Intelligent Oncology 2 (2026) 100039

Contents lists available at ScienceDirect

Intelligent Oncology

journal homepage: www.keaipublishing.com/en/intonc/

Full Length Article

Decision-making performance of large language models vs. human physicians in challenging lung cancer cases: A real-world case-based study

Ning Yang[a,#], Kailai Li[b,#], Baiyang Liu[a], Xiting Chen[c], Aimin Jiang[d], Chang Qi[e], Wenyi Gan[f], Lingxuan Zhu[b], Weiming Mou[b,g], Dongqiang Zeng[h,i], Mingjia Xiao[j], Guangdi Chu[k], Shengkun Peng[l], Hank Z.H. Wong[m], Lin Zhang[n,o], Hengguo Zhang[p], Xinpei Deng[q], Quan Cheng[r,s], Bufu Tang[t], Anqi Lin[b,*], Juan Zhou[a,c,u,*], Peng Luo[b,*]

[a] Department of Oncology, General Hospital of Southern Theater Command, Guangzhou Guangdong 510030, China
[b] Department of Oncology, Zhujiang Hospital of Southern Medical University, Guangzhou Guangdong 510282, China
[c] Graduate School, Guangzhou University of Chinese Medicine, Guangzhou Guangdong 510006, China
[d] Department of Urology, Changhai Hospital, Naval Medical University (Second Military Medical University), Shanghai 200433, China
[e] Department of Clinical Oncology, The University of Hong Kong, Hong Kong SAR 999077, China
[f] Department of Joint Surgery and Sports Medicine, Zhuhai People's Hospital (Zhuhai Clinical Medical College of Jinan University), Zhuhai Guangdong 519000, China
[g] Department of Urology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200080, China
[h] Department of Oncology, Nanfang Hospital, Southern Medical University, Guangzhou Guangdong 510515, China
[i] Cancer Center, The Sixth Affiliated Hospital, South China University of Technology, Guangzhou Guangdong 510006, China
[j] Hepatobiliary Surgery Department, People's Hospital of Quzhou (Quzhou Affiliated Hospital of Wenzhou Medical University), Quzhou Zhejiang 324000, China
[k] Department of Urology, The Affiliated Hospital of Qingdao University, Qingdao Shandong 266003, China
[l] Department of Radiology, Sichuan Provincial People's Hospital/Affiliated Hospital of University of Electronic Science and Technology of China, Chengdu Sichuan 610072, China
[m] Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR 999077, China
[n] School of Public Health and Preventive Medicine, Monash University, Melbourne VIC 3000, Australia
[o] Suzhou Industrial Park Monash Research Institute of Science and Technology, Suzhou Jiangsu 215000, China
[p] College & Hospital of Stomatology, Anhui Medical University, Key Laboratory of Oral Diseases Research of Anhui Province, Hefei Anhui 230032, China
[q] Department of Urology, State Key Laboratory of Oncology in Southern China, Sun Yat-sen University Cancer Center, Guangdong Provincial Clinical Research Center for Cancer, Guangzhou Guangdong 510060, China
[r] Department of Neurosurgery, Xiangya Hospital, Central South University, Changsha Hunan 410008, China
[s] National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha Hunan 410008, China
[t] Department of Radiation Oncology, Zhongshan Hospital Affiliated to Fudan University, Shanghai 200032, China
[u] The First School of Clinical Medicine, Southern Medical University, Guangzhou Guangdong 510515, China

ARTICLE INFO

Keywords:
Large language models
Clinical evaluation
Decision-making
Lung cancer

ABSTRACT

Background: Despite the promise shown by large language models (LLMs) for standardized tasks, their multi-dimensional performance in real-world oncology decision-making remains unevaluated. This study aims to introduce a framework for evaluating LLMs and physician decisions in challenging lung cancer cases.
Methods: We curated 50 challenging lung cancer cases (25 local and 25 published) classified as complex, rare, or refractory. Blinded three-dimensional, five-point Likert evaluations (1–5 for comprehensiveness, specificity, and readability) compared standalone LLMs (DeepSeek R1, Claude 3.5, Gemini 1.5, and GPT-4o), physicians by experience level (junior, intermediate, and senior), and AI-assisted juniors; intergroup differences and augmentation effects were analyzed statistically.
Results: Of 50 challenging cases (18 complex, 17 rare, and 15 refractory) rated by three experts, DeepSeek R1 achieved scores of 3.95 ± 0.33, 3.71 ± 0.53, and 4.26 ± 0.18 for comprehensiveness, specificity, and readability, respectively, positioning it between intermediate (3.68, 3.68, 3.75) and senior (4.50, 4.64, 4.53) physicians. GPT-4o and Claude 3.5 reached intermediate physician–level comprehensiveness (3.76 ± 0.39, 3.60 ± 0.39) but junior-to-intermediate physician–level specificity (3.39 ± 0.39, 3.39 ± 0.49). All LLMs scored higher

* Corresponding authors.
E-mail addresses: smulinanqi0206@i.smu.edu.cn (A. Lin), juanzhn@163.com (J. Zhou), luopeng@smu.edu.cn (P. Luo).
# These authors contributed equally to this work.

https://doi.org/10.1016/j.intonc.2026.100039
Received 22 November 2025; Received in revised form 6 January 2026; Accepted 8 January 2026
Available online 14 January 2026
2950-2616/© 2026 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

on rare cases than intermediate physicians but fell below junior physicians in refractory-case specificity. AI-assisted junior physicians showed marked gains in rare cases, with comprehensiveness rising from 2.32 to 4.29 (84.8%), specificity from 2.24 to 4.26 (90.8%), and readability from 2.76 to 4.59 (66.0%), while specificity declined by 3.2% (3.17 to 3.07) in refractory cases. Error analysis showed complementary strengths, with physicians demonstrating reasoning stability and LLMs excelling in knowledge updating and risk management. *Conclusions:* LLMs performed variably in clinical decision-making tasks depending on case type, performing better in rare cases and worse in refractory cases requiring longitudinal reasoning. Complementary strengths between LLMs and physicians support case- and task-tailored human–AI collaboration.

1. Introduction

Lung cancer, the leading cause of cancer-related mortality worldwide, presents increasingly complex challenges in diagnostic and therapeutic decision-making.[1] In the era of precision oncology, clinicians must integrate multidimensional patient information, including molecular and pathological profiles, radiological presentations, comorbidity management, and continuously evolving evidence.[2,3] While complex scenarios exist across malignancies, lung cancer is particularly prone to difficult-to-diagnose and therapeutically complex cases due to heterogeneous histology and molecular subtypes, frequent overlap of imaging phenotypes, and rapid shifts in biomarker-driven treatment strategies.[4–6] Therefore, accurate diagnosis and individualized management are clinically critical, especially for cases with multisystem involvement, rare pathological subtypes, or drug-resistant recurrent disease, in which decision-making requires interdisciplinary expertise and timely updating of evidence-based recommendations.

