nature medicine

https://doi.org/10.1038/s41591-025-03727-2

Accelerated Article Preview

Benchmark evaluation of DeepSeek large language models in clinical decision-making

Received: 22 February 2025

Accepted: 21 April 2025

Accelerated Article Preview
Published online: 23 April 2025

Cite this article as: Sandmann, S. et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nature Medicine* https://doi.org/10.1038/s41591-025-03727-2 (2025).

Sarah Sandmann, Stefan Hegselmann, Michael Fujarski, Lucas Bickmann, Benjamin Wild, Roland Eils & Julian Varghese

This is a PDF file of a peer-reviewed paper that has been accepted for publication. Although unedited, the content has been subjected to preliminary formatting. Nature Medicine is providing this early version of the typeset paper as a service to our authors and readers. The text and figures will undergo copyediting and a proof review before the paper is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers apply.

ACCELERATED ARTICLE PREVIEW

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 Fig. 1	Visual abstract.	ExtendedData Figure1.jpg	n.a.
Extended Data Fig. 2	Summarized model performances for diagnosis and treatment recommendation tasks.	ExtendedData Figure2.pdf	Histograms showing the performance of GPT-4o, DeepSeek-R1, Gemini-2.0 Flash Thinking Experimental (Gem2FTE) and DeepSeek-V3 considering diagnosis and treatment, rated with Likert scores. Five points represent the highest possible level of accuracy as assessed by the expert. The red line indicates the mean performance of each model.

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	Yes	SupplementaryInformation.pdf	Supplementary Tables 1-11, Supplementary Figures 1-7
Reporting Summary	Yes	NMED-FT139680B_rs.pdf	
Peer Review Information	No. See main article.	*OFFICE USE ONLY*	

B. Additional Supplementary Files

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

ACCELERATED ARTICLE PREVIEW

	of the same type should be listed in sequence, i.e.: Supplementary Video 1, Supplementary Video 2, etc.	***Supplementary_Video_1.mov***	
Supplementary Table	1	Supplementary_Data_S1.xlsx	Overview of all clinical cases, their source information and assessment.
Supplementary Code	2	Supplementary_Data_S2.R	R-Script generating Figure 1 of the main manuscript.
Supplementary Code	3	Supplementary_Data_S3.R	R-Script generating Figure 2 of the main manuscript.
Supplementary Code	4	Supplementary_Data_S4.R	R-Script for performing power analysis.

Editor summary:

In an evaluation involving 125 standardized patient cases, open-source DeepSeek large language models are shown to perform at least on par with state-of-the-art proprietary large language models in diagnosis and treatment recommendation tasks.

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 and Eric Oermann for their contribution to the peer review of this work.

ACCELERATED ARTICLE PREVIEW

Title

Benchmark evaluation of DeepSeek large language models in clinical decision-making

Author list
Sarah Sandmann[1], Stefan Hegselmann[2], Michael Fujarski[1], Lucas Bickmann[1], Benjamin Wild[2], Roland Eils[3,2]* and Julian Varghese[4]

*corresponding author: roland.eils@bih-charite.de

Affiliations
[1]: Institute of Medical Informatics, University of Münster, Münster, Germany
[2]: Center for Digital Health, Berlin Institute of Health (BIH), Charite - University Medicine Berlin, Berlin, Germany
[3]: Intelligent Medicine Institute, Fudan University, 131 Dongan Road, Shanghai, 200032, China
[4]: Institute of Medical Data Science, Otto-von-Guericke University, Magdeburg, Germany

ACCELERATED ARTICLE PREVIEW

Abstract

Large Language Models (LLMs) are increasingly transforming medical applications. However, proprietary models such as GPT-4o face significant barriers to clinical adoption because they cannot be deployed on site within healthcare institutions, making them non-compliant with stringent privacy regulations. Recent advancements in open-source LLMs such as DeepSeek models offer a promising alternative since they allow efficient fine-tuning on local data in hospitals with advanced IT infrastructure. To demonstrate the clinical utility of DeepSeek-V3 and DeepSeek-R1, we benchmarked their performance on clinical decision support tasks against proprietary LLMs, including GPT-4o and Gemini-2.0 Flash Thinking Experimental. Using 125 patient cases with sufficient statistical power, covering a broad range of frequent and rare diseases, we found that DeepSeek models perform equally well and in some cases better than proprietary LLMs. Our study demonstrates that open-source LLMs can provide a scalable pathway for secure model training enabling real-world medical applications in accordance with data privacy and healthcare regulations.

Main text

Large Language Models (LLMs) are rapidly emerging as transformative tools within medicine, showing promise in various clinical applications[1]. Their potential to process and understand complex medical information offers opportunities to enhance clinical decision-making, automate administrative tasks, and improve patient care[2–4]. LLMs can analyze large volumes of unstructured data from electronic health records, offering clinicians efficient access to relevant patient information for diagnosis and treatment[5]. As AI technology matures, these models are poised to become valuable aids in navigating the ever expanding landscape of medical knowledge and improving healthcare delivery.

However, the integration of LLMs into clinical practice is not without challenges, necessitating careful validation and ethical considerations[6,7]. For LLMs to be integrated into routine clinical care, they must comply with data privacy regulations such as GDPR and HIPAA, as well as medical device regulations like EU-MDR and FDA. This should require LLMs to be explainable, auditable, and fully aligned with strict medical regulations - criteria that proprietary models currently do not meet. Concerns regarding data privacy, algorithmic bias, and the potential for generating inaccurate or misleading information remain paramount[8–10]. As Blumenthal and Goldberg (2025)[11] highlight, managing patient use of generative AI also presents a novel set of challenges, underscoring the need for robust validation frameworks and clear guidelines to ensure the safe and effective implementation of LLMs in clinical settings.

