scientific reports

Check for updates

OPEN

A multidimensional benchmarking framework for large language models in oncologic decision making

Mehmet Halici[1,9]✉, Serkan Salturk[2,9]✉, Irem Sayin[3], Burak Ertan[4], Ibrahim Cem Balci[3], Kimia Cepni[1], Tanju Kapagan[5], Cumhur Yildirim[6], Gokmen Umut Erdem[7], Huriye Senay Kiziltan[1], Muhammed Tayyip Kocak[8] & Huseyin Uvet[3]

Large language models (LLMs) are increasingly explored as clinical decision support tools in oncology; however, reliance on isolated metrics has limited the development of multi-dimensional evaluation frameworks. This comparative observational study utilized five stepwise, clinically realistic non-small cell lung cancer scenarios reflecting real-world diagnostic, therapeutic, and follow-up decision-making. Open-ended clinical questions were answered by three LLMs (Gemini 2.5 Pro, GPT-5, and Claude Opus 4.1) via their official APIs and compared with evidence-based reference answers. Model outputs were evaluated using expert-rated clinical accuracy and explainability, alongside operational metrics including cost, response time, and generative efficiency. All dimensions were integrated into an expert-weighted Composite Performance Score (CPS). Across 30 clinical questions, significant inter-model differences were observed for all metrics ($p < 0.001$). GPT-5 achieved the highest accuracy, explainability, and generative efficiency, while Gemini 2.5 Pro demonstrated the lowest cost and Opus 4.1 the fastest response times. Integrated analysis yielded the highest CPS for GPT-5, followed by Gemini 2.5 Pro and Opus 4.1 (Kendall's W = 0.87). A multi-dimensional evaluation framework integrating clinical quality and operational efficiency provides more actionable insights than single metric assessments, enabling pragmatic model selection for oncology practice. Nevertheless, the use of LLMs in this domain should remain clinician-supervised.

Keywords Large language models, Clinical decision support, Oncology, Non-small cell lung cancer, Composite performance score

Rapid medical knowledge expansion and increasing clinical data complexity have positioned LLMs to assume a critical role in healthcare[1]. Owing to their ability to comprehend medical language and process large-scale textual data, LLMs are being progressively adopted in clinical practice, where they support patient–physician communication, facilitate the transformation of unstructured clinical data into actionable information, and contribute to clinically meaningful outcome prediction[2]. Beyond these core applications, the capacity of LLMs to analyze specialized medical content and their potential integration into clinical decision-support systems position them as promising adjunctive tools in contemporary healthcare delivery[3].

These capabilities are particularly relevant in oncology, where multi-disciplinary evaluation and rapidly evolving treatment algorithms are essential[4]. Increasing patient burden and time constraints hinder the integration of multi-dimensional clinical data, thereby intensifying interest in LLM-based decision-support tools[5]. However, despite growing interest and encouraging performance reports, clinically representative and systematic evaluation approaches remain limited, posing challenges for objective comparison across models[6].

[1]Department of Radiation Oncology, Basaksehir Cam and Sakura City Hospital, 34480 Istanbul, Turkey. [2]Department of Informatics, Yildiz Technical University, 34420 Istanbul, Turkey. [3]Department of Mechatronics Engineering, Yildiz Technical University, 34420 Istanbul, Turkey. [4]Department of Computer Engineering, Yildiz Technical University, 34420 Istanbul, Turkey. [5]Department of Medical Oncology, Corlu State Hospital, 59850 Çorlu, Turkey. [6]Department of Radiation Oncology, Cerrahpaşa Faculty of Medicine, Istanbul University-Cerrahpaşa, 34098 Istanbul, Turkey. [7]Department of Medical Oncology, Basaksehir Cam and Sakura City Hospital, 34480 Istanbul, Turkey. [8]Department of Software Engineering, Istanbul Health and Technology University, 34000 Istanbul, Turkey. [9]Mehmet Halici and Serkan Salturk contributed equally to this work. ✉email: mehmethalici95@gmail.com; ssalturk@yildiz.edu.tr

Previous studies evaluating LLM performance have employed heterogeneous methodologies, including board-style multiple-choice questions, tumor board decision comparisons, and analyses of real-world data such as electronic health records, thereby limiting result standardization and cross-study comparability[7]. In contrast, structured evaluation approaches that reflect longitudinal clinical reasoning across diagnostic, therapeutic, and follow-up phases through open-ended, text-based assessments remain scarce[8]. This methodological heterogeneity is further amplified by the rapid proliferation of LLMs and frequent model updates, complicating informed model selection for clinical use[9].

Another notable limitation in the literature is the frequent reliance on isolated metrics, such as accuracy or explainability, to assess LLM performance. While these measures are fundamental, real-world clinical integration requires consideration of complementary operational parameters, including cost, response time, and generative efficiency[10]. Accordingly, recent methodological work advocates multi-dimensional evaluation strategies that extend beyond single-metric assessments[11].

To overcome these methodological challenges, this study employs a structured and consistent evaluation framework based on stepwise oncologic scenarios encompassing the diagnostic, therapeutic, and follow-up phases of patient care. Within this framework, LLMs-generated responses to open-ended clinical questions are systematically compared with reference answers derived from current clinical guidelines. To ensure clinical homogeneity and methodological consistency, all scenarios were restricted to a single disease entity; non–small cell lung cancer (NSCLC) was selected due to its high prevalence, the requirement for multi-disciplinary management, and its rapidly evolving therapeutic landscape[12].

Accordingly, the primary aim of this study is to develop and validate a multi-dimensional, clinically grounded benchmarking framework that integrates clinical accuracy, explainability, and operational efficiency through the CPS, and to demonstrate its utility by evaluating and comparing LLM performance in oncology.

Materials and methods

Study design

This study was designed as an observational and comparative analysis based on comprehensive, stepwise fictional scenarios structured to holistically represent the clinical course of oncologic cases. Each scenario was developed using a standardized template encompassing diagnostic, treatment, and follow-up phases, with two open-ended questions assigned to each step. All questions were presented to the LLMs in a standardized format, and model-generated responses were evaluated against evidence-based reference answers using clinical quality and operational performance metrics, integrated within the CPS framework (Fig. 1). The study was reported in accordance with the TRIPOD-LLM (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis-Large Language Models) guidelines [13].A detailed item-by-item compliance report is provided in Supplementary A.

Clinical scenario development

Five standardized fictional NSCLC scenarios were developed based on real-world experience, covering diverse disease stages and clinical challenges across the diagnostic, therapeutic, and follow-up phases. Each scenario

