Intelligent Oncology 1 (2025) 172–175

Contents lists available at ScienceDirect

Intelligent Oncology

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

Editorial

# DeepSeek AI: Transforming medical AI with cost-efficiency, transparency, and privacy preservation

Wenli Cai

*Chongqing University Cancer Hospital, Chongqing 400030, China*

The rise of DeepSeek marks a significant milestone in the evolution of artificial intelligence (AI), akin to the groundbreaking debut of OpenAI's GPT-3 in 2020. Positioned as a cost-effective yet high-performing alternative to leading large language models (LLMs) like OpenAI's GPT-4, Meta's Llama 3, and Anthropic's Claude 3.5, DeepSeek-V3 has demonstrated remarkable capabilities while substantially reducing training costs.

Unlike conventional LLMs that rely on massive computational resources and proprietary architectures, DeepSeek stands out through comprehensive engineering optimizations, open-source transparency, and an innovative training methodology. This editorial explores the technical innovations, open-source strategy, and transformative potential of DeepSeek, particularly in medical AI, while also examining the challenges and implications of its adoption in healthcare.

## Cost efficiency and training innovations

One of the most remarkable features of DeepSeek-V3 is its cost efficiency. The model was trained using just 2,048 H800 GPUs over two months, totaling 2.788 million GPU hours at a cost of only $5.58 million.[1] By comparison, as shown in Table 1, Llama 3.1 required 30.8 million GPU hours,[a] while the estimated development costs for GPT-4[2] and Claude 3.5 exceeded $100 million.[b]

Despite its significantly lower training cost, DeepSeek-V3/R1 delivers performance on par with GPT-4o/O1 and Claude 3.5 Sonnet while outperforming Llama 3 across multiple benchmarks. This remarkable cost-to-performance ratio challenges the prevailing belief that AI advancements must come at the cost of exponentially increasing computational resources. Instead, DeepSeek demonstrates that engineering optimizations can drive cutting-edge AI innovation without unsustainable expenses.

Another key differentiator is DeepSeek's training methodology. Unlike GPT-3 and GPT-4,[3] which follow a supervised fine-tuning phase before reinforcement learning from human feedback, DeepSeek integrates reinforcement learning from the very foundation of its training. This approach has led to superior reasoning capabilities and greater problem-solving efficiency, further solidifying its innovative edge in AI development.

## Technical advancements

DeepSeek's remarkable cost efficiency is the result of technical innovations that extend beyond algorithmic improvements to engineering-level optimizations, increasing computational efficiency while significantly reducing costs.

*E-mail address:* wenli.cai@cqu.edu.cn.

[a] Meta AI. Introducing Llama 3.1: Our most capable models to date. *Meta AI Blog*. Published July 2024. Accessed March 1, 2025. https://ai.meta.com/blog/meta-llama-3-1/.

[b] Anthropic. Introducing Claude 3.5 Sonnet. *Anthropic News*. Published June 2024. Accessed March 1, 2025. https://www.anthropic.com/news/claude-3-5-sonnet.

https://doi.org/10.1016/j.intonc.2025.03.006

Available online 20 March 2025

2950-2616/© 2025 The Author(s). Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

**Table 1**
Comparison of model size, training cost, and GPU training hours across leading large language models.

| Model | Model size (Parameters) | Training cost (USD) | GPU training hours |
|---|---|---|---|
| **DeepSeek-V3** | 671 billion | $5.58 million | 2.788 million on H800 |
| **Llama 3.1** | 405 billion | $92.4–123.2 million | 30.8 million on H100 |
| **GPT-4** | 1,000–1,800 billion | > $100 million | Approximately 16,000 Nvidia over several months on H100 |
| **Claude 3.5** | 175 billion | Not specified | Not publicly disclosed |

Among these innovations are several optimizations to the Mixture of Experts (MoE) framework originally proposed in 1991.[4] DeepSeek optimizes MoE through dynamic expert routing, sparse activation, and advanced loss-balancing techniques, making it more efficient and scalable for large-scale LLMs. Additionally, it incorporates the Multi-head Latent Attention mechanism,[5] a low-rank approximation for key-value caching that significantly reduces memory requirements in Transformer models without sacrificing performance.

