ACCSCIENCE PUBLISHING

Eurasian Journal of Medicine and Oncology

ORIGINAL RESEARCH ARTICLE

Evaluation of DeepSeek-R1 and its distilled models for performance and cost efficiency in oncology

Xiao Wei[1†], Fangcen Liu[1†], Kai Xin[2*], and Lijing Zhu[2*]

[1]Department of Pathology, Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing, Jiangsu, China

[2]Department of Oncology, Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing, Jiangsu, China

Abstract

Introduction: Malignant tumors represent a significant public health threat, and the integration of artificial intelligence in health care is increasingly becoming a priority. Many oncology institutions are already considering the use of DeepSeek-R1 to assist doctors in making complex medical decisions. However, there remains a lack of sufficient evidence regarding the accuracy, consistency, and cost-efficiency of DeepSeek-R1 and its distilled models in oncology decision-making. This study aims to fill this gap by evaluating the performance and cost-effectiveness of DeepSeek-R1 and its distilled models in oncology, providing critical insights into their potential for clinical integration.

Objectives: This study aimed to systematically evaluate the performance, consistency, and cost-efficiency of the open-source large language model (LLM) DeepSeek-R1 and its distilled variants in the context of oncology decision-making, using a benchmark derived from the MedQA dataset.

Methods: A custom oncology question set containing 1,206 multiple choice questions was curated from MedQA. Seven models, including DeepSeek-R1 and six distilled versions, were evaluated using an automated testing framework. Accuracy, consistency, latency, and token consumption were compared across models. Statistical tests, including McNemar and Wilcoxon signed-rank, were used to assess differences in performance. Questions were also categorized into clinical task types (diagnosis, treatment, triage, and follow-up) for subgroup analysis.

Results: DeepSeek-R1 achieved the highest performance (accuracy: 91.38%; consistency: 90.47%), whereas DeepSeek-R1-Distill-Qwen-32B was the only distilled model to exceed both metrics at the 0.8 threshold (accuracy: 88.72%; consistency: 81.44%). DeepSeek-R1 demonstrated significantly higher accuracy than its distilled counterpart ($p<0.05$), particularly in diagnosis- and treatment-related tasks ($p<0.05$). However, it also exhibited significantly greater latency and token consumption. A Cohen's kappa value of 0.575 indicated moderate agreement between the two models.

Conclusion: DeepSeek-R1 is more suitable for high-stakes oncology tasks requiring high accuracy and consistency, whereas DeepSeek-R1-Distill-Qwen-32B offers a cost-effective alternative for use in outpatient or resource-limited settings. These findings support a task- and resource-adaptive deployment strategy for LLMs in clinical oncology.

Keywords: DeepSeek-R1; Distilled models; Oncology; Performance; Cost efficiency

[†]*These authors contributed equally to this work.*

***Corresponding authors:**
Kai Xin
(kalexin@outlook.com);
Lijing Zhu
(zhulijing@njglyy.com)

Citation: Wei X, Liu F, Xin K, Zhu L. Evaluation of DeepSeek-R1 and its distilled models for performance and cost efficiency in oncology. *Eurasian J Med Oncol*. 2025;9(4):160-167.
doi: 10.36922/EJMO025150097