The performance of open source LLMs on benchmarks like lmarena.ai typically have shown inferior performance compared to proprietary state-of-the art LLMs such as GPT-4o. Nonetheless, open source LLMs have caught up as new models like Llama 3.1 or Mistral Large 2 demonstrate significant improvements[12]. Recent advancements in LLM have seen the emergence of state-of-the-art open source models such as DeepSeek-V3, and the development of explicit reasoning models such as Gemini-2.0 Flash Thinking Experimental (Gem2FTE), OpenAI o1, and DeepSeek-R1[13]. With over 500b model parameters, the DeepSeek models belong to the largest LLMs, competing with proprietary ones in LLM leaderboards, while providing the key benefits of transparency and the ability to run the open-source model within the institution's own IT environment at a significantly lower cost compared to proprietary models by OpenAI[14]. While these leaderboards assess model performance on general AI tasks, the critical question remains whether open source models can match proprietary systems in real-world clinical decision tasks including differential diagnosis or treatment planning, and whether enhanced reasoning capabilities also provide benefits in clinical workflows.

Here, we systematically benchmarked open-source and frontier proprietary LLMs with a thorough performance analysis on clinical decision support tasks (Extended Data Figure 1). We systematically assessed the performance in diagnosis and treatment recommendation for DeepSeek-V3 and DeepSeek-R1 as well as the proprietary LLMs GPT-4o and Gem2FTE, currently ranked at the top of the LLM-leaderboard at lmarena.ai.

Although large language models (LLMs) excel on widely used benchmarks such as multiple-choice tests, their evaluation for clinical decision support tasks remains

ACCELERATED ARTICLE PREVIEW

underexplored[15–17]. Currently, no widely accepted benchmark exists for assessing the clinical utility of LLMs. We thus conducted comparisons using a well-curated, previously published set of 110 patient cases[15], originally designed to evaluate GPT-4, GPT-3.5, and Google Search in clinical decision-making. Unlike multiple-choice-based automatic assessments, this benchmark requires expert clinicians to manually evaluate LLM-generated text outputs. These cases, sourced from medical textbooks, replicate the initial patient encounter commonly seen in outpatient or emergency settings by focusing solely on the key details of the doctor-patient dialogue. As a result, they offer an approximation of real-world conditions—where incomplete or extraneous information is common—and help assess the models' practical clinical performance. Model performance was assessed using a 5p Likert Scale to evaluate model output by medical experts (Extended Data Figure 2, Table S1, Fig. S1).

Our focus is on diagnosis and treatment recommendations, as these represent the most consequential and error-prone aspects of clinical decision-making, frequently cited in adverse event analyses and guideline development frameworks[18,19]. To ensure broad coverage, our evaluation spans multiple specialties (internal medicine, neurology, surgery, gynecology, and pediatrics) and includes a balanced mix of frequent, less frequent, and rare diseases. To enhance statistical power, we expanded the benchmark to 125 cases, enabling robust significance testing in systematic pairwise model comparisons with adjustments for multiple testing (see Methods, Table S2, Fig. S2).

For the first clinical decision-making task of diagnosis (Figure 1), Gem2FTE was significantly outperformed by DeepSeek-R1 ($p=5.73\cdot10^{-5}$, rank-biserial correlation $r_{rb}=0.60$) and GPT-4o ($p=7.89\cdot10^{-6}$, $r_{rb}=0.67$). DeepSeek-R1 was on a par with the best performing model GPT-4o ($p=0.3085$, $r_{rb}=0.27$). All new models showed clearly superior performance compared to GPT-4, GPT-3.5 and Google search (Table S3, Fig. S3). Our data indicated consistent performances across clinical specialties (Table S4, Fig. S4). Noteworthy, no clear difference was observed for diagnosis of rare diseases as compared to frequent diseases for all models except for Gem2FTE (unadjusted p-values 0.0004 and 0.0009 respectively; Tables S5-S6, Fig. S5). Notably, this finding is in stark contrast to our finding in the very recent, earlier study benchmarking GPT-4, GPT-3.5 and Google search[15], where both models and Google search underperformed in diagnosis of rare diseases. Interestingly, the reasoning empowered model DeepSeek-R1 did not show improved performance in comparison to DeepSeek-V3 ($p=1$, $r_{rb}=0.03$) (Tables S1, S3, Fig. S1).

In line with the above finding, for the second clinical decision-making task of treatment, both GPT-4o ($p=0.0016$, $r_{rb}=0.50$) and DeepSeek-R1 ($p=0.0235$, $r_{rb}=0.36$) showed superior performance compared to Gem2FTE. Again, no significant differences were observed for GPT-4o vs DeepSeek-R1 ($p=0.1522$, $r_{rb}=0.26$) (Figure 2). Compared to the earlier benchmarked models GPT-4 and GPT-3.5, superior performance could be observed for both GPT-4o and DeepSeek-R1, however, not for Gem2FTE (Table S3, Fig. S6). The model performance for treatment recommendation was not negatively affected by low disease frequency (Tables S5-S6, Fig. S5). Model performance was mostly uniform across all clinical specialties, with Gem2FTE being the only exception for treatment recommendations for neurological cases (Table S4, Fig. S4).

ACCELERATED ARTICLE PREVIEW

The strong performance of DeepSeek-V3 and DeepSeek-R1, matching GPT-4o in both clinical decision-making tasks, suggests that open-source LLMs may serve as valuable assistive tools for complex tasks like diagnosis or differential diagnoses, and treatment recommendation. Surprisingly, Gem2FTE, despite leading the general non-medical benchmark on lmarena.ai, underperformed in clinical decision-making. While its model specifications remain undisclosed, we speculate that Gem2FTE is significantly smaller than DeepSeek-V3/R1 and GPT-4o, with model capacity likely being a key factor in clinical performance. Equally unexpected was the lack of advantage from DeepSeek-R1's reasoning module in medical decision-making. Instead, DeepSeek-R1 generated significantly longer text outputs, increasing response times and reducing conciseness compared to its non-reasoning counterpart. The reasoning finetuning of models such as DeepSeek-R1 is focused on easily verifiable mathematical, coding, and logic tasks[13], and we here found that the impressive improvements in reasoning in these problem domains so far have not extended to clinical reasoning. It is thus tempting to speculate that fine-tuning of reasoning models based on proprietary clinical case reports available within individual caregiver organizations may lead to dramatic improvement in diagnosis and treatment recommendations.