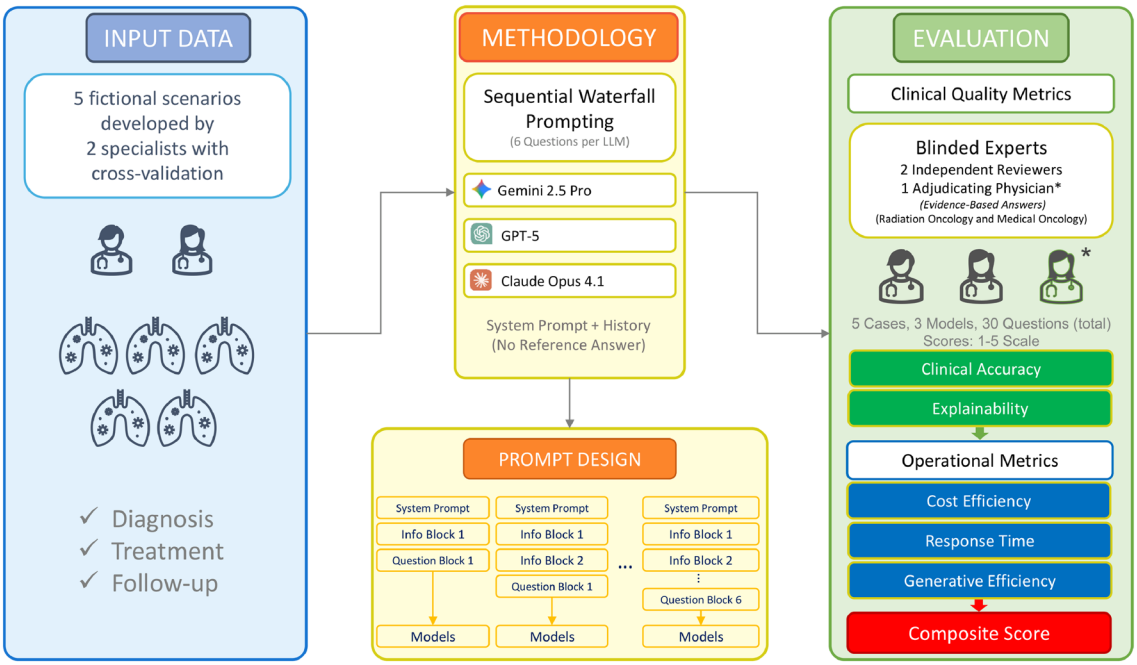

Fig. 1. Study workflow for multi-dimensional evaluation of LLMs in clinical oncology scenarios.

was structured in a stepwise manner and incorporated comprehensive clinical, radiological, pathological, and molecular data. An overview of the stepwise clinical case scenarios is presented in Supplementary B.

Scenarios were developed in English by a radiation oncologist and a medical oncologist (each with>10 years of clinical experience) using a standardized clinical assessment template (Supplementary C). A multi-stage cross-validation protocol was applied: scenarios were initially drafted by each specialist, independently reviewed and revised by the other for clinical accuracy and consistency, and subsequently refined by the original author under the oversight of the coordinating first author.

Question framework and evidence-based answers

Each step included two open-ended questions designed to enable a comprehensive evaluation of the corresponding clinical process. These questions were specifically structured to assess different clinical competencies of LLMs.

Diagnostic phase:

- Selection of appropriate diagnostic modalities to establish and confirm the diagnosis
- Cancer staging and/or risk stratification in accordance with current clinical guidelines

Treatment phase:

- Identification of the correct indication and selection of the appropriate treatment modality
- Planning details of the selected treatment modality in accordance with current clinical guidelines

Follow-up and complication management phase:

- Planning of patient follow-up according to current clinical guidelines
- Management of early and late treatment-related complications

The experts who developed the scenarios also prepared the correct answers to these questions in the form of a clear, concise, and evidence-based reference answer key, grounded in up-to-date and internationally accepted guidelines for non–small cell lung cancer[14–21]. This evidence-based reference answers set was used as the benchmark for evaluating the responses generated by the LLMs.

Large language models and prompting procedure

The stable versions (between August-December 2025) of Google Gemini 2.5 Pro (gemini-2.5-pro, GA release June 17, 2025), OpenAI, GPT-5 (gpt-5-2025-08-07), and Anthropic Claude Opus 4.1 (claude-opus-4-1-20250805) were used to evaluate responses generated for the clinical scenarios[22–24]. Model selection was based on the "longer-query leaderboard" results of the independent benchmarking platform lm-arena.ai, which aggregates human preference evaluations across diverse long-form prompts during the study period. While open-domain human preference does not indicate clinical accuracy, this criterion was utilized strictly to identify models with the highest baseline capacity for long-context management and sustained logical reasoning, which are prerequisites for processing complex, multi-step clinical scenarios.By adopting this approach, we aimed to ensure that observed performance differences reflected the models' intrinsic reasoning capabilities rather than technical constraints.

All models were accessed through their official API interfaces, and parameters that could limit generative performance were set to their maximum allowable levels. For Gemini 2.5 Pro, no explicit reasoning budget cap was imposed; for GPT-5, the *effort* parameter was set to *high*; and for Opus 4.1, a reasoning token budget of 4,096 and an output limit of 32,000 tokens were used. These configurations allowed the models to perform detailed reasoning without being constrained by abbreviated response strategies. Notably, the temperature setting was left unmodified, and the stochastic–deterministic balance was therefore not manually controlled.

The prompt structure followed a sequential, phase-based prompting strategy aligned with stepwise clinical reasoning, which we term the Sequential Waterfall Prompting (SWP) strategy. This approach partitions the clinical process into consecutive phases and ensures that, at each phase, the model generates new reasoning based solely on previously verified information. The phases were structured as diagnosis, treatment planning, and follow-up. At each phase, the model was instructed to produce both a concise definitive answer and a longer, explanatory response providing the underlying rationale.

The prompt structure consisted of three components:

1. System prompt (S): a high-level instruction that defines the model's expert role and response protocol. This prompt remains fixed across all phases.
2. kth question (Q_k): the query that specifies the cognitive objective of phase k.
3. Knowledge base (Σ_{k-1}): the cumulative set of verified knowledge blocks collected from previous phases.

At phase i, the process adds a verified knowledge block B_i. Accordingly, the cumulative verified knowledge available before phase k is defined as shown in Eq. (1).

$$\Sigma_{k-1} = \bigcup_{i=0}^{k-1} B_i \tag{1}$$

Importantly, generated responses are not included in Σ_{k-1}. This design allows verified information to propagate across phases while preventing the carry-over of potentially erroneous intermediate reasoning chains. As a result, each phase restarts its reasoning process while still leveraging the accumulated verified knowledge.

Within this tripartite structure, the model receives the following composite input at each phase, as formalized in Eq. (2):

$$A_k = \mathrm{LLM}(S, Q_k, \Sigma_{k-1}) \tag{2}$$

This design mitigates uncontrolled growth of the model's context window, enables stepwise reasoning that restarts at each phase, and preserves both logical continuity through Σ_{k-1} and contextual isolation by excluding prior responses throughout the entire process.

Performance metrics

Clinical quality performance metrics

The performance of LLMs was evaluated using an expert-based assessment approach. Given that the clinical scenarios consisted of open-ended questions and the responses were generated in a free-text format, automated metrics based on text similarity were considered insufficient to accurately reflect clinical correctness. Accordingly, all responses were independently and blindly evaluated by two experts, a radiation oncologist and a medical oncologist, each with more than 15 years of clinical experience. The evaluation was based on direct comparison with gold-standard, evidence-based reference answers developed in accordance with current international guidelines for each clinical step.

1. Accuracy: The level of agreement between the model's concise and definitive response and the predefined evidence-based reference answer.
2. Explainability: The degree to which the response provides robust clinical justification, logical consistency, and real-world clinical applicability.

For both dimensions, a 5-point Likert scale was applied to evaluate the responses, the detailed criteria of which are provided in Supplementary D. The final score for each item was calculated as the mean of the scores assigned by the two experts. In cases of substantial disagreement between evaluators, responses were reviewed by a third radiation oncology expert with more than 20 years of clinical experience, and rescoring was performed accordingly.

Operational metrics

For each model, performance evaluation included two primary token-based metrics: the number of input tokens and the number of output tokens.

The input token count was recorded based on the input token value reported in the provider's API usage output and comprised the total input delivered to the model at each phase, including the system prompt, the phase-specific question, and the cumulative verified knowledge blocks. The output token count was also calculated using the total output token value reported by each provider's API.