Received: April 8, 2025

Revised: April 17, 2025

Accepted: May 6, 2025

Published online: June 3, 2025

1. Introduction

Large language models (LLMs) are a class of foundational models trained on vast text corpora, capable of generating coherent, context-aware responses and engaging in complex reasoning tasks across multiple domains.[1,2] Since the release of GPT-3 in 2020, LLMs have increasingly demonstrated transformative potential across industries, especially in domains demanding extensive domain-specific reasoning, such as law, education, and health care.[3,4] In particular, the health-care sector has witnessed a surge of interest in leveraging LLMs to support clinical decision-making, optimize workflows, and enhance patient communication.[5] Despite significant advances, a major bottleneck to the broader adoption of LLMs in medicine is the closed-source nature and high computational costs associated with frontier models such as GPT-4 or Claude 3.[6] These limitations hinder flexibility, fine-tuning, and integration in resource-constrained settings such as small clinics or low- and middle-income countries. In contrast, open-source LLMs provide a promising alternative, offering transparency, modifiability, and cost-effective deployment. Among these, DeepSeek-R1, released under the Massachusetts Institute of Technology license in January 2025, has drawn attention for its performance in complex tasks such as code generation, mathematical reasoning, and natural language understanding.[7] Unlike traditional models relying heavily on supervised fine-tuning, DeepSeek-R1 adopts reinforcement learning methods for end-to-end training, thereby demonstrating the feasibility of direct reinforcement learning for large-scale natural language understanding.[8,9] Furthermore, its distillation into smaller variants, such as DeepSeek-R1-Distill-Qwen-32B, offers lightweight alternatives suitable for institutions with limited infrastructure, supporting broader accessibility and real-world utility.[10,11] Model distillation plays a vital role in the democratization of artificial intelligence (AI) by transferring knowledge from a larger "teacher" model to a smaller "student" model, achieving considerable reductions in computational cost without substantial loss of performance.[12] This is especially crucial in health-care environments where timely response and real-time deployment are essential, such as emergency triage systems or outpatient consultations.[13]

In the domain of oncology, where diagnostic complexity and treatment planning are highly nuanced, the use of LLMs has the potential to revolutionize decision support systems. Malignant tumors remain among the leading causes of death worldwide, and oncologists frequently rely on multimodal data and evolving clinical guidelines to make informed decisions.[14] The integration of LLMs such as DeepSeek-R1 could assist clinicians in rapidly synthesizing knowledge, identifying treatment pathways, and updating protocols based on the latest evidence. However, limited empirical research has been conducted on the real-world utility of LLMs in oncology-specific contexts. Existing evaluations often focus on general medical question-answering, which may not reflect the domain complexity or clinical risk associated with oncological decision-making.[15] To bridge this gap, we created a benchmark test set specifically for oncology, derived from the MedQA dataset, and conducted a systematic evaluation of DeepSeek-R1 and its six distilled variants.

In addition, the ongoing integration of the Internet of Things in clinical settings further amplifies the relevance of LLMs. Wearable sensors, remote patient monitoring, and smart infusion systems generate large volumes of data requiring real-time interpretation. LLMs can serve as intelligent intermediaries, contextualizing these data streams to support dynamic clinical decision-making. This synergy holds particular promise for decentralized or rural care models, where access to human specialists may be limited.[16,17] Beyond clinical utility, there is growing recognition of LLMs' potential in medical education. In oncology training, LLMs can serve as virtual tutors, simulate multidisciplinary case discussions, or provide instant feedback on clinical reasoning exercises.[18] Their ability to digest, summarize, and explain complex literature also supports continuing education for practicing clinicians. As health-care systems increasingly embrace digital transformation, the dual role of LLMs in clinical practice and pedagogy becomes evident. Therefore, the present study addresses a critical gap by conducting a comprehensive, multidimensional evaluation of DeepSeek-R1 and its distilled variants using a domain-specific oncology benchmark. In contrast to prior research that primarily reports top-1 accuracy, we assess not only accuracy and consistency but also latency and token usage to capture model performance from both clinical and computational perspectives. In addition, we stratify performance by clinical task types, including diagnosis, treatment, triage, and follow-up, to identify when and where each model may be most applicable. While traditional benchmark evaluations often focus on single-point accuracy, real-world clinical tasks demand a broader and more nuanced set of evaluation metrics. Factors such as interpretability, response time, and robustness to input variation are critical in medical use cases. Clinical decision-making also frequently involves ambiguous queries, domain-specific terminology, and multi-turn interactions. These scenarios require models to demonstrate contextual understanding and longitudinal reasoning – capabilities that remain underrepresented in current evaluation frameworks. Such challenges are not unique to any one

region but are shaped by local clinical practices, data environments, and documentation styles. In the context of the Chinese health-care system, the rapid digitization of medical records, combined with linguistic and institutional diversity, presents both opportunities and complexities for LLM deployment. For example, the coexistence of multiple clinical coding systems, diverse expression styles in free-text medical notes, and the growing use of bilingual documentation can all affect model performance and adaptability. Despite the increasing availability of open-source models, systematic evaluations in non-English, domain-specific settings, such as Chinese oncology, remain limited. Our study seeks to bridge this gap by applying a task-stratified, resource-aware framework that reflects both the clinical demands and infrastructural realities of a diverse health-care environment. In doing so, we aim to offer practical guidance for integrating LLMs into real-world workflows across varying institutional settings.

