7.85 ± 1.81	F=0.49	0.744
GQS, mean ± SD	3.28 ± 1.04	3.50 ± 1.00	3.30 ± 0.86	3.10 ± 1.07	3.15 ± 1.14	3.35 ± 1.14	F=0.47	0.758
ARI, M (Q1, Q3)	19.24 (17.46, 20.35)	17.37 (17.00, 19.00)	19.98 (18.66, 21.42)	18.82 (17.02, 20.13)	18.65 (17.80, 19.66)	20.14 (19.86, 21.41)	χ^2=24.24	<0.001
FRES, M (Q1, Q3)	14.00 (2.00, 23.25)	22.00 (15.25, 32.00)	3.50 (0.00, 20.00)	13.50 (8.25, 26.50)	17.00 (4.00, 23.00)	8.50 (1.50, 16.50)	χ^2=11.00	0.027
GFOG, M (Q1, Q3)	17.20 (15.78, 18.50)	15.40 (14.02, 16.85)	18.05 (16.30, 19.52)	17.70 (16.35, 18.72)	17.35 (15.00, 18.40)	17.70 (16.93, 18.42)	χ^2=18.56	<0.001
FKGL, M (Q1, Q3)	16.89 (15.50, 18.50)	15.50 (15.01, 17.38)	18.41 (15.60, 19.27)	16.41 (15.36, 17.57)	16.70 (15.70, 18.02)	17.77 (16.86, 18.49)	χ^2=14.21	0.007
CL, M (Q1, Q3)	18.66 (16.66, 20.36)	16.87 (15.24, 19.02)	19.37 (17.39, 21.42)	17.66 (16.64, 20.15)	17.80 (16.52, 19.43)	19.80 (18.84, 20.63)	χ^2=16.16	0.003
SMOG, M (Q1, Q3)	14.68 (13.81, 15.75)	13.55 (12.87, 13.97)	15.62 (14.39, 16.43)	14.45 (13.86, 15.45)	14.54 (13.64, 15.28)	15.23 (15.04, 15.90)	χ^2=24.29	<0.001
LW, M (Q1, Q3)	51.00 (47.00, 54.00)	52.00 (48.75, 56.00)	45.00 (42.00, 50.25)	52.00 (50.75, 55.00)	53.00 (50.00, 56.00)	49.00 (46.00, 51.25)	χ^2=23.42	<0.001

3.5 Correlation analysis

C-PEMAT-P and GQS were moderately positively correlated (r=0.57, p<0.001): responses rated as more suitable for patient education also tended to receive higher overall quality ratings. C-PEMAT-P was not significantly correlated with any of the seven readability indices in the original dataset. Correlations between GQS and readability were also weak. GQS was negatively correlated with FRES (r=-0.21, p=0.041) and positively correlated with CL (r=0.27, p=0.008); correlations with ARI, GFOG, FKGL, SMOG, and LW were not statistically significant.

The readability indices were strongly intercorrelated. ARI correlated positively with FKGL (r=0.93), CL (r=0.79), and GFOG (r=0.72), whereas FRES correlated negatively with CL (r=-0.91), FKGL (r=-0.88), and GFOG (r=-0.80); all of these correlations had p<0.001. LW correlated positively with FRES (r=0.64) and negatively with ARI (r=-0.68), FKGL (r=-0.70), and CL (r=-0.66), again with p<0.001. These patterns suggest that readability formulas capture a shared dimension of surface-text complexity, while patient-education suitability and overall quality are related but distinct attributes.

4. Discussion

4.1 Principal findings

Model choice affected patient-education quality, while readability varied by both model and question topic. GPT-5 had the highest C-PEMAT-P and GQS scores, but the model with the highest quality scores was not the easiest to read across all indices. The relationship between quality and readability was weak: C-PEMAT-P and GQS were moderately correlated, whereas neither quality measure showed a consistent association with the readability indices. Patient-education suitability, overall quality, and linguistic complexity should therefore be treated as distinct, complementary dimensions.

