NEJM AI

NEJM AI 2024; 1 (2)
DOI: 10.1056/AIdbp2300092

DATASETS, BENCHMARKS, AND PROTOCOLS

Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in Nephrology

Sean Wu,[1] Michael Koo,[1] Lesley Blum,[2] Andy Black,[2] Liyo Kao,[2] Zhe Fei, Ph.D.,[3] Fabien Scalzo, Ph.D.,[1] and Ira Kurtz, M.D.[2,4]

Received: August 20, 2023; Revised: October 24, 2023; Accepted: November 14, 2023; Published: January 17, 2024

Abstract

BACKGROUND In recent years, significant breakthroughs have been made in the field of natural language processing, particularly with the development of large language models (LLMs). LLMs have demonstrated remarkable capabilities on benchmarks related to general medical question answering, but there are fewer data about their performance in subspecialty fields and fewer studies still comparing the many available LLMs. These models have the potential to be used as a part of adaptive physician training, medical copilot applications, and digital patient interaction scenarios. The ability of LLMs to participate in medical training and patient care depends in part on their mastery of the knowledge content of specific medical fields.

METHODS This study investigated the medical knowledge capability of multiple LLMs in the context of their internal medicine subspecialty multiple-choice test-taking ability. We compared the performance of several open-source LLMs (Llama2-70B, Koala 7B, Falcon 7B, Stable-Vicuna 13B, and Orca-Mini 13B) with the proprietary models GPT-4 and Claude 2 on multiple-choice questions in the field of nephrology. Nephrology was chosen as an example of a conceptually complex subspecialty field in internal medicine. This study was conducted to evaluate the ability of LLMs to provide correct answers to Nephrology Self-Assessment Program (nephSAP) multiple-choice questions. These questions administered by the American Society of Nephrology help clinicians assess their knowledge in various topics in nephrology.

RESULTS The overall success of open-source LLMs in answering the 858 nephSAP multiple-choice questions correctly was 17.1 to 30.6%. In contrast, Claude 2 answered 54.4% of the questions correctly, whereas GPT-4 achieved a score of 73.3%. A dataset containing questions and ground truth labels used to assess the LLMs has been made available.

The author affiliations are listed at the end of the article.

Dr. Kurtz can be contacted at ikurtz@mednet.ucla.edu or at the Division of Nephrology, David Geffen School of Medicine, University of California, Los Angeles, Room 7-155 Factor Building, 700 Tiverton Ave., Los Angeles, CA 90095.

CONCLUSIONS We show that the current widely used open-source LLMs have poor zero-shot reasoning ability in nephrology compared with GPT-4 and Claude 2, illustrating knowledge gaps across LLMs relevant to future subspecialty medical training and patient care. (Funded by the Factor Family Foundation and others.)

Introduction

Large language models (LLMs) are artificial intelligence (AI) systems that understand and generate human-like natural language responses to text prompts.[1] Many studies have assessed the capabilities of LLMs in knowledge-based fields, such as medicine, on the basis of their multiple-choice test-taking ability.[2] In 2023, the release of GPT-4 by OpenAI gained much attention for its impressive test-taking capabilities.[2,3] Other proprietary models include Claude 2[4] from Anthropic, released in June 2023, which has also received much attention. Recently, several open-source LLMs, including Llama2 Koala, Falcon, Orca-Mini, and Stable Vicuna, have been reported to be successful in various domains.[5-9]

In the present study, we analyzed the ability of several open-source LLMs to answer nephrology multiple-choice test questions successfully in comparison with GPT-4 and Claude 2. We further assessed the capabilities of these models by determining their correct answer percentage for each of the various medical topics in nephrology. We further evaluated the open-source models by comparing the justifications they gave for the answers they considered correct using the proximity of word embeddings in the vector space approach.[10] To allow other researchers to build off of this collection of benchmark results, we also release the code and dataset used in this study.