2. Methods

2.1. Data source

This cross-sectional study was conducted following the Strengthening the Reporting of Observational Studies in Epidemiology guidelines. We used data from the publicly available MedQA database, which provides a standardized set of medical questions designed to evaluate AI models. The dataset spans a variety of medical topics, including diagnosis, treatment, and clinical decision support. To create an oncology-specific benchmark test set, we focused on keywords such as "cancer," "sarcoma," and "malignant tumor," resulting in a total of 1,206 multiple choice questions.

2.2. LLM access

We developed an automated testing framework to assess the performance of DeepSeek-R1 and its distilled models by processing and evaluating their answers to the questions in the benchmark set. Each model was evaluated three times to assess its performance, and the results were stored for analysis. Due to the instability of the official DeepSeek server, we accessed the SiliconFlow and DeepInfra platform, which provide AI model integration services, through their application programming interface (API) to ensure reliable and consistent evaluations. For simplicity, we referred to some distilled models using abbreviated names (e.g., "DeepSeek-R1-32B" for "DeepSeek-R1-Distill-Qwen-32B").

2.3. Statistical analysis

Accuracy was defined as the proportion of correct answers among the total evaluation cases, while consistency was calculated as the proportion of responses that were consistent across all three evaluations. To compare model performance, we used the McNemar test to assess statistical significance in accuracy differences between pairs of models. This test is particularly effective for binary classification problems.

Cohen's kappa coefficient (κ) was used to determine the level of agreement between two models. Kappa values were categorized as follows: 0 – 0.20 (no to slight agreement), 0.21 – 0.40 (fair agreement), 0.41 – 0.60 (moderate agreement), 0.61 – 0.80 (substantial agreement), and 0.81 – 1.00 (almost perfect agreement).

To investigate differences in latency and token consumption between models, we applied the Shapiro–Wilk test to check for normality and the Wilcoxon signed-rank test to evaluate significant differences. All statistical tests were two-tailed, with a significance threshold set at 0.05. Statistical analyses were conducted using Python 3.8, and visualizations were created using Prism Graph 10.4.0.

3. Results

Performance was evaluated based on two key metrics: accuracy and consistency. Figure 1 summarizes the evaluation results. DeepSeek-R1-Distill-Qwen-1.5B achieved an accuracy of 24.96% and a consistency of 9.28%. DeepSeek-R1-Distill-Qwen-7B showed an accuracy of 38.47% and a consistency of 24.61%. DeepSeek-R1-Distill-Llama-8B exhibited an accuracy of 40.96% and a consistency of 55.97%. DeepSeek-R1-Distill-Qwen-14B performed with an accuracy of 83.50% and a consistency of 72.41%. DeepSeek-R1-Distill-Qwen-32B demonstrated an accuracy of 88.72% and a consistency of 81.44%. DeepSeek-R1-Distill-Llama-70B showed an accuracy of 84.99% and a consistency of 68.85%. DeepSeek-R1 exhibited the highest performance, with an accuracy of 91.38% and a consistency of 90.47%. Only DeepSeek-R1-Distill-Qwen-32B and DeepSeek-R1 achieved both accuracy and consistency above 0.8 and were thus selected for further detailed comparison.

The results of the McNemar test revealed a statistically significant difference in accuracy between DeepSeek-R1 and DeepSeek-R1-Distill-Qwen-32B (p=0.0011), as shown in Table 1. The Cohen's kappa value was 0.575, indicating moderate agreement.

