Journal of Hepatocellular Carcinoma

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

A Comparative Study of Five Large Language Models' Response for Liver Cancer Comprehensive Treatment

Deyuan Zhong, Yuxin Liang, Hong-Tao Yan, Xinpei Chen, Qinyan Yang, Shuoshuo Ma, Yuhao Su, YaHui Chen, Xiaolun Huang, Ming Wang

Department of Liver Transplantation Center and HBP Surgery, Sichuan Clinical Research Center for Cancer, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, School of Medicine, University of Electronic Science and Technology of China, Chengdu, People's Republic of China

Correspondence: Xiaolun Huang; Ming Wang, Email huangxiaolun@med.uestc.edu.cn; wangming0610@163.com

Introduction: Large language models (LLMs) are increasingly used in healthcare, yet their reliability in specialized clinical fields remains uncertain. Liver cancer, as a complex and high-burden disease, poses unique challenges for AI-based tools. This study aimed to evaluate the comprehensibility and clinical applicability of five mainstream LLMs in addressing liver cancer–related clinical questions.

Methods: We developed 90 standardized questions covering multiple aspects of liver cancer management. Five LLMs—GPT-4, Gemini, Copilot, Kimi, and Ernie Bot—were evaluated in a blinded fashion by three independent hepatobiliary experts. Responses were scored using predefined criteria for comprehensibility and clinical applicability. Overall group comparisons were conducted using the Fisher–Freeman–Halton test (for categorical data) and the Kruskal–Wallis test (for ordinal scores), followed by Dunn's post-hoc test or Fisher's exact test with Bonferroni correction. Inter-rater reliability was assessed using Fleiss' kappa.

Results: Kimi and GPT-4 achieved the highest proportions of fully applicable responses (68% and 62%, respectively), while Ernie Bot and Copilot showed the lowest. Comprehensibility was generally high, with Kimi and Ernie Bot scoring over 98%. However, none of the LLMs consistently provided guideline-concordant answers to all questions. Performance on professional-level questions was significantly lower than on common-sense ones, highlighting deficiencies in complex clinical reasoning.

Conclusion: LLMs demonstrate varied performance in liver cancer–related queries. While GPT-4 and Kimi show promise in clinical applicability, limitations in accuracy and consistency—particularly for complex medical decisions—underscore the need for domain-specific optimization before clinical integration.

Trial Registration: Not applicable.

Keywords: large language models, liver cancer, clinical applicability, ChatGPT, medical chatbot

Introduction

The expanding role of artificial intelligence (AI) in clinical decision-making and patient care has drawn increasing attention.[1] AI chatbots, functioning as virtual assistants on digital platforms, allow users to engage in natural language conversations and access a broad range of services.[2] Advances in natural language processing (NLP) have extended their application across disease prevention, diagnostics, therapeutics, and patient support.[3,4] Early research has yielded auspicious results, underscoring the transformative potential of AI chatbots in the healthcare landscape.[5] Notably, Large Language Models (LLMs) have experienced marked evolution in comparison to conventional NLP models. These models leverage self-supervised learning methodologies and are trained on expansive textual datasets, thereby enabling the generation of human-esque textual responses. Consequently, the role of LLMs in the dissemination of patient information has emerged as a domain of extensive scholarly inquiry.[6]

Received: 15 April 2025
Accepted: 8 August 2025
Published: 20 August 2025

Unlike retrieval-based medical chatbots that rely on curated databases, LLMs have leveraged deep learning technologies to become promising instruments for demystifying complex medical information.[7] A case in point is ChatGPT, which is predicated on the Generative Pre-trained Transformer 3.5 (GPT-3.5) and GPT-4 models, utilizing a Transformer architecture and Reinforcement Learning from Human Feedback (RLHF).[8] Capable of generating conversational text, ChatGPT also engages in language translation, diverse content creation, and provides responses across a spectrum of subjects. It has even exhibited performance that eclipses human experts in the Massive Multitask Language Understanding (MMLU) evaluation. Attaining graduate-level proficiency in multiple domains, ChatGPT has demonstrated its substantial potential as an ancillary tool.[9] Inspired by this success, major companies have released their own models, including Google's Gemini, Microsoft's Copilot, Moonshot AI's Kimi, and Baidu's Ernie Bot.