4.2 Model-dependent differences in patient-education quality

GPT-5 had the highest C-PEMAT-P and GQS scores, suggesting that its responses were more often judged clearly organized, suitable for patient education, and of greater overall informational value. Lung-cancer LLM studies have likewise reported differences among models in readability, information quality, and patient-centered presentation [31-33]. The comparator studies show a similar pattern. In skin-cancer patient education, GPT-5 had the highest GQS despite relatively small between-model differences in C-PEMAT-P [34]; in pertussis education, GPT-5 also ranked highest on quality-related measures [35]; and in autoimmune hepatitis, ChatGPT performed best on GQS and C-PEMAT among the systems tested [37]. Model-dependent differences in structure and patient-oriented presentation may therefore persist across disease areas.

Our data do not identify the mechanisms behind these differences. Medical LLM studies and the comparator studies suggest that variation may reflect training data, post-training alignment, safety constraints, model architecture, or platform-specific response strategies [23,24,28,29,35,37]. Because we did not examine internal model mechanisms or control hidden generation settings, these explanations remain hypotheses. In practice, the products should not be treated as interchangeable; clinicians should evaluate the specific system and configuration used to generate patient-facing material.

4.3 Topic-related differences in SCLC readability

Readability varied across SCLC content domains even though C-PEMAT-P and GQS were relatively stable across topics. Clinical symptoms and signs were generally easier to read, whereas standardized diagnosis and treatment and follow-up/prognosis management were more complex across several indices. This pattern is clinically plausible because treatment and monitoring questions require terms related to chemoradiotherapy, systemic therapy, imaging, toxicity monitoring, recurrence, and prognosis [3,4,8-19]. Similar topic effects have been reported elsewhere: diagnosis- and treatment-related pertussis content was harder to read than basic disease or symptom information [35], and autoimmune-hepatitis content on disease mechanisms and diagnostic evaluation required higher reading levels [37].

A single generic instruction to 'explain the disease' may not produce equally accessible material throughout the patient care pathway. Symptom questions can often be answered in more familiar language, whereas treatment and prognosis information may need deliberate simplification, shorter information units, explicit definitions of technical terms, and clearer prioritization of what patients most need to know. Topic-adapted prompting may therefore be more appropriate than assuming a fixed readability profile for any model.

4.4 Information quality and readability are not interchangeable

C-PEMAT-P and GQS were moderately correlated, but both had only weak associations with readability. Zhang et al. reported a similar separation between quality and readability in skin-cancer patient education [34]. Wang et al. also found that informational completeness, actionability, and readability did not change in parallel in chronic heart-failure patient education [36]. More broadly, easy-to-read language alone does not guarantee adequate understanding, good actionability, or better health outcomes [21,22]. The present SCLC data show the same distinction: Wenxin Yiyan produced relatively easier-to-read text across several indices but had the lowest mean GQS, whereas GPT-5 had the highest quality score without being the easiest to read on every metric.

Readability formulas mainly reflect word length, sentence length, syllable burden, and related surface-text features [21]. C-PEMAT-P also evaluates information organization and actionable presentation [38,39], whereas GQS reflects experts' overall judgment of information quality [40]. A shorter, simpler answer may therefore be easy to read but incomplete; a more detailed answer may receive a higher quality score while increasing reading burden. These dimensions need separate evaluation because improving one metric does not ensure improvement in the others.

4.5 Implications for SCLC patient education

Our findings support a supervised rather than autonomous role for LLMs in SCLC patient education. When structured educational material with relatively high perceived quality is the priority, GPT-5 may provide a useful starting point; outputs from other systems may require

more editing to improve completeness or practical usability. Regardless of platform, treatment-, staging-, and follow-up-related content should be checked for unnecessary jargon and complex sentence structures. Prompt templates may request plain language, short paragraphs, definitions of technical terminology, and clear action-oriented recommendations. This supervised approach is consistent with medical LLM research, lung-cancer-specific evaluations, and comparator work in chronic heart failure [23-33,36].