We also evaluated the two models in terms of latency and token consumption. Figure 2 illustrates that DeepSeek-R1-Distill-Qwen-32B had a median token consumption of 12.00 (ranging from 6.00 to 185.00) per question, whereas DeepSeek-R1 consumed a median of 40.00 tokens (ranging from 24.00 to 249.00). In terms of latency, DeepSeek-R1-

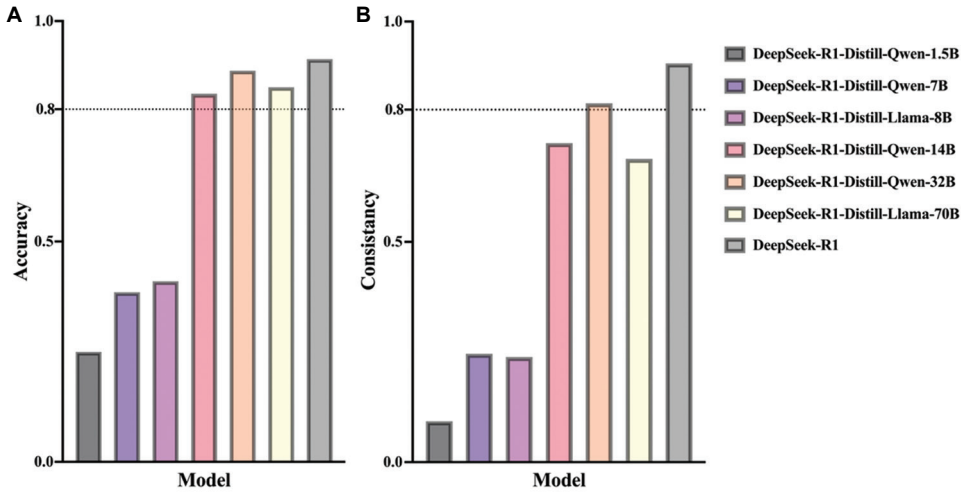

Figure 1. Performance of the DeepSeek-R1 serious model. (A) Accuracy: the proportion of correct responses out of the total number of evaluation cases. (B) Consistency: the proportion of responses with consistent conclusions across three evaluations, relative to the total number of evaluation cases. Models are typically required to achieve high accuracy (>0.8) and strong consistency (>0.8).

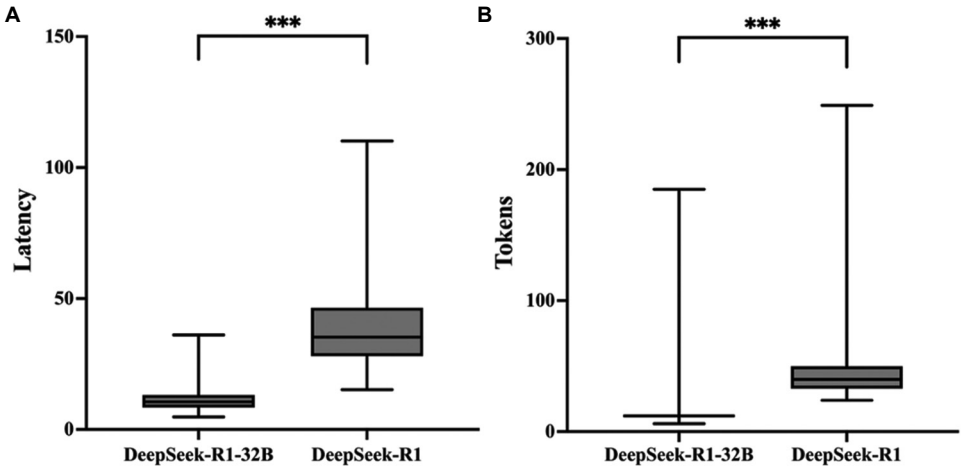

Figure 2. Token and time consumption. (A) Latency: the time (ms) taken by the model to process an input and generate a response. (B) Tokens: the individual units of text processed by a language model, used to measure the amount of input and output.
Notes: *p<0.05, **p<0.01, ***p<0.001

Table 1. Confusion matrix comparing DeepSeek-R1-32B and DeepSeek-R1 across all tasks

p=0.0011	DeepSeek-R1-32B		
	Correct (%)	Incorrect (%)	Total (%)
DeepSeek-R1			
Correct	1,040 (86.24)	62 (5.14)	1,102 (91.38)
Incorrect	30 (2.49)	74 (6.14)	104 (8.62)
Total	1,070 (88.72)	136 (11.28)	1,206 (100)

Notes: A statistically significant difference was observed between the two models (p<0.05) according to the McNemar test. The Cohen's kappa value was 0.575, indicating moderate agreement.