However, in real-world settings, clinical decision-making ability varies substantially across physicians with different levels of experience; junior physicians confronting challenging lung cancer cases often lack interdisciplinary breadth and depth of experiential expertise. Currently, clinical decisions commonly rely on traditional literature retrieval and multidisciplinary team consultation; however, both approaches have inherent limitations. Nonsystematic searches may generate fragmented or nonactionable evidence, increasing the risk of deviation from optimal pathways. Multidisciplinary team consultations, although valuable, face practical challenges, including limited accessibility, scheduling delays, variable recommendations across institutions, inconsistent quality control, and increased costs.[7,8] These inefficiencies and uncertainties are especially prominent in primary care institutions,[9] underscoring the urgent need for intelligent assistance tools to support accurate diagnosis and management in difficult-to-diagnose and therapeutically complex lung cancer cases.

Large language models (LLMs) have demonstrated transformative potential in medicine. Prior studies have assessed their performance on medical licensure examinations, standardized case analyses, and consultation-style tasks, indicating that leading models, such as GPT-4o and Claude 3.5, can match or even surpass physicians in knowledge breadth and basic reasoning.[10–15] However, emerging clinical translation research suggests that the value of LLMs may extend beyond question answering to workflow-embedded applications in real clinical settings. For example, a recent study evaluated prompt-engineered ChatGPT systems' ability to generate electronic medical records during lung nodule screening and integrated the approach into practical tools (e.g., WeChat mini-programs) to support pre-visit documentation and potentially improve workflow efficiency and diagnostic processes.[16] Despite such progress, a systematic review by Bedi et al.[17] identified critical gaps. Of the 519 studies published from 2022 to 2024, only 5% evaluated LLMs using real-world clinical cases, while the most common tasks were answering licensure examination questions (44.5%) and making diagnoses (19.5%), offering limited insight into performance in complex clinical scenarios. Regarding evaluation dimensions, most studies (95.4%) focused on answer accuracy, with few providing systematic, multidimensional assessments of decisional comprehensiveness, degree of individualization, or communication effectiveness. Furthermore, there remains a lack of a well-established, comprehensive, and widely adopted multidimensional correspondence framework to relate LLM decision-making capabilities to oncology physicians at different seniority levels. Pathways for optimizing human–AI collaborative decision-making also remain underexplored, constraining translation to clinical practice.[18–20]

This study establishes a standardized assessment system for real-world challenging lung cancer cases from both local and online sources by employing a rigorous, multidimensional, blinded evaluation design to systematically compare differences in clinical decision-making capabilities between four mainstream LLMs (DeepSeek R1, Claude 3.5, Gemini 1.5, and GPT-4o) and human physicians at different levels (junior, intermediate, and senior). In addition, it introduces an AI-assisted junior physician group to examine the augmentation potential of human–machine collaboration. This article is presented in accordance with the TRIPOD-LLM reporting guidelines for studies using LLMs.[21]

2. Methods

2.1. Dataset construction and participants

2.1.1. Dataset construction

For this study, we constructed a retrospective clinical case dataset that included real-world cases and published case reports. Two thoracic oncology specialists, each with over 15 years of clinical experience and over 10 years of experience in academic medical writing, selected local cases from the lung cancer treatment database of the General Hospital of Southern Theater Command (2020–2024) according to predefined criteria for challenging cases. They also retrieved lung cancer case reports published in the *Journal of Thoracic Oncology* (2014–2024).

The inclusion criteria were as follows: (1) pathologically confirmed lung cancer; (2) at least one challenging feature, such as multisystem involvement, complex comorbidities, rare pathological subtypes, uncommon driver mutations, distinctive clinical phenotypes, or nonresponse to standard therapy or acquired resistance; and (3) availability of complete decision-chain data, including imaging, pathology, laboratory tests, and treatment response records. Cases with incomplete data or those deemed insufficiently challenging upon expert review were excluded.

This study was approved by the institutional ethics committee of General Hospital of Southern Theater Command (Approval No.: NZLLKZ2025023) and was conducted in accordance with the ethical principles of the Declaration of Helsinki. As all study data were thoroughly deidentified, the ethics committee waived the requirement to obtain informed consent.

The two thoracic oncology specialists curated and standardized the selected cases by compiling key clinical information (patient history, physical examination findings, imaging results, and laboratory tests) to develop case-based questionnaires simulating diverse clinical scenarios. The respondents were asked to propose next-step diagnostic and therapeutic strategies for patients or their families. Notably, imaging and pathological findings were provided only as text and without images.

Case selection and questionnaire design were conducted independently of the expert panel responsible for subsequent evaluation, and the two thoracic oncology specialists did not participate in scoring or data analysis. In addition, the same two thoracic oncology specialists

verified the accuracy and consistency of the forward and back translations between the Chinese and English study materials.

2.1.2. Respondents

The respondents consisted of LLMs and human physicians. The LLM group comprised four widely used models: DeepSeek R1, Claude 3.5, Gemini 1.5, and GPT-4o. The physician group consisted of two junior physicians (2 and 3 years of clinical experience, respectively), two intermediate physicians (7 years each), and two senior medical oncologists (14 and 15 years, initially) from tertiary hospitals. The senior medical oncologist with 15 years of clinical experience withdrew during the study for personal reasons and was replaced by another senior oncologist with 13 years of experience, who completed the questionnaire independently. All physicians were instructed not to consult the literature or use LLMs during the response process.

An AI-assisted junior physician group was also established, which consisted of two additional junior oncologists with 2 and 3 years of clinical experience, respectively, who formulated diagnostic and therapeutic decisions with assistance from GPT-4o. These physicians were not involved in dataset processing and remained blinded to the evaluation dimensions and scoring criteria during the response process.

2.1.3. Expert review panel

The expert review panel consisted of three board-certified thoracic oncology specialists, each with over 15 years of clinical experience. The panel led the development and finalization of the study prompt and independently evaluated the responses from all participants. In the final study phase, the panel synthesized human–AI error patterns through collaborative discussion.

2.2. Scoring dimensions and criteria

The expert review panel collaboratively developed a structured scoring framework for evaluating the clinical decision outputs of both physicians and LLMs based on an evidence-oriented, patient-centered approach aligned with the study objectives. The framework assessed three dimensions: (1) comprehensiveness of diagnostic and therapeutic strategies, (2) specificity of treatment planning (degree of individualization), and (3) readability of clinical recommendations (clarity and communicative effectiveness). Each dimension was rated on a five-point Likert scale (1 = very poor, 5 = excellent) using detailed operational definitions (Supplementary Table 1). Inter-rater reliability was assessed in a pilot test, with an intraclass correlation coefficient (ICC) of above 0.75.

2.3. Study design

The study was conducted in three phases. The first phase involved decision generation, during which the research team distributed standardized case questionnaires to all participants. The bilingual Chinese–English questionnaires were initially translated using DeepL (https://www.deepl.com/), then iteratively edited for grammar and terminology using GPT-4o over three rounds, and finally verified by the two aforementioned bilingual experts to ensure semantic equivalence between the two versions (Supplementary Table 2).

The LLM group received the English questionnaires via application programming interfaces (APIs) that enabled batch input and structured, table-based output. All LLMs were accessed via their official APIs and run under each platform's default inference settings to reflect typical real-world clinical use. Specifically, GPT-4o was queried via the OpenAI API under default settings, with sampling parameters such as temperature and top-p preset by the platform; Claude 3.5 via the Anthropic API under its default configuration; Gemini 1.5 via Google AI Studio under default inference settings; and DeepSeek R1, as an open-source model, via its official API under default settings.