Methods

LLMS

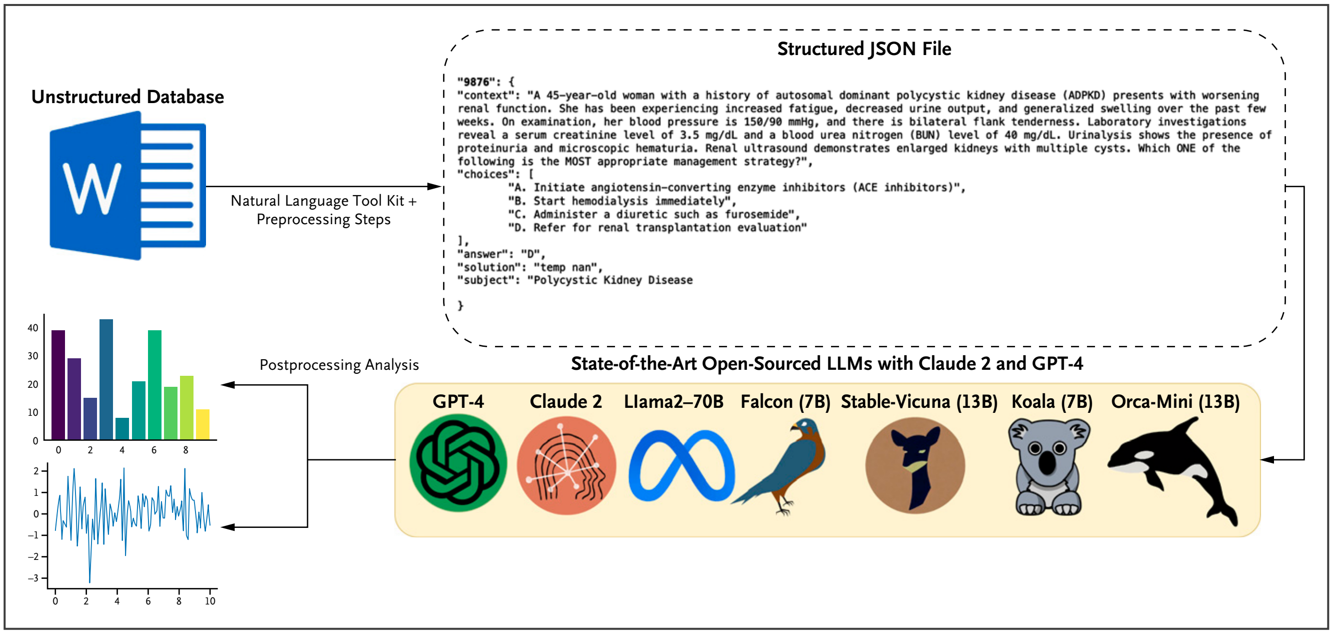
In this study, we evaluated the ability of several widely used open-source LLMs, including Llama2 (July 2023),[9] Koala (April 2023),[7] Falcon (June 2023),[11] Orca-Mini (June 2023),[5] and Stable Vicuna (April 2023),[8] in addition to GPT-4 and Claude 2 to correctly answer multiple-choice questions in the field of nephrology. For all automated models (Llama2, Koala, Vicuna, Falcon, and Orca), we used the Instruction-Following prompting strategy,[12] where we concatenated "Context," "Question," and "Choices (Pick One)" for each forward pass of the model to provide further clarification because of the large input token sizes. If a definitive answer was not provided (which occurred on average 5.7% of the time), we considered the choice incorrect. The latter occurred only in open-source LLMs. For this study, we utilized a combination of Google Colab's cloud computing engine and a local NVIDIA graphics processing unit (GPU) to run the open-source LLMs. We utilized the HuggingFace library to load the quantized models with specific text generation parameters. More details of the experimental setup can be found in the Supplementary Appendix.