Distill-Qwen-32B had a median of 10.54 ms (ranging from 4.81 ms to 36.09 ms) per question, compared to DeepSeek-R1's median latency of 35.29 ms (ranging from 15.30 ms to 110.18 ms). The Shapiro–Wilk test showed that both token consumption and latency did not follow a normal distribution (p<0.001). The Wilcoxon signed-rank test revealed statistically significant differences between the two models in both token consumption and latency (p<0.001).

To improve the clinical relevance of the evaluation, we categorized each question in the oncology subset of MedQA into one of four task types: diagnosis, treatment,

triage, and follow-up. This classification was based on manually curated keyword patterns that reflect the underlying clinical intent of each question:

(i) Diagnosis questions focused on identifying disease types, staging, pathological features, or clinical manifestations (e.g., keywords: "stage," "type," "symptom," "most likely," and "commonly seen")
(ii) Treatment questions addressed therapeutic decisions, medication choices, or intervention strategies (e.g., "treatment," "first-line," "chemotherapy," and "response")
(iii) Triage questions involved decisions regarding urgency, hospitalization, or referral (e.g., "emergency," "admission," and "immediate management")
(iv) Follow-up questions assessed monitoring strategies, disease progression, or prognostic outcomes (e.g., "follow-up," "monitor," "recurrence," and "prognosis").

Classification was performed in three iterative rounds:

(i) An initial rule-based mapping using core medical keywords
(ii) A refinement stage to capture task-specific phrases not included in round one
(iii) A final review phase incorporating expanded synonyms and clinical terminology, based on expert heuristics.

Questions that did not clearly align with any category were labeled as uncertain and excluded from task-specific performance analyses. After the final round, approximately 71.6% of questions (864/1206) were successfully assigned to a clinical task type (Figure 3A). The majority of questions are related to diagnosis, followed by treatment, with fewer instances of follow-up and triage tasks. In terms of answer accuracy, DeepSeek-R1 significantly outperformed DeepSeek-R1-32B in diagnosis and treatment-related tasks (McNemar test, $p<0.05$; Figure 3B and Table 2).

4. Discussion

Accuracy and consistency are fundamental metrics when LLMs are used for domain-specific tasks. For routine tasks, an accuracy and consistency of at least 0.8 is generally considered acceptable. Smaller distilled models, such as DeepSeek-R1-1.5B, 7B, and 8B, generally exhibit lower accuracy and consistency than their larger counterparts. Although DeepSeek-R1-Distill-Qwen-14B and DeepSeek-R1-Distill-Llama-70B show promising accuracy and consistency, they still fall short of the 0.8 threshold required for basic tasks. While model distillation significantly improves computational efficiency, it inevitably introduces some measurable degree of performance degradation. A key limitation is the reduced model capacity, which constrains the ability to capture complex hierarchical relationships and nuanced medical knowledge. In addition, distilled models often lose the multi-step reasoning chains necessary for handling ambiguous or high-stakes clinical tasks, compressing deep logical processes into simpler patterns. Generalization may also suffer, as the distillation process can overfit the student model to the specific structure of the teacher's outputs, impairing adaptability to novel, noisy, or multisource inputs. Furthermore, optimization during distillation typically focuses on output alignment rather than deep internal feature preservation, weakening the model's ability to extract subtle clinical cues. If the distillation training data lack sufficient diversity, rare but clinically important edge cases may be poorly represented, reducing robustness in real-world oncology practice. Finally, the compact parameter space increases the risk of hallucination, where the model generates plausible but inaccurate responses more frequently. Taken together, these limitations highlight that while distilled models offer practical advantages, careful consideration is needed before deploying them for complex clinical reasoning tasks that demand high precision and reasoning depth. Despite these challenges, DeepSeek-R1-Distill-Qwen-32B performed exceptionally well in our testing, surpassing