Each input consisted of case information, followed by the following fixed prompt: "Please first analyze the case and then inform the patient or family in detail about the next-step diagnostic and therapeutic strategy." The prompt was drafted and iteratively refined by the expert review panel through repeated internal reviews using multiple LLMs, and the final version was prespecified to meet the study task requirements and promote standardized, comparable outputs across models. Three independent queries were administered to each model, with the conversation history reset before each query to ensure that outputs were independent and free of contextual carryover. All responses were recorded in a standardized format, including model identification and full-text outputs.

All physicians completed the Chinese version of the questionnaire under supervised conditions across 3–5 sessions. Participants in the AI-assisted group used GPT-4o as an aid when formulating their diagnostic and therapeutic decisions. The time spent by each group was not recorded.

The second phase involved blinded scoring. The English responses from the LLM group were translated into Chinese using the previously verified process. Before scoring, all responses underwent standardized translation and style normalization (e.g., harmonized paragraph structure, bullet/number formatting, and terminology conventions) and were independently scored by three native Chinese-speaking experts under fully blinded conditions. Reviewers were not informed whether a response originated from a physician or an LLM, minimizing potential bias related to language, formatting, or source attribution.

The third phase involved data analysis. After confirming acceptable inter-rater reliability among the three reviewers, we used a modified majority-rule consensus approach to integrate scores: identical ratings were accepted as is; when two of three ratings agreed, the majority rating was used; and when all three differed, the lowest score was adopted as a conservative approach. The final dataset was standardized within groups, using the mean values of three responses for the LLM group and those of paired physicians for the human and AI-assisted groups, before statistical analysis.

2.4. Statistical analysis

The sample size was determined through a two-sided power analysis based on scoring consistency from preliminary experiments involving five local and five publicly available cases. Using G*Power 3.1, the calculation was performed with $\alpha = 0.05$, test power $(1 - \beta) = 0.80$, and an observed effect size (Cohen's d) of 0.82 derived from the pilot data. The results indicated that at least 46 samples were required to detect effects of comparable magnitudes. To account for potential data loss, the final target sample size was set at 50 cases.

Inter-rater consistency was evaluated using the ICC, interpreted according to Cicchetti's criteria,[22] with ICC values below 0.40 considered poor, between 0.41 and 0.59 fair, between 0.60 and 0.74 good, and between 0.75 and 1.00 excellent. Statistical analyses were performed in Python (version 3.8) using the scipy.stats and statsmodels packages. Data normality was tested using the Shapiro–Wilk test, and homogeneity of variance using Levene's test. When parametric assumptions were met, independent-samples t tests were used for between-group comparisons; otherwise, Mann–Whitney U tests were applied. Effect sizes were calculated using Cohen's d for parametric data and Cliff's delta for nonparametric data. The Benjamini–Hochberg correction was applied to control the false discovery rate for multiple comparisons. Statistical significance was set at $P < 0.05$, 0.01, and 0.001, and data are presented as mean ± standard deviation (SD) unless otherwise specified.

3. Results

3.1. Dataset

Of 562 lung cancer cases recorded in the General Hospital of Southern Theater Command database between January 2020 and

January 2024, 119 were initially identified as challenging. After three rounds of expert review, 60 cases were excluded for insufficient complexity and 34 for incomplete decision-chain data, leaving 25 local challenging cases. Similarly, of 258 lung cancer case reports published in the *Journal of Thoracic Oncology* over the previous decade, 89 were identified as challenging, of which 64 were excluded after expert evaluation, yielding 25 published cases.

These 50 cases were standardized, translated, and verified by two thoracic oncology specialists to develop a bilingual (Chinese–English) clinical questionnaire. This questionnaire was administered to the physicians, AI-assisted physicians, and LLMs, yielding 20 completed decision worksheets. Each worksheet contained responses for all 50 cases, resulting in 1000 case-level responses in total (20 × 50). All responses underwent blinded three-dimensional Likert scale scoring by three independent experts. Accordingly, each case-level response received three scores (one per expert), yielding 3000 scores in total (1000 × 3). For analysis, a consensus score for each response was derived by combining the three expert scores. The study workflow is shown in Figure 1.

3.2. Inter-rater reliability analysis

To reduce rater bias and enhance cross-group comparability, ratings from three independent experts were integrated using a modified majority-consensus approach to derive a consensus score for each response. To address response heterogeneity, we averaged the three independent outputs for each LLM and calculated the mean scores for predefined physician-pair combinations in the human and AI-assisted groups. Inter-rater reliability, assessed using ICCs across the three dimensions, was excellent, with comprehensiveness (ICC = 0.81), specificity (ICC = 0.77), and readability (ICC = 0.87) all meeting Cicchetti's criterion (ICC ≥0.75). Inter-rater reliability remained excellent across decision-maker groups (junior, intermediate, senior, AI-assisted junior, and the four LLMs), with ICCs