In the digital health era, patients increasingly seek medical information online and engage in self-diagnosis through search engines.[10] With wider access to electronic medical records, the accuracy and reliability of online content play a vital role in patient self-management.[11] However, LLMs are trained on heterogeneous internet sources with inconsistent quality, which can lead to misleading or incorrect outputs—particularly in complex fields such as liver cancer.[12] Users without medical expertise may struggle to assess the credibility of LLM-generated content.[13] Furthermore, models may produce "hallucinations"—plausible yet factually incorrect statements—due to insufficient domain-specific training.[14] Although ChatGPT and similar tools have shown promise in answering liver disease questions, relying on a single model may limit both scope and reliability. Few studies have systematically compared the readability and quality of responses generated by diverse LLMs in high-stakes specialties like oncology.[15]

To address this gap, we systematically evaluated the performance of five representative LLMs—GPT-4, Gemini, Copilot, Kimi, and Ernie Bot—on clinically relevant questions related to liver cancer. A total of 90 questions were compiled based on national and international guidelines and categorized into professional and common-sense domains. Responses were independently assessed by liver cancer experts along two key dimensions: comprehensibility (clarity and readability) and clinical applicability (alignment with evidence-based guidelines and relevance to clinical decision-making). Through this comparative analysis, we aimed to elucidate the potential benefits and risks associated with the use of LLMs in addressing common inquiries related to liver diseases. Preliminary findings revealed considerable variation in performance across the five models, with GPT-4 and Kimi demonstrating stronger clinical applicability, while Kimi and Ernie Bot performed better in terms of response comprehensibility. Furthermore, this study seeks to explore the capacity of these models to enhance patient health literacy and facilitate more effective communication between patients and healthcare providers.

Methods

Study Setting and Ethical Statement

This study was conducted from April 1st to June 20th, 2024, in the Hepatobiliary Surgery Department of the affiliated Cancer Hospital of University of Electronic Science and Technology of China. Ethical review board approval was not required for this research due to its non-invasive nature, which involved no human data, animal models, or biological samples. To minimize potential bias, we implemented precautionary measures, including the clearance of all private browser data. Prior to initiating searches, we ensured a thorough deletion of browser data and established distinct accounts for interactions with each AI chatbot, ensuring clear differentiation between them. Each query was conducted on a separate chat page to ensure the independence of the queries and to optimize the analysis process. Additionally, all response results were archived for subsequent evaluation of readability and quality.

Question Development and Prompt Design

In this study, two hepatobiliary surgeons with senior academic titles and over 15 years of experience in the comprehensive management of liver cancer collaboratively designed a set of 90 clinical questions covering the full continuum of liver cancer care, including prevention, treatment, and follow-up. These questions were constructed based on authoritative guidelines issued by the American Association for the Study of Liver Diseases (AASLD), the European

Association for the Study of the Liver (EASL), and the Chinese Society of Hepatology of the Chinese Medical Association.

To ensure semantic precision and clinical fidelity, all prompts were written and submitted in Chinese, which is the primary language of clinical communication and documentation in the target healthcare setting. While some source materials were derived from international guidelines originally written in English, standardized Chinese phrasing was used to avoid distortion introduced through translation and to ensure alignment with terminology familiar to Chinese clinicians.[16] This approach also facilitates linguistically fair evaluation across multilingual and Chinese-pretrained models, allowing for a realistic simulation of model performance in native usage contexts.

To ensure transparency and reproducibility, all prompts were designed following a standardized single-turn format, in which each question was posed independently, without model-specific instructions, system messages, or contextual cues. The questions were open-ended but focused, simulating realistic patient or clinician inquiries, and did not include multiple-choice or binary yes/no structures. Identical prompts were submitted to all five LLMs, ensuring consistency in evaluation across models. Each question was clearly categorized as either professional or common-sense, based on evidence-based operational definitions. Professional questions required expert-level clinical reasoning through: (a) interpretation of ≥2 interdependent biomarkers or imaging parameters (eg, AFP kinetics combined with PIVKA-II trends), (b) application of stage-specific therapeutic algorithms (eg, second-line treatment selection for BCLC stage C HCC), and (c) integration of advanced multimodal therapies (eg, optimal timing for TACE combined with immune checkpoint inhibitors). Common-sense questions focused on health literacy domains: (a) evidence-based prevention protocols (eg, HBV vaccination schedules), (b) standardized diagnostic preparation (eg, pre-procedural instructions for contrast-enhanced CT), and (c) daily self-management guidance (eg, sodium restriction in ascites management). These questions were further divided into four domains respectively: screening and diagnosis, routine treatment decision-making, liver cancer transformation therapy, neoadjuvant and postoperative adjuvant therapy decision-making, systemic treatment decision-making, and an overview with screening, treatment methods, postoperative care and medication guidance, prognosis, and psychological support.

LLM Interaction Process

From May 1st to June 1st, the research team systematically input these search terms into various LLMs, including GPT4 (https:// chat.Openai.com/), Gemini (https://deepmind.google/technologies/gemini/), Copilot (https://copilot.microsoft.com/), Kimi (https://kimi.moonshot.cn/) and Ernie Bot (https://yiyan.baidu.com/), preserving the exact sequence of the original searches. Each question received a response, and after each answer, the system cleared the traces to prepare for the subsequent query. To ensure raters could not identify the characteristics of different LLMs, all answers were copied into a generic table, removing any LLM-specific identifiers.