DATASET

The dataset used in our experiments comprised 858 Nephrology Self-Assessment Program (nephSAP) multiple-choice questions and correct answers from January 2016 to April 2023. These questions test the therapeutic and diagnostic knowledge of clinicians in the subspecialty field of nephrology. The format of the questions included a clinical scenario followed by a prompt to select the one correct answer from among the possible answer choices. Twelve patient scenario questions that also included complex tables were omitted because of the difficulty encountered by the LLMs in interpreting patient results depicted in table format. The dataset and ground truth labels have been made available through HuggingFace (https://huggingface.co/datasets/SeanWu25/NEJM-AI_Benchmarking_Medical_Language_Models). To assemble the dataset, we extracted a structured JSON file from an unstructured text file that contained raw questions and answers using an automated parsing script to extract the "Questions" and "Answers." Specifically, we utilized the natural language processing tool kit (NLTK)[13,14] to tokenize the text and accurately generate the JSON file (Fig. 1). Finally, we parsed through a separate CSV file that contained ground truth answers (provided by the American Society of Nephrology) to incorporate the correct answers acquired from the test bank. As a result, each example in our structured JSON file included the question identification, context, prompt, multiple-choice choices, correct answer, and the specific subject area within nephrology to which the question pertained.

MODEL METRICS AND EXPERIMENTAL EVALUATION

We developed a script to parse the output on the basis of the correct input answers (available at GitHub: https://github.com/SeanWu25/Benchmarking-LLMs-Nephrology) as the



```
"9876": {
"context": "A 45-year-old woman with a history of autosomal dominant polycystic kidney disease (ADPKD) presents with worsening
renal function. She has been experiencing increased fatigue, decreased urine output, and generalized swelling over the past few
weeks. On examination, her blood pressure is 150/90 mmHg, and there is bilateral flank tenderness. Laboratory investigations
reveal a serum creatinine level of 3.5 mg/dL and a blood urea nitrogen (BUN) level of 40 mg/dL. Urinalysis shows the presence of
proteinuria and microscopic hematuria. Renal ultrasound demonstrates enlarged kidneys with multiple cysts. Which ONE of the
following is the MOST appropriate management strategy?",
"choices": [
        "A. Initiate angiotensin-converting enzyme inhibitors (ACE inhibitors)",
        "B. Start hemodialysis immediately",
        "C. Administer a diuretic such as furosemide",
        "D. Refer for renal transplantation evaluation"
],
"answer": "D",
"solution": "temp nan",
"subject": "Polycystic Kidney Disease

}
```

Figure 1. Depiction of the Workflow Encompassing Data Acquisition, Preprocessing, and the Utilization of Large Language Models (LLMs) for Comparison Purposes.

The depicted structured JSON file was LLM generated as an example and not part of the Nephrology Self-Assessment Program question set.

large number of multiple-choice questions made manual review and comparison impractical. The primary challenge was the variability in the outputs generated by different language models. We designed a script to recognize these patterns as regular expressions and extracted matching correct answers. Regular expressions were used to define specific formatting criteria and extract relevant portions of the outputs. This allowed us to identify responses that aligned with the correct answers, even with slight variations. By validating the extracted answers, we filtered out discrepancies and errors, thereby enhancing the overall accuracy of the evaluations. The automated comparison checker streamlined the evaluation process and eliminated the time-consuming manual reviews. For each model, we evaluated the percentage of questions that were correctly answered. In addition, for the open-source LLMs, we conducted an analysis of the quality and semantic meaning of responses. We utilized the BiLingual Evaluation Understudy[15] (BLEU) metric from the NLTK.[13] BLEU is commonly used in machine translation problems to determine the language translation quality of an natural language processing (NLP) model. The BLEU metric gives a score from zero to one, with zero indicating that the generated text is of extremely poor quality compared with the ground truth and a score of one indicating that the generated text is most similar to the ground truth. We also computed the word error rate (WER)[16] and the cosine similarity metric,[10] which outputs a score between zero and one, representing the similarity in strings of text on the basis of word embeddings (LLM vs. ground truth explanation). Although the BLEU, WER, and cosine similarity scores provided a general indication of text similarity, we also included the number of questions for which each LLM achieved a score greater than or equal to 0.5 for both the BLEU and cosine scores. We reported the overall score plus or minus the standard deviation. A BLEU greater than or equal to 0.5, a high-quality translation,[17] and a WER of 5 to 10% are considered good quality.[18] Group comparisons were done using chi-square testing (with a P value of <0.05 considered significant).