Token-based cost calculations were performed according to the current official pricing policies published by each model provider[25–27]. Publicly available prices per 1 million input tokens and per 1 million output tokens were used as references and were linearly applied to the actual input and output token counts recorded in the study. Using this approach, the token-based cost required for each model to generate the same clinical scenarios was calculated (Eq. 3).

$$\mathrm{Cost} = \left(\frac{\mathrm{Input\ Tokens}}{10^6}\right) P_{\mathrm{in}} + \left(\frac{\mathrm{Output\ Tokens}}{10^6}\right) P_{\mathrm{out}} \tag{3}$$

Generative efficiency was defined as the ratio of output tokens to input tokens (output tokens / input tokens) for each model and was analyzed as a quantitative performance metric.

The response time for each model was measured manually. The response time was defined as the interval between the timestamp at which the request was submitted and the completion of the response of the model, and the mean response time was calculated for each phase.

Construction of the composite performance score

After deriving clinical quality and operational performance metrics, a Composite Performance Score (CPS) was developed to integrate heterogeneous performance indicators with differing scales and directional properties into a single standardized measure. To ensure comparability across metrics with different units and ranges, all values were normalized to a 0–1 interval using min–max normalization.

For metrics in which higher values indicate better performance, including accuracy, explainability, and generative efficiency, normalization was performed using in Eq. (4):

$$X'_{qi} = \frac{X_{qi} - \min(X_q)}{\max(X_q) - \min(X_q)} \tag{4}$$

Conversely, for metrics in which lower values indicate better performance, total cost and response time, inverse min–max normalization was applied using Eq. (5):

$$X'_{qi} = \frac{\max(X_q) - X_{qi}}{\max(X_q) - \min(X_q)} \tag{5}$$

In the min–max normalization process, minimum and maximum values were calculated separately for each question based on metric values obtained from all evaluated models for that specific question.

To derive metric weights, prior to all analyses, five oncology specialists who were among the study authors independently rated the relative importance of each performance dimension in real-world clinical decision-making on a 1–10 scale, guided by a standardized question, without access to any model performance data or preliminary results. The standardized scoring form and operational definitions used by the experts are detailed in Supplementary E. The transition from raw expert scores to final weighting coefficients followed a structured mathematical protocol, including global min-max normalization and per-metric aggregation. The complete weight derivation formula and procedural steps are documented in Supplementary F

Accordingly, the CPS for each question *q* and model *i* was calculated as a weighted linear combination of normalized metrics, as shown in Eq. (6):

$$CPS_{qi} = w_1 A'_{qi} + w_2 E'_{qi} + w_3 C'_{qi} + w_4 S'_{qi} + w_5 GE'_{qi} \tag{6}$$

Here, A' denotes normalized accuracy, E' explainability, C' total cost, S' response speed (defined as inverse-normalized response time), and GE' generative efficiency, while w_1 through w_5 represent expert-derived weighting coefficients.

Statistical analysis

The distributional properties of all continuous variables were assessed using the Shapiro-Wilk test. Because the same set of clinical questions was evaluated across all models, inter-model comparisons were performed within a repeated-measures framework using the Friedman test, and effect sizes were reported using Kendall's W. For metrics showing statistical significance in the Friedman test, pairwise comparisons were conducted using the Wilcoxon signed-rank test with Bonferroni correction.

Model responses were independently and blindly evaluated by two clinical experts using Likert-scale ratings. Inter-rater agreement was primarily assessed using the Quadratic Weighted Kappa (QWK) for ordinal outcomes. To support interpretation under skewed rating distributions, Gwet's AC2 (quadratic weights) was additionally reported as a prevalence-robust agreement statistic.

An a priori power analysis was conducted using G*Power (version 3.1.9.4)[28] assuming a repeated-measures within-subject design with three measurements, a medium-to-large effect size (f = 0.35), an alpha level of 0.05, a correlation among repeated measures of 0.5, and a nonsphericity correction factor of 1.0. Under these assumptions, a minimum sample size of 23 questions was required to achieve 95% statistical power.

Responses showing $\geq$ 2-point discrepancy were re-evaluated by a third adjudicator, and final scores were determined following this additional review. All statistical analyses were performed with the `R` software and a two-sided p value < 0.05 was considered statistically significant for all analyses.

Results

Overall data structure and input token usage

Each clinical case was structured into three sequential phases-diagnosis, treatment, and follow-up-resulting in six phase-specific questions per case and a total of 30 questions across the study. For each step of the clinical scenarios, corresponding phase-specific input texts were provided to the models. Within this repeated-measures framework, a post hoc power analysis was conducted based on the observed study design parameters (effect size $f = 0.30$, significance level $\alpha = 0.05$, three repeated measurements, correlation among repeated measures of 0.5, and a nonsphericity correction factor of 1.0). This analysis demonstrated that the inclusion of 30 questions yielded high achieved statistical power ($1 - \beta = 0.95$) for detecting between-model differences.

Given the length and complexity of the phase-specific inputs, the median number of input tokens processed per case was 6,175 for Gemini 2.5 Pro, 6,054 for GPT-5, and 7,488 for Opus 4.1. Across all five clinical cases, total input token usage amounted to 30,875 tokens for Gemini 2.5 Pro, 30,270 tokens for GPT-5, and 37,440 tokens for Opus 4.1.

Clinical response quality

Inter-rater agreement, assessed using the QWK, was 0.351 for accuracy (p = 0.000196) and 0.124 for explainability (p = 0.184), a pattern consistent with the well-known kappa paradox observed in skewed rating distributions. Given the strong clustering of ratings at the upper end of the 5-point scale, we additionally reported the prevalence-robust Gwet's AC2, which indicated high agreement for both accuracy (AC2 = 0.912, 95% CI 0.871–0.954) and explainability (AC2 = 0.817, 95% CI 0.718–0.917). Because mean scores were similar across both metrics and score distributions showed limited variance, agreement was assessed based on score differences. Evaluations showing a difference of two points or more were identified in 4.5% of accuracy assessments and 9% of explainability assessments.Final scores were generated by averaging the reviewers' ratings after these discrepancies were resolved. Score distributions for accuracy and explainability are presented in Fig. 2 respectively.

A statistically significant difference in accuracy was observed among the three models (p < 0.001). The mean accuracy scores were 4.87 ± 0.26 for GPT-5, 4.65 ± 0.49 for Opus 4.1, and 4.42 ± 0.67 for Gemini 2.5 Pro. The overall effect size of the three-way comparison was moderate (Kendall's W = 0.35). In pairwise analyses, GPT-5 demonstrated a strong superiority over Gemini 2.5 Pro (r = 0.63). Opus 4.1 showed a moderate advantage over