Expert Reviewers and Evaluation Criteria

Questions were randomly assigned to raters for evaluation. Question allocation employed stratified block randomization. While stratified randomization ensured group balance, residual confounding from unmeasured complexity factors cannot be fully excluded. Each answer was reviewed by three additional liver cancer experts not involved in the study to verify their accuracy. To ensure rigorous evaluation, three independent liver cancer experts were selected through a standardized process: First, candidates were required to meet three criteria: 1) Hold senior clinical titles (Professor or Chief Physician) in hepatobiliary surgery or oncology; 2) Have ≥10 years of subspecialty experience in hepatocellular carcinoma management, with demonstrated expertise in at least two of the following areas: liver transplantation, interventional therapy, targeted/systemic therapy, or multidisciplinary treatment (MDT); 3) Publish ≥5 peer-reviewed articles on liver cancer in the past 5 years. From 12 qualified candidates, we ultimately selected: Expert A: Director of Hepatobiliary Surgery at a cancer center, specializing in surgical resection and conversion therapy for advanced HCC; Expert B: Lead medical oncologist at a liver transplant center, focused on systemic therapies and post-transplant management; Expert C: Professor of interventional radiology with particular expertise in TACE/HAIC and bridging therapies. All evaluators completed conflict-of-interest disclosures and underwent standardized training using 10 sample questions to ensure consistent application of evaluation criteria prior to formal assessment.

Based on whether the answers aligned with clinical guidelines and practices, they were classified as: fully applicable (Strict adherence to ≥2 authoritative guidelines (AASLD, EASL, CMA); Inclusion of essential parameters (BCLC staging, Child-Pugh classification, biomarker cutoffs); Actionable recommendations with precise therapeutic specifications (eg, sorafenib 400 mg bid); Absence of clinically significant inaccuracies), partially applicable (Addresses primary clinical intent but omits ≥1 critical parameter; Provides guideline-concordant principles without operational details; Contains minor inaccuracies not impacting therapeutic decisions), or not applicable (Direct contradictions with guideline-endorsed protocols; Major factual errors (eg, contraindicated drug combinations); Non-specific recommendations lacking HCC relevance). Moreover, assessing the comprehensibility of text based on its complexity and clarity involves evaluating how easily a reader can understand and interpret the content, each answer was also categorized as: easy to understand (≤8th-grade reading level (Flesch-Kincaid), <1 unexplained jargon/100 words, structured presentation, and <5% ambiguous terms) or difficult to understand (Requires postgraduate-level comprehension (Flesch-Kincaid >12), ≥3 unexplained abbreviations/100 words, or >10% non-quantified descriptors). Based on these assessments, we calculated the percentage of scores in each category. By majority vote, we summarized the proportion of classifications for each LLM response by the three physicians, ensuring the objectivity and accuracy of the evaluation. To assess the consistency among reviewers, Fleiss' Kappa was calculated for each dimension (clinical applicability and interpretability) across all items rated by the three experts.

Expert Consent and Data Protection

All participating experts provided written informed consent through institutional academic collaboration agreements, explicitly outlining: 1) Role Definition: Advisory capacity limited to evaluating anonymized LLM outputs against clinical standards, with no access to patient data or involvement in clinical decisions; 2) Data Protection: Implementation of dual anonymization through metadata removal and data pseudonymization; 3) Conflict Management: Mandatory disclosure of financial/non-financial interests related to AI or hepatology innovations within 3 years, followed by blinded evaluation protocols; 4) Voluntary Participation: Unrestricted withdrawal rights without academic or professional penalty, accompanied by immediate disassociation of participant-linked data from the study dataset.

Statistical Analysis

All data were analyzed using SPSS (IBM Corporation) and R (R Core Team, 2022). For binary variables such as response comprehensibility, an overall comparison across the five models was conducted using Fisher's exact test for r×c tables (Freeman–Halton extension). Pairwise comparisons were performed using Fisher's exact test with Bonferroni correction. For ordinal outcomes like clinical applicability scores, Kruskal–Wallis tests were used for overall comparisons, followed by Dunn's post hoc tests with Bonferroni adjustment. Differences between professional and common-sense questions were assessed using the Mann–Whitney *U*-test. Inter-rater agreement was evaluated using Fleiss' Kappa for both outcome dimensions. A p-value < 0.05 was considered statistically significant.

Results

Inter-Rater Reliability

To ensure consistency in expert evaluations, inter-rater agreement was assessed using Fleiss' Kappa statistic. The results indicated substantial agreement for both clinical applicability ($\kappa = 0.754$, $p < 0.001$) and response comprehensibility ($\kappa = 0.671$, $p < 0.001$), supporting the robustness and reliability of the three-reviewer evaluation framework.