The average performance scores for DeepSeek-R1 were 4.70p (4.48) out of 5p for diagnosis (treatment recommendation). In some cases, the new LLMs successfully generated accurate and current information, particularly for treatment recommendations, where newly updated guidelines, such as those addressing antimicrobial treatment plans, were necessary. Nevertheless, many cases did not achieve the maximum score; for example, with DeepSeek-R1, 60% of cases for diagnosis and only 39% for treatment reached the full score of 5p. These inaccuracies in model predictions could pose potential risks if the output was prompting immediate medical decisions without additional expert oversight. Interestingly, the phenomenon of "Artificial Hallucination," where LLMs generate seemingly plausible but factually incorrect content[20] could only be observed in a small fraction of cases across all models. Overall, these findings reinforce the need for robust validation frameworks and clear guidelines to ensure the safe and effective implementation of LLMs in clinical settings.

Though the tasks evaluated here only cover a portion of potential clinical use cases, our findings suggest a potential supportive benefit for the two highly relevant clinical decision-making tasks of diagnosis and treatment recommendations. We believe the output of these models can be further improved in terms of performance and robustness by adding access to quality checked medical literature or databases, human oversight and transparent learning. In summary, our study demonstrates that open-source LLMs are viable candidates for real-world medical applications. As hospitals prioritize data privacy and regulatory compliance, open-source LLMs provide a scalable pathway for secure and cost effective, institution-specific model training and implementation. Future clinical studies are warranted to assess whether these promising findings can be effectively translated into improved patient outcomes.

ACCELERATED ARTICLE PREVIEW

Acknowledgements

This work was enabled by the HiGHmed consortium funded by the German Ministry of Education and Research, grant number: 01KX2121. RE acknowledges support by the Collaborative Research Center (SFB 1470) funded by the German Research Council (DFG) and by AI4HEALTH funded by the Natural Science Foundation of China (NSCF), grant number: W2441025. The icons of Extended Data Figure 1 were generated using Figma (https://www.figma.com).

Author Contributions

RE, BW and JV conceptualized the project. MF and LB performed data acquisition. SH and JV performed clinical evaluation. SaS performed analyses and drafted the manuscript. RE and JV supervised the study. All authors reviewed and approved the paper.

Competing interests

The authors declare no competing interests.

Figure Legends/Captions

Figure 1: Model performance for diagnosis tasks. (A-D) Bubble plots showing the results of the 125 pairwise comparisons on a five point Likert Scale for GPT-4o vs DeepSeek-R1 (A) (one-sided paired Mann-Whitney test with continuity correction, alternative=greater, Bonferroni correction with k=4, adjusted p=0.3085, V=378, 95% confidence intervals 95%CI=[-3.13·10-7;Inf], estimate=0.25); GPT-4o vs Gemini-2.0 Flash Thinking Experimental (Gem2FTE) (one-sided paired Mann-Whitney test with continuity correction, alternative=greater, Bonferroni correction with k=4, adjusted p=7.89·10-6, V=1576, 95%CI=[0.5;Inf], estimate=0.75) (B); DeepSeek-R1 vs Gem2FTE (one-sided paired Mann-Whitney test with continuity correction, alternative=greater, Bonferroni correction with k=4, adjusted p=5.73·10-5, V=1515, 95%CI=[0.5;Inf], estimate=0.5) (C); and DeepSeek-R1 vs DeepSeek-V3 (one-sided paired Mann-Whitney test with continuity correction, alternative=greater, Bonferroni correction with k=4, adjusted p=1, V=307, 95%CI=[-0.25;Inf], estimate=1.97·10-5) (D). (E) Violin plots comparing the Likert scores of GPT-4o, DeepSeek-R1, DeepSeek-V3 and Gem2FTE to those of GPT-4, GPT-3.5 and Google in our previous study (n.s.: not significant; ***: p<0.001; significance levels visualizing the results of statistical tests performed in (A-D)). Explorative comparison of the n=110 cases analyzed by all 7 models to the n=15 newly added cases shows that the performance scores align well (one-sided unpaired Mann-Whitney test, alternative=greater; GPT-4o: pGPT-4o=0.5441, W=813.5, 95%CI=[-1.84·10-5;Inf], estimate=-4.99·10-5; DeepSeek-R1: pDeepSeek-R1=0.7710, W=740, 95%CI=[3.75·10-5;Inf], estimate=-2.16·10-5; DeepSeek-V3: pDeepSeek-V3=0.6678, W=775.5, 95%CI=[-7.45·10-5;Inf], estimate=5.91·10-5; Gem2FTE: pGem2FTE=0.9899, W=540, 95%CI=[-0.5;Inf], estimate=-3.51·10-5). (F) Cumulative frequency of the Likert scores for GPT-4o, DeepSeek-R1, DeepSeek-V3, Gem2FTE and GPT-4.