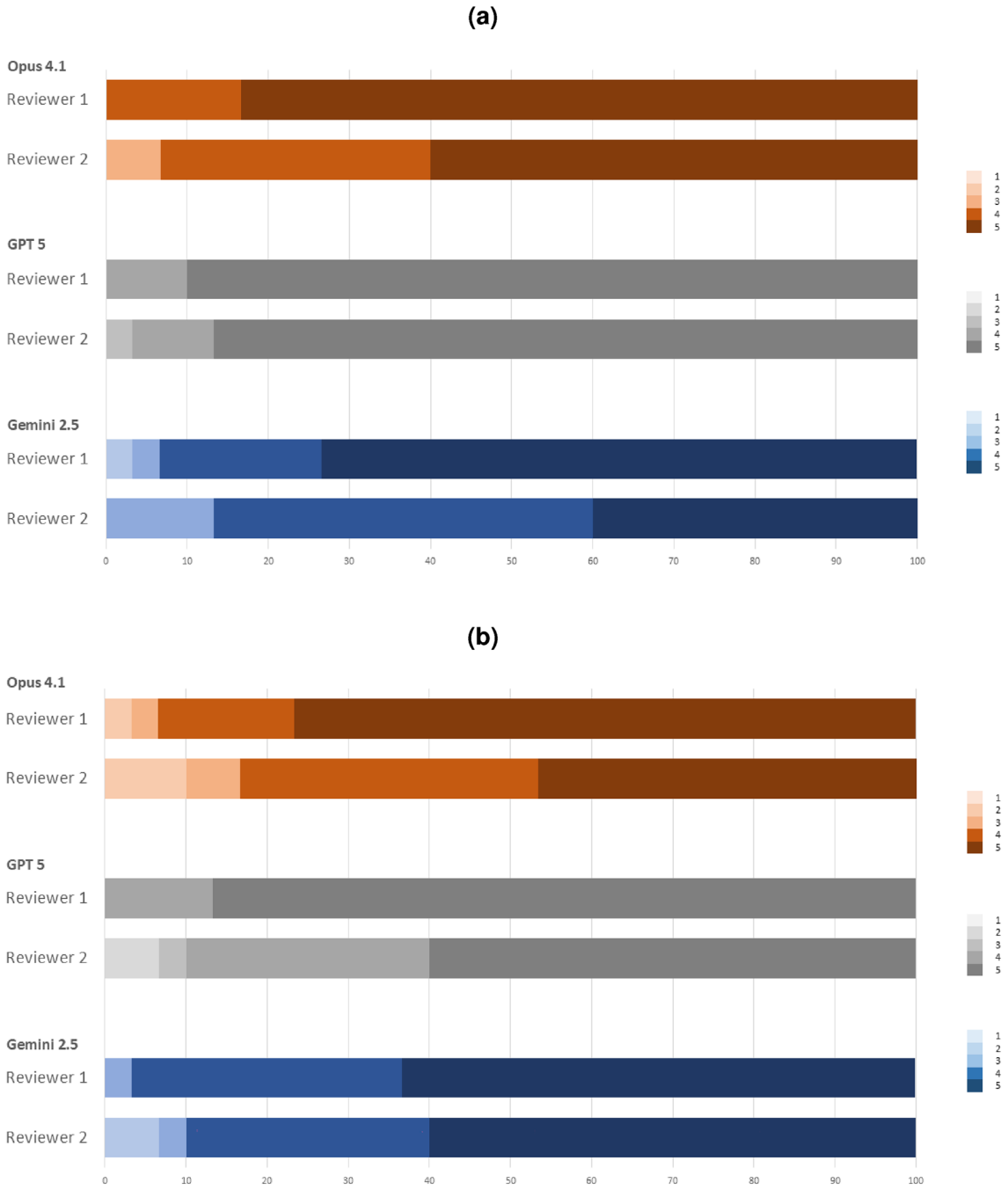

Fig. 2. Distribution of reviewer scores across models for (**a**) accuracy and (**b**) explainability. Stacked bars show the proportion percentages of five-point Likert scores assigned independently by two reviewers, with color gradients indicating score distributions.

Gemini 2.5 Pro (r = 0.45), and a moderate but statistically significant difference was also observed between Opus 4.1 and GPT-5 (r = 0.49).

Explainability scores also differed significantly among the models ($p < 0.001$). The mean explainability scores were 4.85 ± 0.40 for GPT-5, 4.50 ± 0.60 for Gemini 2.5 Pro, and 4.38 ± 0.70 for Opus 4.1. The overall effect size of the three-way comparison was moderate (Kendall's W = 0.33). In pairwise comparisons, GPT-5 showed a moderate advantage over Gemini 2.5 Pro (r = 0.50). The difference between Gemini 2.5 Pro and Opus 4.1 was small (r = 0.22), whereas the comparison between GPT-5 and Opus 4.1 revealed a marked and statistically significant difference (r = 0.61) (Table 1).

	Gemini 2.5 Pro mean ± SD	GPT-5 mean ± SD	Opus 4.1 mean ± SD	*p* value	Kendall's W	Pairwise comparison	Effect size (r)
Accuracy	4.42 ± 0.67	4.87 ± 0.26	4.65 ± 0.49	< 0.001	0.35	GPT-5 > Gemini 2.5 Pro	0.63
						Opus 4.1 > Gemini 2.5 Pro	0.45
						GPT-5 > Opus 4.1	0.49
Explainability	4.50 ± 0.60	4.85 ± 0.40	4.38 ± 0.70	< 0.001	0.33	GPT-5 > Gemini 2.5 Pro	0.50
						Gemini 2.5 Pro > Opus 4.1	0.22
						GPT-5 > Opus 4.1	0.61

Table 1. Comparison of clinical response quality metrics (accuracy and explainability) across models.

	Gemini 2.5 Pro Median (IQR)	GPT-5 median (IQR)	Opus 4.1 median (IQR)	*p* value	Kendall's W	Pairwise comparison	Effect size (r)
Cost	0.0288 (0.01)	0.0681 (0.02)	0.3995 (0.02)	< 0.001	1.00	Opus 4.1 > GPT-5	0.85
						GPT-5 > Gemini 2.5 Pro	0.87
						Opus 4.1 > Gemini 2.5 Pro	0.87
Response Time	31.20 (10.78)	123.48 (57.31)	28.06 (8.04)	< 0.001	0.73	GPT-5 > Gemini 2.5 Pro	0.87
						Gemini 2.5 Pro > Opus 4.1	0.48
						GPT-5 > Opus 4.1	0.87
Generative Efficiency	2.67 (1.43)	6.91 (3.70)	4.32 (1.36)	< 0.001	0.94	GPT-5 > Gemini 2.5 Pro	0.87
						Opus 4.1 > Gemini 2.5 Pro	0.87
						GPT-5 > Opus 4.1	0.86

Table 2. Comparison of cost, response time, and generative efficiency across models.

Operational model performance

A statistically significant difference in total cost was observed among the models (Friedman test, $p < 0.001$). Gemini 2.5 Pro showed the lowest cost (median 0.0288, IQR 0.01), followed by GPT-5 with intermediate values (median 0.0681, IQR 0.02), while Opus 4.1 exhibited the highest cost (median 0.3995, IQR 0.02). The consistency of model ranking was very high (Kendall's W = 1.00), and pairwise comparisons revealed very large effect sizes across all model pairs (r = 0.85–0.87).

Response time also differed significantly among the models (Friedman test, $p < 0.001$). Opus 4.1 was the fastest model (median 28.06 s, IQR 8.04), followed by Gemini 2.5 Pro (median 31.20 s, IQR 10.78), whereas GPT-5 showed substantially longer response times (median 123.48 s, IQR 57.31). Ranking consistency was high (Kendall's W = 0.73). In pairwise post hoc analyses (Wilcoxon signed-rank test), effect sizes were very large for GPT-5–Gemini 2.5 Pro and Opus 4.1–GPT-5 comparisons (r = 0.87), while a moderate effect size was observed for the Opus 4.1–Gemini 2.5 Pro comparison (r = 0.48).

Significant differences were also observed in generative efficiency among the models (Friedman test, $p < 0.001$). GPT-5 demonstrated the highest efficiency (median 6.91, IQR 3.70), followed by Opus 4.1 (median 4.32, IQR 1.36), while Gemini 2.5 Pro showed the lowest efficiency (median 2.67, IQR 1.43). The overall ranking consistency was high (Kendall's W = 0.94), and all pairwise post hoc comparisons (Wilcoxon signed-rank test) demonstrated very large effect sizes (r = 0.86–0.87) (Table 2). Figure 3 shows the cost, efficiency and response time comparison of utilized LLMs.