Comprehensibility of LLM Responses

The overall comprehensibility of LLM-generated answers demonstrated high interpretability across all five models. Ernie Bot achieved the highest interpretability rate (100%, 90/90), followed by Kimi (99%, 89/90), Gemini (96%, 86/90), Copilot (93%, 84/90), and GPT-4 (91%, 82/90) (Table 1). Statistical analysis revealed significant differences in comprehensibility among the models ($p = 0.007$). Pairwise comparisons further indicated that Ernie Bot significantly

Table 1 Comprehensibility of Responses From LLMs

LLMs	Number	Comprehensibility	Incomprehensibility	*p*
GPT4	90	82(91)	8(9)	0.007
Gemini	90	86(96)	4(4)	
Copilot	90	84(93)	6(7)	
Kimi	90	89(99)	1(1)	
Ernie Bot	90	90(100)	0	

Abbreviations: LLM, large language model; Numbers in parentheses represent percentages.

outperformed Copilot ($p = 0.038$) and GPT-4 ($p = 0.011$), while Kimi also showed a significant advantage over GPT-4 ($p = 0.040$) (Figure 1). No other pairwise differences were statistically significant.

Clinical Applicability Evaluation

The clinical applicability of the responses generated by the five LLMs was evaluated based on a three-level rating scale: fully applicable, partially applicable, and not applicable. The distribution of scores across 90 questions revealed notable differences among models. GPT-4 yielded 62% (56/90) fully applicable, 31% (28/90) partially applicable, and 7% (6/90) not applicable responses. Gemini demonstrated 58% (52/90) fully applicable, 31% (28/90) partially applicable, and 11% (10/90) not applicable. Copilot showed a less favorable distribution, with 44% (40/90) fully applicable, 36% (32/90) partially applicable, and 20% (18/90) not applicable. Kimi performed the best in this regard, achieving 68% (61/90) fully applicable, 26% (23/90) partially applicable, and only 7% (6/90) not applicable. In contrast, Ernie Bot had 33% (30/90)

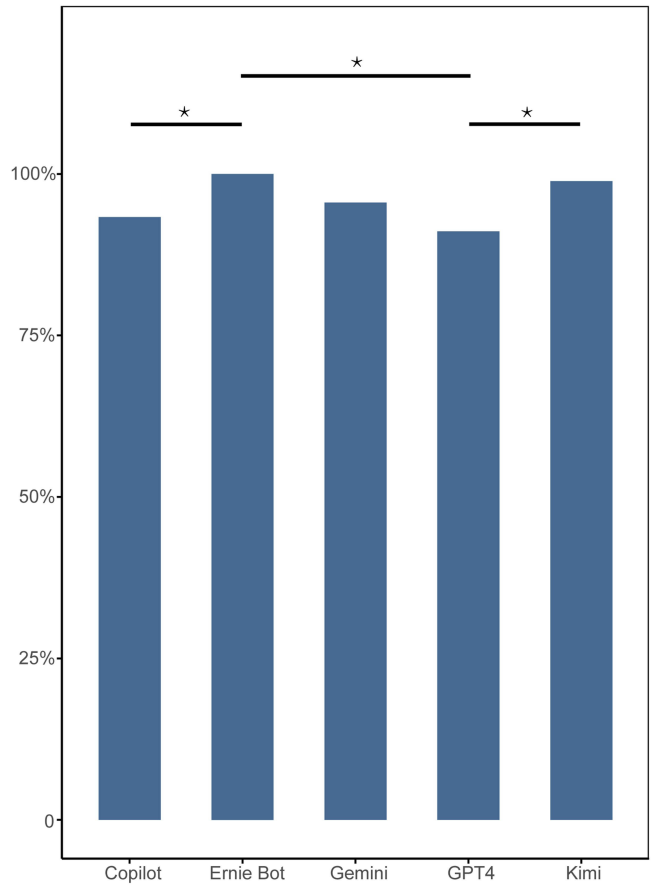

Figure 1 Statistical Variances in Comprehensibility Among LLM Responses. Bar plot comparing the comprehensibility rates of responses generated by different LLMs. The y-axis represents the percentage of comprehensible answers. Statistical significance: *p < 0.05.

https://doi.org/10.2147/JHC.S531642

Table 2 Clinical Applicability of Responses From LLMs

LLMs	Number	Complete Applicability	Partial Applicability	Inapplicability	Z	*p*
GPT4	90	56(62)	28(31)	6(7)	29.324	<0.001
Gemini	90	52(58)	28(31)	10(11)		
Copilot	90	40(44)	32(36)	18(20)		
Kimi	90	61(68)	23(26)	6(7)		
Ernie Bot	90	30(33)	45(50)	15(17)		

Notes: Numbers in parentheses represent percentages.

fully applicable, 50% (45/90) partially applicable, and 17% (15/90) not applicable. Statistical analysis revealed significant differences in clinical applicability among the five LLMs ($Z = 29.324$, $p < 0.001$; Table 2). Post hoc pairwise comparisons using Dunn's test showed that Kimi's clinical applicability scores were significantly higher than those of Copilot ($p = 0.007$) and Ernie Bot ($p < 0.001$). Additionally, both GPT-4 and Gemini demonstrated significantly higher clinical applicability than Ernie Bot ($p < 0.001$ and $p = 0.019$, respectively) (Figure 2).