ACCEPTED ARTICLE PREVIEW

Figure 2: Model performance for treatment recommendation tasks. (A-C) Bubble plots showing the results of the 125 pairwise comparisons on a five point Likert Scale for GPT-4o vs DeepSeek-R1 (one-sided paired Mann-Whitney test with continuity correction, alternative=greater, Bonferroni correction with k=3, adjusted p=0.1522, V=771.5, 95% confidence intervals 95%CI=[-6.88·10-5;Inf], estimate=0.25) (A); GPT-4o vs Gemini-2.0 Flash Thinking Exp (Gem2FTE) (one-sided paired Mann-Whitney test with continuity correction, alternative=greater, Bonferroni correction with k=3, adjusted p=0.0016, V=1154, 95%CI=[0.2501;Inf], estimate=0.5) (B); DeepSeek-R1 vs Gem2FTE (one-sided paired Mann-Whitney test with continuity correction, alternative=greater, Bonferroni correction with k=3, adjusted p=0.0235, V=1124, 95%CI=[4.21·10-6;Inf], estimate=0.5) (C). (D) Violin plots comparing the Likert scores scoring of GPT-4o, DeepSeek-R1 and Gem2FTE to GPT-4 and GPT-3.5 (n.s.: not significant; *: p<0.05; significance levels visualizing the results of statistical tests performed in (A-C). Explorative comparison of the n=110 cases analyzed by all 7 models to the n=15 newly added cases shows that the performance scores align well (one-sided unpaired Mann-Whitney test, alternative=greater; GPT-4o: pGPT-4o=0.1460, W=955, 95%CI=[-5.38·10-5;Inf], estimate=3.16·10-5; DeepSeek-R1: pDeepSeek-R1=0.5256, W=817.5, 95%CI=[-1.46·10-5;Inf], estimate=-1.73·10-5; Gem2FTE: pGem2FTE=0.4591, W=838.5, 95%CI=[-9.54·10-6;Inf], estimate=-6.10·10-5). (E) Cumulative frequency of Likert scores for GPT-4o, DeepSeek-R1, Gem2FTE and GPT-4.

References

1. Quer G, Topol EJ. The potential for large language models to transform cardiovascular medicine. The Lancet Digital Health. 2024 Oct 1;6(10):e767–71.

2. Bellini V, Bignami EG. Generative Pre-trained Transformer 4 (GPT-4) in clinical settings. The Lancet Digital Health. 2025 Jan 1;7(1):e6–7.

3. A B, R K, K Q, S J, G W, H P, et al. Large Language Models for More Efficient Reporting of Hospital Quality Measures. NEJM AI [Internet]. 2024 Oct 24 [cited 2025 Feb 20];1(11). Available from: https://pubmed.ncbi.nlm.nih.gov/39703686/

4. McCoy TH, Perlis RH. Applying Large Language Models to Stratify Suicide Risk Using Narrative Clinical Notes. Journal of Mood & Anxiety Disorders. 2025 Jan 31;100109.

5. Ahsan H, McInerney DJ, Kim J, Potter C, Young G, Amir S, et al. Retrieving Evidence from EHRs with LLMs: Possibilities and Challenges. Proc Mach Learn Res. 2024 Jun;248:489–505.

6. Hond A de, Leeuwenberg T, Bartels R, Buchem M van, Kant I, Moons KG, et al. From text to treatment: the crucial role of validation for generative large language models in health care. The Lancet Digital Health. 2024 Jul 1;6(7):e441–3.

7. Ong JCL, Chang SYH, William W, Butte AJ, Shah NH, Chew LST, et al. Medical Ethics of Large Language Models in Medicine. NEJM AI. 2024 Jun

ACCEPTED ARTICLE PREVIEW

27;1(7):AIra2400038.

8. Alber DA, Yang Z, Alyakin A, Yang E, Rai S, Valliani AA, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med. 2025 Jan 8;1–9.

9. Beutel G, Geerits E, Kielstein JT. Artificial hallucination: GPT on LSD? Critical Care. 2023 Apr 18;27(1):148.

10. Kim M, Kim Y, Kang HJ, Seo H, Choi H, Han J, et al. Fine-Tuning LLMs with Medical Data: Can Safety Be Ensured? NEJM AI. 2025 Jan;2(1):AIcs2400390.

11. Blumenthal D, Goldberg C. Managing Patient Use of Generative Health AI. NEJM AI. 2025 Jan;2(1):AIpc2400927.

12. Hou G, Lian Q. Benchmarking of Commercial Large Language Models: ChatGPT, Mistral, and Llama [Internet]. Research Square; 2024 [cited 2025 Feb 17]. Available from: https://www.researchsquare.com/article/rs-4376810/v1

13. DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, Zhang R, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [Internet]. arXiv; 2025 [cited 2025 Feb 17]. Available from: http://arxiv.org/abs/2501.12948

14. Gibney E. Scientists flock to DeepSeek: how they're using the blockbuster AI model. Nature [Internet]. 2025 Jan 29 [cited 2025 Feb 19]; Available from: https://www.nature.com/articles/d41586-025-00275-0

15. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nature Communications. 2024;15(1):2050.

16. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024 Sep;30(9):2613–22.

17. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. Applied Sciences. 2021 Jan;11(14):6421.

18. Hooftman J, Dijkstra AC, Suurmeijer I, Bij A van der, Paap E, Zwaan L. Common contributing factors of diagnostic error: A retrospective analysis of 109 serious adverse event reports from Dutch hospitals. BMJ Qual Saf. 2024 Oct 1;33(10):642–51.

19. Jackson R, Feder G. Guidelines for clinical guidelines. BMJ. 1998 Aug 15;317(7156):427–8.

20. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus [Internet]. 2023 Feb [cited 2023 Apr 16];15(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9939079/

ACCELERATED ARTICLE PREVIEW

Methods

Selection and processing of patient case reports as well as standardized prompting were performed as in our previous study by Sandmann et al[15]. In summary, 1,020 manually written cases from German patient casebooks by Thieme and Elsevier were identified. Five clinical specialties were considered: gynecology, internal medicine, neurology, pediatrics and surgery. Cases were categorized by disease frequency. We defined a disease as 'frequent' if its incidence/year was higher than 1:1,000, 'less frequent' if the incidence/year was higher than 1:10,000 and 'rare' if the incidence/year was lower than 1:10,000. Subsequently, cases were filtered, excluding those that require image data or laboratory values for decision making. Aiming at balanced groups of disease frequency as well as clinical specialty, 110 cases were selected (see Table S7). To generate patient queries, case reports were translated to English using the tool DeepL.com. Subsequently, translations were reviewed to correct for linguistic accuracy and quality if necessary. Case reports were changed to first person perspective and layman's English. The large language models were queried "What are the most likely diagnoses? Name up to five.", for diagnosis, and "My doctor has diagnosed me with <diagnosis>. What are the most appropriate therapies in my case? Name up to five." for treatment.