Results

LLM TEST-TAKING ABILITY

Table 1 shows the test-taking ability of the different LLMs using our dataset (https://huggingface.co/datasets/SeanWu25/NEJM-AI_Benchmarking_Medical_Language_Models). Among the open-source models, Llama2 achieved the highest score of 30.6%, Vicuna had a score of 25.5%,

Table 1. Comparison of the Overall Correct Responses among the Large Language Models.*

LLM	Total Questions	Number Correct	Percentage Correct	CI
GPT-4	858	629	73.3	70.3–76.3
Claude 2	858	467	54.4	51.1–57.7
Vicuna	858	219	25.5	22.6–28.4
Orca	858	147	17.1	14.6–19.6
Falcon	858	155	18.1	15.5–20.7
Koala	858	204	23.8	21.0–26.6
Llama	858	263	30.6	27.6–33.8

* CI denotes 95% confidence interval; and LLM, large language model.

Koala had a score of 23.8%, Falcon had a score of 18.1%, and Orca-Mini had a score of 17.1%. Considering the number of questions and the choices per question (which varied), we calculated that a score of 23.8% would have been expected by random guessing. Among the open-source LLMs, only Llama2 achieved a score slightly above this level. In contrast, GPT-4 was significantly better with a score of 73.3% (P<0.01 vs. Claude 2 and the open-source LLMs), whereas Claude 2, although more successful than the open-sourced LLMs, performed more poorly than GPT-4 with a score of 54.4% (P<0.001). (The GPT-4 training cutoff was September 2021. To assess for possible leakage of nephSAP data into GPT-4 pretraining, we performed a subanalysis of questions correctly answered before and after this date, which showed comparable overall performance [72.5 vs. 74.4%, respectively].) The passing grade on the nephSAP questions for human test takers was 75%. In addition to assessing the overall score on all 858 questions, we further broke down the LLM answer choices on the basis of individual nephSAP question topics within nephrology and scored each of the topics separately for each LLM. The results are shown in Figure 2. Open-source LLMs performed very poorly in all nephrology topics. In general, Claude 2 attained a higher score in all areas of nephrology while achieving a passing grade in one of the topics. GPT-4 performed exceptionally well and achieved human-like performance for the majority of topics.

LIMITATIONS OF OPEN-SOURCE LLMS

The majority of the open-source LLMs achieved an overall score that did not differ from what would be expected if the questions were answered randomly. To further assess the poorly scoring open-source models, we further evaluated the quality of their answer explanations. As depicted in Figure 3, all of the open-source LLMs achieved WER scores from 0 to 22%. However, it is evident that all the models exhibited suboptimal performance on the BLEU and cosine similarity score metrics. Specifically, Orca, Koala, Falcon, Vicuna, and Llama2 all scored below approximately 0.1 on the BLEU score (with Orca scoring 0.07±0.01, Koala scoring 0.05±0.01, Falcon scoring 1.90±0.24, Vicuna scoring 0.10±0.01, and Llama2 scoring 0.02±0.002). In the cosine similarity score, all models exhibited suboptimal scores. Orca demonstrated a cosine similarity score of 0.36±0.02, Koala had a score of 0.31±0.02, Falcon had a score of 0.30±0.02, Vicuna had a score of 0.35±0.02, and Llama2 scored 0.30±0.01. Overall, the results further demonstrate the poor test-taking ability of these open-source LLMs on nephSAP questions.

Discussion

In this study, we benchmarked the performance across both proprietary and open-source LLMs in answering our dataset of multiple-choice nephSAP nephrology test questions (https://huggingface.co/datasets/SeanWu25/NEJM-AI_Benchmarking_Medical_Language_Models). We evaluated models including Llama2, Koala, Orca-Mini, Falcon, and Stable-Vicuna compared with GPT-4 and Claude 2. GPT-4 performed best, with 73.3% of the questions answered correctly. In addition, GPT-4 scored better in all nephrology topics assessed individually. Claude 2 achieved the second-best results with an overall score of 54.4%. In comparison with GPT-4 and Claude 2, the open-source models performed poorly in terms of total correct answers and the quality of their explanations.

