nature medicine

https://doi.org/10.1038/s41591-025-03726-3

Accelerated Article Preview

Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning

Received: 6 February 2025

Accepted: 21 April 2025

Accelerated Article Preview
Published online: 23 April 2025

Cite this article as: Tordjman, M. et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. *Nature Medicine* https://doi.org/10.1038/s41591-025-03726-3 (2025).

Mickael Tordjman, Zelong Liu, Murat Yuce, Valentin Fauveau, Yunhao Mei, Jerome Hadjadj, Ian Bolger, Haidara Almansour, Carolyn Horst, Ashwin Singh Parihar, Amine Geahchan, Anis Meribout, Nader Yatim, Nicole Ng, Phillip Robson, Alexander Zhou, Sara Lewis, Mingqian Huang, Timothy Deyer, Bachir Taouli, Hao-Chih Lee, Zahi A. Fayad & Xueyan Mei

1. Extended Data

Figure or Table # Please group Extended Data items by type, in sequential order. Total number of items (Figs. + Tables) must not exceed 10.	**Figure/Table title** One sentence only	**Filename** Whole original file name including extension. i.e.: Smith_ED_Fig1.jpg	**Figure/Table Legend** If you are citing a reference for the first time in these legends, please include all new references in the main text Methods References section, and carry on the numbering from the main References section of the paper. If your paper does not have a Methods section, include all new references at the end of the main Reference list.
Extended Data Table. 1	Number of parameters for each model according to Azure.	Extended_Data_Tables-1.jpg	
Extended Data Table. 2	Prompts used for the different tasks	Extended_Data_Tables-2.jpg	
Extended Data Table. 3	Comparison of the average scores for the linguistic evaluation of the “Findings to Impression” task on the 2 databases (private database of reports and MIMIC) based on the evaluation of 2 independent radiologists for each database.	Extended_Data_Tables-3.jpg	Number of tokens per summary provided by the model and ratio of the number of tokens for the Findings section prompted/ Impression output.
Extended Data Table. 4	RECIST 1.1 performances for the 3 LLMs per category of response based on the consensus between the 3 human experts (average of the 3 runs)	Extended_Data_Tables-4.jpg	CR= Complete Response; PR= Partial Response; SD= Stable Disease; PD= Progressive Disease

Extended Data Table. 5	Text-based case evaluation using 5-point Likert score (adapted from the R-IDEA score) for the reasoning evaluation of Complex NEJM Cases and Management evaluation of the BMJ cases; and accuracy for the NEJM and private database of multiple choice questions	Extended_Data_Tables-5.jpg	
Extended Data Table. 6	BERTscore (F1) between the 3 runs of each model for test-retest repeatability.	Extended_Data_Tables-6.jpg	
Extended Data Table. 7	Fleiss kappa between the 3 runs of each model for test-retest repeatability.	Extended_Data_Tables-7.jpg	
Extended Data Table. 8	Inter-model Cohen kappa based on the first run of each model for USMLE, RECIST, Medicilline questions, and NEJM diagnostic cases.	Extended_Data_Tables-8.jpg	

1. Supplementary Information:

A. PDF Files

Item	Present?	Filename Whole original file name including extension. i.e.: Smith_SI.pdf. The extension must be .pdf	A brief, numerical description of file contents. i.e.: *Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.*
Supplementary Information	Choose an item.		
Reporting Summary	Yes	NMED-FT139399B_RS.pdf	
Peer Review Information	Choose an item.	*OFFICE USE ONLY*	

B. Additional Supplementary Files

Type	Number Each type of file (Table, Video, etc.) should be numbered from 1 onwards. Multiple files of the same type should be listed in sequence, i.e.: Supplementary Video 1, Supplementary Video 2, etc.	Filename Whole original file name including extension. i.e.: *Smith_ Supplementary_Video_1 .mov*	Legend or Descriptive Caption Describe the contents of the file

3. Source Data

Parent Figure or Table	Filename Whole original file name including extension. i.e.: *Smith_SourceData_Fig1.xls*, or *Smith_ Unmodified_Gels_Fig1.pdf*	Data description i.e.: Unprocessed western Blots and/or gels, Statistical Source Data, etc.

Editor summary:

The open-source DeepSeek large language model showed variable performance relative to two leading models when benchmarked on four different medical tasks, with relatively strong reasoning capabilities but similar or weaker relative performance on other tasks, such as summarization of imaging reports.

Editor recognition statement:

Primary Handling Editors: Lorenzo Righetto, Michael Basson and Saheli Sadanand, in collaboration with the Nature Medicine team.

Peer Review:

Nature Medicine thanks Jie Yang, Ahmed Alaa and Kirk Roberts for their contribution to the peer review of this work.

Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning

Mickael Tordjman[1,2,*], Zelong Liu[1,*], Murat Yuce[1,2], Valentin Fauveau[1], Yunhao Mei[3], Jerome Hadjadj[4,5], Ian Bolger[1,2], Haidara Almansour[6], Carolyn Horst[7], Ashwin Singh Parihar[8], Amine Geahchan[1,2], Anis Meribout[1], Nader Yatim[9], Nicole Ng[10], Phillip Robson[1,2], Alexander Zhou[1], Sara Lewis[1,2], Mingqian Huang[1,2], Timothy Deyer[11,12], Bachir Taouli[1,2], [#], Hao-Chih Lee[1,2,#], Zahi A. Fayad[1,2,13,#], Xueyan Mei[1,2,13,#]

1. Biomedical Engineering and Imaging Institute, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA
2. Department of Diagnostic, Molecular and Interventional Radiology, Mount Sinai Hospital, New York, NY, 10029, USA
3. Erasmus University Rotterdam, Rotterdam, Netherlands
4. Sorbonne Universite, Service de Medecine Interne, Hopital Saint Antoine, AP-HP, Paris, France
5. Center for Human Genetics and Genomics, New York University Grossman School of Medicine, New York, NY
6. Department of Diagnostic and Interventional Radiology, Tuebingen University Hospital, Tuebingen, Germany
7. School of Biomedical Engineering and Imaging Sciences, King's College, London, UK
8. Mallinckrodt Institute of Radiology, Washington University School of Medicine, Saint Louis, MO, USA
9. Department of Immunology and Immunotherapy, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA
10. Department of Pulmonary, Critical Care and Sleep Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA
11. East River Medical Imaging, New York, NY, 10021, USA
12. Department of Radiology, Cornell Medicine, New York, NY, 10065, USA
13. Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, 10029, New York, NY, USA

[*]These authors contributed equally to this work.

[#]These authors jointly supervised this manuscript.

Corresponding authors:

Hao-Chih Lee: hao-chih.lee@mssm.edu; Zahi A. Fayad: zahi.fayad@mssm.edu;
Xueyan Mei: xueyan.mei@mssm.edu. Corresponding address: 1470 Madison Avenue,
New York, New York, 10029, USA.