This study did not evaluate factual accuracy or guideline concordance, so higher C-PEMAT-P or GQS scores should not be interpreted as evidence of clinical correctness. Studies in pertussis and autoimmune hepatitis included guideline-based concordance or clinical-intent assessments, suggesting that future SCLC research could add this dimension to the present framework [35,37]. Because SCLC is a high-risk malignancy and treatment and surveillance recommendations change over time, future studies should compare LLM responses directly with current SCLC guidelines and expert-verified key clinical points [3,4].

4.6 Strengths and limitations

This study used identical questions across five mainstream LLMs, covered five clinically relevant SCLC domains, evaluated seven readability indices, and included independent expert assessment with C-PEMAT-P and GQS. The correlation analysis also kept readability and patient-education quality as separate constructs, consistent with recent LLM benchmarking studies and validated patient-education assessment frameworks [31-40].

Each model answered each question only once, so stochastic variation across repeated generations could not be quantified. The fixed set of 20 questions cannot capture all patient concerns, individualized clinical contexts, or follow-up questions in multi-turn conversations. Model versions, hidden routing, web-search behavior, and default platform settings may change over time, limiting reproducibility [23,28,29]. C-PEMAT-P and GQS primarily assess presentation characteristics and expert-perceived quality rather than factual accuracy, guideline concordance, citation validity, or clinical safety [27,38-40]. Readability formulas estimate surface-text complexity rather than actual patient comprehension, and their absolute interpretation may depend on language context [21,22]. The analyses provided in the study workbook treated the models as conventional independent groups despite reuse of the same question set across models; adding a repeated-measures sensitivity analysis before formal submission would strengthen statistical inference. The study also did not test whether optimized prompts, clinician editing, or patient co-design could improve information quality and readability simultaneously.

5. Conclusion

The five mainstream LLMs differed in patient-education suitability, overall quality, and readability when answering common SCLC questions. GPT-5 had the highest C-PEMAT-P and GQS scores, whereas Wenxin Yiyan produced comparatively easier-to-read text across several readability indices but had lower overall quality scores. Readability also varied by clinical topic: diagnosis/treatment and follow-up/prognosis content was generally more complex than symptom-related information. The weak correlations between quality scores and readability indicate that these dimensions should be evaluated separately. LLMs may support SCLC patient education, but their outputs should be adapted to the topic, simplified for the intended audience, and reviewed by clinical professionals before patient-facing use.

Abbreviations

LLMs
large language models
SCLC
small cell lung cancer
C- PEMAT-P
Patient Education Materials Assessment Tool for Printable Materials
GQS
Global Quality Score
ARI
Automated Readability Index
FRES
Flesch Reading Ease Score
GFOG
Gunning Fog Index
FKGL
Flesch-Kincaid Grade Level

CL
Coleman-Liau Index
SMOG
Simple Measure of Gobbledygook
LW
Linsear Write

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets for this study are available from the corresponding author (email: limo@dmu.edu.cn) upon reasonable request.

Competing interests

The authors declare no competing interests.

Funding

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

Authors' contributions

Xiaowei Liu: conceptualization, methodology, investigation, project administration, and writing – original draft. Mo Li.: investigation, data curation, formal analysis, and writing – original draft. Ge Sun: investigation, resources, and validation.Jiawen Liu.: investigation and data curation.

Acknowledgements

We would like to acknowledge the hard and dedicated work of all the staff who implemented the intervention and evaluation components of the study.