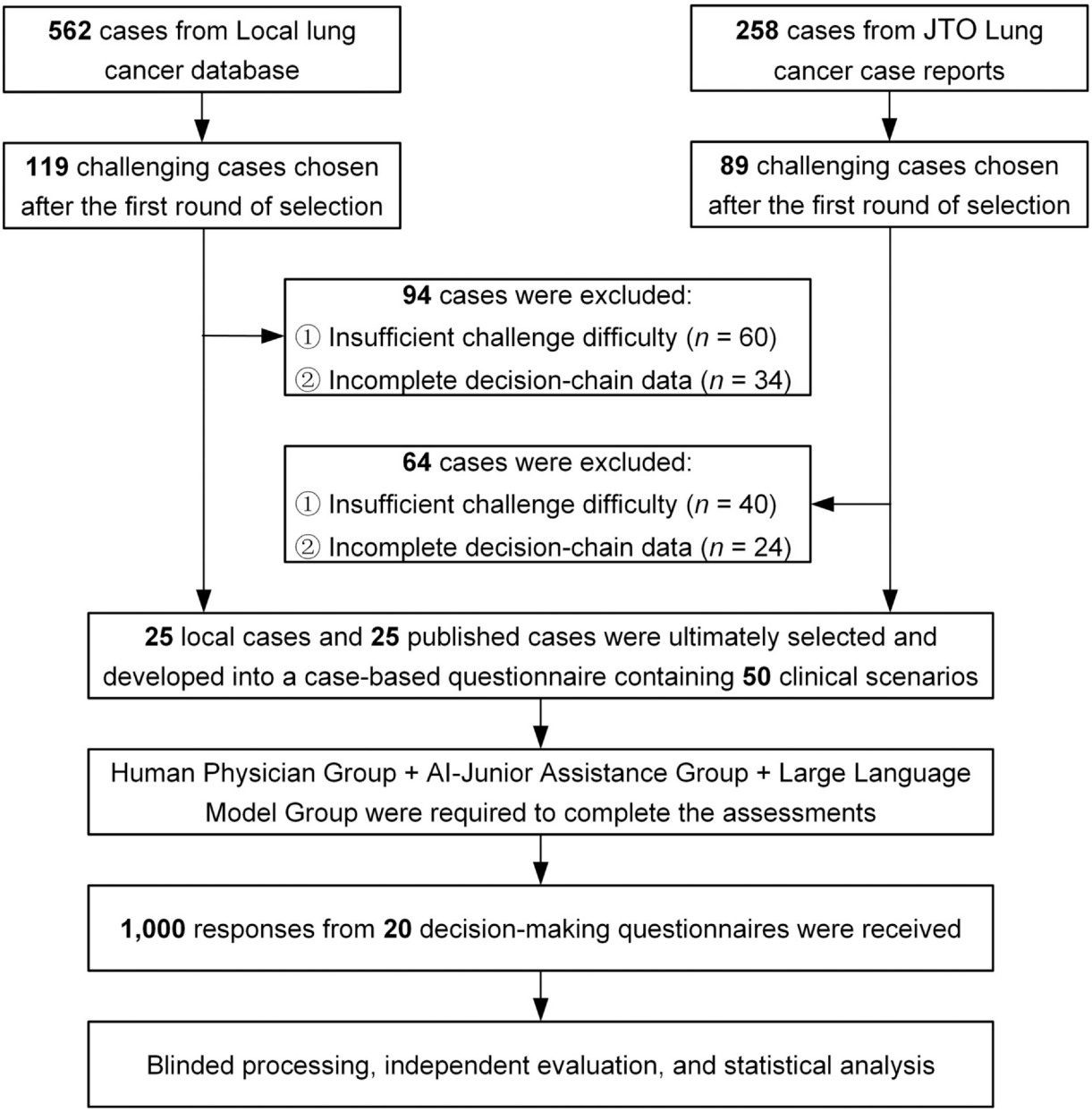


Figure 1. Study workflow from case curation to statistical analysis. Abbreviation: JTO: *Journal of Thoracic Oncology*.

consistently at least 0.75 (Figure 2A), supporting the robustness of the rating system.

3.3. Overall three-dimensional performance

A three-dimensional rubric was applied for cross-group comparisons (Figure 2B). Human physicians showed an experience-associated increase across junior, intermediate, and senior levels in comprehensiveness (2.64 → 3.68 → 4.50), specificity (2.67 → 3.68 → 4.64), and readability (2.93 → 3.75 → 4.53; one-way ANOVA, $P < 0.001$). Of the LLMs, DeepSeek R1 showed a relatively balanced profile (comprehensiveness 3.95 ± 0.33, specificity 3.71 ± 0.53, readability 4.26 ± 0.18) numerically between intermediate and senior physicians. GPT-4o and Claude 3.5 reached the intermediate range in comprehensiveness (3.76 ± 0.39 and 3.60 ± 0.39) but were lower in specificity (3.39 ± 0.39 and 3.39 ± 0.49), suggesting performance closer to the junior-to-intermediate range. Gemini 1.5 exceeded the intermediate benchmark in readability (3.90 ± 0.29), while comprehensiveness (3.41 ± 0.38) and specificity (3.17 ± 0.49) remained lower. The AI-assisted junior group exceeded the intermediate benchmark across all dimensions (comprehensiveness 4.07 ± 0.27, specificity 3.69 ± 0.60, readability 4.39 ± 0.21).

3.4. Three-dimensional performance across case types

Fifty cases were categorized by their core challenging features into three types: complex ($n = 18$), involving multisystem disease or intricate comorbid interactions; rare ($n = 17$), characterized by uncommon histologies, low-frequency driver mutations, or atypical metastatic patterns; and refractory ($n = 15$), exhibiting primary or secondary resistance to first-line targeted therapy, immune checkpoint inhibitors, or platinum-based chemotherapy. The case grouping is detailed in Supplementary Table 2.

As shown in Figure 3A, the physician groups exhibited a consistent experience-dependent gradient across all case categories, supporting the consistency of the stratification. Multidimensional visualization (Figure 3B–3C) showed that physicians at all levels performed lowest in rare cases (comprehensiveness 2.32–4.26; specificity 2.24–4.44; readability 2.76–4.35) compared with complex (2.61–4.58; 2.61–4.78; 2.92–4.69) and refractory cases (3.00–4.67; 3.17–4.70; 3.10–4.53), possibly reflecting

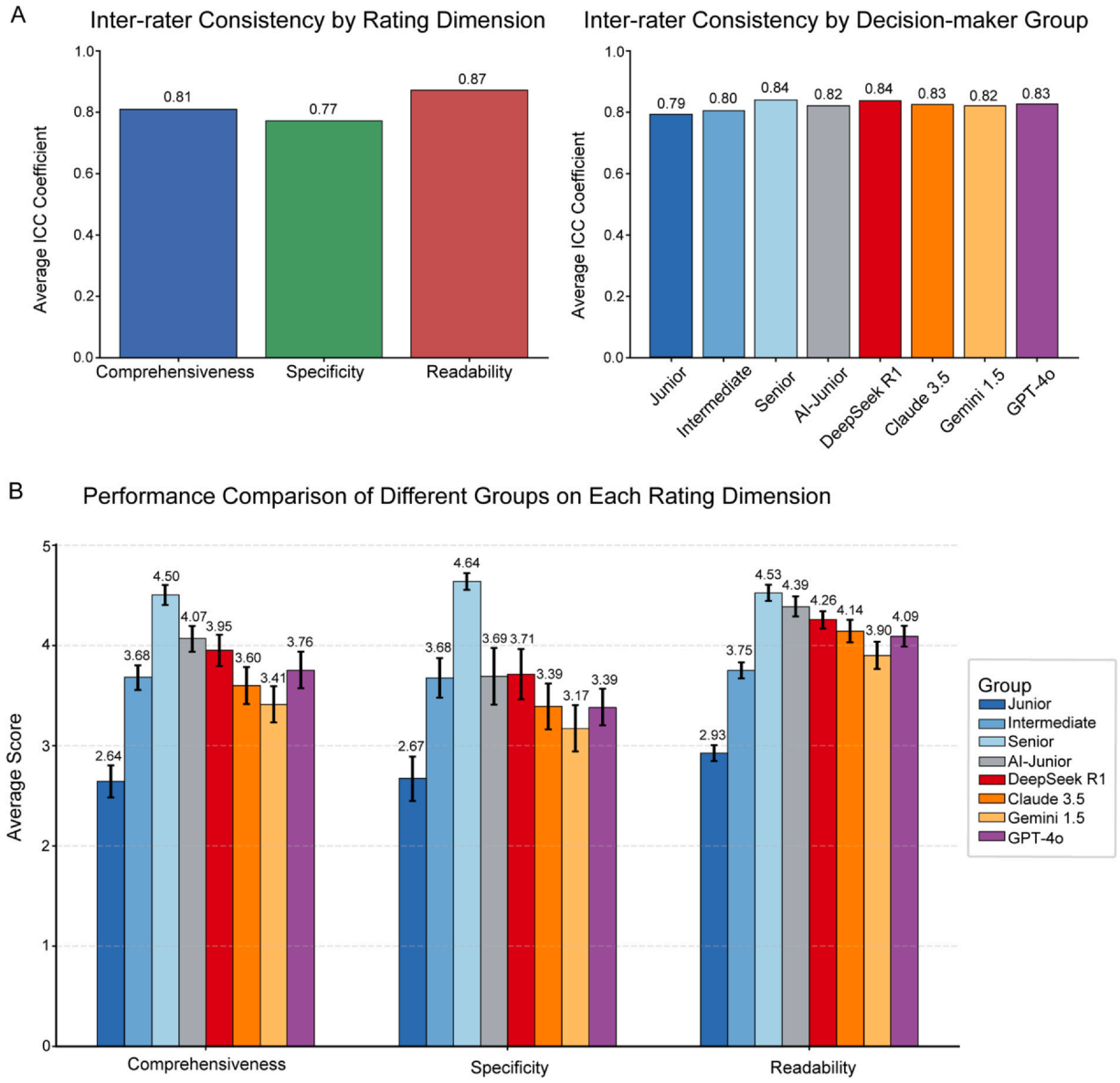


Figure 2. Inter-rater reliability and multidimensional performance across decision-maker groups. A, Inter-rater consistency (ICC) across the three rating dimensions; B, Performance of each group by dimension (comprehensiveness, specificity, and readability).