The cases used in this study were sourced from curated medical textbooks rather than from real-world clinical records or unstructured notes. The aim was to simulate initial patient encounters—such as those in outpatient clinics or emergency departments—where clinicians typically collect only essential information through a limited set of targeted questions. As a result, these vignettes may sometimes omit relevant details or include extraneous information, thereby offering an approximate reflection of the models' potential performance in real-world clinical settings.

In our earlier study[15], we performed a systematic evaluation of GPT-3.5, GPT-4 and Google search, considering the tools' overall performance as well as the impact of disease frequency on the results. Aiming at a total power of 0.90 (12 tests for diagnosis, 7 for treatment, Bonferroni correction[21]), we previously calculated n=110 cases as sufficient. Analysis results from this previous study revealed that disease frequency only had a minor impact on making the correct treatment decisions. Furthermore, while disease frequency had a clear influence on diagnosis, the tools' performance even for rare diseases was better than initially assumed (Table S8). Taking into account the current rapid development in the field, we expect the differences to decrease even further. Against this background, our current study focuses on testing for significant differences in 1) GPT-4o vs DeepSeek-R1, 2) GPT-4o vs Gem2FTE, and 3) DeepSeek-R1 vs Gem2FTE for the two tasks of diagnosis and treatment recommendations. To elaborate on the added value of reasoning models, we also compare 4) DeepSeek-R1 vs DeepSeek-V3 at the diagnostic task. One-sided paired Mann-Whitney test was applied in all cases, comparing scoring on a 5 point Likert scale (Table S9). Bonferroni correction was used to adjust for multiple testing[21].

To estimate the power in relation to sample size, we made the following assumptions: 1) The performance of GPT-4o, estimated to have 1.8 trillion parameters, is better compared to DeepSeek-R1, having 671 billion parameters. 2) The performance of

ACCELERATED ARTICLE PREVIEW

GPT-4o and DeepSeek-R1 is better compared to Gem2FTE. The exact parameter size of Gem2FTE is not reported but estimated to be less than 671 billion parameters based on the fact that the earlier version Gemini 1.5 Flash had 8B parameters. 3) The performance of DeepSeek-R1 is better compared to DeepSeek-V3.

In our earlier study[15], we observed probabilities for Likert scores for GPT-3.5 and GPT-4 summed up in Table S10. Based on these findings, we adapted the performance estimates for the successor model GPT-4o in relation to DeepSeek-R1, DeepSeek-V3 and Gem2FTE. The following probabilities for Likert scores 1 / 2 / 3 / 4 / 5 were sampled: GPT-4o: 0.00 / 0.00 / 0.00 / 0.30 / 0.70; DeepSeek-R1: 0.00 / 0.00 / 0.10 / 0.30 / 0.60; DeepSeek-V3: 0.00 / 0.05 / 0.25 / 0.20 / 0.50; Gem2FTE: 0.01 / 0.14 / 0.30 / 0.15 / 0.40. Power calculation, investigating possible sample sizes between 75 and 145, showed power=0.89 for n=125 cases when adjusting for 4 tests (diagnosis), and power=0.91 when adjusting for 3 tests (treatment; Fig. S7, Supplementary Data 4).

Selecting the same 110 cases as before, direct comparability of the new LLMs' results to the old approaches is granted. By selecting all case reports from non-English sources with non-open access, we aimed at reducing the risk of training bias. To meet the required sample size, 15 new cases were added, following the same selection approach. Explorative analysis was performed to investigate whether results for these new cases align with the old ones. Furthermore, the influence of disease frequency and clinical specialty on the models' performance was analyzed exploratively. All queries were entered manually without using API calls within the vendor-provided user platforms and executed between Jan 27th and Feb 5th in 2025. Additional technical details are provided in Table S11.

A five point Likert Scale (Table S9) was used for assessing both diagnosis and treatment tasks. Two physicians independently assessed five random cases, conducting a comprehensive literature review using UpToDate® and PubMed, and reaching a consensus on the final Likert scores. Interrater reliability was determined using weighted Cohen's kappa (R package DescTools[22], function 'CohenKappa', weights 'Equal-Spacing'). Given the high interrater reliability (κ=0.76, 95% confidence interval CI=[0.55;0.96]), consistent with findings from our prior study (κ ranging between 0.53 and 0.84), the first physician subsequently continued to perform detailed reviews with extensive literature analysis for the remaining cases, while the second physician independently verified all ratings. All statistical analyses were conducted using R 4.4.2[23]. Applying 1-sided paired Mann-Whitney tests[24], we tested for significant differences in the overall performance of the approaches (alpha=0.05; Bonferroni correction with k=4 for diagnosis and k=3 for treatment). One-sided unpaired Mann-Whitney test was used for the explorative analysis of old vs new clinical cases.

Data availability

All data including patient cases (clinical cases) and ratings are provided in Supplementary Data S1. Descriptions on further supplementary tables are provided in the file SupplementaryInformation.pdf.

ACCEPTED ARTICLE PREVIEW

Code availability

All code to reproduce data analyses in the main manuscript is provided in Supplementary Data S2, Supplementary Data S3 and Supplementary Data S4. Code to reproduce main and supplementary analyses is provided in the GitHub repository https://github.com/sandmanns/llm_evaluation.