Abstract:

DeepSeek is a newly introduced large language model (LLM) designed for enhanced reasoning, but its medical-domain capabilities have not yet been evaluated. This study assessed the capabilities of three LLMs— DeepSeek-R1, ChatGPT-o1, and Llama 3.1-405B—in performing four different medical tasks: answering questions from the United States Medical Licensing Examination (USMLE), interpreting and reasoning based on text-based diagnostic and management cases, providing tumor classification according to RECIST 1.1 criteria, and providing summaries of diagnostic imaging reports across multiple modalities. In the USMLE test, the performance of DeepSeek-R1(accuracy=0.92) was slightly inferior to that of ChatGPT-o1(accuracy=0.95; p=0.04) but better than that of Llama 3.1-405B (accuracy=0.83; $p<10^{-3}$). For text-based case challenges, DeepSeek-R1 performed similarly to ChatGPT-o1 (accuracy of 0.57 vs 0.55; p=0.76 and 0.74 vs 0.76; p=0.06, using New England Journal of Medicine and Medicilline databases, respectively). For RECIST classifications, DeepSeek-R1 also performed similarly to ChatGPT-o1 (0.73 vs 0.81; p=0.10). Diagnostic reasoning steps provided by DeepSeek were deemed more accurate than those provided by ChatGPT and Llama 3.1-405B (average Likert score of 3.61, 3.22, and 3.13, respectively, p=0.005 and $p<10^{-3}$). However, summarized imaging reports provided by DeepSeek-R1 exhibited lower global quality than those provided by ChatGPT-o1 (5-point Likert score: 4.5 vs 4.8; $p<10^{-3}$**).** This study highlights the potential of DeepSeek-R1 LLM for medical applications but also underlines areas needing improvements.

Main:

Large Language Models (LLMs) have emerged as transformative tools in a variety of domains, exhibiting remarkable capabilities in natural language processing, knowledge synthesis, and task automation[1]. In healthcare, these models are being leveraged for applications such as medical diagnosis, clinical decision support, and imaging interpretation. Their ability to process vast amounts of medical knowledge and generate contextually relevant insights has made them increasingly invaluable assets for researchers and clinicians. Notable LLMs include proprietary models like ChatGPT (OpenAI), open-access models such as Llama (Meta), and emerging models like DeepSeek (DeepSeek Artificial Intelligence Co.).

DeepSeek-R1, a 671B parameter LLM released on January 20th 2025, represents a new generation of open-source LLMs exhibiting competitive accuracy in third-party evaluations[2,3]. However, despite the broad applicability of all models, differences in their training and optimization may affect performance for specific tasks. For example, DeepSeek-R1 aims for better reasoning using reinforcement learning, which may be optimally suited for medical interpretation. While established models like ChatGPT[4] and Llama[5] have set performance benchmarks, DeepSeek-R1's release raises important questions about tailoring comparisons within the context of specified applications, such as medical interpretation tasks**.**

This study evaluated DeepSeek-R1 against two comparable leading LLMs, ChatGPT-o1 and Llama 3.1-405B (**Extended Data Table 1**), on four medical domain tasks to benchmark their medical reasoning and processing capabilities. These four tasks consisted of: **(1)** answering United States Medical Licensing Examination (USMLE) questions, **(2)** clinical diagnosis and management based on cases from the New England Journal of Medicine (NEJM), British Medical Journal (BMJ), and Medicilline (not available online), **(3)** summarizing the "Findings" section of imaging reports into an appropriate "Impression" based on cases from the MIMIC-III dataset[6] and another private dataset, and **(4)** performing RECIST 1.1 tumor treatment response classification based on two consecutive reports (private dataset) from consecutive CT scans in cancer patients (**Figure 1**). Datasets and prompts used across all tasks are summarized in **Figure 1 and Extended Data Table 2**; the inclusion of private datasets was important for addressing data leak. These tasks were selected to assess the models' abilities in retrieving medical knowledge, reasoning on complex case analysis and management options, providing imaging report classification, and summarization.

Additionally, we investigated intra-model variability by evaluating each model three times with the same inputs with analysis of outputs utilizing BERTscore[7] for text outputs and Fleiss' kappa[8] for categorial outputs. BERTscore assesses the semantic similarity between different text outputs by measuring how closely the meaning aligns across

repeated model generations. Fleiss' kappa evaluates the agreement between repeated categorical outputs across multiple evaluations, with values ranging from 0 (no agreement) to 1 (perfect agreement).

Results and Discussion

Performance of the three LLMs on the **USMLE question dataset**[9] (323 questions) varied, with DeepSeek-R1 achieving an overall accuracy of 0.92 [95%CI: 0.90-0.95], ChatGPT-o1 an accuracy of 0.95 [95%CI: 0.92-0.98], and Llama 3.1-405B an accuracy of 0.83 [95%CI: 0.78-0.87] (**Figure 2**). DeepSeek-R1 achieved accuracy slightly inferior to ChatGPT-o1 when combining all three Step examinations ($p=0.04$) and better than Llama 3.1-405B ($p<10^{-3}$).

We evaluated the 3 models on **diagnostic and management tasks.** We first used the **NEJM Case Challenges**[10] (50 cases) to assess diagnostic capabilities through both multiple-choice questioning, reasoning steps, and open-ended differential diagnosis generation.

The pooled accuracy of DeepSeek-R1 for these text-based challenges was 0.57 [95%CI: 0.49-0.64] compared to ChatGPT-o1 with an accuracy of 0.55 [95%CI: 0.47-0.62] ($p=0.76$) and Llama 3.1-405B with an accuracy of 0.47 [95%CI: 0.40-0.55] ($p=0.16$) when provided multiple-choices.

When provided the same NEJM cases without answer choices, accuracy dropped to 0.36 [95%CI: 0.24-0.50] for DeepSeek-R1, 0.32 [95%CI: 0.21-0.46] for ChatGPT-o1 ($p=0.23$), and 0.26 [95%CI: 0.16-0.40] for Llama 3.1-405B ($p=0.77$). Average accuracy on providing 3 differential diagnoses were 0.53 [95%CI: 0.39-0.66] for DeepSeek-R1, 0.57 [95%CI: 0.43-0.69] for ChatGPT-o1, and 0.49 [95%CI: 0.35-0.62] for Llama 3.1-405B (**Extended Data Table 3**).

Additionally, a 5-point Likert scale (1: poor quality; 5: excellent) was used to evaluate the models' clinical reasoning on NEJM cases, assessed by two physicians who are specialized in internal medicine. Evaluation of the generated reasoning yielded scores of 3.61 for DeepSeek-R1, 3.22 for ChatGPT-o1 ($p=0.005$), and 3.13 for Llama 3.1-405B ($p<10^{-3}$).

Then, the **BMJ Endgames**[11] were used to evaluate the models' clinical management capabilities. These were also quantified with 5-point Likert scores by two physicians, with scores of 4.58 for DeepSeek-R1, compared to 4.28 for ChatGPT-o1 ($p<10^{-3}$), and 4.14 for Llama 3.1-405B ($p<10^{-3}$).