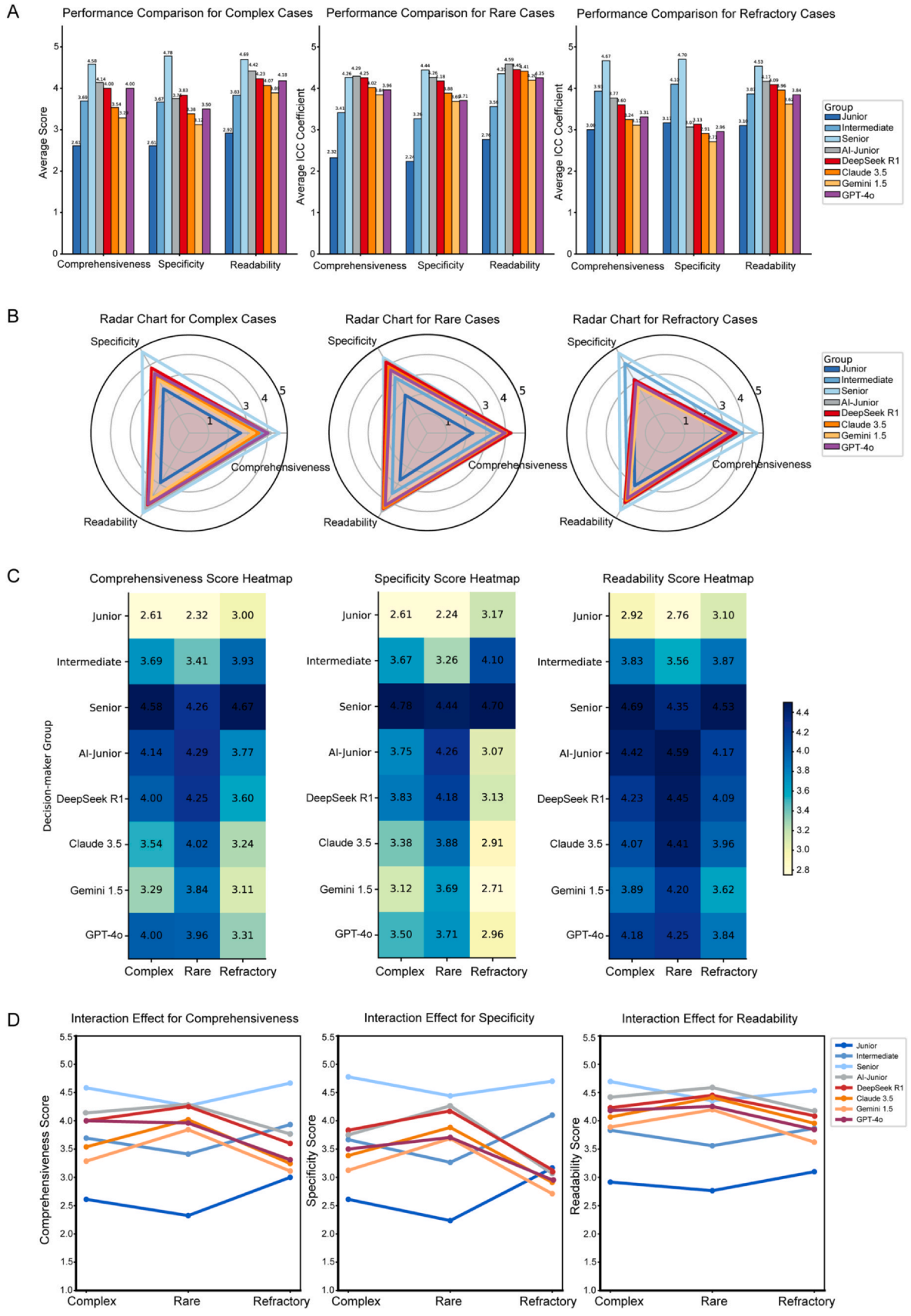


Figure 3. Heterogeneity of three-dimensional performance across case-type strata (complex, rare, and refractory). A, Dimension-wise score distributions by case type; B, Radar profiles of three-dimensional performance across case types; C, Heatmap of three-dimensional scores across case types; D, Interaction effects between decision-maker groups and case types.

cognitive constraints under closed book conditions. In contrast, all four LLMs achieved rare case scores exceeding those of the intermediate physicians. However, in refractory cases requiring longitudinal clinical reasoning, all models performed worse and specificity fell below that of the junior physicians. Interaction analysis (Figure 3D) suggested complementary performance patterns between physicians and LLMs across case types. These cross-type divergences may reflect differences in human versus model decision processes and suggest that linking real-world challenging-case typologies to task design could help optimize human–AI collaboration.

3.5. Analysis of assistive enhancement effects

The effects of LLM assistance for the junior physicians are shown in Figure 4A–4C. Across case types, AI assistance yielded varying degrees of improvement. In rare cases, comprehensiveness, readability, and specificity increased from 2.32 to 4.29 (1.97; 84.9% relative increase), from 2.76 to 4.59 (1.823; 66.3%), and from 2.24 to 4.26 (2.02; 90.2%), respectively. Complex cases showed consistent but smaller gains: comprehensiveness 1.53 (58.6%), readability 1.50 (51.4%), and