References

1. Megyesfalvi Z, Gay CM, Popper H, et al. Clinical insights into small cell lung cancer: tumor heterogeneity, diagnosis, therapy, and future directions. CA Cancer J Clin. 2023;73(6):620–52. 10.3322/caac.21785.
2. Rudin CM, Brambilla E, Faivre-Finn C, Sage J. Small-cell lung cancer. Nat Rev Dis Primers. 2021;7(1):3. 10.1038/s41572-020-00235-0.
3. Ganti AKP, Loo BW Jr, Bassetti M, et al. Small Cell Lung Cancer, Version 2.2022, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw. 2021;19(12):1441–64. 10.6004/jnccn.2021.0058.
4. Dingemans AMC, Früh M, Ardizzoni A, et al. Small-cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2021;32(7):839–53. 10.1016/j.annonc.2021.03.207.
5. Nicholson AG, Chansky K, Crowley J, et al. The International Association for the Study of Lung Cancer Lung Cancer Staging Project: proposals for the revision of the clinical and pathologic staging of small cell lung cancer in the forthcoming eighth edition of the TNM classification for lung cancer. J Thorac Oncol. 2016;11(3):300–11. 10.1016/j.jtho.2015.10.008.
6. Rudin CM, Poirier JT, Byers LA, et al. Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. Nat Rev Cancer. 2019;19(5):289–97. 10.1038/s41568-019-0133-9.
7. Gay CM, Stewart CA, Park EM, et al. Patterns of transcription factor programs and immune pathway activation define four major subtypes of SCLC with distinct therapeutic vulnerabilities. Cancer Cell. 2021;39(3):346–e3607. 10.1016/j.ccell.2020.12.014.

8. Rossi A, Di Maio M, Chiodini P, et al. Carboplatin- or cisplatin-based chemotherapy in first-line treatment of small-cell lung cancer: the COCIS meta-analysis of individual patient data. J Clin Oncol. 2012;30(14):1692–8. 10.1200/JCO.2011.40.4905.
9. Horn L, Mansfield AS, Szczęsna A, et al. First-line atezolizumab plus chemotherapy in extensive-stage small-cell lung cancer. N Engl J Med. 2018;379(23):2220–9. 10.1056/NEJMoa1809064.
10. Paz-Ares L, Dvorkin M, Chen Y, et al. Durvalumab plus platinum-etoposide versus platinum-etoposide in first-line treatment of extensive-stage small-cell lung cancer (CASPIAN): a randomised, controlled, open-label, phase 3 trial. Lancet. 2019;394(10212):1929–39. 10.1016/S0140-6736(19)32222-6.
11. Cheng Y, Han L, Wu L, et al. Effect of first-line serplulimab vs placebo added to chemotherapy on survival in patients with extensive-stage small cell lung cancer: the ASTRUM-005 randomized clinical trial. JAMA. 2022;328(12):1223–32. 10.1001/jama.2022.16464.
12. Wang J, Zhou C, Yao W, et al. Adebrelimab or placebo plus carboplatin and etoposide as first-line treatment for extensive-stage small-cell lung cancer (CAPSTONE-1): a multicentre, randomised, double-blind, placebo-controlled, phase 3 trial. Lancet Oncol. 2022;23(6):739–47. 10.1016/S1470-2045(22)00224-8.
13. Turrisi AT 3rd, Kim K, Blum R, et al. Twice-daily compared with once-daily thoracic radiotherapy in limited small-cell lung cancer treated concurrently with cisplatin and etoposide. N Engl J Med. 1999;340(4):265–71. 10.1056/NEJM199901283400403.
14. Faivre-Finn C, Snee M, Ashcroft L, et al. Concurrent once-daily versus twice-daily chemoradiotherapy in patients with limited-stage small-cell lung cancer (CONVERT): an open-label, phase 3, randomised, superiority trial. Lancet Oncol. 2017;18(8):1116–25. 10.1016/S1470-2045(17)30318-2.
15. Cheng Y, Spigel DR, Cho BC, et al. Durvalumab after chemoradiotherapy in limited-stage small-cell lung cancer. N Engl J Med. 2024;391(14):1313–27. 10.1056/NEJMoa2404873.
16. Slotman B, Faivre-Finn C, Kramer G, et al. Prophylactic cranial irradiation in extensive small-cell lung cancer. N Engl J Med. 2007;357(7):664–72. 10.1056/NEJMoa071780.
17. Takahashi T, Yamanaka T, Seto T, et al. Prophylactic cranial irradiation versus observation in patients with extensive-disease small-cell lung cancer: a multicentre, randomised, open-label, phase 3 trial. Lancet Oncol. 2017;18(5):663–71. 10.1016/S1470-2045(17)30230-9.
18. Trigo J, Subbiah V, Besse B, et al. Lurbinectedin as second-line treatment for patients with small-cell lung cancer: a single-arm, open-label, phase 2 basket trial. Lancet Oncol. 2020;21(5):645–54. 10.1016/S1470-2045(20)30068-1.
19. Ahn MJ, Cho BC, Felip E, et al. Tarlatamab for patients with previously treated small-cell lung cancer. N Engl J Med. 2023;389(22):2063–75. 10.1056/NEJMoa2307980.
20. Sarna L, Ganley BJ. A survey of lung cancer patient-education materials. Oncol Nurs Forum. 1995;22(10):1545–50. PMID:8577622.
21. Rooney MK, Santiago G, Perni S, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. J Patient Exp. 2021;8:2374373521998847. 10.1177/2374373521998847.
22. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med. 2011;155(2):97–107. 10.7326/0003-4819-155-2-201107190-00005.
23. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–40. 10.1038/s41591-023-02448-8.
24. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80. 10.1038/s41586-023-06291-2.
25. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–96. 10.1001/jamainternmed.2023.1838.
26. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. 2023;6(10):e2336483. 10.1001/jamanetworkopen.2023.36483.
27. Chen A, Chen DO. Accuracy of chatbots in citing journal articles. JAMA Netw Open. 2023;6(8):e2327647. 10.1001/jamanetworkopen.2023.27647.
28. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA. 2023;330(9):866–9. 10.1001/jama.2023.14217.
29. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthc (Basel). 2023;11(6):887. 10.3390/healthcare11060887.
30. 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. 10.1371/journal.pdig.0000198.