When evaluated on the **Medicilline database**[12] of medical cases with multiple-choice questions (with multiple possible correct answers), DeepSeek-R1 achieved an accuracy of 0.74 [95%CI: 0.68-0.80], comparable to ChatGPT-o1 with an accuracy of 0.76

[95%CI: 0.68-0.80] (p=0.06) and better than Llama 3.1-405B with an accuracy of 0.66 [95%CI: 0.59-0.72] (p=0.01).

Four independent radiologists (two for the MIMIC dataset and two for the private dataset) performed clinical and linguistic evaluation of the radiological impressions generated by models from the "Findings" section of radiological reports using a 5-points Likert score. Average scores were 4.6 for DeepSeek-R1, 4.8 for ChatGPT-o1, and 4.5 for Llama 3.1-405B on a private dataset (100 reports) and 4.3, 4.8, and 4.6, respectively, on the MIMIC-III dataset (100 reports)[6] (**Extended Data Table 4**). The two evaluators consistently gave lower scores to Llama 3.1-405B and DeepSeek-R1 compared to ChatGPT-o1 ($p<10^{-3}$). Additionally, the mean evaluation scores of the management recommendation of Deepseek-R1 and Llama 3.1-405B on both datasets were significantly lower than ChatGPT-o1 (3.8 and 3.9 vs 4.6 respectively, $p<10^{-3}$ for the 2 comparisons), with subsequent lower scores for harmlessness. The number of words per output for the MIMIC (average initial report: 187 words) and private reports (average initial report: 222 words) were 128 and 138 for DeepSeek-R1, 116 and 131 for ChatGPT-o1, and 151 and 169 for Llama 3.1-405B, demonstrating limited conciseness for DeepSeek-R1 and Llama 3.1-405B.

Accuracy of the models for **RECIST 1.1 tumor treatment response classification** (compared to consensus evaluation by human experts as the gold standard) **on two consecutive imaging reports** of 100 cancer patients was 0.73 [95%CI: 0.63-0.80] for DeepSeek-R1, compared to 0.72 [95%CI: 0.63-0.80] for Llama 3.1-405B (p=0.83) and 0.81 [95%CI: 0.72-0.88] for ChatGPT-o1 (p=0.10) (**Extended Data Table 5**).

We evaluated model output reproducibility by performing 3 runs of each task for all models. On the USMLE dataset, Fleiss' kappa was 0.96 for DeepSeek-R1, 0.98 for ChatGPT-o1, and 0.95 for Llama 3.1-405B. On the Medicilline dataset, Fleiss' kappa was 0.61 for DeepSeek-R1, 0.93 for ChatGPT-o1, and 0.87 for Llama 3.1-405B. On the RECIST dataset, Fleiss' kappa was 0.80 for DeepSeek-R1, compared to 0.87 for ChatGPT-o1 and 0.96 for Llama 3.1-405B. BERTScore F1-scores ranged from 0.82-0.93 for DeepSeek, 0.84-0.93 for ChatGPT-o1, and 0.93-0.97 for Llama 3.1-405B on NEJM case reasoning, BMJ management reasoning, and report summarization tasks with text outputs (**Extended Data Table 6** and **Extended Data Table7**).

Inter-model kappa ranged from 0.52-0.77 between DeepSeek-R1 and Llama 3.1-405B, 0.60-0.91 between Deepseek-R1 and ChatGPT-o1, and 0.44-0.77 between ChatGPT-o1 and Llama 3.1-405B on USMLE, Medicilline multiple-choice questions, NEJM multiple-choice, and RECIST classification (**Extended Data Table 8**).

Our results offer initial insights into the capabilities and limitations of DeepSeek-R1 compared to two major LLMs, ChatGPT-o1 and Llama 3.1-405B, across a range of medical tasks.

Our evaluation demonstrates notable variability among LLMs when assessed on the USMLE dataset. ChatGPT-o1 exhibited superior performance with the highest overall accuracy (0.95), significantly outperforming both DeepSeek-R1 (0.92; p=0.04) and Llama 3.1-405B (0.83; $p<10^{-3}$). While DeepSeek-R1 was closely comparable to ChatGPT-o1, its slightly lower accuracy suggests potential room for improvement in medical knowledge recall and application. In contrast, Llama 3.1-405B demonstrated markedly inferior accuracy, highlighting considerable differences in clinical knowledge capabilities across models.

The models showed mixed diagnostic and clinical reasoning capabilities across multiple medical evaluation tasks. On diagnostic assessments using NEJM Case Challenges, DeepSeek-R1 performed comparably to ChatGPT-o1 in both multiple-choice (0.57 vs. 0.55, p=0.76) and open-ended scenarios (0.36 vs. 0.32, p=0.23), but both clearly struggled without answer options, underscoring persistent challenges in open-ended differential diagnosis tasks. Despite the advantage of DeepSeek-R1 in reasoning scores (Likert 3.61 vs. ChatGPT-o1's 3.22; p=0.005), clinical management evaluations through the BMJ Endgames further highlighted DeepSeek-R1's strength (Likert 4.58), significantly outperforming both ChatGPT-o1 and Llama 3.1-405B. Additionally, both DeepSeek-R1 (accuracy 0.74) and ChatGPT-o1 (accuracy 0.76; p=0.06) demonstrated strong and comparable performance on Medicilline's multiple-choice questions, highlighting their suitability for tasks requiring general diagnostic and management knowledge.

Radiological impression generation assessments revealed consistent superiority of ChatGPT-o1 in terms of clinical and linguistic quality (average Likert scores 4.8 vs. 4.3-4.6; $p<10^{-3}$). ChatGPT-o1 also excelled significantly in formulating clinical management recommendations (average 4.6 vs. 3.8-3.9 for DeepSeek-R1 and Llama 3.1-405B; $p<10^{-3}$). Notably, ChatGPT-o1 demonstrated superior conciseness, generating shorter but clearer outputs compared to DeepSeek-R1 and Llama 3.1-405B.

In tumor response classification tasks based on imaging reports (RECIST 1.1), ChatGPT-o1 again demonstrated superior accuracy (0.81), though not significantly surpassing DeepSeek-R1 (0.73, p=0.10) and Llama 3.1-405B (0.72, p=0.83). This suggests a moderate performance ceiling for current models in complex oncologic response evaluation tasks, emphasizing a critical area for targeted refinement.

Our reproducibility analysis underscores substantial differences in model stability across tasks. While all models demonstrated high reproducibility on the USMLE dataset, significant variability emerged in the Medicilline and RECIST datasets, particularly for DeepSeek-R1 (Fleiss' kappa 0.61 and 0.80, respectively), compared to ChatGPT-o1 (0.93 and 0.87) and Llama 3.1-405B (0.87 and 0.96). These differences highlight the necessity of rigorous reproducibility testing when deploying LLMs clinically. Inter-model

agreement analyses (Fleiss' kappa ranging from 0.44 to 0.91) revealed varying degrees of consensus, indicating that despite similarities, models leverage distinct internal reasoning pathways. These results illustrate the potential for leveraging complementary strengths through ensemble approaches or hybrid model designs to improve clinical reliability.