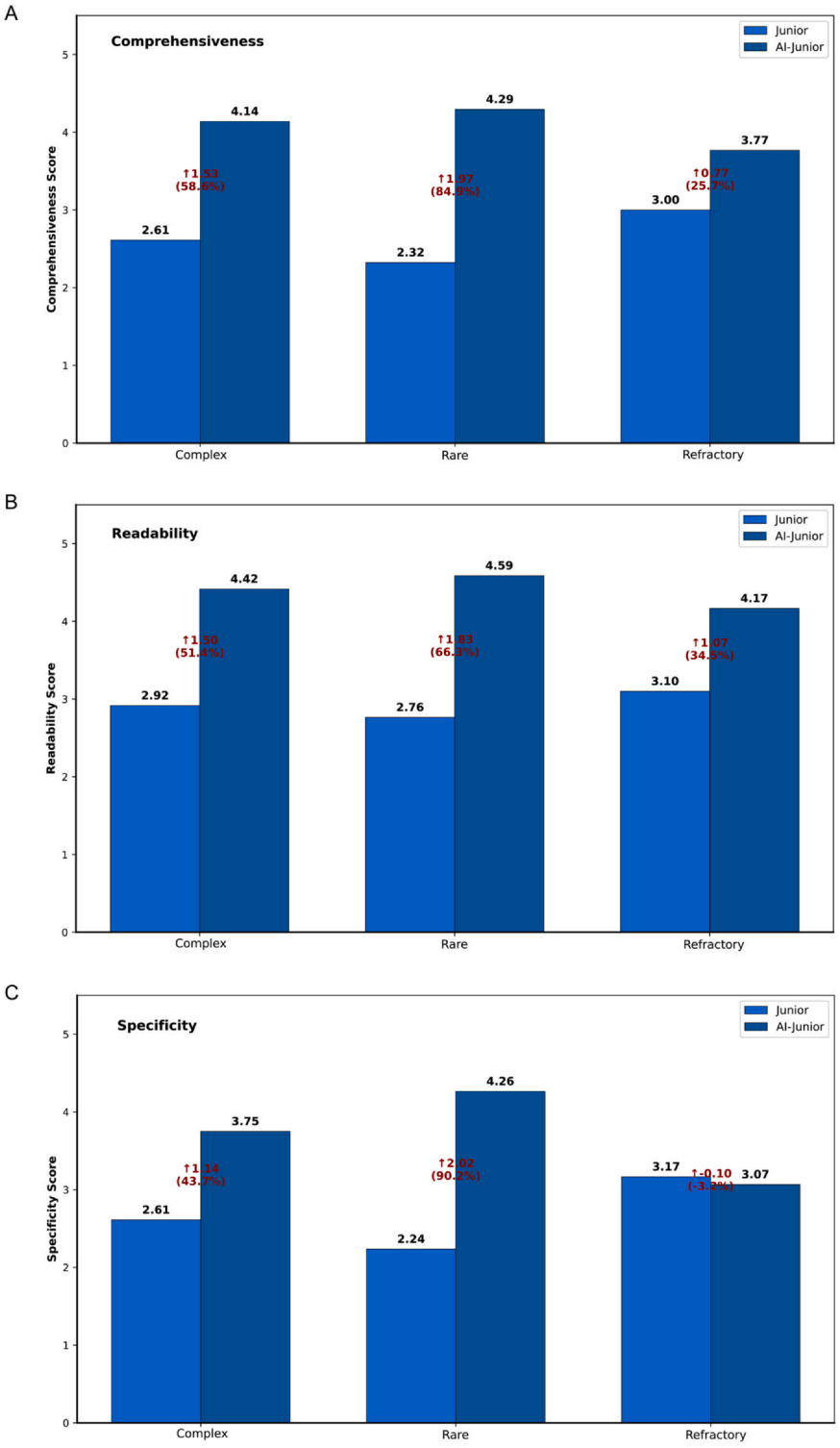


Figure 4. Assistive gains from large language models across case types. A, Comprehensiveness; B, Readability; C, Specificity.

specificity 1.14 (43.7%). In refractory cases, comprehensiveness and readability increased modestly (0.77, 25.7%; 1.07, 34.5%), whereas specificity declined slightly, from 3.17 to 3.07 (−0.10; −3.2%).

3.6. Human–machine error landscape and synergy

Four error categories comprising 12 patterns were identified across GPT-4o and intermediate-level clinicians in lung cancer decision-making. Knowledge-based errors refer to the inaccurate use of medical knowledge, including knowledge-authenticity errors, delayed adoption of updated guidance, and misalignment between the strength of evidence and clinical recommendations. Logical errors refer to breaks in the reasoning chain, such as incomplete differential diagnosis, incorrect causal attribution, and improper sequencing of clinical priorities. Individualized adaptation errors refer to the failure to tailor decisions to patient-specific constraints, such as organ function reserve, comorbidities, and practical considerations. Risk control errors refer to insufficient anticipation and mitigation of safety risks, including inadequate risk identification, missing preventive measures, and a lack of dynamic monitoring (Supplementary Table 3A).

Comparative profiling indicated greater LLM susceptibility to knowledge-authenticity errors (e.g., referencing nonexistent agents or studies), whereas physicians more frequently exhibited knowledge-timeliness gaps and misalignment with the evidence hierarchy (Figure 5, Supplementary Table 3B). In reasoning-intensive settings, physicians outperformed LLMs in causal attribution and organ function–based treatment adjustment, consistent with a more robust integration of clinicopathologic context and physiologic reserve into decision-making. In contrast, LLMs showed fewer omissions in differential diagnosis and greater consistency in preventive and dynamic risk-monitoring execution. Given the limited sample size and descriptive statistical nature of this analysis, these findings should be interpreted as indicative associations rather than definitive causal inferences.

4. Discussion

AI in oncology has progressed in several complementary directions. Deep learning approaches have been widely applied to extract clinically meaningful information from medical images (e.g., computed tomography and whole-slide pathology), enabling tasks such as histologic subtype classification, risk stratification, and identification of prognostically relevant patterns in lung adenocarcinoma.[23–25] In parallel, LLMs can work with unstructured clinical narratives by synthesizing heterogeneous patient data into decision-oriented text, thereby positioning them as a potential workflow-facing layer for clinical decision support.

This study developed a multidimensional evaluation system using 50 challenging lung cancer cases to compare LLMs with physicians across experience levels and to examine potential human–AI complementarity. Overall, advanced LLMs matched or exceeded intermediate-level physicians in comprehensiveness, specificity, and readability, with the clearest advantage in rare cases. Performance declined in refractory or relapsed scenarios requiring longitudinal reasoning, in which LLMs generally underperformed junior physicians. AI assistance produced the largest gains in rare cases and smaller gains in complex and refractory cases, with a slight decrease in specificity in the refractory subgroup. Error-pattern profiling indicated complementary strengths: LLMs more consistently provided up-to-date knowledge, structured differential diagnoses, and standardized risk alerts, whereas physicians more reliably integrated clinicopathologic context and physiologic reserve, supporting case- and task-tailored collaboration.

We used real-world challenging cases as evaluation benchmarks, addressing the limitations of prior research that relied on standardized exams or simulated scenarios.[15,17,26–28] Compared with simplified, artificially constructed settings, real cases embed contextual complexity, fragmented information, and decision risks, thereby more closely reflecting model performance in complex clinical environments.[29,30] Recognizing concerns that the direct use of published cases may