31. Richlitzki C, Mansoorian S, Käsmann L, et al. Assessing ChatGPT's educational potential in lung cancer radiotherapy from clinician and patient perspectives: content quality and readability analysis. JMIR Cancer. 2025;11:e69783. 10.2196/69783.
32. Song C, Zhao CY, Yang T, et al. Transformation of lung cancer patient education paradigm driven by large language models: a multidimensional performance study. Int J Surg. 2026;112(1):1881–5. 10.1097/JS9.0000000000003503.
33. Önal Ö, Temiz Bekce S. A cross-sectional study of large language models in lung cancer information delivery: readability, quality, and patient-centred evaluation. Healthc (Basel). 2026;14(12):1769. 10.3390/healthcare14121769.
34. Zhang Y, Wang L, Zhang W, Lan W. Decoupled quality and readability in skin cancer education from large language models. Front Public Health. 2026;14:1777577. 10.3389/fpubh.2026.1777577.
35. Zhong XF, Liu R. Benchmarking the readability, quality, and educational suitability of large language models in communicating pertussis. Front Public Health. 2026;14:1799204. 10.3389/fpubh.2026.1799204.
36. Wang Z, Li X, Ma C, Zhang Z. Challenges of using generative AI for patient education in chronic heart failure: an evaluation of content quality, readability, and actionability in cross-platform LLM-generated texts. Front Public Health. 2026;14:1801829. 10.3389/fpubh.2026.1801829.
37. Li H, Lin Y, Lv L. Performance evaluation of mainstream large language models in autoimmune hepatitis patient education: a comparative study of readability, quality, and reliability. Front Public Health. 2026;14:1805848. 10.3389/fpubh.2026.1805848.
38. Shoemaker SJ, Wolf MS, Brach C. 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. 2014;96(3):395–403. 10.1016/j.pec.2014.05.027.
39. Shan Y, Ji M, Dong Z, Xing Z, Wang D, Cao X. The Chinese Version of the Patient Education Materials Assessment Tool for Printable Materials: translation, adaptation, and validation study. J Med Internet Res. 2023;25:e39808. 10.2196/39808.
40. Bernard A, Langille M, Linton S, et al. Identification and assessment of online patient education resources for maternal-fetal medicine. J Obstet Gynaecol Can. 2007;29(12):969–73.