Moreover, DeepSeek optimizes its hardware utilization to enhance computational efficiency even further. It leverages parallel thread execution instructions, integrates NVLink support, and adopts the 8-bit floating point mixed-precision training framework.[1]

Another defining feature of DeepSeek is its transparent reasoning processes.[6] Unlike conventional LLMs that often function as "black boxes", producing results without revealing their underlying logic, DeepSeek provides a clear, interpretable chain of reasoning. This transparency builds trust, facilitates more effective debugging, and supports continuous model optimization. This level of transparency makes DeepSeek particularly valuable in critical fields like medicine, law, and scientific research.

## Open-source approach and industry impact

DeepSeek's decision to open-source its models, parameters, training data, and documentation represents a major departure from the industry trend of closed AI ecosystems.[c] This move has significant implications for the democratization of AI technology, lowering barriers for researchers, startups, and enterprises seeking to develop AI-driven applications. By challenging proprietary AI models, DeepSeek disrupts the long-standing knowledge asymmetry that has favored industry giants, such as OpenAI and Google. Meta's chief AI scientist, Yann LeCun, has highlighted DeepSeek's success as a victory for open-source AI rather than a geopolitical competition between China and the United States.[d]

The open-source release has gained traction among AI hardware manufacturers, notably including Huawei and Moore Threads, accelerating the development of alternative AI computing ecosystems and reducing dependency on closed-source AI infrastructure. Furthermore, DeepSeek's low application programming interface costs, ranging from $0.1–1 per million tokens compared to GPT-4's $5–15 per million tokens, pose a serious disruption to the AI market. Competitors may be forced to either lower prices or enhance performance to maintain their customer base. By offering an affordable, high-performance alternative and fostering an open innovation environment, DeepSeek broadens access to AI, promoting inclusivity and leveling the playing field in the AI community.

## Transformative potential in healthcare

DeepSeek AI has the potential to transform the industry by making advanced AI capabilities more affordable and accessible. One of its most compelling advantages is its ability to deliver high-performance AI solutions at a fraction of the cost of existing models. This cost reduction could eliminate financial barriers that have historically prevented smaller hospitals and research institutions from adopting cutting-edge AI technology. By democratizing access to sophisticated AI models, DeepSeek could level the playing field in healthcare, ensuring equitable access to state-of-the-art AI-driven medical solutions.

Beyond affordability, the open-source nature of DeepSeek makes it an ideal solution for privacy-preserving clinical applications. Unlike proprietary LLMs that require patient data to be transmitted to third-party servers, DeepSeek allows local deployment within hospitals and research institutions. This ensures that healthcare institutions retain full control over sensitive medical data. This approach enhances compliance with data privacy regulations, mitigates security risks, and fosters greater trust among healthcare professionals and patients. By balancing high performance with strong privacy safeguards, DeepSeek represents a significant advancement in the ethical and responsible integration of AI into clinical practice.

Another defining feature of DeepSeek is its transparent reasoning process, which provides an interpretable chain of logic for AI-driven decisions. Techniques such as attention visualization and intermediate output logging allow clinicians to see the key medical image features, lab results, or clinical notes that influenced the model's conclusions. By enabling clinicians to trace AI-generated insights step by step, DeepSeek builds trust in AI-assisted medical decision-making. This enhanced explainability fosters stronger human-AI collaboration, allowing doctors to refine diagnoses, explore alternative hypotheses, and make more confident clinical decisions. By bridging the gap between AI and medical expertise, DeepSeek enhances the safety, reliability, and acceptance of AI in clinical environments.

[c] Chen X. AI start-up DeepSeek to open-source 5 code repositories next week for 'full transparency'. *South China Morning Post*. Published February 2025. Accessed March 1, 2025. https://www.scmp.com/tech/big-tech/article/3299597/ai-start-deepseek-open-source-5-code-repositories-next-week-full-transparency.

[d] Wodecki B. Meta's AI chief: DeepSeek proves AI progress isn't about chips. *Capacity*. Published January 2025. Accessed March 1, 2025. https://www.capacitymedia.com/article/metas-ai-chief-deepseek-proves-ai-progress-isnt-about-chips.

The potential applications of DeepSeek AI in healthcare are vast. In real-time diagnostics, it can analyze MRI, CT, or X-ray images to detect tumors, fractures, and other anomalies with high precision. In wearable health monitors, it can continuously track vital signs such as heart rate, oxygen levels, and blood glucose, issuing immediate alerts when abnormalities are detected. These capabilities enable early disease detection, more personalized treatment plans, and improved patient outcomes. Moreover, DeepSeek can assist healthcare professionals by providing second opinions, reducing workloads, and improving decision-making accuracy. In medical research, it accelerates the discovery of new treatments by analyzing large-scale patient datasets, genomic information, and clinical trials, bringing us closer to breakthroughs in personalized medicine and advanced therapeutics.