This study has several limitations. First, the use of publicly available datasets raises the possibility of inclusion in the original training data for models. The oldest diagnostic challenge used in this study was published in 2016. In order to balance the needs of reproducibility of LLM research using publicly available databases while mitigating data leak in training sets, we included one private dataset of medical cases with multiple-choice questions (Medicilline, not available online), and two private datasets of radiological reports for the summarization and classification tasks. Additionally, datasets included the most recent cases to further decrease the risk of data leak.

Second, it is important to acknowledge that no imaging tasks were included for consistency of comparison of these non-multimodal models. DeepSeek-R1 lacks support for image-based tasks, while DeepSeek-Janus, an older and smaller model, supports image interpretation. Similarly, ChatGPT-o1 and Llama 3.1-405B don’t allow image interpretation. Future studies could incorporate multimodal models with direct image analysis capabilities once they become available.

Third, our analysis was restricted to Azure API-based models for consistency; thus, performance might differ from the same LLMs when accessed through their native applications or other deployment environments. Additionally, these platforms do not provide the exact parameters used, and the online applications may use data shared for their future training, raising questions on their use in research settings.

Fourth, reasoning evaluation of NEJM cases, management evaluation of BMJ cases, and quality evaluation of generated Impressions are subject to human variability. All the tasks were performed independently by blinded evaluators (the evaluators were not aware of which output corresponded to which LLM).

Fifth, the constant and rapid evolution of the LLMs may limit the validity of the current results when updated versions of these models are released.

In conclusion, DeepSeek-R1 models exhibit both strengths and limitations compared to ChatGPT-o1 and Llama 3.1-405B in general medical knowledge and clinical reasoning. Its open-source nature offers a solution for structured medical tasks. Its reasoning capabilities are of great interest, but summarization and repeatability issues exist. Future studies may focus on enhancing clinical reasoning and open-ended diagnostic capabilities of large language models, especially in complex clinical scenarios where

current models exhibit limitations. Investigating methods to increase reproducibility, consistency, and leveraging complementary strengths would also be beneficial.

Acknowledgments:

We thank Guillaume Zagury and Médicilline editions for sharing their dataset. We thank Dr. Venturelli, Dr. Chiche, Dr. Beaziz, and Dr. Lejoyeux for their help updating these questions based on the latest medical knowledge. This project is supported by the Eric and Wendy Schmidt AI in Human Health Fellowship, a program of Schmidt Sciences. M.T. is supported by the French Society of Radiology and the French Musculoskeletal Imaging Society. Figure 1 was created in BioRender.

Author contribution Statement:

Concept and design: MT, ZL, HCL, ZAF, XM.
Acquisition, analysis or interpretation of data: MT, ZL, MY, VF, YM, JH, IB, HA, CH, ASP, AG, AM, NY, NN, PR, AZ, SL, MH, TD, BT, HCL, ZAF, XM.
Drafting of the paper: MT, ZL, MY, VF, AM, IB, SL, BT, HCL, ZAF, XM.
Critical revision of the paper and final draft: All authors.

Competing Interest Statement:

T.D. is the managing partner of RadImageNet LLC and a paid consultant to GEHC and AirsMedical. X.M. is a paid consultant to RadImageNet LLC. The other authors declare no competing interests.

Figure Captions:

Figure 1: Overview of this study comparing three Large Language Models in performing the four indicated medical tasks. Abbreviations: pt: point, PD: Progressive Disease, PR: Partial Response, N: Number.

Figure 2: Comparative performance of three LLMs for a variety of medical tasks. The performance of the three LLMs (ChatGPT-o1, DeepSeek-R1 and Llama 3.1-405B) was evaluated in the following medical tasks: United States Medical Licensing Examination (USMLE) questions, diagnostic questions (in the NEJM and Medicilline databases), reasoning (in the BMJ and NEJM databases), RECIST 1.1 classification from radiological reports, and report summarization (in MIMIC-III and private datasets). For USMLE questions, diagnostic challenges, and RECIST 1.1 classification, bar plots for each model represent the mean of the average output score from each model over three runs for each model, with error bars indicating the standard deviation of the average output scores across those runs. The average output score is the accuracy of the model for each respective task as determined from the average outcome of the task over the number of patient datasets included for each task, respectively. In the USMLE questions, Step 1 assesses basic science, Step 2 evaluates clinical knowledge, and Step 3 tests patient management skills. For RECIST 1.1 classification, evaluations are shown for individual classes: complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD), as well as overall model performance. For the reasoning evaluation using the BMJ and NEJM databases, the bar plots for each model indicate the mean of the average Likert score from two human reviewers, and the error bars represent the standard deviation between their scores. The average Likert score for one human reader is the average rating of the accuracy or appropriateness of the LLM-based output by the human reader over all patient datasets used in the respective task. For report summarization, the radar charts present model performance on report summarization in the MIMIC-III and private datasets across eight criteria, including scientific terminology, coherence, lack of bias, harmlessness, comprehensiveness, specific diagnosis, differential diagnosis, and management recommendations. The numbers shown on the radar charts (e.g., 0, 3.5, and 5.0) correspond to values on a 5 points Likert score ranging from 0 to 5. The average 5-point Likert score over all patient datasets evaluated for each task and for both human readers is shown on respective spokes.

References:

1. The Lancet Digital Health, null. Large language models: a new chapter in digital health. *Lancet Digit. Health* **6**, e1 (2024).
2. Gibney, E. Scientists flock to DeepSeek: how they're using the blockbuster AI model. *Nature* (2025) doi:10.1038/d41586-025-00275-0.
3. Conroy, G. & Mallapaty, S. How China created AI model DeepSeek and shocked the world. *Nature* (2025) doi:10.1038/d41586-025-00259-0.
4. OpenAI. GPT-4 Technical Report. Preprint at http://arxiv.org/abs/2303.08774 (2023).
5. Grattafiori, A. *et al.* The Llama 3 Herd of Models. Preprint at https://doi.org/10.48550/arXiv.2407.21783 (2024).
6. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
7. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: Evaluating Text Generation with BERT. Preprint at https://doi.org/10.48550/arXiv.1904.09675 (2020).
8. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**, 378–382 (1971).
9. United States Medical Licensing Examination dataset. https://www.usmle.org/exam-resources.
10. The New England Journal of Medicine. Case challenges. https://www.nejm.org/case-challenges.
11. BMJ Endgames. https://www.bmj.com/education/endgames.
12. Medicilline dataset. http://www.medicilline.com/13-externe-ecni.