Figures

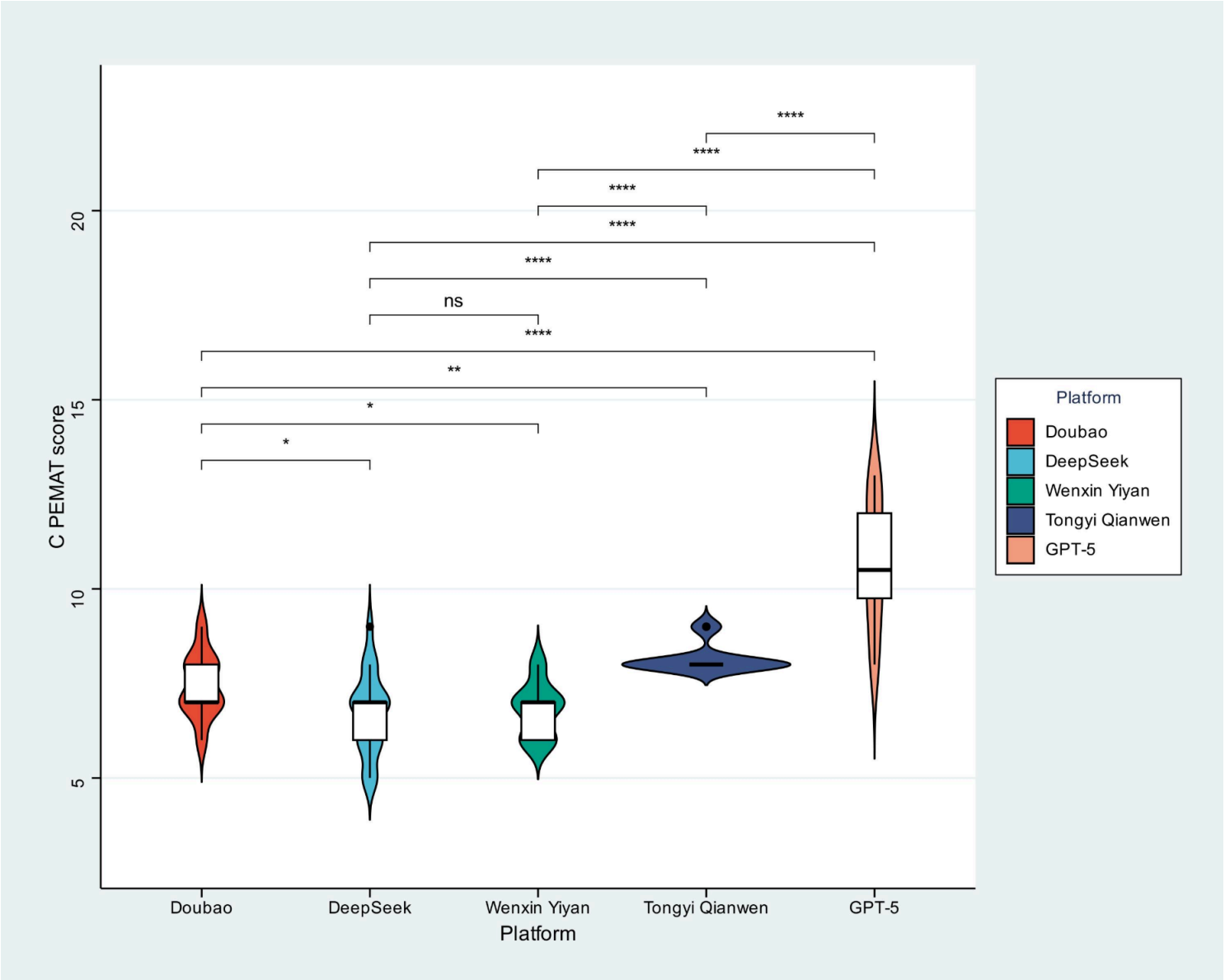


Figure 1

Distribution of C-PEMAT-P scores across the five LLM platforms. Significance annotations are reproduced from the original analysis figure.