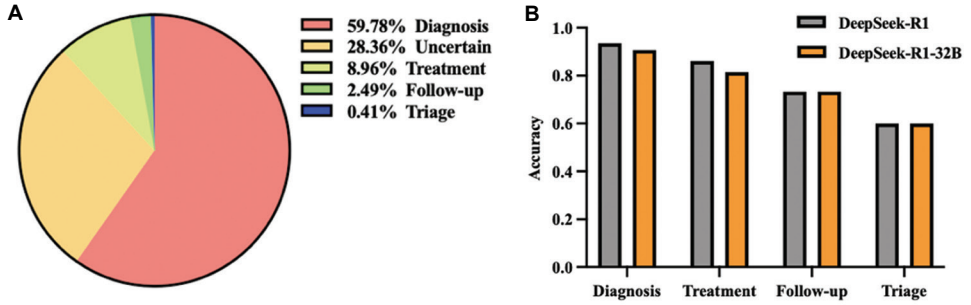

Figure 3. Distribution of task types and model performance comparison. (A) Distribution of questions by clinical task type. The majority of questions relate to diagnosis, followed by treatment, with fewer instances of follow-up and triage tasks. (B) Accuracy comparison between DeepSeek-R1 and DeepSeek-R1-32B across the four clinical task types. DeepSeek-R1 consistently outperformed its distilled version in all categories, particularly in diagnosis and treatment.

Table 2. Confusion matrix comparing DeepSeek-R1 and DeepSeek-R1-32B on diagnosis and treatment tasks

p=0.0025	DeepSeek-R1-32B		
	Correct (%)	Incorrect (%)	Total (%)
DeepSeek-R1			
Correct	620 (74.79)	44 (5.31)	664 (80.10)
Incorrect	19 (2.29)	146 (17.61)	165 (19.90)
Total	639 (77.08)	190 (22.92)	829 (100.00)

Note: A statistically significant difference was observed between the two models ($p<0.05$) according to the McNemar test.

even DeepSeek-R1-Distill-Llama-70B in both accuracy and consistency. This superior performance may be attributed to several factors, including model architecture, distillation strategies, the appropriateness of training data, and computational resources. While its accuracy is slightly lower than that of DeepSeek-R1, it offers markedly reduced latency and token consumption. These characteristics suggest that DeepSeek-R1-Distill-Qwen-32B may be more efficient for scenarios where computational resources are limited and ultra-high accuracy is not critical. In contrast, DeepSeek-R1 remains preferable for high-stakes medical tasks that demand greater consistency and precision. However, this comes at the cost of higher latency and computational requirements, which may pose practical constraints in resource-limited settings.

The kappa value measures the degree of agreement between two models. For DeepSeek-R1 and DeepSeek-R1-Distill-Qwen-32B, the kappa value was 0.575, indicating a moderate level of agreement. This suggests that while both models are capable of providing similar answers in many instances, there are still notable differences in their decision-making processes. Such discrepancies may arise from variations in how the models process and interpret input data, reflecting the intrinsic trade-offs between accuracy and consistency. The moderate kappa value also highlights that the models' responses are not entirely interchangeable, underscoring the importance of considering both accuracy and consistency when choosing the right model for medical decision-making tasks. In clinical contexts, even moderate inconsistencies may have profound consequences, potentially leading to different diagnostic or treatment decisions. Therefore, outputs from different models should not be considered interchangeable, especially in oncology, where precision is critical. This study also demonstrates that DeepSeek-R1 achieved significantly higher accuracy than its distilled counterpart, DeepSeek-R1-32B, in diagnosis- and treatment-related tasks (McNemar test, $p<0.05$). Given the central role of diagnosis and therapeutic decision-making in oncology, this finding highlights the greater potential of DeepSeek-R1 in supporting high-stakes clinical reasoning. While distilled models such as DeepSeek-R1-32B may offer advantages in speed and deployment cost, they may be less reliable for tasks requiring complex, multistep reasoning or the integration of nuanced medical information. As such, DeepSeek-R1 may be more suitable for use in clinical decision support systems where accuracy and robustness are critical, especially in diagnostic interpretation and treatment planning. Given that incorrect responses could lead to diagnostic errors or suboptimal treatment recommendations, LLMs should be deployed as decision support tools under appropriate clinical supervision.