13. Suh, P. S. *et al.* Comparing Large Language Model and Human Reader Accuracy with New England Journal of Medicine Image Challenge Case Image Inputs. *Radiology* **313**, e241668 (2024).
14. Zhang, L. *et al.* Constructing a Large Language Model to Generate Impressions from Findings in Radiology Reports. *Radiology* **312**, e240885 (2024).
15. Laurent, G., Craynest, F., Thobois, M. & Hajjaji, N. Automatic Classification of Tumor Response From Radiology Reports With Rule-Based Natural Language Processing Integrated Into the Clinical Oncology Workflow. *JCO Clin. Cancer Inform.* **7**, e2200139 (2023).
16. Penny, P., Bane, R. & Riddle, V. Advancements in AI Medical Education: Assessing ChatGPT's Performance on USMLE-Style Questions Across Topics and Difficulty Levels. *Cureus* **16**, e76309 (2024).

Online Methods

This retrospective study, using anonymized diagnostic imaging reports, was approved by the local institutional review board of the Icahn School of Medicine at Mount Sinai (IRB protocol number GCO# 20-2199). The requirement to obtain individual informed consent was waived.

Large Language Models

LLMs can process and generate text that mirrors human-like conversation. This task is achieved by predicting what word (token) should come next given an input, allowing them to generate coherent, contextually relevant responses for diverse tasks, such as question answering, summarization, and report generation in the medical field[17]. In this study, DeepSeek-R1 was evaluated in comparison with two major LLMs, ChatGPT-o1 (proprietary model) and Llama 3.1-405B (open-access model), on various medical tasks. All models were implemented using standardized computational resources on an Azure Cloud infrastructure (Azure OpenAI Service, Azure AI Service, Azure AI Foundry), following each LLM provider's guidelines and instructions for Azure deployment. By leveraging Azure's consistent computational environment, we ensured uniform experimental conditions, enabling reproducibility and rigorous comparative evaluation. All LLMs evaluated in our study are instruct-version models.

ChatGPT-o1[18], developed by OpenAI, and released in September 2024, is optimized for enhanced reasoning capability. It demonstrates strong performance across various domains, including the medical field[19].

LLama3.1[5], developed by Meta, is a powerful open-source LLM. In this study, the Llama model was applied using its default parameters with temperature of 1.0 and top_p of 1.0.

Deepseek-R1 has been developed to use large-scale post-training reinforcement learning to improve the reasoning of LLMs and has demonstrated superior performance against GPT-o1 in various benchmark datasets[20].

Study Design

This comparative study evaluated the performance of these three LLMs—DeepSeek-R1, ChatGPT-o1 and Llama 3.1-405B—on four distinct medical tasks. The models are summarized in **Extended Data Table 1**. Each of the different tasks was chosen to determine the capability of these models in different contexts: i) to answer medical multiple-choice questions, ii) to produce text-based diagnostic reasoning and management options in complex clinical case scenarios, iii) to perform classification, and iv) to perform summarization. No image interpretation was evaluated since these models are not multimodal. Additionally, to investigate the robustness of the models we

assessed the test-retest repeatability of the outputs of each model for each task from three separate queries with the same prompts. All of the prompts used are available in **Extended Data Table 2**.

Task 1: USMLE Questions

Three hundred and twenty-three USMLE-style questions were used for evaluation. These included questions from Step 1 (95 questions), Step 2 (106 questions), and Step 3 (122 questions) of the examination, comprising the entire publicly available database (https://www.usmle.org/exam-resources), excluding the questions involving image interpretation. Step 1 questions focus on foundational sciences (e.g. biochemistry, pathology), Step 2 questions emphasize clinical knowledge, and Step 3 questions require reasoning for patient management.

Each LLM was prompted with identical questions using standardized formats, including the question stem and answer options. The models provided their final answers without additional clarifications or follow-ups. The accuracy of each model was the proportion of correct responses.

Task 2: Diagnostic and Management Case Challenges

Fifty NEJM "Case Challenges" (the entire online available database) were selected to evaluate the models' interpretive skills, in a similar manner to previous studies[13,21]. These challenges included complex clinical scenarios with text and tables of biological values. These cases typically require integration of patient history, laboratory findings, description of imaging examinations, and clinical reasoning to arrive at a diagnosis.

The models were tested in two ways. Firstly, the LLM was prompted to select the correct diagnosis from a list of multiple-choices provided in the prompt. Data provided in the prompt included patient history, symptoms, and all available data provided in the Case Challenge. The model was also prompted to provide the reasoning associated with the answer. Secondly, the LLM was prompted to provide the most probable diagnosis and two differential diagnoses based in an open-ended answer format, based on the same information (but not including the multiple-choice answers provided in the first test). For each test, the accuracy of each model was the proportion of correct answers. For the reasoning portion of each test, the responses of the LLMs were evaluated on their ability to use inductive, deductive, and probabilistic reasoning to progressively narrow down diagnosis possibilities based on the available clinical history, physical exam observations, and test results. Independently, two Internal Medicine specialists evaluated the pertinency of the elements evaluated, their interpretation, and how the combination of the clinical examination and different tests were handled to achieve the diagnosis on a 5-point Likert scale (1: poor quality; 5: excellent). We adapted the score from the R-IDEA score[22]. However, we did not use it in this case

because we specifically asked the models to provide differential diagnoses. The expert reviewers also evaluated the accuracy of the LLM-derived differential diagnoses based on the number of plausible differential diagnoses provided in the open-ended questions: one point was awarded for each plausible diagnosis up to a maximum of 3 (0-3/3).

Fifty BMJ “Endgames” which included clinical management questions were also used to evaluate the management capabilities of the models. These cases usually evaluate the image interpretation, diagnostic skills, and management capabilities of the readers with open-ended questions. We provided each model with the text prompts and final diagnosis (as image analysis is needed to reach final diagnosis) and prompted questions about the management of these conditions (as provided in the original cases). The clinical management plans proposed by the models were evaluated by two physicians using a 5-point Likert score based on the different therapeutic options proposed, the harmfulness of the treatment, and the essential measures associated with medical treatment (1: poor management; 5: excellent management). The management suggestions provided in the BMJ cases were used as the reference standard for adjudication by the physicians.

Finally, 200 questions evaluating diagnostic and management capabilities were extracted from a private French database (ECN-Integrale, from Medicilline Editions). This series of textbooks contain training material for medical students and residents but are not available online, preventing leakage into the training datasets of the LLMs.. These multiple-choice questions allowed multiple correct answers. Correct answers were rated Questions were translated into English, then updated and verified by experts in the different fields, including one who is involved in student training who has more than 10 years of experience. The LLMs were provided with the questions in text format and prompted to select the correct answers. Correct answers were rated scored as 1, answers that contained one mistake among the multiple-choices was scored as 0.5, and answers that contained 2 or more mistakes was scored as 0. The performance of each model was the sum of the scores divided by the maximum possible total scores.