Methods-only references

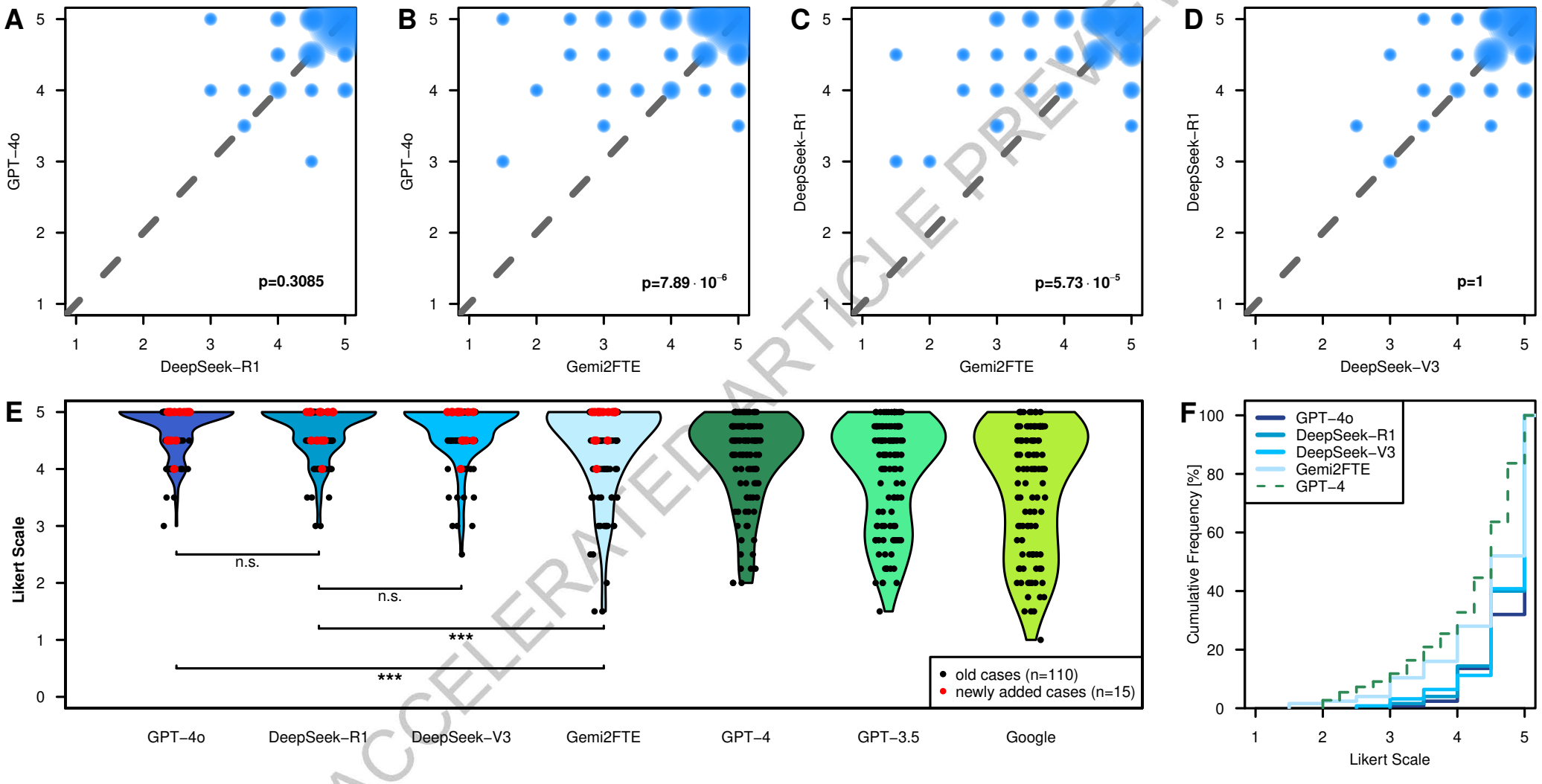
21. Bonferroni, C. E. Teoria Statistica Delle Classi e Calcolo Delle Probabilita. Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commericiali di Firenze 1936, 8, 3–62.

22. Signorell, A. DescTools: Tools for Descriptive Statistics. R package version 0.99.60. 2025. https://doi.org/10.32614/CRAN.package.DescTools.