Impact of Question Type on Model Performance

Further stratified analysis revealed that LLMs performed significantly better on common-sense questions compared to professional, domain-specific ones. Among the 41 professional questions, the distribution was 47% fully applicable, 36%

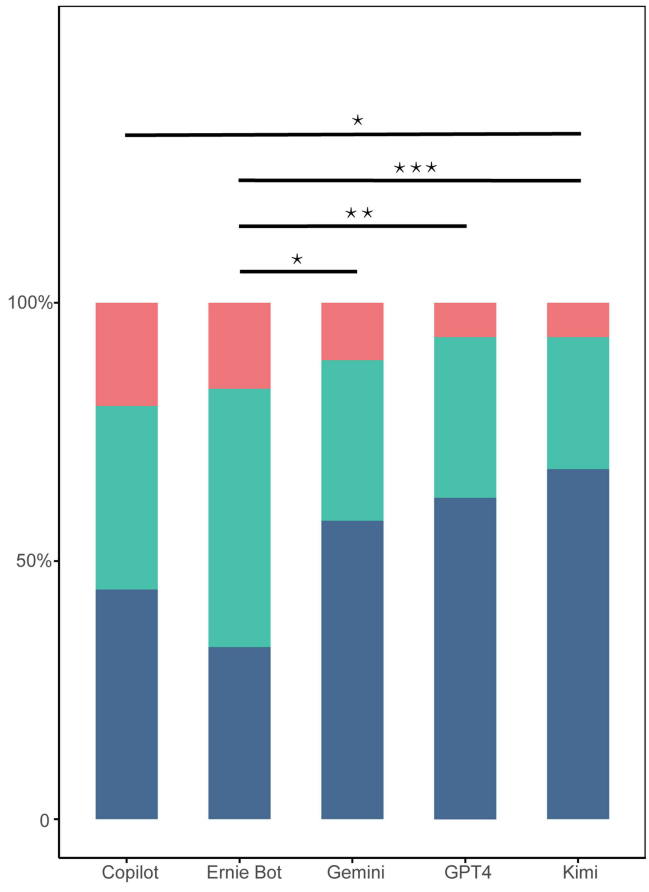

Figure 2 Statistical Variances in Clinical Applicability Among LLM Responses. Bar plot showing the proportions of completely, partially, and inapplicable responses across LLMs in clinical scenarios. The y-axis indicates percentage distribution. Dark blue indicates completely applicable answers, Green indicates partially applicable answers, Red indicates inapplicable answers. Statistical significance: *p < 0.05; **p < 0.01; ***p < 0.001.

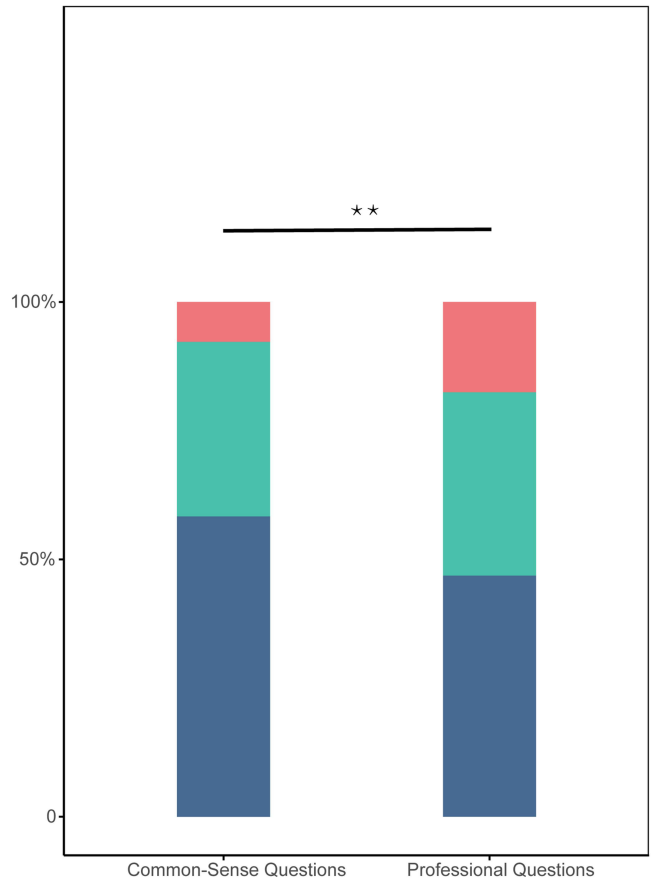

Figure 3 Statistical Differences in the Performance of LLMs on Professional and Common-Sense Questions. Stacked bar chart depicting the comparative applicability of LLM responses in two question categories: professional vs common-sense. Each bar is subdivided into complete, partial, and inapplicable responses. Dark blue indicates completely applicable answers, Green indicates partially applicable answers, Red indicates inapplicable answers. Statistical significance: **p < 0.01.

partially applicable, and 18% not applicable. For the 49 common-sense questions, these proportions were 58%, 34%, and 8%, respectively. This difference was statistically significant ($Z = –2.997$, $p = 0.003$; Figure 3, Table 3).