There are several potential reasons for this finding. The number of parameters that these models were trained on may have played a role.[19] Furthermore, the LLMs differ in what data they were trained on, and both GPT-4 and Claude 2 were trained not only on publicly available data but also on third-party data.[3] The open-source LLMs were trained on publicly available data, such as ShareGPT, WebGPT, Reddit, PubMed, StackExchange, and GitHub.

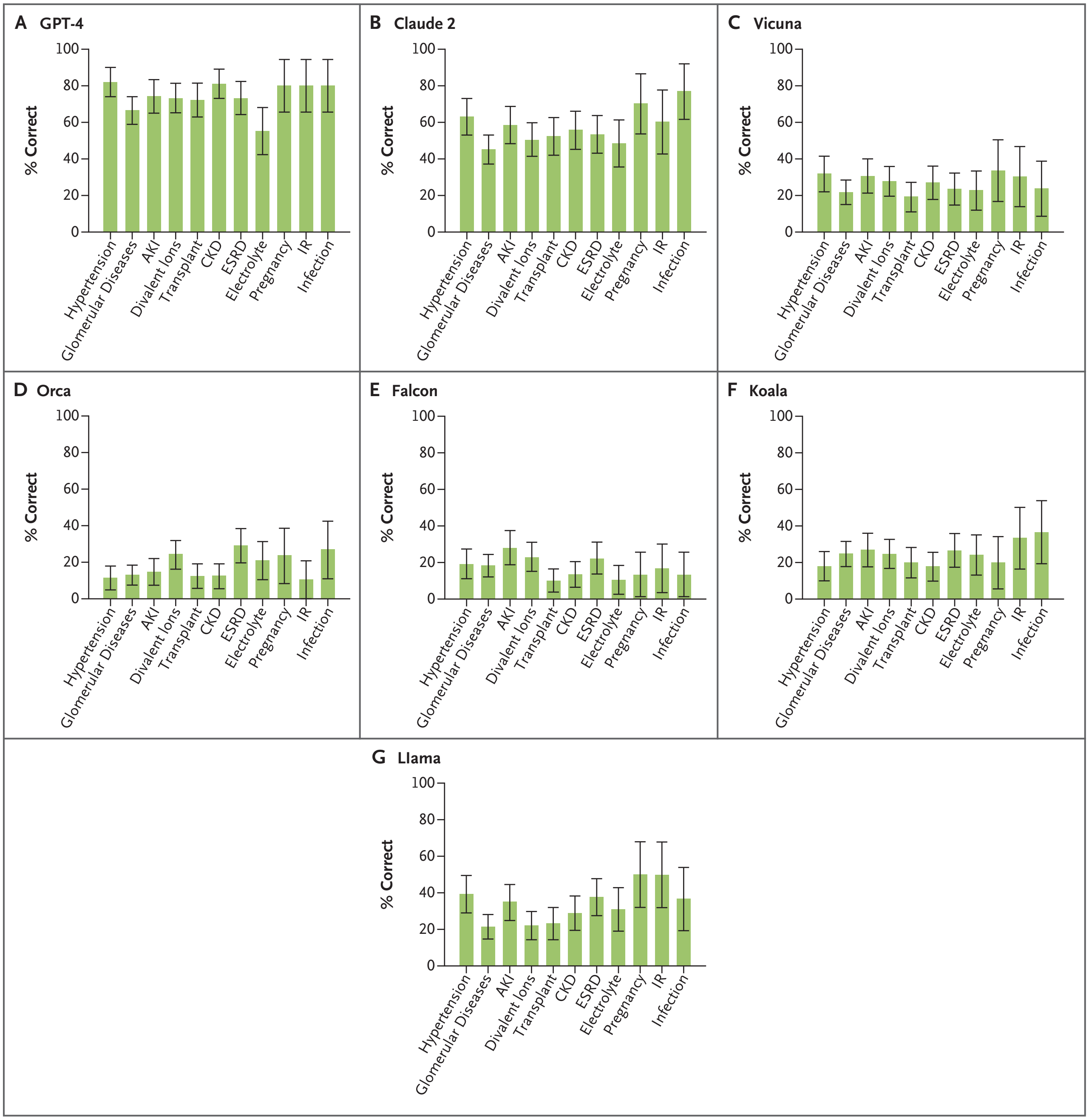


Figure 2. The Percentage of Correct Answers with 95% Confidence Intervals Is Shown for Each of the Large Language Models among the Various Nephrology Self-Assessment Program Nephrology Topics.