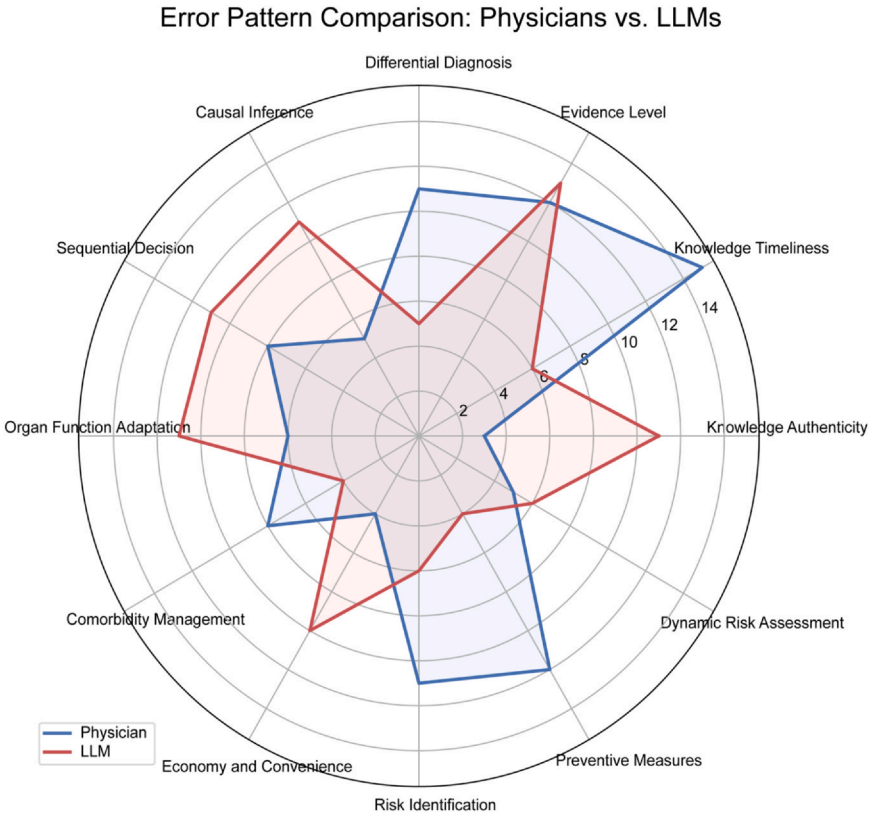


Figure 5. Radar chart of complementary error domains between physicians and large language models. Abbreviation: LLMs: large language models.

overestimate performance due to potential information leakage,[31,32] this study implemented a three-stage process to mitigate bias: (1) substantive reconstruction of case published in *Journal of Thoracic Oncology* by adjusting demographics, key laboratory values, and treatment timelines while preserving the core clinical logic, thereby reducing verbatim retrievability; (2) strict de-identification of local cases by removing 12 categories of identifiers (e.g., imaging numbers and visit dates); and (3) LLM-assisted semantic rephrasing of case narratives to reduce memorization effects while preserving medical meaning and avoiding the introduction of new clinical content. This dual-source, triple-protection design was intended to retain the ecological validity of real cases while reducing the likelihood of memorized responses to public data.[33]

Previous assessments of LLMs' clinical efficacy have often relied on single-dimension accuracy metrics (e.g., diagnostic correctness or exam pass rates), with limited multidimensional evaluation or benchmarking against physicians at different experience levels, making it difficult to situate LLMs in clinical practice.[17,34–36] Zhang et al.[37] proposed a multidimensional framework for radiology report evaluation (e.g., consistency, correctness, comprehensiveness, and harmlessness), but its dimensions are not directly transferable to clinical decision assessment and show conceptual overlap, whereas decision evaluation benefits from clearly differentiated domains. The five-dimension framework by Tam et al.[38] (e.g., information quality, comprehension reasoning, and expression style) captures core elements of clinical applications yet insufficiently addresses human–machine collaborative evaluation needs. Building on prior work and study requirements, we constructed a three-dimensional evaluation system for medical decision-making (comprehensiveness, specificity, and readability) with a multi-level physician benchmarking design. The value of this framework lies in shifting the focus from "whether LLMs replace physicians" to "how to optimize human–AI collaboration." As Selman et al.[39] emphasized, "Artificial intelligence should continue to be viewed as an evolving resource rather than a substitute for clinical judgment."

This study found that LLM decision-making performance varied by case type. For rare conditions (e.g., low-frequency driver mutations), LLMs achieved higher scores, likely reflecting broad static knowledge coverage and efficient learning from limited exemplars.[40,41] However, performance was lower in refractory or relapsed cases, which may relate to limitations in longitudinal clinical reasoning, including synthesis of treatment history, interpretation of resistance mechanisms, and individualized regimen adjustment.[42] For example, in "Local 3," the model recommended investigational agents and generated hallucinated targeted drugs for PIK3CA/FGFR3/CCND1 after EGFR-TKI resistance. In "Local 4," it did not recognize primary ROS1 resistance and suggested continuing ROS1-TKI therapy. These observations suggest the need to assess model applicability by case type and to apply added caution when consulting model outputs for refractory or relapsed cases.

The human–AI complementarity observed in this study provides pragmatic directions for optimizing clinical decision support. Our error pattern analysis suggests "misaligned but complementary" strengths between LLMs and clinicians, indicative of the concept of hybrid augmented intelligence,[43] and supports collaborative designs that combine rapid evidence synthesis with real-time clinical verification. With integrated AI-based standardized knowledge support and physician-led individualized decision-making, both efficiency and decision quality may improve.

This study has several limitations. First, the sample size was modest (50 cases), the scope was limited to a single cancer type, and both the physician cohort and the set of evaluated LLMs were relatively small, which may limit generalizability. Second, we relied on simulated, text-based decision scenarios rather than real clinical workflows, which may overestimate performance in dynamic care settings and do not directly test longitudinal reasoning. Third, we did not systematically explore model parameters; instead, we evaluated each model under standardized settings, so our results reflect typical real-world use rather than peak performance under optimal tuning. Fourth, although the review was blinded, residual bias related to the AI-generated writing style may still have influenced ratings. Accordingly, the findings are hypothesis-generating and should be interpreted as associations rather than causal evidence. Future studies should expand disease coverage, adopt more dynamic evaluation frameworks, and include prospective real-world validation.

5. Conclusions

LLM performance in clinical decision tasks varied by case type, being better in rare cases and worse in refractory cases requiring longitudinal reasoning. Complementary strengths between LLMs and physicians support case- and task-tailored human–AI collaboration. These hypothesis-generating findings warrant broader disease coverage, dynamic context-aware evaluation, and prospective real-world validation.

CRediT authorship contribution statement

Hengguo Zhang: Writing – review & editing, Writing – original draft, Methodology, Data curation. **Xinpei Deng:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Chang Qi:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Wenyi Gan:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Quan Cheng:** Writing – review & editing, Writing – original draft, Supervision, Data curation. **Lingxuan Zhu:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Bufu Tang:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Weiming Mou:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Anqi Lin:** Writing – review & editing, Writing – original draft, Supervision, Software, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Dongqiang Zeng:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Juan Zhou:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Mingjia Xiao:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Peng Luo:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Ning Yang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Guangdi Chu:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Shengkun Peng:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Kailai Li:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Data curation, Conceptualization. **Hank Z.H. Wong:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Baiyang Liu:** Writing – review & editing, Writing – original draft, Software, Methodology, Data curation. **Lin Zhang:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Xiting Chen:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Aimin Jiang:** Writing – review & editing, Writing – original draft, Methodology, Data curation.

Ethical approval and consent to participate

This study was approved by the institutional ethics committee (Approval No.: NZLLKZ2025023) and strictly adhered to the ethical standards outlined in the Declaration of Helsinki. Given that all research data underwent a thorough deidentification process, the ethics committee waived the requirement for informed consent from the subjects.

Funding

Not applicable.