Task 3: Findings to Impression

The “Findings” section of 200 imaging reports, 100 from an outpatient radiology facility in the New York Metropolitan area and 100 from MIMIC[6] reports (each with 20 examples of X-rays, ultrasound, CT-scan, MRI and PET/CT), were provided to the 3 LLMs for summarization. Four readers (radiologists with 3, 4, 6 and 7 years of experience) each evaluated 100 Impressions generated by these LLMs (2 radiologists evaluated the MIMIC dataset and 2 the local dataset). The readers were blinded to which model provided the outputs. Nine domains were evaluated with a 5-point Likert score based on previous studies evaluating similar tasks[14,23]. The average number of words and the ratio of the number of words in the LLM-derived Impression and the number of words in

the original Findings Impression/Findings were also evaluated for each model to assess the summarization capabilities.

Task 4: RECIST evaluation

Reports from two consecutive oncological chest-abdomen-pelvis CT-scans were provided to each model which was subsequently prompted to classify the progression of the disease according to RECIST 1.1[24]. No baseline or NADIR were available in these reports. The models were prompted to provide a text summary. Each model evaluated data from 100 patients from an outpatient radiology facility in New York. Two radiologists with 3 and 5 years of experience in cancer imaging independently evaluated the classification. An expert radiologist with more than 10 years of experience resolved any discrepancy. All evaluators were blinded to the LLM providing the output. The accuracy of the LLM-derived summary compared to the Radiologists' interpretation was assessed.

Test-Retest Reproducibility

To evaluate the reproducibility of each model, we performed a test-retest analysis by prompting each model 3 times with the same prompts for each task. Depending on the tasks, the Fleiss kappa score was used for categorical/multiple-choice questions (like USMLE, NEJM accuracy, Medicilline, and RECIST), the Cohen kappa score was used for comparison between two runs of 2 models, and the BERTscore was used for text outputs (reasoning outputs, management options, and Impression summarizations).

Statistical Analysis

SPSS v28.0 was used for statistical analysis. Descriptive statistics were used to summarize accuracy metrics or linguistic evaluation (for Task 3) for each model across tasks. Confidence intervals were calculated using the Wilson score method. McNemar tests were performed for the comparisons between 2 models for paired-sample analysis and Mann-Whitney U test for independent samples. A p-value less than 0.05 was considered significant. Variability between the 3 runs of each model was analyzed using Fleiss's kappa score for multiple-choice questions, Cohen's kappa score for comparison between 2 runs/models,) for multiple-choice questions and the BERTscore for text outputs.

Data Availability

The USMLE dataset is available at https://www.usmle.org/exam-resources. The NEJM Case Challenges are available at https://www.nejm.org/case-challenges. The BMJ Endgames are available at https://www.bmj.com/specialties/endgames. The MIMIC-III dataset can be requested at https://physionet.org/content/mimiciii/1.4/. The Médicilline dataset (http://www.medicilline.com/) of multiple-choice questions for clinical diagnosis

and management translated in English is available upon request for private/research use only, after agreement of Medicilline. Mediciline data requests should be sent to Mickael.Tordjman@mssm.edu. In response to the inquiry, the timeframe for responding to requests is approximately within 2 weeks of the request. The radiological reports used for summarization and RECIST classification are not available due to privacy issues.

Code Availability

All studies were conducted using Azure OpenAI service (ChatGPT), Azure AI Foundry (Llama), and Azure AI service (DeepSeek). The algorithms used in this study are based on the official guidelines provided by the developers of the Large Language Models evaluated—DeepSeek, ChatGPT, and Llama. The implementation followed these guidelines strictly to ensure consistency and reproducibility of results. As the study did not involve the development of new code but rather the application of existing, officially provided algorithms, specific source codes referenced are proprietary and maintained by their respective developers.

References:

17. Zhao, W. X. *et al.* A Survey of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2303.18223 (2024).
18. Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/.
19. Temsah, M.-H., Jamal, A., Alhasan, K., Temsah, A. A. & Malki, K. H. OpenAI o1-Preview vs. ChatGPT in Healthcare: A New Frontier in Medical AI Reasoning. *Cureus* **16**, e70640 (2024).
20. DeepSeek-AI *et al.* DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. Preprint at https://doi.org/10.48550/arXiv.2501.12948 (2025).
21. Tenner, Z. M., Cottone, M. C. & Chavez, M. R. Harnessing the open access version of ChatGPT for enhanced clinical opinions. *PLOS Digit. Health* **3**, e0000355 (2024).

22. Cabral, S. *et al.* Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. *JAMA Intern. Med.* **184**, 581 (2024).
23. Park, J., Oh, K., Han, K. & Lee, Y. H. Patient-centered radiology reports with generative artificial intelligence: adding value to radiology reporting. *Sci. Rep.* **14**, 13218 (2024).
24. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer Oxf. Engl. 1990* **45**, 228–247 (2009).

1- USMLE questions

USMLE	*Accuracy*

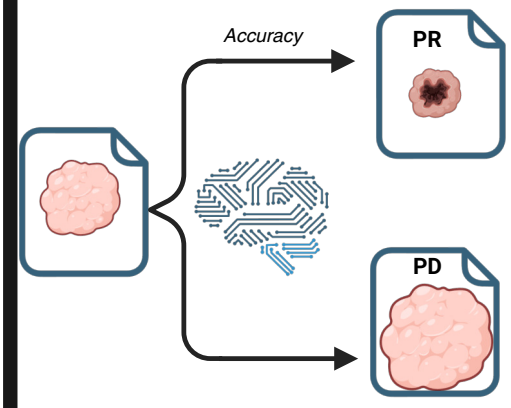
2- Text-based cases

The New England Journal of Medicine	• *Accuracy* • *Differential diagnoses* • *Reasoning (5 pt-Likert)*
BMJ	*Management (5 pt-Likert)*
Medicilline	*Accuracy*