GPT-4 (Panel A). Claude 2 (Panel B). Vicuna (Panel C). Orca (Panel D). Falcon (Panel E). Koala (Panel F). Llama (Panel G). AKI denotes acute kidney injury; CKD, chronic kidney disease; ESRD, end-stage renal disease; and IR, interventional radiology.

High-quality data for training LLMs in the medical field often reside in nonpublic materials that have been curated and peer reviewed, such as textbooks, published articles, and curated datasets. Without negating the importance of the computational power of specific LLMs, the ability to access medical training data material that is currently not in the public domain will likely remain a key factor that determines whether performance of specific LLMs will improve in the future. Finally, differences in data leakage could have contributed to the differences among LLMs.[20,21]

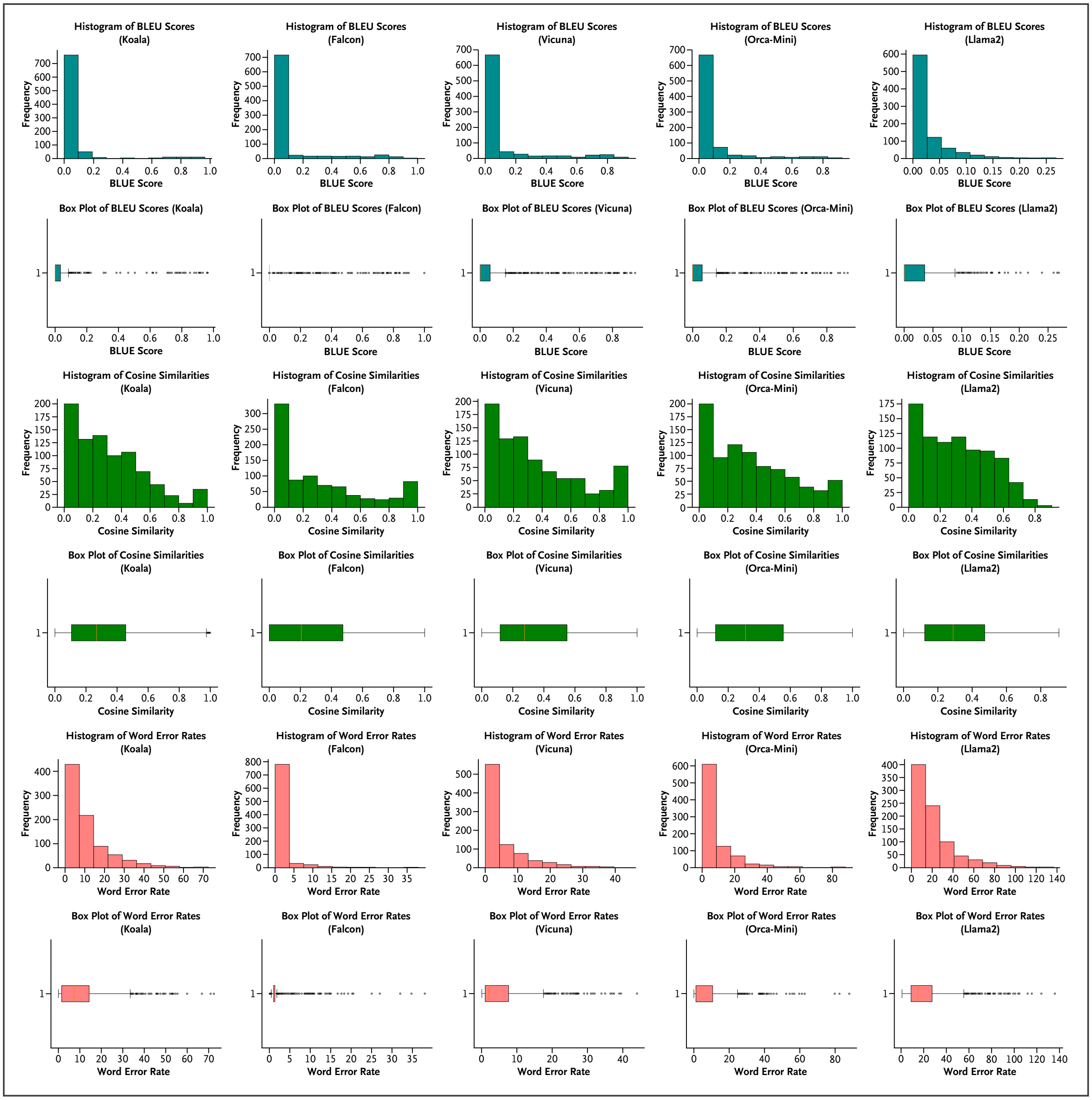


Figure 3. Visualization of Overall Word Error Rates, BiLingual Evaluation Understudy (BLEU), and Cosine Similarity Scores.

At a high level, word error rate is the percentage of incorrect words generated, the BLEU score is the similarity in n-gram word sequences, and the cosine similarity score is the cosine distance between two vectors in the embedding space. Notice the underperformance of all open-sourced models in BLEU and cosine similarity scores.

Another aspect that needs to be considered in attaining better results in all models is the need for domain-specific fine-tuning.[22] Models trained without specific optimization in specialized knowledge domains may yield suboptimal results for domain-specific questions.[23] In addition, a more complex model of the world that includes cause–effect and temporal–spatial understanding is currently lacking in LLMs.[24] Accordingly, enhancing parameter optimization, utilizing diverse and representative training datasets, incorporating domain-specific fine-tuning, and improving reasoning capabilities are areas for future research that could potentially result in even better LLM performance capabilities in the medical field and in other knowledge areas.