Subgroup Analysis by Clinical Domain

In the analysis of professional question subdomains, Kimi consistently demonstrated high performance, with >50% of its responses rated as fully applicable across all four sub-domains. In screening and diagnostic decision-making, as well as routine treatment planning, GPT-4, Gemini, and Kimi all exceeded the 50% fully applicable threshold, whereas Copilot and Ernie Bot lagged behind. Kimi showed marked superiority in the domain of liver cancer transformation therapy, neoadjuvant, and adjuvant treatment decision-making, with 86% of its responses rated as fully applicable, compared to Ernie Bot's 14%. Similarly, in systemic treatment decisions, GPT-4 (43%), Gemini (57%), Copilot (57%), and Kimi (57%) outperformed Ernie Bot (14%) (Table 4).

Table 3 Performance of LLMs on Professional and Common-Sense Questions

Category	Num	GPT4			Gemini			Copilot			Kimi			Ernie bot		
		CA	PA	IA	CA	PA	IA	CA	PA	IA	CA	PA	IA	CA	PA	IA
professional questions	41	22(54)	14(34)	5(12)	22(54)	15(37)	4(10)	16(39)	12(29)	13(32)	27(66)	11(27)	3(7)	9(22)	21(51)	11(27)
common-sense questions	49	34(69)	14(29)	1(2)	30(61)	13(27)	6(12)	24(49)	20(41)	5(10)	34(69)	12(24)	3(6)	21(43)	24(49)	4(8)

Note: Numbers in parentheses represent percentages.
Abbreviations: CA, represents Complete Applicability; PA, represents Partial Applicability; IA, represents Inapplicability; Num, denotes the Total Number of Questions.

https://doi.org/10.2147/JHC.S531642

Table 4 Performance of LLMs in the Domain of Professional Questions

Domain	Num	GPT4			Gemini			Copilot			Kimi			Ernie bot		
		CA	PA	IA	CA	PA	IA	CA	PA	IA	CA	PA	IA	CA	PA	IA
A	8	4(50)	4(50)	0	4(50)	3(38)	1(13)	3(38)	1(13)	4(50)	5(63)	1(13)	2(25)	1(13)	4(50)	3(38)
B	19	12(63)	5(26)	2(11)	12(63)	6(32)	1(5)	8(42)	6(32)	5(26)	12(63)	6(32)	1(5)	6(32)	10(53)	3(16)
C	7	3(43)	2(29)	2(29)	2(29)	4(57)	1(14)	1(14)	3(43)	3(43)	6(86)	1(14)	0	1(14)	4(57)	2(29)
D	7	3(43)	3(43)	1(14)	4(57)	2(29)	1(14)	4(57)	2(29)	1(14)	4(57)	3(43)	0	1(14)	3(43)	3(43)

Notes: A = Screening and Diagnosis; B = Routine Treatment Decision-Making; C = Liver Cancer Transformation Therapy; D = Neoadjuvant, Postoperative Adjuvant Therapy Decision-Making, and Systemic Treatment Decision-Making. Numbers in parentheses represent percentages.

Table 5 Performance of LLMs in the Domain of Common-Sense Questions

Domain	Num	GPT4			Gemini			Copilot			Kimi			Ernie bot		
		CA	PA	IA	CA	PA	IA	CA	PA	IA	CA	PA	IA	CA	PA	IA
a	11	6(55)	5(45)	0	5(45)	3(27)	3(27)	2(18)	7(64)	2(18)	5(45)	5(45)	1(9)	1(9)	8(73)	2(18)
b	7	7(100)	0	0	5(71)	2(29)	0	5(71)	2(29)	0	6(86)	1(14)	0	6(86)	1(14)	0
c	21	14(67)	6(29)	1(5)	13(62)	5(24)	3(14)	13(62)	5(24)	3(14)	14(67)	5(24)	2(10)	10(48)	10(48)	1(5)
d	10	7(70)	3(30	0	7(70)	3(30)	0	4(40)	6(60)	0	9(90)	1(10)	0	4(40)	5(50)	1(10)

Notes: a = Overview with Screening; b = Treatment Methods; c = Postoperative Care and Medication Guidance; d = Prognosis and Psychological Support. Numbers in parentheses represent percentages.

In the common-sense question subgroup, GPT-4 maintained >50% fully applicable responses across all four subdomains and reached 100% in the treatment methods category. In the overview and screening subgroup, GPT-4 (55%), Gemini (45%), and Kimi (45%) performed better than Copilot (18%) and Ernie Bot (9%). Regarding postoperative care and medication guidance, all models except Ernie Bot (48%) exceeded 60%. In the domain of prognosis and psychological support, Kimi led with a 90% fully applicable rate, followed by GPT-4 and Gemini (70%), while Copilot and Ernie Bot both had 40%. Despite these observed numerical differences, no statistically significant differences were found in the fully applicable rates between LLMs within each common-sense subdomain (Table 5).