Declaration of Competing Interest

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

Acknowledgments

We acknowledge Sogen Biotechnology, Zhengzhou, Henan Province, and the SolvingLab team for their technical support and assistance.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.intonc.2026.100039.

References

1. Meyer ML, Peters S, Mok TS, et al. Lung cancer research and treatment: global perspectives and strategic calls to action. *Ann Oncol.* 2024;35(12):1088–1104.
2. Daly R, Hetherington K, Hazell E, et al. Precision medicine is changing the roles of healthcare professionals, scientists, and research staff: learnings from a childhood cancer precision medicine trial. *J Pers Med.* 2023;13(7):1033.
3. Kenny K, Broom A, Page A, et al. A sociology of precision-in-practice: the affective and temporal complexities of everyday clinical care. *Socio Health Illn.* 2021;43(9):2178–2195.
4. Al Zaidi M, Wright GM. Locally advanced non-small cell lung cancer: the place of specialist thoracic surgery in the multidisciplinary team. *Transl Lung Cancer Res.* 2020;9(4):1680–1689.
5. Barr T, Ma S, Li Z, Yu J. Recent advances and remaining challenges in lung cancer therapy. *Chin Med J.* 2024;137(5):533–546.
6. Cheng J, Xiao M, Meng Q, et al. Decoding temporal heterogeneity in NSCLC through machine learning and prognostic model construction. *World J Surg Oncol.* 2024;22(1):156.
7. Zasada M, Harris J, Groothuizen J, et al. Investigating the efficiency of lung multidisciplinary team meetings-a mixed methods study of eight lung multi-disciplinary teams. *Cancer Med.* 2023;12(8):9999–10007.
8. Brown GTF, Bekker HL, Young AL. Quality and efficacy of Multidisciplinary Team (MDT) quality assessment tools and discussion checklists: a systematic review. *BMC Cancer.* 2022;22(1):286.
9. Patel P, Bradley SH, McCutchan G, Brain K, Redmond P. What should the role of primary care be in lung cancer screening? *Br J Gen Pract.* 2023;73(733):340–341.
10. 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.
11. Schmidl B, Hütten T, Pigorsch S, et al. Assessing the role of advanced artificial intelligence as a tool in multidisciplinary tumor board decision-making for recurrent/metastatic head and neck cancer cases - the first study on ChatGPT-4o and a comparison to ChatGPT-4.0. *Front Oncol.* 2024;14:1455413.
12. Deng L, Wang T, Zhang Y, et al. Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2. *Int J Surg.* 2024;110(4):1941–1950.
13. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. *JAMA Netw Open.* 2023;6(11):e2343689.
14. Yau JY, Saadat S, Hsu E, et al. Accuracy of prospective assessments of 4 large language model chatbot responses to patient questions about emergency care: experimental comparative study. *J Med Internet Res.* 2024;26:e60291.
15. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. *Nat Med.* 2025;31(3):943–950.
16. Ding H, Xia W, Zhou Y, et al. Evaluation and practical application of prompt-driven ChatGPTs for EMR generation. *npj Digit Med.* 2025;8(1):77.
17. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA.* 2025;333(4):319–328.
18. Giuffrè M, Kresevic S, You K, et al. Systematic review: the use of large language models as medical chatbots in digestive diseases. *Aliment Pharmacol Ther.* 2024;60(2):144–166.
19. Yang X, Li T, Su Q, et al. Application of large language models in disease diagnosis and treatment. *Chin Med J.* 2025;138(2):130–142.
20. Steyvers M, Kumar A. Three challenges for AI-assisted decision-making. *Perspect Psychol Sci.* 2024;19(5):722–734.
21. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med.* 2025;31(1):60–69.
22. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess.* 1994;6(4):284–290.
23. Ding H, Feng Y, Huang X, et al. Deep learning-based classification and spatial prognosis risk score on whole-slide images of lung adenocarcinoma. *Histopathology.* 2023;83(2):211–228.
24. Feng Y, Ding H, Huang X, et al. Deep learning-based detection and semi-quantitative model for spread through air spaces (STAS) in lung adenocarcinoma. *npj Precis Oncol.* 2024;8(1):173.
25. Ding H, Xia W, Zhang L, et al. CT-based deep learning model for invasiveness classification and micropapillary pattern prediction within lung adenocarcinoma. *Front Oncol.* 2020;10:1186.
26. Staal J, Hooftman J, Gunput STG, et al. Effect on diagnostic accuracy of cognitive reasoning tools for the workplace setting: systematic review and meta-analysis. *BMJ Qual Saf.* 2022;31(12):899–910.
27. Omega A, Wijaya Ramlan AA, Soenarto RF, Heriwardito A, Sugiarto A. Assessing clinical reasoning in airway related cases among anesthesiology fellow residents using Script Concordance Test (SCT). *Med Educ Online.* 2022;27(1):2135421.
28. Lai H, Liu J, Bai C, et al. Language models for data extraction and risk of bias assessment in complementary medicine. *npj Digit Med.* 2025;8(1):74.
29. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA.* 2023;330(1):78–80.
30. Nori H, King N, McKinney S.M, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:230313375*. Published online 2023.
31. Rajore T, Chandran N, Sitaram S, et al. Truce: Private benchmarking to prevent contamination and improve comparative evaluation of llms. *arXiv preprint arXiv:240300393*. Published online 2024.
32. Yang X, Li T, Wang H, et al. Multiple large language models versus experienced physicians in diagnosing challenging cases with gastrointestinal symptoms. *npj Digit Med.* 2025;8(1):85.
33. Wang B, Chen W, Pei H, et al. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. *arXiv preprint arXiv:230611698*. Published online 2023.
34. Wu SH, Tong WJ, Li MD, et al. Collaborative enhancement of consistency and accuracy in US diagnosis of thyroid nodules using large language models. *Radiology.* 2024;310(3):e232255.
35. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine.* 2023;95:104770.
36. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J Med Internet Res.* 2023;25:e48659.
37. Zhang L, Liu M, Wang L, et al. Constructing a large language model to generate impressions from findings in radiology reports. *Radiology.* 2024;312(3):e240885.
38. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digit Med.* 2024;7(1):258.
39. Selman F, Obletz K, Vismara V, Putko R, Perry NPJ. Editorial commentary: artificial intelligence and language learning models can be improved by curated input of medical training data but still face the limitations of available literature and require continued human oversight. *Arthroscopy.* 2025;41(8):2758–2760.
40. Abdullahi T, Singh R, Eickhoff C. Learning to make rare and complex diagnoses with generative ai assistance: qualitative study of popular large language models. *JMIR Med Educ.* 2024;10:e51391.
41. Li T, Shetty S, Kamath A, et al. CancerGPT for few shot drug pair synergy prediction using large pretrained language models. *npj Digit Med.* 2024;7(1):40.
42. Bentzen SM. Artificial intelligence in health care: a rallying cry for critical clinical research and ethical thinking. *Clin Oncol (R Coll Radio).* 2025;41:103798.
43. Strieth-Kalthoff F, Szymkuć S, Molga K, Aspuru-Guzik A, Glorius F, Grzybowski BA. Artificial intelligence for retrosynthetic planning needs both data and expert knowledge. *J Am Chem Soc.* 2024. https://doi.org/10.1021/jacs.4c00338 Published online April 10.