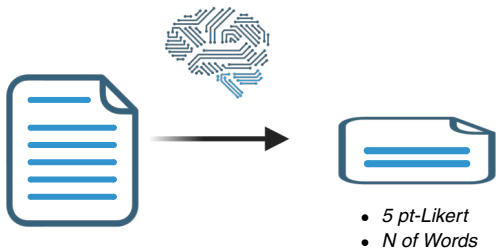
OpenAI

deepseek

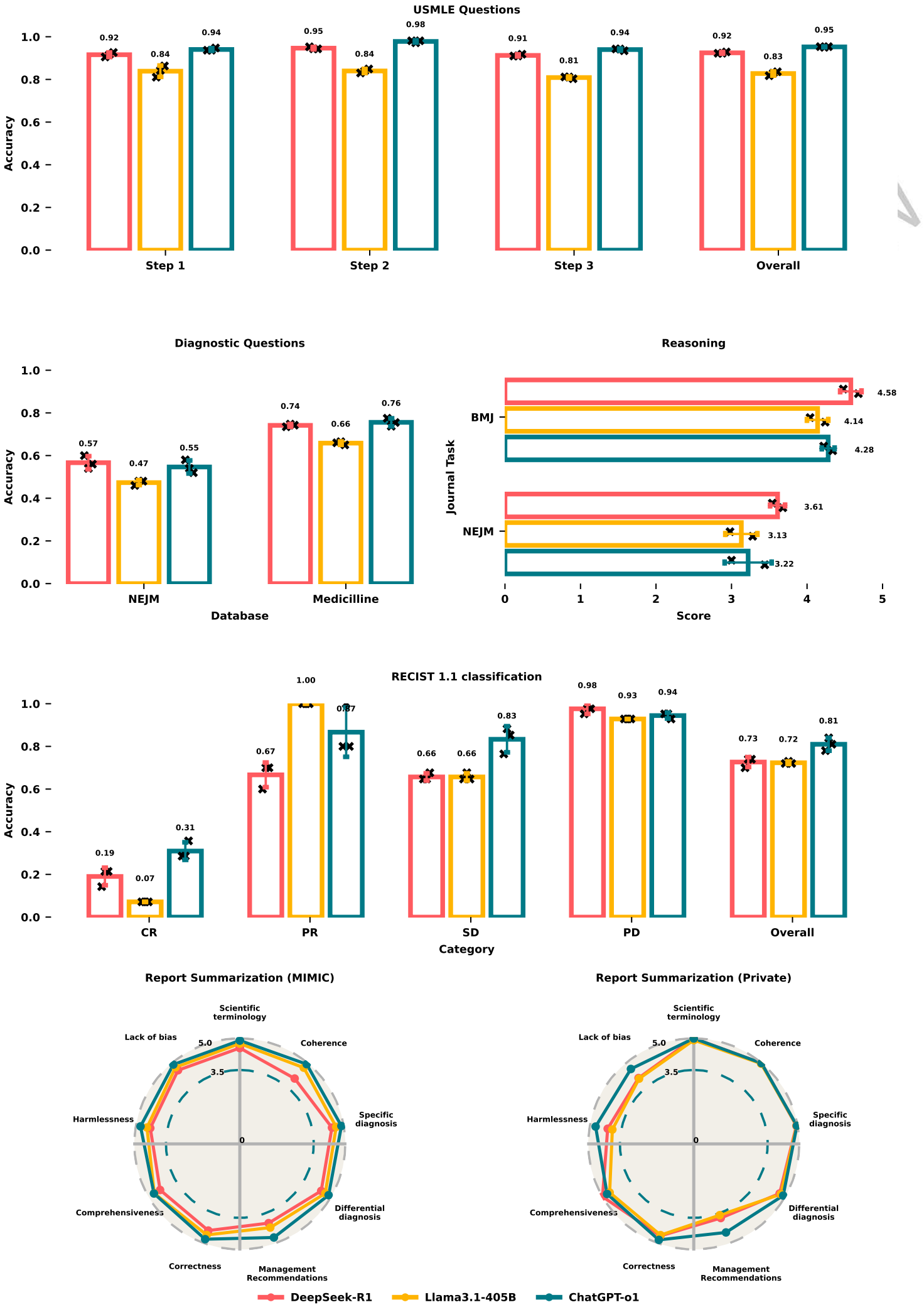
3- RECIST 1.1 classification



4- Imaging Report Summarization



- *5 pt-Likert*
- *N of Words*


USMLE Questions
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.92
0.84
0.94
0.95
0.84
0.98
0.91
0.81
0.94
0.92
0.83
0.95
Step 1
Step 2
Step 3
Overall
Diagnostic Questions
Accuracy
0.57
0.47
0.55
0.74
0.66
0.76
NEJM
Medicilline
Database
Reasoning
Journal Task
BMJ
NEJM
4.58
4.14
4.28
3.61
3.13
3.22
0
1
2
3
4
5
Score
RECIST 1.1 classification
Accuracy
0.19
0.07
0.31
0.67
1.00
0.87
0.66
0.66
0.83
0.98
0.93
0.94
0.73
0.72
0.81
CR
PR
SD
PD
Overall
Category
Report Summarization (MIMIC)
Report Summarization (Private)
Scientific terminology
Coherence
Specific diagnosis
Differential diagnosis
Management Recommendations
Correctness
Comprehensiveness
Harmlessness
Lack of bias
5.0
3.5
0
DeepSeek-R1
Llama3.1-405B
ChatGPT-o1

Models	Parameters	Context length
ChatGPT-o1	N/A	128K
Llama3.1-405B	405B	128K
Deepseek-R1	671B	128K

Tasks	Prompts
USMLE/Medicilline	[Question] + [Options]
Text-based cases and Reasoning	[Text description] + [Options]+ “Please provide the reasoning steps”
Differential	'"Please provide the most likely diagnosis and the 2 additional most likely differential diagnoses.'"
Management	'" You are a physician. Below is a medical case: What is the management of this condition?'"
Findings to impression summarization	""You are a radiologist, summarize the findings in this report into an impression containing the clinically significant findings and recommend any following up radiologic testing, if needed"" + [Findings]
RECIST classifications	"""As a radiologist, utilize the Response Evaluation Criteria in Solid Tumors (RECIST) 1.1 to evaluate cancer treatment responses. The RECIST guidelines classify tumor responses into four categories: Complete Response (CR): The tumor is no longer detectable. Partial Response (PR): The tumor size has reduced by at least 30%. Stable Disease (SD): The tumor size remains unchanged, or does not meet the criteria for PR or PD. Progressive Disease (PD): The tumor size has increased by at least 20% or at least one new lesion appeared. Your task is to analyze radiology reports from two separate visits for the same patient and classify the tumor's response according to the RECIST criteria. Only respond with your classification result, choose one from CR, PR, SD or PD. Radiology Report 1: [Report1] Radiology Report 2: [Report2] """

	DeepSeek-R1	**Llama 3.1-405B**	**ChatGPT-o1**
BMJ Management	4.58	4.14	4.28
NEJM Reasoning	3.53	3.14	3.11
NEJM Accuracy (MCQ)	0.57 [CI: 0.49-0.64]	0.47 [CI: 0.40-0.55]	0.55 [CI: 0.47-0.62]
NEJM Accuracy (Open-ended)	0.36 [CI: 0.24-0.50]	0.26 [CI: 0.16-0.40]	0.32 [CI: 0.21-0.46]
NEJM Differential	0.53 [CI: 0.39-0.66]	0.49 [CI: 0.35-0.62]	0.57 [CI: 0.43-0.69]
Medicilline Database	0.74 [CI: 0.68-0.80]	0.66 [CI: 0.59-0.72]	0.76 [CI: 0.68-0.80]

	Deepseek R1		**Llama 3.1-405B405B**		**ChatGPT-o1**	
	Private	*MIMIC*	*Private*	*MIMIC*	*Private*	*MIMIC*
Scientific terminology	5.0	4.5	4.9	4.8	5.0	4.9
Coherence	5.0	4.0	5.0	4.7	5.0	4.9
Specific diagnosis	4.9	4.4	5.0	4.6	5.0	4.9
Differential diagnosis	4.7	4.5	4.8	4.7	4.9	4.9
Management Recommendations	3.7	4.0	3.6	4.2	4.5	4.7
Correctness	4.6	4.4	4.6	4.6	4.8	4.8
Comprehensiveness	5.0	4.4	4.6	4.7	4.7	4.7
Harmlessness	4.1	4.3	3.9	4.5	4.7	4.8
Lack of bias	4.1	4.6	4.0	4.7	4.6	4.9
Mean score	**4.6**	**4.3**	**4.5**	**4.6**	**4.8**	**4.8**
Word Count	138	128	169	151	131	116
Ratio Impression/ Findings Word	**0.62**	**0.68**	**0.76**	**0.81**	**0.59**	**0.62**