This study has several limitations. First, due to funding constraints, we were unable to conduct expert manual evaluations or perform repeated large-scale inference using API-based models. As a result, the current evaluation may not fully capture the stochastic nature of LLM outputs, nor does it assess critical aspects such as reasoning quality, semantic nuance, or hallucination risks. Second, while our task-type classification covered four clinical domains – diagnosis, treatment, triage, and follow-up – the number of questions in the triage and follow-up categories was relatively small. This limited sample size may be insufficient to comprehensively assess model performance in these specific clinical contexts. In addition, the method used to estimate token consumption in our automated testing framework does not reflect the true tokenization mechanisms of each model and should be interpreted as an approximation. However, this limitation was applied consistently across all models, ensuring fairness in cost-related comparisons. Finally, as the test set was derived from a publicly available dataset, there remains a possibility of data leakage, which could inflate model performance if prior exposure to similar content occurred. Future studies should incorporate expert review, increased sampling diversity, and real-world clinical scenarios to improve the robustness and generalizability of LLM evaluations.

Beyond model accuracy and latency, real-world clinical deployment requires careful consideration of institutional constraints and operational priorities. In high-volume cancer centers, a hybrid strategy may be optimal – using high-performing models such as DeepSeek-R1 for diagnosis and treatment decisions, whereas leveraging distilled models for triage or routine follow-ups. Such task-specific routing can reduce resource strain while preserving quality of care. In addition, cost-efficiency is crucial in low- and middle-income countries or rural clinics, where limited hardware makes lightweight models more viable. Beyond resource constraints, another important consideration is the generalizability of LLMs trained on benchmark datasets. While MedQA provides a structured framework, real

clinical queries often contain ambiguous, multi-layered, or incomplete information. The robustness of DeepSeek-R1 and its distilled versions in such scenarios remains to be validated. Future research should incorporate noisy, real-world clinical notes, integrate multimodal inputs (e.g., radiology reports or laboratory data), and explore few-shot or retrieval-augmented learning to improve adaptability. Rather than selecting a single optimal model, institutions may benefit from adopting a multimodal orchestration strategy, wherein different LLMs are assigned to distinct clinical tasks based on their strengths. For instance, high-capacity models can be prioritized for diagnosis or treatment planning, while faster, lightweight models can handle triage, patient education, or administrative queries. This modular approach to LLM deployment enhances flexibility and may improve scalability in diverse healthcare environments. Furthermore, expanding future evaluations to include multilingual tasks, real-world physician prompts, and cross-institutional data will be critical for validating model generalizability. As many institutions operate under different documentation styles, regulatory policies, and patient demographics, collaborative benchmarking across centers would yield more representative findings and promote responsible AI development. Finally, the rapid iteration of LLMs presents both opportunities and challenges. While newer model versions may offer performance gains, continual updates necessitate repeated validation, version tracking, and compatibility checks with existing health-care information technology systems. Developing standardized evaluation pipelines and updating protocols will be key to ensuring that model upgrades do not compromise safety or interoperability. Ethical and regulatory concerns also merit attention. The interpretability of LLMs, their susceptibility to hallucinations, and their potential biases pose risks if deployed without oversight. Therefore, human-in-the-loop architectures, traceable response pathways, and continual monitoring systems should be established before model deployment in clinical workflows. As health-care systems embrace AI, balancing innovation with accountability is essential for sustainable integration.

5. Conclusion

DeepSeek-R1 demonstrates superior accuracy and consistency compared to its distilled counterparts, particularly in complex oncology tasks such as diagnosis and treatment planning. Although it requires greater computational resources and longer response time, its performance makes it a suitable choice for large oncology centers. Conversely, DeepSeek-R1-Distill-Qwen-32B provides a cost-effective alternative with acceptable performance for smaller institutions. The findings of this study highlight that model deployment should be tailored to institutional needs and task complexity. Given the potential for error in current generative AI systems, these models should be implemented as clinical decision support tools rather than autonomous decision-makers.