Composite performance score

The Composite Performance Score (CPS) used in this study was defined based on five core metrics representing clinical quality and operational efficiency. The weighting coefficients applied in CPS calculation were 0.33 for accuracy, 0.26 for explainability, 0.16 for cost-efficiency, 0.2 for generative efficiency, and 0.05 for response time. The distribution of expert-assigned importance scores for each metric and the resulting normalized weighting coefficients are presented in Fig. 4. These weights were subsequently applied to compute model-specific CPS values.

CPS values differed significantly across models ($p < 0.001$), with a very high overall effect size (Kendall's W = 0.874). Median CPS values were highest for GPT-5 at 0.839 (IQR 0.065), followed by Gemini 2.5 Pro at 0.741 (IQR 0.186) and Opus 4.1 at 0.596 (IQR 0.152), indicating a clear ranking among the models. In pairwise comparisons, GPT-5 demonstrated a strong advantage over both Gemini 2.5 Pro (r = 0.84) and Opus 4.1 (r = 0.87). In addition, Gemini 2.5 Pro showed a moderate-to-high advantage over Opus 4.1 (r = 0.69) (Table 3).

To assess the robustness of the model rankings, a sensitivity analysis was performed under alternative weighting schemes (e.g., equal weights, accuracy-dominant, and operationally-focused profiles). The results, which confirm the stability of the hierarchical ranking, are presented in Supplementary G.

In the radar plot based on normalized metrics without weighting, the models exhibited heterogeneous performance profiles across individual metrics (Fig. 5). In contrast, after applying expert-derived weighting coefficients, the radar plot generated, demonstrates that these metric-level differences are integrated into a composite profile, with GPT-5 achieving the highest overall weighted performance across the combined metrics.

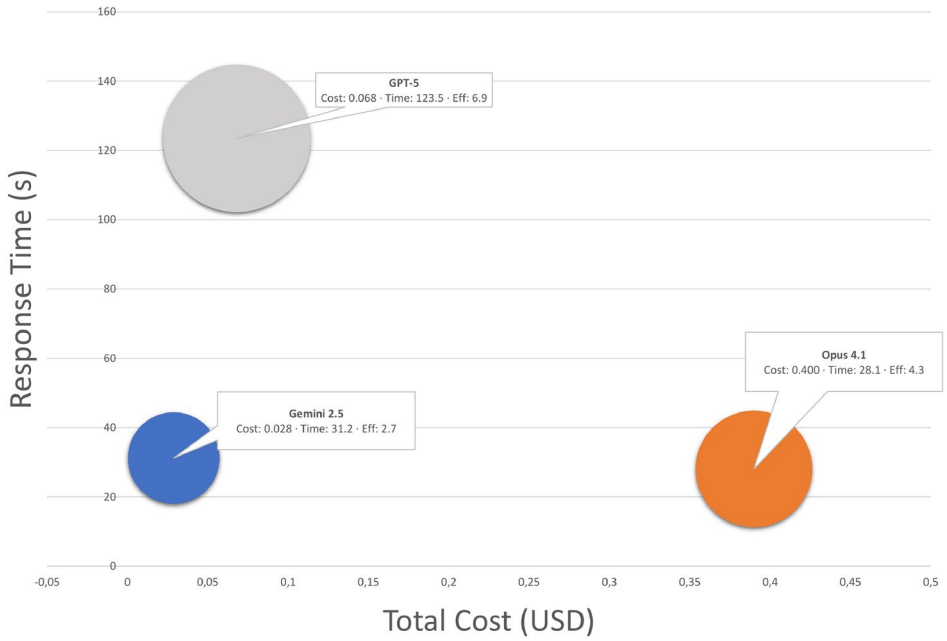

Fig. 3. Cost per question (X-axis)–time (Y-axis)–generative efficiency (bubble size) map illustrating the multi-dimensional relationship among three LLMs.

Discussion

The growing interest in deploying LLMs within oncology has been driven by their potential to support complex clinical reasoning and streamline decision-making across multiple stages of patient care. Parallel to this expanding interest, academic research in this area has also increased, giving rise to a wide variety of evaluation approaches and study designs. Although the majority of previous studies have primarily focused on response accuracy, operational parameters such as cost, response time, and generative efficiency have generally been treated as secondary outcomes[29]. In this context, evaluation frameworks that sufficiently capture the requirements of clinical integration remain limited. To address this gap, the present study evaluates LLM performance using a multi-dimensional framework aligned with clinical practice.

Given that clinical accuracy and explainability are regarded by physicians as key determinants of clinical utility, these two dimensions were emphasized and evaluated with particular priority in our analyses[30]. All evaluated models demonstrated high overall clinical accuracy, supporting the growing body of evidence that contemporary LLMs are capable of generating clinically acceptable oncology-related responses. Although GPT-5 achieved the highest mean accuracy score, the accuracy rates observed for Gemini 2.5 Pro, Opus 4.1, and GPT-5 (approximately 88%, 93%, and 97%, respectively) indicate that all models reached a generally satisfactory performance threshold.

Nevertheless, direct comparison of accuracy outcomes with previous studies is inherently limited due to substantial heterogeneity in question sources, evaluation methodologies, and clinical contexts[29]. For instance, a study using medical oncology board examination questions reported higher output quality for Claude 3.5 Sonnet compared with GPT-4o and Gemini 1.5[31]. In contrast, a larger synthesis encompassing ten examination systems demonstrated that GPT-o1 achieved approximately 13.4% higher accuracy than Claude 3 Opus[32]. Despite these model-specific differences, the literature consistently indicates progressive improvements in higher-generation LLMs across both clinical and examination-based settings relative to earlier model versions[33].

Beyond accuracy, explainability emerged as a key discriminator of model behavior. Higher explainability scores, particularly for GPT-5, indicate that structured and detailed reasoning better supports clinicians' critical appraisal of model-generated recommendations. In line with this, Savage et al. reported that approximately 82% of GPT-5's correct responses to open-ended diagnostic questions included logically coherent and clinically interpretable reasoning[34]. Indeed, explainability is of critical importance in clinical decision support contexts, as transparent justifications may reduce the "black box" perception attributed to LLMs and thereby support safer human-AI collaboration[35].

One of the principal strengths of this study lies in the systematic evaluation of operational performance metrics alongside clinical quality measures. In real-world clinical environments, decision support systems are expected not only to generate accurate and interpretable outputs but also to do so within acceptable time and cost constraints[36]. In our study, cost evaluation was conducted based on publicly available USD-denominated pricing policies per one million input and output tokens. By submitting all queries through each model's official API, a standardized and comparable cost analysis across models was ensured[37]. Opus 4.1 was the most expensive model, at approximately USD 75 per one million tokens; although it generated fewer output tokens, this did not offset its high per-token cost. Conversely, Opus 4.1 produced shorter, optimized responses and demonstrated faster response times than other models. Although seemingly negligible at the level of individual queries, these