	DeepSeek-R1	**Llama 3.1-405B405B**	**ChatGPT-o1**
CR	0.21 (3/14)	0.07 (1/14)	0.29 (4/14)
PR	0.70 (7/10)	1 (10/10)	0.90 (9/10)
SD	0.65 (22/34)	0.65 (22/34)	0.82 (28/34)
PD	1 (42/42)	0.93 (39/42)	0.95 (40/42)
Overall Accuracy	**0.74 (74/100)**	**0.72 (72/100)**	**0.81 (81/100)**

	USMLE (n=323)	RECIST (n=100)	Medicilline (n=200)	NEJM accuracy (n=50)
DeepSeek-R1 vs Llama 3.1 405B	0.77	0.61	0.53	0.59
DeepSeek-R1 vs Chatgpt-o1	0.91	0.68	0.69	0.60
Llama 3.1 405B vs Chatgpt-o1	0.77	0.66	0.57	0.44

	Private Report	MIMIC	BMJ	NEJM Reasoning	NEJM Differential Diagnoses
DeepSeek-R1	0.909	0.906	0.849	0.876	0.873
Llama 3.1 405B	0.962	0.961	0.940	0.936	0.945
Chatgpt-o1	0.919	0.925	0.892	0.865	0.944

	USMLE	RECIST	Medicilline
DeepSeek-R1	0.962	0.803	0.611
Llama 3.1 405B	0.948	0.958	0.929
Chatgpt-o1	0.976	0.875	0.870

nature portfolio

Corresponding author(s): Zahi Fayad, Xueyan Mei, and Hao-Chih Lee

Last updated by author(s): Apr 17, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used in data collection.
Data analysis	We used DeepSeek-R1 (DeepSeek), LLaMA 3.1 405B (Meta), and ChatGPT-o1(OpenAI)'s official scripts to evaluate the model performance on Azure Clouds. The Azure services used in this study are: https://ai.azure.com/explore/models/Meta-Llama-3.1-405B-Instruct/version/1/registry/azureml-meta?tid=77e89d61-570f-43b0-b9e4-634f462e34b8#details https://azuremarketplace.microsoft.com/en/marketplace/apps/metagenai.meta-llama-3-1-405b-instruct-offer?tab=overview https://ai.azure.com/explore/models/DeepSeek-R1/version/1/registry/azureml-deepseek?tid=77e89d61-570f-43b0-b9e4-634f462e34b8 SPSS v28.0 was used for statistical analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The USMLE dataset is available at https://www.usmle.org/exam-resources. The NEJM Case Challenges are available at https://www.nejm.org/case-challenges. The BMJ Endgames are available at https://www.bmj.com/specialties/endgames. The MIMIC-III dataset can be requested at https://physionet.org/content/mimiciii/1.4/. The Médicilline dataset (http://www.medicilline.com/) of multiple-choice questions for clinical diagnosis and management translated in English is available upon request for private/research use only, after agreement of Medicilline. Mediciline data requests should be sent to Mickael.Tordjman@mssm.edu. In response to the inquiry, the timeframe for responding to requests is approximately within 2 weeks of the request. The radiological reports used for summarization and RECIST classification are not available due to privacy issues.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	The NEJM and BMJ case challenges cover patients with diverse sex and gender. The imaging reports of the private dataset used for summarization and RECIST classification have a balanced sex distribution.
Reporting on race, ethnicity, or other socially relevant groupings	Race, ethnicity, or other socially relevant groupings were not relevant to the objectives or analyses of this study and therefore were not collected or analyzed. This study focused solely on the comparative performance of the Large Language Models across different medical tasks without considering these social variables.
Population characteristics	The population characteristics in this study are related to the challenges presented in the New England Journal of Medicine and the BMJ Endgames. These sources provided the basis for the tasks assigned to the evaluated Large Language Models, focusing on their application in diverse and complex analytical settings.
Recruitment	The reports of the private dataset were randomly selected from the Radiology Imaging System between 2010 and 2023 for the summarization task. Two cancer imaging reports of chest-abdomen-pelvis CT-scans were selected consecutively for the RECIST classification task. For the case challenges of the medical journals, we included all available cases from NEJM case challenges and the most recent cases from BMJ Endgames including a management section.
Ethics oversight	This study was approved by the Institutional Review Board of the Icahn School of Medicine at Mount Sinai to use the anonymized diagnostic imaging reports, in accordance with the institution's Federalwide Assurances (FWA00005656, FWA00005651), under IRB protocol number GCO# 20-2199-00001-01.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The publicly available datasets used in this study include NEJM Case Challenges published up to March 5, 2025, the 50 most recent BMJ Endgames management cases up to the same date, and the full set of USMLE questions (excluding image-based cases), all sourced from online resources. For the MIMIC-III dataset, we randomly sampled 20 cases per imaging modality (CT, MRI, PET, X-ray, and ultrasound) to ensure balanced coverage across modalities while maintaining feasibility for manual evaluation and computational analysis. For report summarization, we also randomly sampled 20 cases per imaging modality (CT, MRI, PET, X-ray, and ultrasound) from an additional private dataset. For RECIST 1.1 classification, two consecutive cancer imaging reports of 100 patients were randomly selected from clinical notes. 200 questions were randomly chosen for the Medicilline QA task. These sample sizes were chosen to provide adequate representation for each task while ensuring that performance evaluation could be conducted effectively within the scope of this study.
Data exclusions	This study strictly focuses on text-based evaluations of LLM capabilities. Image-based cases from NEJM, BMJ, and USMLE were excluded during the study design because they require multi-modal version of LLMs.

Replication	Each model was evaluated three times per task on Azure Cloud, with all runs completing successfully.
Randomization	The imaging reports were randomly selected from the RIS system between 2010 and 2023. Two cancer imaging reports of chest-abdomen-pelvis CT-scans were selected consecutively for the RECIST classification task. The "Findings" section of 200 imaging reports, 100 from an outpatient radiology facility in the New York Metropolitan area and 100 from MIMIC reports (each with 20 examples of X-rays, ultrasound, CT-scan, MRI and PET/CT reports), were provided to the three LLMs for summarization. Since our study evaluates the performance of LLMs in summarizing medical reports rather than assessing clinical interventions or comparing patient outcomes, random allocation or covariate control typically applied in interventional studies is not relevant here.
Blinding	The human readers participating in evaluation of the report summarization generated by the LLMs were blinded from which model provided the outputs.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*