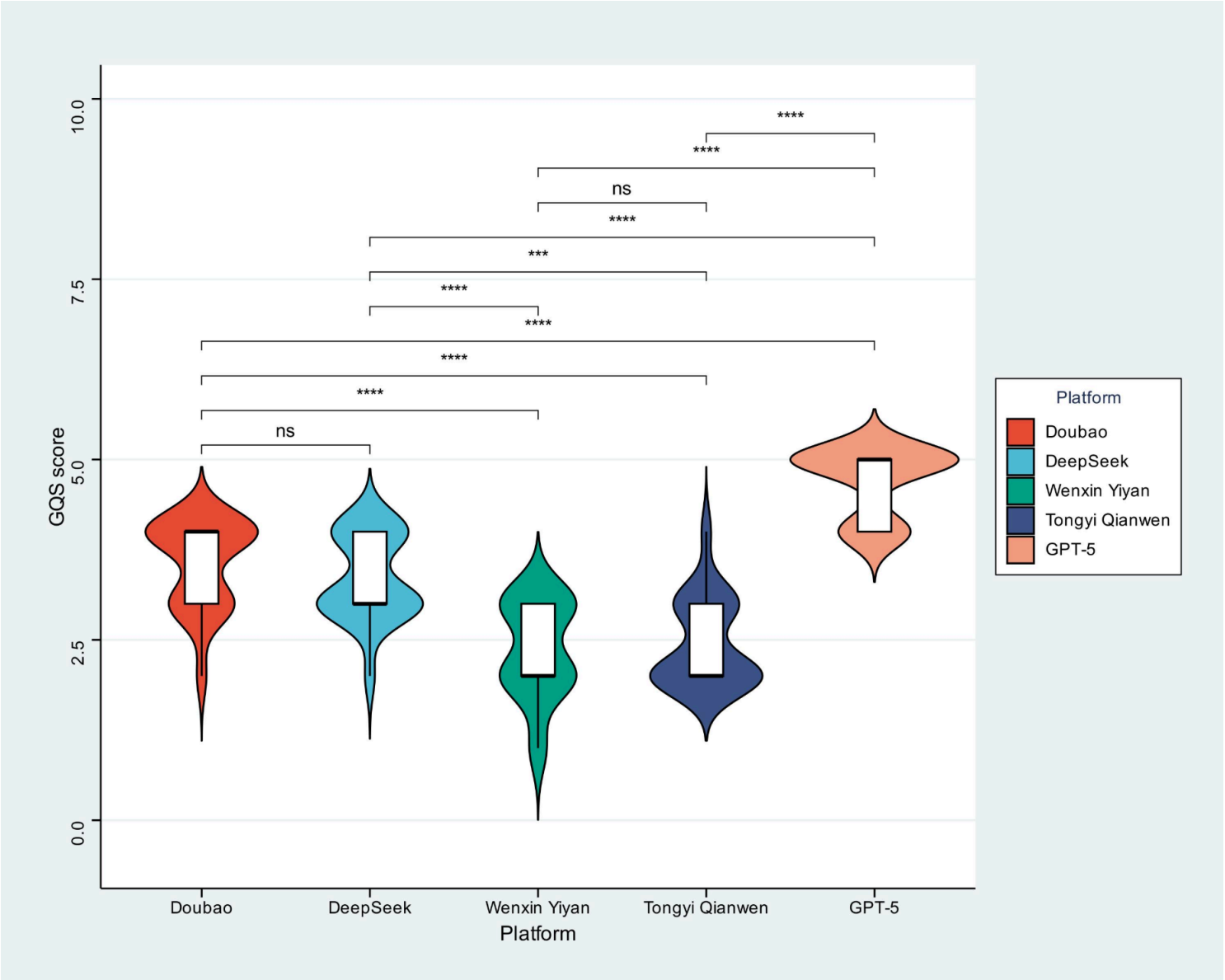


Figure 2

Distribution of Global Quality Score (GQS) values across the five LLM platforms. Significance annotations are reproduced from the original analysis figure.