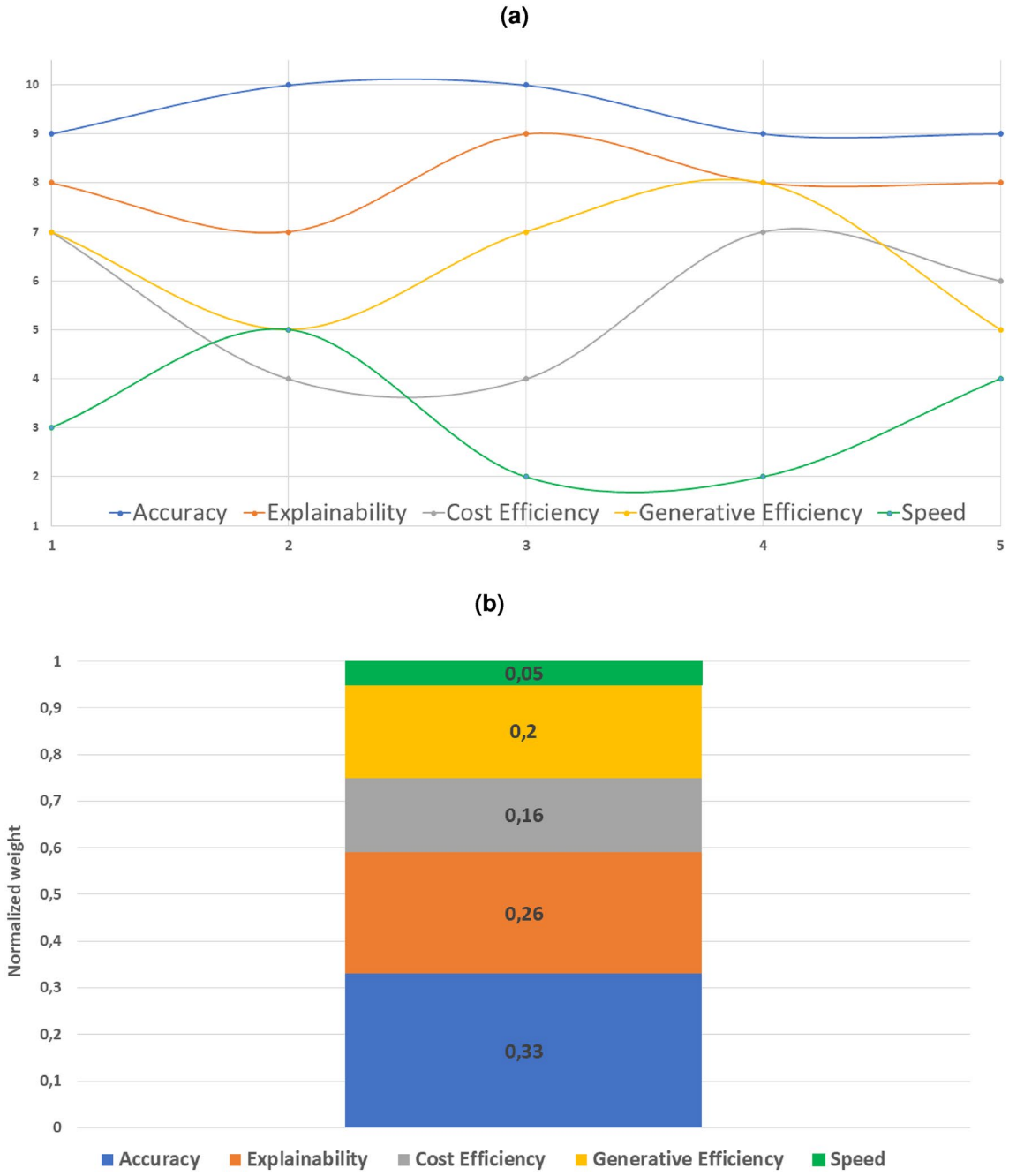

Fig. 4. (**a**) Expert-assigned importance scores for the five metrics included in the CPS, (**b**) normalized weight coefficients for each metric, scaled to sum to 1, used in CPS calculation.

	Gemini 2.5 Pro	GPT-5	Opus 4.1	p value	Kendall's W	Pairwise comparison	Effect size (r)
CPS	0.741 (0.186)	0.839 (0.065)	0.596 (0.152)	< 0.001	0.874	GPT-5 > Gemini 2.5 Pro	0.84
						GPT-5 > Opus 4.1	0.87
						Gemini 2.5 Pro > Opus 4.1	0.69

Table 3. Inter-model comparison of CPS assessed by Friedman test and Wilcoxon post hoc analysis.

latency differences, limited to seconds may become cumulatively meaningful in high-frequency, sequential clinical use scenarios, thereby exerting a nontrivial influence on model preference[38].

The generative efficiency metric offered additional insight into model behavior. The output-to-input token ratio captures differences in content density and explanatory richness. GPT-5 showed higher productivity, which

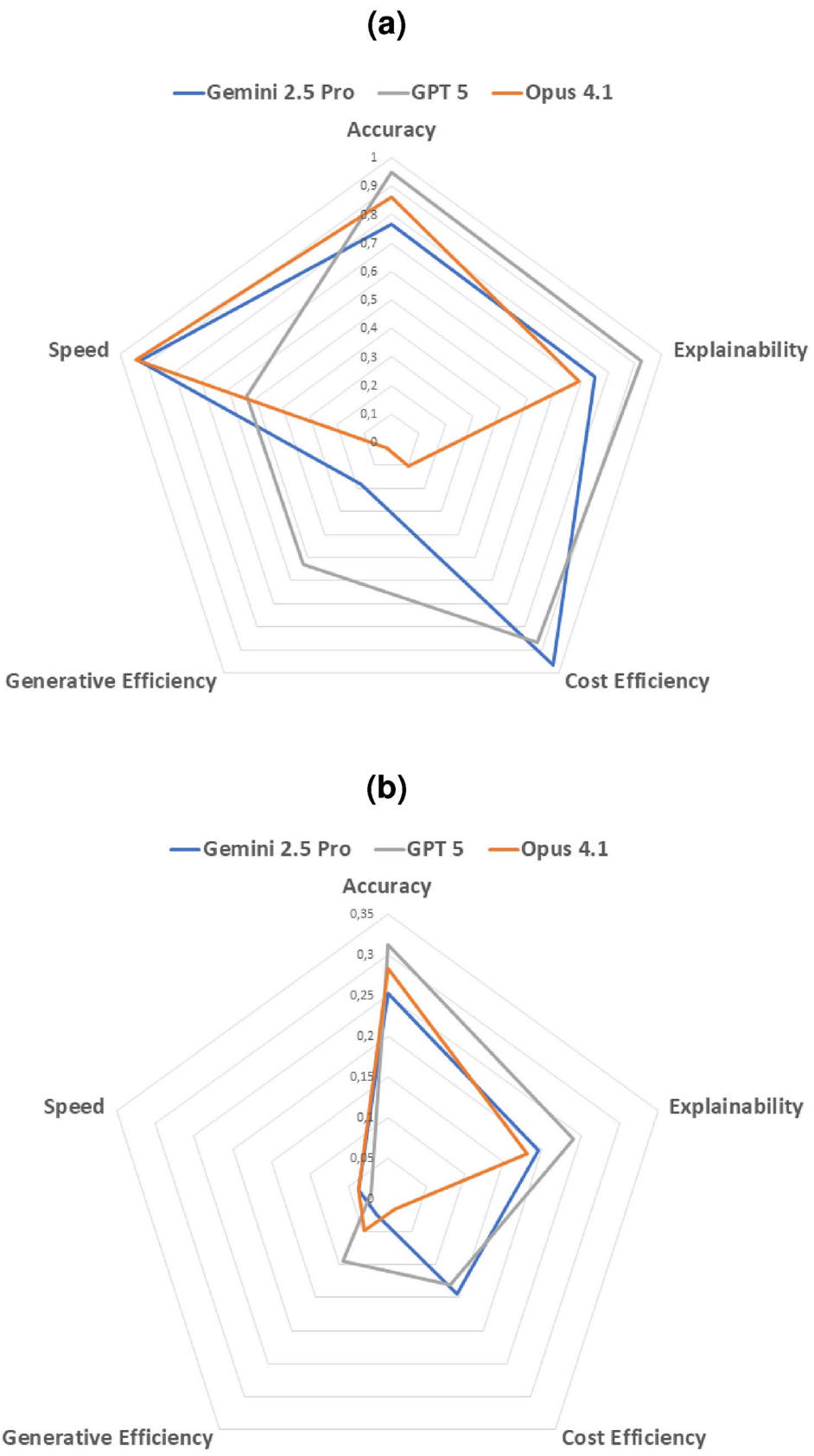

Fig. 5. Radar plots illustrating model performance profiles based on (**a**) min–max normalized metrics and (**b**) metrics after normalization followed by multiplication with weighting coefficients.

correlated with superior explainability, suggesting that more detailed responses improve interpretability at the cost of increased time and expense. Thus, generative efficiency can be viewed as an intermediary metric linking operational efficiency and clinical explainability[39].