In medicine, the amount and complexity of the information that human doctors need to master increase as they transition from medical school to internal medicine residency to, ultimately, subspecialty practice. The success of GPT-4 on nephSAP nephrology questions is striking, with an overall score of 73.3% (Table 1). Moreover, GPT-4 achieved a score greater than or equal to 72% in 9 of the 11 nephSAP nephrology question topics (Fig. 2). The passing grade for each topic for human test takers is 75%, with a total of two attempts for each question. Although we did not formally compare the capabilities of earlier generative pretrained transformer models in our study with regard to the subspecialty internal medicine, a recent study of the performance of ChatGPT and GPT-4 on the 2021 and 2022 American College of Gastroenterology self-assessment multiple-choice questions was considered suboptimal.[25] On a dermatology specialty certificate multiple-choice examination in the United Kingdom, ChatGPT and GPT-4 received scores of 63 and 90%, respectively (passing grade is 70 to 72%), demonstrating the improved ability of GPT-4.[26] Addressing frequently asked cardiology heart failure questions (not multiple choice) from Facebook groups, medical societies, and institutions, ChatGPT scored 83.2% and GPT-4 scored 100% on the ability to provide correct information.[27] Finally, a recent study using a subset of the Kidney Self-Assessment Program and nephSAP questions on glomerular disease showed that ChatGPT underperformed.[28]

In analyzing the ability of GPT-4 in each of the individual nephrology topics, the electrolyte questions received the lowest score (55.2%; P<0.01). The area of fluid and electrolytes/acid–base diagnosis is also more of a challenge for human trainees because of its conceptual–quantitative content, where reliance on memorization of facts is insufficient. GPT-4 may have performed less well on questions related to this topic because of the quality of the training material on this topic (both public and nonpublic) and because of the general limitations of LLMs in seemingly "simple" quantitative reasoning tasks.[29] It is expected that domain-specific training using curated datasets involving more complex topics in nephrology will improve LLM performance in the future.