Discussion

In this study, we conducted a comprehensive evaluation of five LLMs—GPT-4, Gemini, Copilot, Kimi, and Ernie Bot—focusing on their comprehensibility and clinical applicability in liver cancer scenarios. Through a randomized, blinded evaluation by three liver cancer experts, we ensured methodological rigor and minimized subjective bias. This study offers an early comparative exploration of these five models in a liver cancer–specific clinical context, contributing valuable perspectives on their differential capabilities and areas for improvement.

While all five LLMs exhibited a fundamental capacity to address liver cancer–related clinical questions, their performance varied notably across evaluation dimensions. In terms of comprehensibility, Kimi and Ernie Bot demonstrated superior results, whereas GPT-4 and Kimi achieved higher ratings in clinical applicability, indicating their promising potential for deployment in healthcare settings. This comparative evaluation elucidated the diverse competencies of these models and highlighted GPT-4, Kimi, and Ernie Bot as particularly capable of generating accurate and clinically relevant information on liver cancer. Continued evaluation will be essential to further enhance both the comprehensibility and clinical applicability of these tools, thereby supporting their safe and effective integration into medical practice.[10] Notably, inter-rater reliability was substantial for both assessed domains ($\kappa = 0.754$ for clinical applicability; $\kappa = 0.671$ for comprehensibility), lending credibility to the consistency of expert evaluations.

Despite encouraging results, our analysis uncovered multiple clinically meaningful errors. For instance, Ernie Bot frequently suggested outdated or inappropriate systemic regimens, such as recommending TACE monotherapy for BCLC stage C patients—contrary to global guidelines. Similarly, Copilot occasionally misidentified immunotherapy indications,

suggesting immune checkpoint inhibitors for early-stage patients without indication, or confusing adjuvant and neoadjuvant settings. GPT-4, while relatively accurate overall, occasionally hallucinated drug combinations (eg, pairing lenvatinib with chemotherapy in a curative setting), which are not supported by evidence or trial data. These observations emphasize that LLMs, even when seemingly accurate, may propagate dangerous clinical misinformation if not carefully scrutinized.

The observed performance differences stem from several technical factors. First, disparities in training data breadth and source quality play a central role. GPT-4 benefits from a broader corpus, frequent user feedback, and fine-tuning with expert input, which likely enhance its clinical reasoning and applicability.[17] Kimi's strong performance may reflect its better alignment with regional clinical guidelines, suggesting that geographic and cultural relevance of training data can influence domain-specific accuracy.[18]

In contrast, Copilot and Ernie Bot may lack fine-tuning on medical literature, relying instead on web-based general corpora that omit important clinical nuance. Furthermore, the internal architecture and prompt interpretation mechanisms vary significantly across models. While some models handle structured medical questions well, they struggle with ambiguous or compound queries—highlighting limitations in contextual understanding and long-form reasoning.[10,19] Moreover, most LLMs lack the capability to assess temporal dynamics, such as changes in liver cancer staging systems or recent drug approvals, unless updated regularly.[19]

Our results may help inform several possible avenues for refining LLMs in clinical applications. First, designing prompts specifically for medical use -for example, requesting responses based on current liver cancer guidelines or providing sources from recognized standards- may help reduce hallucinations and enhance response accuracy.[20] Secondly, domain-specific fine-tuning using peer-reviewed literature, clinical pathways, and real-world data can bolster applicability and reduce spurious associations.[21] Third, integrating basic logic verification tools or clinical guidance checks may help LLMs avoid offering outdated, contradictory, or clinically inappropriate suggestions.[22] Notably, differences between comprehensibility and clinical applicability scores—such as GPT-4's relatively lower interpretability—highlight the need for balanced design. A model that provides correct advice may still fail if users cannot easily understand or apply its recommendations, especially in high-risk clinical environments.

Although GPT-4 and Kimi showed promise in clinical applicability, no model achieved perfect performance. This underscores the need for cautious use of LLMs as decision-support tools rather than standalone advisors. Given that patients are increasingly using LLMs to interpret medical information, clinicians may need to proactively engage with these tools to correct misconceptions, enhance shared decision-making, and reduce anxiety and confusion.[23,24] Ultimately, LLMs may act as bridges between professional expertise and patient comprehension, but they must be closely monitored and critically appraised. Future development should prioritize integration of multi-source medical knowledge, inclusion of expert judgments, and continual updating with regulatory and clinical trial data.[25] In parallel, international efforts are needed to develop benchmarking frameworks for LLM performance across clinical specialties. Additionally, assessing the reproducibility of LLM outputs across multiple sessions may offer valuable insights into model stability, which will be critical for their safe and reliable integration into clinical workflows.[26,27]

In summary, while LLMs hold transformative potential for healthcare communication and clinical education, their safe and effective deployment depends on continuous validation, domain-specific optimization, and alignment with expert-driven standards. Our study contributes foundational evidence supporting this goal within the field of liver cancer and lays groundwork for future AI-clinician-patient collaboration.