A key innovative aspect of this study is the introduction of the CPS, a metric that consolidates diverse performance dimensions into a single weighted index. In contrast to earlier methods that relied on isolated metrics, the CPS uses expert-defined weighting coefficients to capture the relative clinical significance of each measure. Accordingly, GPT-5 demonstrated the highest overall performance across all evaluated scenarios, driven by its strong results in clinically critical domains prioritized by experts, including accuracy and explainability. As shown in Figure 5, when each metric is examined independently, every model exhibits particular strengths in distinct performance domains, supporting context-specific model selection. Yet, once these metrics are systematically prioritized and integrated through weighting coefficients, GPT-5 stands out across most performance dimensions even prior to the final CPS aggregation. This visual trend is corroborated by the numerical outcomes derived from the CPS calculation, underscoring CPS's value not only as a ranking mechanism but also as a comprehensive decision-support tool for context-aware model selection[40].

One of the primary reasons LLMs have not yet achieved the expected level of trust for clinical use is the lack of methodological standardization in evaluation processes. Substantial heterogeneity in question types, clinical contexts, and analytical approaches limits cross-model comparability and generalizability[29].

In this context, the evaluation of open-ended responses according to predefined clinical quality standards constituted a core methodological component of the study. To reduce evaluator bias, the assessment process was structured around evidence-based reference answers derived from current international guidelines, and the clinical meaning of each Likert scale level was explicitly defined. Initially, the inherently subjective nature of open-ended and ordinal evaluations, together with a tendency toward generous scoring by reviewers, led to a kappa paradox, resulting in relatively low QWK values. However, the observation that disagreements of two or more points accounted for less than 10% across both metrics suggested that these differences were largely attributable to terminological nuances. Consequently, the analysis was extended to Gwet's AC2 statistic, which is less sensitive to prevalence imbalance and rater scoring tendencies[41].

The high agreement levels observed using this metric support the consistency of the evaluation process. In addition to human expert reviewers, vector-based text comparison methods were considered to enhance inter-rater agreement[42]. However, because model responses varied in depth relative to evidence-based reference answers, raising concerns that text-based similarity methods might misclassify clinically accurate but more comprehensive responses as semantically discordant, clinical accuracy and appropriateness were ultimately evaluated by independent human reviewers.

Unlike many prior studies, prompts were submitted exclusively via each model's official API, avoiding user interface related bias. Within this framework, we opted to keep the models' API configurations at optimal settings that reflect the peak performance capacity of each model at the time. Our objective was to comparatively evaluate their true potential in clinical reasoning processes, rather than confining the models to a restricted token budget. Furthermore, to more closely simulate real-world clinical processes and optimize cost consumption, a SWP approach was adopted, whereby new clinical information was incrementally added to each preceding scenario[37].

This study was designed to evaluate LLMs within a clinically grounded and methodologically controlled framework. Methodological rigor was supported through multidisciplinary planning, dual expert review, and the use of evidence-based reference answers, while clinically weighted operational metrics were integrated into a CPS to enable a broader assessment of model utility than clinical accuracy alone.

Several limitations should be considered when interpreting these findings. Although both a priori and post hoc power analyses suggested that the number of clinical questions was sufficient to detect meaningful between-model differences, inclusion of a wider range of models could further strengthen real-world relevance. The focus on a single cancer type provided a consistent clinical context and enhanced internal validity; however, despite incorporating diverse clinical problem structures[43], this approach necessarily limited the scope of evaluated decision pathways, particularly with respect to surgical decision-making. Another limitation is the exclusive use of the SWP strategy without ablation against standard single-turn or Chain-of-Thought approaches. While specifically designed to simulate stepwise clinical decision-making, this longitudinal structure inherently demands robust long-context management. To address this gap, future studies should include comparative analyses of different prompting strategies to systematically evaluate their impact on model performance. In addition, reliance on two expert raters and a 5-point Likert scale may have constrained the granularity of the assessments and introduced residual subjectivity. The exclusive use of fictional scenarios, while enabling standardized evaluation, cannot fully replicate real-world clinical complexity including comorbidities and atypical presentations; regulatory and data-security constraints precluded the use of real patient data, and consequently direct comparison with actual clinician decisions was not possible, limiting external validity.

Also, several methodological constraints must be explicitly acknowledged to ensure transparency, aligning with our TRIPOD-LLM compliance report. First, the expert-derived weighting coefficients used for the CPS were assigned by co-authors who possessed deep contextual knowledge of the study's objectives. While this ensured clinical relevance, it introduces a potential source of bias compared to an evaluation by an entirely independent, blinded external panel. However, to mitigate hindsight bias, it is crucial to note that these weighting coefficients were established strictly prior to any data analysis, ensuring that the co-authors were completely blinded to the models' performance outcomes. Second, to evaluate each model at its absolute maximum reasoning potential, model-specific API configurations were applied (e.g., operating without a reasoning budget cap for Gemini 2.5 Pro, utilizing 'effort=high' for GPT-5, and defining a specific reasoning token budget for Opus 4.1). While this provider-tailored standardization ensured each model performed at peak capacity, it inevitably limits direct, token-for-token comparability across models. Additionally, during the experimental phase, the API for Opus 4.1 did not natively report exact output token counts. Consequently, the output token values and subsequent

cost calculations for this specific model were based on standardized estimations. This approximation may subtly skew the precise generative efficiency and operational cost comparisons involving Opus 4.1. Third, due to the stochastic nature of large language models and the inability to manually control internal temperature settings for these specific advanced reasoning APIs, complete reproducibility of the exact lexical outputs cannot be fully guaranteed, even though robust semantic and logical consistency was observed across multiple iterations.

In conclusion, this study demonstrates that evaluating LLMs using a holistic, clinically grounded, and multi-dimensional framework yields more actionable and meaningful insights than reliance on isolated performance metrics. Although LLMs hold substantial promise as clinical decision support tools, they should be deployed under clinician supervision rather than as autonomous decision-makers. Accordingly, the primary contribution of this study is the evaluation framework itself rather than the specific model rankings, which may shift as LLMs continue to evolve. Future research should focus on expanding standardized evaluation frameworks, incorporating broader clinical domains, and validating performance using real-world clinical data.

Data availability

The datasets generated during the current study are available from the corresponding author on reasonable request.

Code availability

The custom code implementing the Sequential Waterfall Prompting workflow and the model evaluation pipeline is publicly available at https://github.com/burakertann/Sequential-Waterfall-Prompting.