Acknowledgments

We extend our sincere thanks to the providers of the MedQA dataset, whose standardized medical question set was crucial for this study. We also gratefully acknowledge the DeepSeek team for providing the open-source models, which significantly contributed to the progress of this research.

Funding

None.

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Conceptualization: Kai Xin
Formal analysis: Xiao Wei, Fangcen Liu
Methodology: Kai Xin
Writing–original draft: Xiao Wei, Fangcen Liu
Writing–review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The dataset and code are publicly available at https://github.com/roubaokai/DeepSeek_Oncology_Evaluation/tree/main.

References

1. Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33: 1877-1901.

   doi: 10.48550/arXiv.2005.14165

2. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Pennsylvania: Association for Computational Linguistics; Vol. 1. 2019. p. 4171-4186.

doi: 10.48550/arXiv.1810.04805

3. Kaplan J, McCandlish S, Henighan T, *et al*. Scaling laws for neural language models. arXiv; 2020.

   doi: 10.48550/arXiv.2001.08361

4. Stiennon N, Ouyang L, Wu J, *et al*. Learning to summarize with human feedback. In: *Advances in Neural Information Processing Systems*. Vol. 33. Cambridge: MIT Press; 2020. p. 3008-3021.

   doi: 10.48550/arXiv.2009.03125

5. Liao H. Deepseek large-scale model: Technical analysis and development prospect. J Comput Sci Electr Eng. 2025;7(1):33-37.

   doi: 10.61784/jcsee3035

6. De Carvalho GP, Sawanobori T, Horii T. Data-driven motion planning: A survey on deep neural networks, reinforcement learning, and large language model approaches. IEEE Access. 2025;13:52195-52245.

   doi: 10.1109/ACCESS.2025.3552225

7. Guo D, Yang D, Zhang H, *et al*. *Deepseek-R1: Incentivizing Reasoning Capability in LLMs Via Reinforcement Learning*. arXiv. China: DeepSeek; 2025.

   doi: 10.48550/arXiv.2501.12948

8. Gou J, Yu B, Maybank SJ, Tao D. Knowledge distillation: A survey. *Int J Comput Vis*. 2021;129(6):1789-1819.

   doi: 10.1007/s11263-021-01453-z

9. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv. 2015.

   doi: 10.48550/arXiv.1503.02531

10. Shimizu H, Nakayama KI. Artificial intelligence in oncology. *Cancer Sci*. 2020;111(5):1452-1460.

    doi: 10.1111/cas.14377

11. Mulita F, Verras GI, Anagnostopoulos CN, Kotis K. A smarter health through the internet of surgical things. *Sensors (Basel)*. 2022;22(12):4577.

    doi: 10.3390/s22124577

12. Sanh V, Debut L, Chaumond J, Wolf T. *DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter*. [arXiv Preprint]; 2019.

    doi: 10.48550/arXiv.1910.01108

13. Esmaeilzadeh P. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: A perspective for healthcare organizations. *Artif Intell Med*. 2024;151:102861.

    doi: 10.1016/j.artmed.2024.102861

14. Sung H, Ferlay J, Siegel RL, *et al*. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209-249.

    doi: 10.3322/caac.21660

15. Huang Y, Tang K, Chen M, Wang B. *A Comprehensive Survey on Evaluating Large Language Model Applications in the Medical Industry*. [arXiv Preprint]; 2024.

    doi: 10.48550/arXiv.2404.15777

16. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56.

    doi: 10.1038/s41591-018-0300-7

17. Jiang L, Wu Z, Xu X, *et al*. Opportunities and challenges of artificial intelligence in the medical field: current application, emerging problems, and problem-solving strategies. *J Int Med Res*. 2021;49(3):03000605211000157.

    doi: 10.1177/03000605211000157

18. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: Applications and implications. *JMIR Med Educ*. 2023;9:e50945.

    doi: 10.2196/50945