Data Sharing Statement

The datasets generated and/or analyzed during the current study are not publicly available but are available from the corresponding author, Dr. Ming Wang, upon reasonable request.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising, or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article

has been submitted; and agree to be accountable for all aspects of the work. Deyuan Zhong, Yuxin Liang and Hong-Tao Yan should be regarded as co-first authors.

Funding

There is no funding to report.

Disclosure

The authors declare that they have no competing interests.

References

1. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and google bard on a neurosurgery oral boards preparation question bank. *Neurosurgery*. 2023;93(5):1090–1098. doi:10.1227/neu.0000000000002551
2. Pérez-Soler S, Juárez-Puerta S, Guerra E, Lara J. Choosing a chatbot development tool. *IEEE Software*. 2021;38(4):94–103. doi:10.1109/MS.2020.3030198
3. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616(7956):259–265. doi:10.1038/s41586-023-05881-4
4. Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. *JAMA*. 2023;329(16):1349–1350. doi:10.1001/jama.2023.5321
5. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589–596. doi:10.1001/jamainternmed.2023.1838
6. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New Engl J Med*. 2023;388(13):1233–1239. doi:10.1056/NEJMsr2214184
7. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. *Lancet Digital Health*. 2023;5(4):e179–e81. doi:10.1016/S2589-7500(23)00048-1
8. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, LA, USA: Curran Associates Inc.; 2022. p. Article2011.
9. Nature Medicine. Will ChatGPT transform healthcare? *Nature Med*. 2023;29(3):505–506. doi:10.1038/s41591-023-02289-5
10. 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. doi:10.1016/j.ebiom.2023.104770
11. Doshi R, Amin K, Khosla P, Bajaj S, Chheang S, Forman H. *Utilizing Large Language Models to Simplify Radiology Reports: A Comparative Analysis of Chatgpt-3.5, Chatgpt-4.0*. Google Bard, and Microsoft Bing; 2023.
12. Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. *Front Artificial Intell*. 2023;6:1166014. doi:10.3389/frai.2023.1166014
13. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023;614(7947):224–226. doi:10.1038/d41586-023-00288-7
14. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. *Nature*. 2023;614(7947):214–216. doi:10.1038/d41586-023-00340-6
15. Şahin MF, Ateş H, Keleş A, et al. Responses of five different artificial intelligence chatbots to the top searched queries about erectile dysfunction: a comparative analysis. *J Med Systems*. 2024;48(1):38. doi:10.1007/s10916-024-02056-0
16. Yao Z, Duan L, Xu S, Chi L, Sheng D. Performance of large language models in the non-english context: qualitative study of models trained on different languages in chinese medical examinations. *JMIR Med Info*. 2025;13:e69485. doi:10.2196/69485
17. Fan L, Li L, Ma Z, Lee S, Yu H, Hemphill L. A bibliometric review of large language models research from 2017 to 2023. *arXiv*. 2023;2023:1.
18. Ai M. Kimi technical whitepaper. 2024. Available from: https://moonshot.ai/kimi-whitepaper.pdf. Accessed August 14, 2025.
19. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the Chat-GPT model. *Res Square*. 2023;2023:rs–3.
20. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–180. doi:10.1038/s41586-023-06291-2
21. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs google bard. *Radiology*. 2023;307(5):e230922. doi:10.1148/radiol.230922
22. Gilbert S, Kather JN, Hogan A. Augmented non-hallucinating large language models as medical information curators. *Npj Digital Med*. 2024;7(1):100. doi:10.1038/s41746-024-01081-0
23. Lee TC, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B. ChatGPT answers common patient questions about colonoscopy. *Gastroenterology*. 2023;165(2):509–11.e7. doi:10.1053/j.gastro.2023.04.033
24. Mezrich JL, Jin G, Lye C, Yousman L, Forman HP. Patient electronic access to final radiology reports: what is the current standard of practice, and is an embargo period appropriate? *Radiology*. 2021;300(1):187–189. doi:10.1148/radiol.2021204382
25. Doshi RH, Bajaj SS, Krumholz HM. ChatGPT: temptations of progress. *Am J Bioethics*. 2023;23(4):6–8. doi:10.1080/15265161.2023.2180110
26. Desai A, Abdelhamid M, Padalkar NR. What is reproducibility in artificial intelligence and machine learning research?*AI Magazine*. 2025;46(2):e70004.
27. Guimarães A, Magalhães J, Martins B, et al. A Reproducibility Study on Consistent LLM Reasoning for Natural Language Inference over Clinical Trials. In: Hauff C, Macdonald C, Jannach D, editors. *Advances in Information Retrieval*. Cham: Springer Nature Switzerland; 2025:48–63.