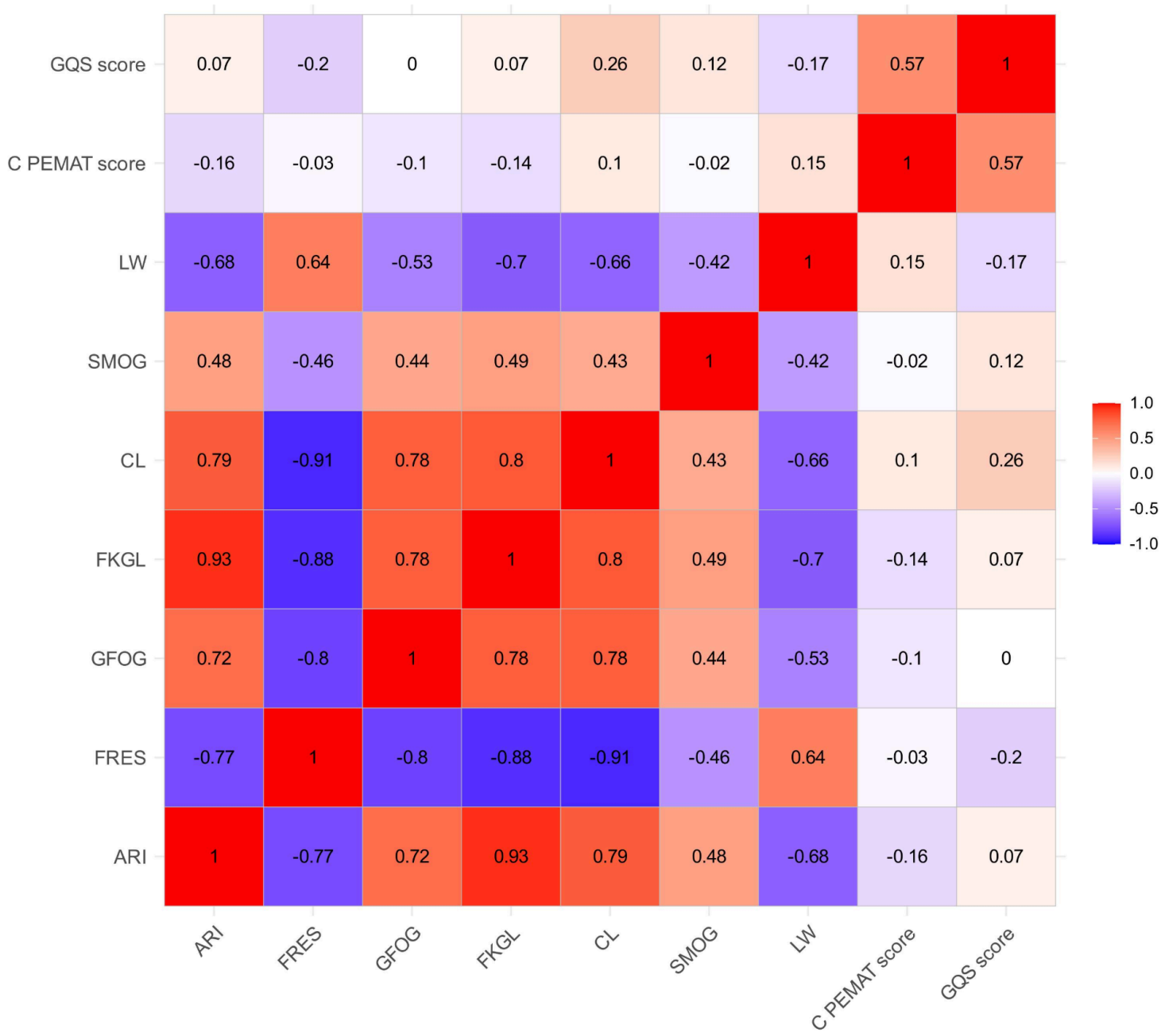


Figure 3

Pearson correlation heatmap of readability indices, C-PEMAT-P, and GQS across the 100 LLM-generated responses.