Received: 28 December 2025; Accepted: 2 July 2026
Published online: 21 July 2026

References

1. Huang, J. et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *NPJ Digit. Med.* **7**, 106 (2024).
2. Gholipour, M., Khajouei, R., Amiri, P., Hajesmaeel Gohari, S. & Ahmadian, L. Extracting cancer concepts from clinical notes using natural language processing: A systematic review. *BMC Bioinform.* **24**, 405 (2023).
3. Klang, E. et al. A strategy for cost-effective large language model use at health system-scale. *NPJ Digit. Med.* **7**, 320 (2024).
4. Gaber, F. et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digit. Med.* **8**, 263 (2025).
5. Benary, M. et al. Leveraging large language models for decision support in personalized oncology. *JAMA Netw. Open* **6**, e2343689–e2343689 (2023).
6. Khan, M. P. & O'Sullivan, E. D. A comparison of the diagnostic ability of large language models in challenging clinical cases. *Front. Artif. Intell.* **7**, 1379297 (2024).
7. Rydzewski, N. R. et al. Comparative evaluation of LLMS in clinical oncology. *NEJM AI* **1**, AIoa2300151 (2024).
8. Wong, E. ESMO guidance on the use of large language models in clinical practice (ELCAP). *Ann. Oncol.* (2025).
9. Abbas, A., Rehman, M. S. & Rehman, S. S. Comparing the performance of popular large language models on the national board of medical examiners sample questions. *Cureus* **16** (2024).
10. Dong, J. et al. Cost-efficient knowledge-based question answering with large language models. *Adv. Neural Inf. Process. Syst.* **37**, 115261–115281 (2024).
11. Tam, T. Y. C. et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit. Med.* **7**, 258 (2024).
12. Brown, E. D. Precision oncology in non-small cell lung cancer: A comparative study of contextualized ChatGPT models. *Cureus* **17** (2025).
13. Gallifant, J. et al. The tripod-LLM reporting guideline for studies using large language models. *Nat. Med.* **31**, 60–69 (2025).
14. Society, B. T. et al. BTS guidelines for the investigation and management of pulmonary nodules. https://www. brit-thoracic. org.uk/quality-improvement/guidelines/pulmonary-nodules/. Accessed 1 Dec 2021 (2015).
15. Riely, G. J. NCCN guidelines® insights: Non-small cell lung cancer, version 7.2025: Featured updates to the NCCN guidelines®. *J. Natl. Comprehens. Cancer Netw.* **23**, 354–362 (2025).
16. Le Rhun, E. et al. EANO-ESMO clinical practice guidelines for diagnosis, treatment and follow-up of patients with brain metastasis from solid tumours. *Ann. Oncol.* **32**, 1332–1347 (2021).
17. Gondi, V. et al. Radiation therapy for brain metastases: An astro clinical practice guideline. *Pract. Radiat. Oncol.* **12**, 265–282 (2022).
18. Postmus, P. et al. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **28**, iv1–iv21 (2017).
19. Zer, A. Early and locally advanced non-small-cell lung cancer: ESMO clinical practice guideline for diagnosis, treatment and follow-up. *Ann. Oncol.* (2025).
20. Nestle, U. et al. ESTRO ACROP guidelines for target volume definition in the treatment of locally advanced non-small cell lung cancer. *Radiother. Oncol.* **127**, 1–5 (2018).
21. Daly, M. E. et al. Management of stage III non-small-cell lung cancer: Asco guideline. *J. Clin. Oncol.* **40**, 1356–1384 (2022).
22. Google Cloud. Gemini 2.5 Pro Model Documentation. Vertex AI Documentation for Gemini 2.5 Pro. https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro (2025).
23. OpenAI. GPT-5 System Card. System Card for GPT-5 (API and Reasoning Variants). https://openai.com/tr-TR/index/gpt-5-system-card/. Accessed Dec 2025 (2025).
24. Anthropic. Claude Opus 4.1. Model Announcement and Details for Claude Opus 4.1. https://www.anthropic.com/news/claude-opus-4-1 (2025).
25. https://openai.com/tr-tr/api/pricing/. Aug-Nov 2025 (2025).
26. https://ai.google.dev/gemini-api/docs/models?hl=tr#gemini-2.5-pro. Aug-Nov 2025 (2025).
27. https://claude.com/pricing#api. Aug-Nov 2025 (2025).
28. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191 (2007).
29. Hao, Y. et al. Large language model integrations in cancer decision-making: A systematic review and meta-analysis. *NPJ Digit. Med.* **8**, 450 (2025).

30. Chow, R. et al. The accuracy of artificial intelligence ChatGPT in oncology examination questions. *J. Am. Coll. Radiol.* **21**, 1800–1804 (2024).
31. Erdat, E. C. & Kavak, E. E. Benchmarking LLM chatbots' oncological knowledge with the Turkish Society of Medical Oncology's annual board examination questions. *BMC Cancer* **25**, 197 (2025).
32. Kasagga, A. Performance of ChatGPT and large language models on medical licensing exams worldwide: A systematic review and network meta-analysis with meta-regression. *Cureus* **17** (2025).
33. Chen, D., Avison, K., Alnassar, S., Huang, R. S. & Raman, S. Medical accuracy of artificial intelligence chatbots in oncology: A scoping review. *Oncologist* **30**, oyaf038 (2025).
34. Savage, T., Nayak, A., Gallo, R., Rangan, E. & Chen, J. H. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digit. Med.* **7**, 20 (2024).
35. Mesinovic, M., Watkinson, P. & Zhu, T. Explainability in the age of large language models for healthcare. *Commun. Eng.* **4**, 128 (2025).
36. Burns, M. L. et al. Generative AI costs in large healthcare systems, an example in revenue cycle. *NPJ Digit. Med.* **8**, 579 (2025).
37. Nori, H. Sequential diagnosis with language models. arXiv preprint arXiv:2506.22405 (2025).
38. Chitty-Venkata, K. T. LLM-inference-bench: Inference benchmarking of large language models on AI accelerators. In *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1362–1379 (IEEE, 2024).
39. Du, Z., Kang, H., Han, S., Krishna, T. & Zhu, L. Ockbench: Measuring the efficiency of LLM reasoning. arXiv preprint arXiv:2511.05722 (2025).
40. Liang, P. Holistic evaluation of language models. arXiv preprint: arXiv:2211.09110 (2022).
41. Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* **61**, 29–48 (2008).
42. Yalamanchili, A. et al. Quality of large language model responses to radiation oncology patient care questions. *JAMA Netw. Open* **7**, e244630–e244630 (2024).
43. Rahsepar, A. A. et al. How AI responds to common lung cancer questions: ChatGPT versus google bard. *Radiology* **307**, e230922 (2023).

Author contributions

M.H. contributed to Conceptualization, Methodology, Resources, Formal Analysis, and Writing - Original Draft. S.S. contributed to Conceptualization, Methodology, Software, Resources, Formal Analysis, Writing - Original Draft, Visualization, and Validation. I.S. contributed to Writing - Original Draft and Statistical Analysis. B.E. contributed to Methodology and Software. I.C.B. contributed to Conceptualization and Review & Editing. K.C. contributed to Clinical Quality Assessment and Scoring of Model Responses. T.K. contributed to Clinical Scenario Development and Preparation of Evidence-Based Reference Answers. C.Y. contributed to Clinical Scenario Development and Preparation of Evidence-Based Reference Answers. G.U.E. contributed to Clinical Quality Assessment and Scoring of Model Responses. H.S.K. contributed to Adjudication and Secondary Evaluation of Discrepant Ratings. M.T.K. contributed to Review & Editing. H.U. contributed to Review & Editing and Supervision.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

The study was based entirely on fictional clinical scenarios and did not involve real patient data; therefore, formal ethical approval was not required.

Informed consent

Documented informed consent was obtained from all oncology specialists who contributed as domain experts to the preparation of the study questions and the evaluation of the Large Language Models' responses, in accordance with the journal's policy.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-61195-1.

Correspondence and requests for materials should be addressed to M.H. or S.S.