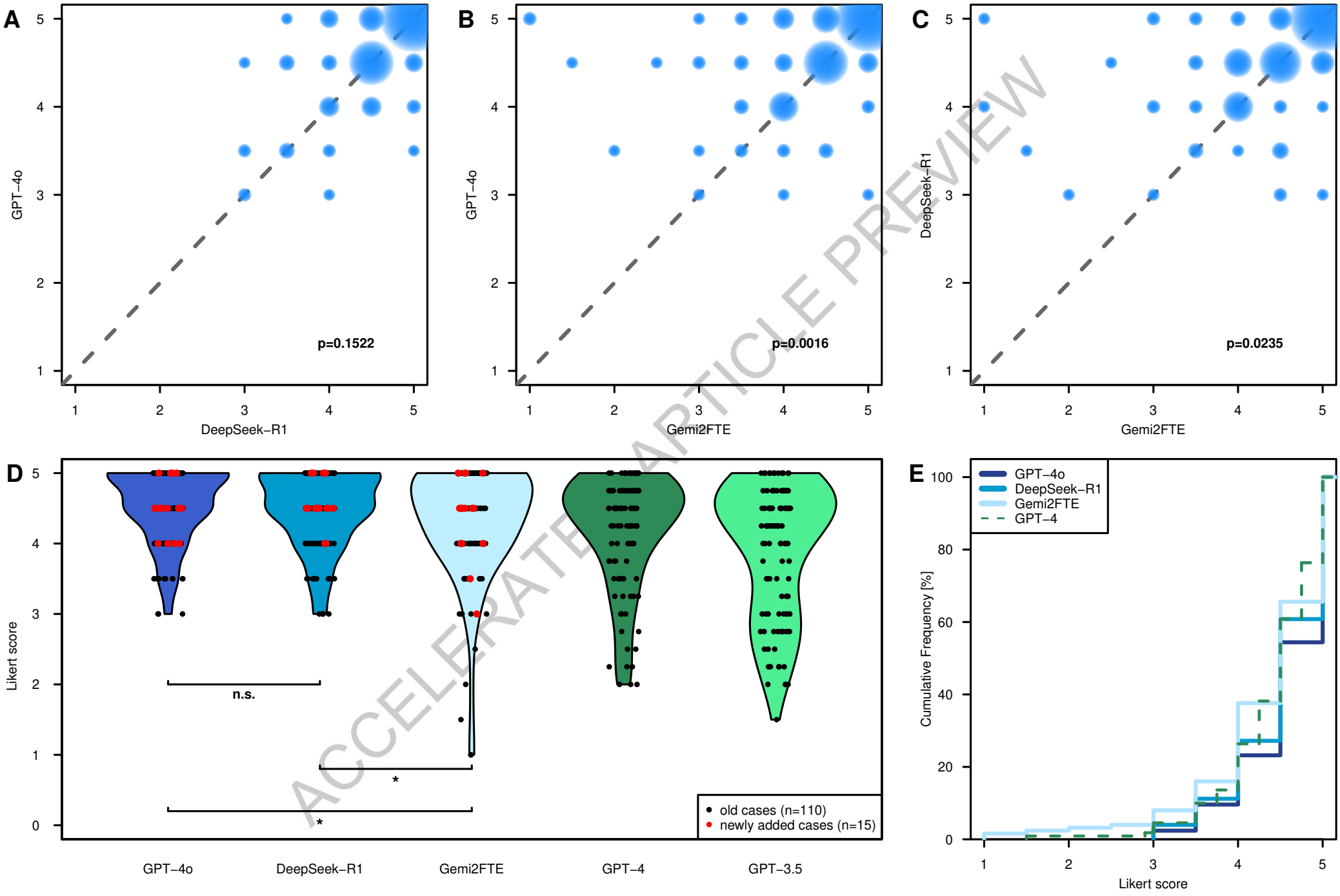
23. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/ (accessed 2025-04-01).

24. Mann, H. B.; Whitney, D. R. On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. The Annals of Mathematical Statistics 1947, 18 (1), 50–60. https://doi.org/10.1214/aoms/1177730491.