## Limitations of DeepSeek

Despite its impressive capabilities, DeepSeek-V3 has notable limitations compared to leading LLMs like GPT-4 and Claude. One major drawback is its current restriction to text-only inputs, lacking support for multimodal capabilities such as image and voice processing.

Performance-wise, DeepSeek-V3 exhibits slightly slower response times, with an initial latency of 1.1 s, trailing behind GPT-4o (0.74 s) and Claude (0.76 s). Its token generation speed is also lower, producing 60 tokens per second, compared to GPT-4o's 109 and Claude's 85. Furthermore, DeepSeek-V3's input/output context window is currently capped at 128 K/32 K tokens, which falls short of GPT-4o's 200 K/100 K token capacity,[e] thus constraining its ability to manage extremely long and complex conversations.

While DeepSeek-V3 is highly optimized for cost-efficiency and strong text performance, these limitations highlight areas for potential improvement. Future iterations will need to address these constraints to compete more effectively in an increasingly multimodal and high-performance AI landscape.

## Challenges and ethical considerations

Despite its promising potential, DeepSeek AI faces critical challenges, particularly in data security and regulatory validation, especially in regions like Europe and the United States. As a China-developed AI model, concerns over data privacy, compliance, and security are amplified, particularly in the healthcare sector, where strict regulations such as the General Data Protection Regulation and the Health Insurance Portability and Accountability Act govern patient confidentiality. The deployment of foreign-developed AI models in Western healthcare systems is likely to undergo rigorous scrutiny to ensure compliance with local legal, ethical, and cyber security standards, safeguarding patient data from potential risks.

For DeepSeek AI to gain widespread acceptance, seamless integration into global healthcare systems must align with existing regulatory frameworks. Policymakers, health institutions, and AI developers must collaborate to establish clear guidelines addressing key concerns such as data ownership, model validation, accountability, and bias mitigation. Without robust oversight and clear regulatory policies, the adoption of DeepSeek AI in a clinical environment could face resistance, potentially undermining patient trust and safety. The need for transparent governance, ethical AI deployment, and rigorous validation processes is paramount to ensuring AI-driven healthcare remains both secure and responsible.

## Conclusions: a future of smarter, more affordable healthcare

DeepSeek AI's unprecedented combination of cost-efficiency, privacy-preserving architecture, transparent reasoning, and broad medical applications has the potential to revolutionize clinical decision-making, elevate patient care, and accelerate medical breakthroughs. By making cutting-edge AI more accessible, it paves the way for a future where advanced healthcare is no longer a privilege but a universal standard.

Yet, its path forward is not without challenges. Concerns over data security, regulatory compliance, and ethical deployment loom large, particularly as AI becomes increasingly integrated into sensitive healthcare environments.

The ultimate question remains: Will DeepSeek AI be the catalyst for a new era of smarter, more affordable healthcare, or will these challenges stall its momentum? While the answer is still unfolding, one certainty prevails—AI's role in healthcare has never been more critical, and the future it promises has never been more exciting.

## CRediT authorship contribution statement

**Wenli Cai:** Conceptualization, Writing – original draft, Writing – review & editing.

## Declaration of Competing Interest

Wenli Cai is an editorial board member for Intelligent Oncology and was not involved in the editorial review or the decision to publish this article. The author declares no conflicts of interest.

[e] DeepSeek-R1 Model Comparison. *DocsBot*. Accessed March 1, 2025. https://docsbot.ai/models/compare/deepseek-r1/o1.

## References


1. Liu A, Feng B, Xue B, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:241219437*. Published online 2024.
2. Adler S, Agarwal S, Ahmad L, et al. GPT-4 technical report. *arXiv preprint arXiv:230308774*. Published online 2024.
3. Bai Y, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:220405862*. Published online 2022.
4. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive Mixtures of Local Experts. *Neural Comput.* 1991;3(1):79–87.
5. Liu A, Feng B, Wang B, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:240504434*. Published online 2024.
6. Guo D, Yang D, Zhang H, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:250112948*. Published online 2025.