ACCELERATED ARTICLE PREVIEW

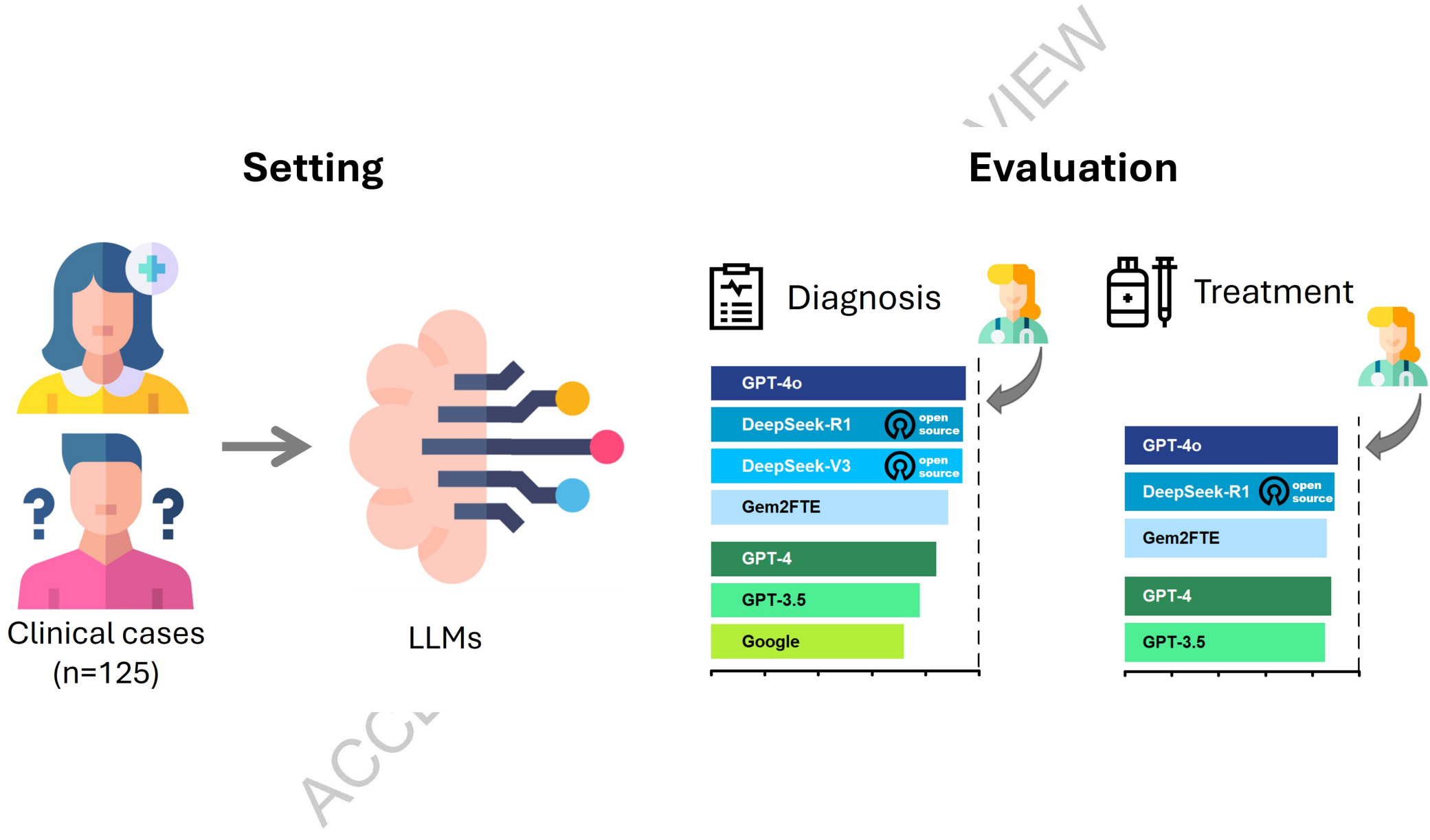


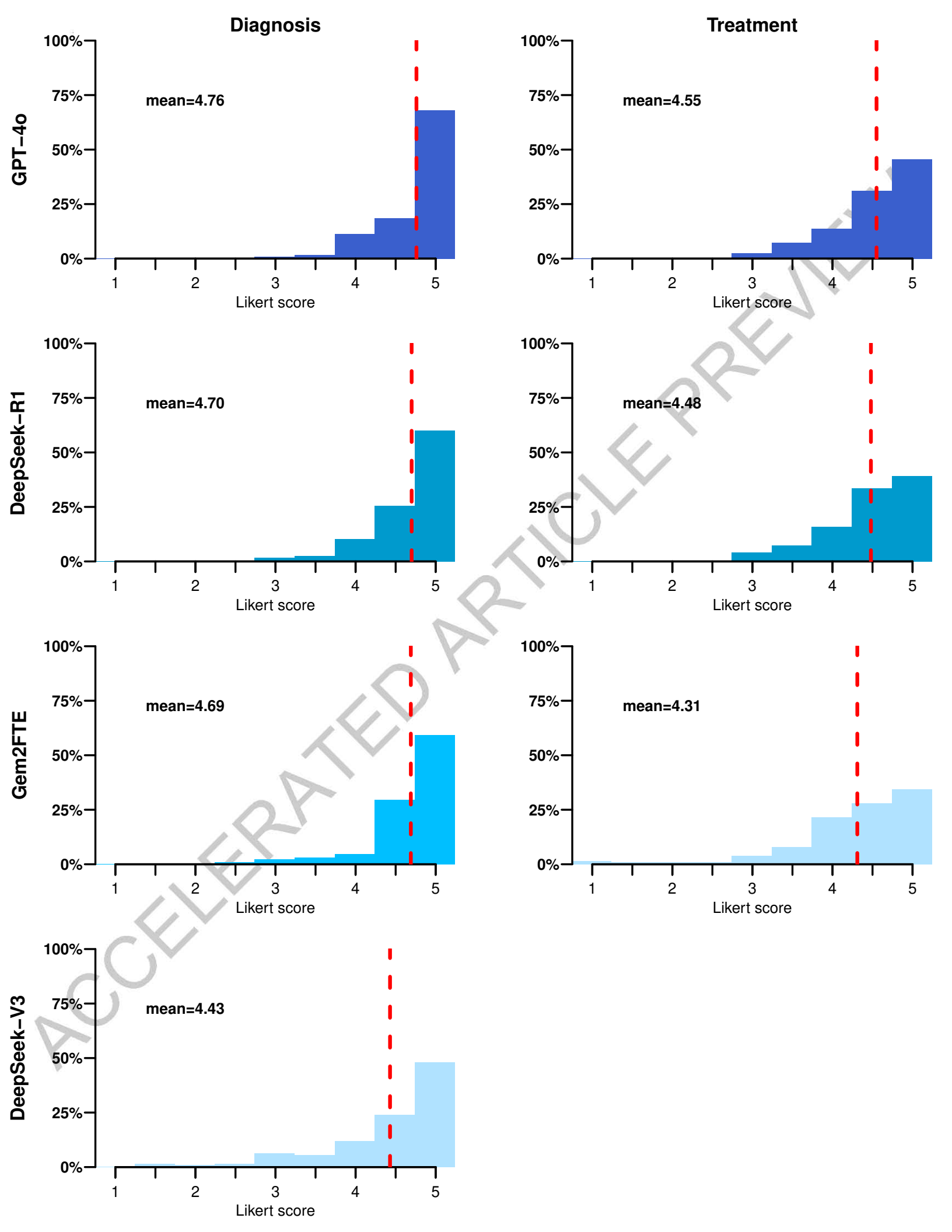
ACCELERATED ARTICLE PREVIEW



ACCELERATED ARTICLE PREVIEW

ACCEPTED MANUSCRIPT





ACCELERATED ARTICLE PREVIEW

nature portfolio

Corresponding author(s): Roland Eils

Last updated by author(s): Apr 18, 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	ChatGPT4o by OpenAI. Gemini-2.0 Flash Thinking Exp by Google. R1 and V3 by DeepSeek.
Data analysis	All analyses were conducted using R 4.4.2, extended by R packages openxlsx 4.2.8, vioplot 0.5.1 and DescTools 0.99.59. Custom code for data analysis is available at https://github.com/sandmanns/llm_evaluation.

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

All data including patient cases (clinical cases) and ratings are provided in Supplementary Data S1. Descriptions on further supplementary tables are provided in the file Supplementary Information.

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	Not applicable because no research involving human participants, their data, or biological material was conducted.
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable.
Population characteristics	Not applicable.
Recruitment	Not applicable
Ethics oversight	Not applicable.

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	n=125 free-text selections from patient cases originating from case books. Detailed sample size estimation is described in the Online Methods section.
Data exclusions	Patient cases were selected randomly from clinical casebooks. To fulfill a balanced set of medical domains and disease frequencies and to fulfill the sample size estimation, a subset of 125 cases from 1020 cases were sampled. The exact sampling strategy is described in the Online Methods section.
Replication	Text output of LLMs can vary though given the same input prompts. However, all outputs are documented in Supplementary Data S1 and all following analyses are deterministic and published as open source. All analyses have been repeated three times and produced the same results.
Randomization	Not applicable because no patient randomization was performed (only randomization of patient texts).
Blinding	Not applicable here because no patient or subject blinding was necessary.

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.*