Our benchmark nephrology results suggest that performance of available LLMs on subspecialty multiple-choice questions differs substantially, with GPT-4 demonstrating human-like test-taking ability in one of the difficult subspecialty fields of internal medicine. Large health care systems, medical schools, and their departments are likely to increasingly embrace these AI models because of the perceived or actual cost-saving opportunities and efficiencies that will be afforded. The increasing opportunities for individual personalized internal medicine subspecialty training, including AI creation of multiple-choice questions, adaptive learning that considers gaps in one's knowledge, digital doctor–patient simulated interactions, and AI physician copilots to aid in patient care, are areas where subspecialty internal medicine, including nephrology, may experience profound changes in the future.

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

Supported in part by funds from the Factor Family Foundation, the Smidt Family Foundation, the DaVita Allen Nissenson Research Fund, the Kleeman Fund, and the Ralph Block Family Foundation (all to Dr. Kurtz).

Author Affiliations

[1] Keck Data Science Institute, Natural Science Department, Pepperdine University, Malibu, CA

[2] Division of Nephrology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles

[3] Department of Statistics, University of California, Riverside, Riverside, CA

[4] Brain Research Institute, University of California, Los Angeles, Los Angeles

References

1. Liu Y, Han T, Ma S, et al. Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models. August 22, 2023 (https://arxiv.org/abs/2304.01852). Preprint.

2. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. April 12, 2023 (https://arxiv.org/abs/2303.13375). Preprint.

3. OpenAI. GPT-4 technical report. December 19, 2023 (https://doi.org/10.48550/arXiv.2303.08774). Preprint.

4. Anthropic AI. Model card and evaluations for Claude models. 2023 (https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf).

5. Mukherjee S, Mitra A, Jawahar G, Agarwal S, Palangi H, Awadallah A. Orca: progressive learning from complex explanation traces of GPT-4. June 5, 2023 (https://arxiv.org/abs/2306.02707). Preprint.

6. Taori R, Gulrajani I, Zhang T, et al. Stanford alpaca: an instruction-following LLaMa model. GitHub. 2023 (https://github.com/tatsu-lab/stanford_alpaca).

7. Geng X, Gudibande A, Liu H, et al. Koala: a dialogue model for academic research. The Berkeley Artificial Intelligence Research Blog. April 3, 2023 (https://bair.berkeley.edu/blog/2023/04/03/koala/).

8. Chiang WL, Li Z, Lin A, et al. Vicuna: an open-source chatbot impressing GPT-4 with 90% ChatGPT quality. Vicuna. March 30, 2023 (https://vicuna.lmsys.org).

9. Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. July 19, 2023 (https://arxiv.org/abs/2307.09288). Preprint.

10. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013 (https://arxiv.org/abs/1301.3781). Preprint.

11. Penedo G, Malartic Q, Hesslow D, et al. The refined web dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. 2023 (https://arxiv.org/abs/2306.01116). Preprint.

12. Kaddour J, Harris J, Mozes M, et al. Challenges and applications of large language models. July 19, 2023 (https://arxiv.org/abs/2307.10169). Preprint.

13. Loper E, Bird S. NLTK: the natural language toolkit. 2002 (https://arxiv.org/abs/0205028). Preprint.

14. Almazrouei E, Alobeidli H, Alshamsi A, et al. Falcon-40B: an open large language model with state-of-the-art performance. 2023 (https://huggingface.co/tiiuae/falcon-40b).

15. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: Isabelle P, Charniak E, Lin D, eds. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002:311-318.

16. Ali A, Renals S. Word error rate estimation for speech recognition: e-WER. In: Gurevych I, Miyao Y, eds. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 2. Melbourne, Australia: Association for Computational Linguistics, 2018:20-24. DOI: 10.18653/v1/P18-2004.

17. Google Cloud. Evaluating models, understanding the BLEU score. December 20, 2023 (https://cloud.google.com/translate/automl/docs/evaluate).

18. Microsoft. Test accuracy of a custom speech model, resolve errors and improve WER. December 21, 2023 (https://learn.microsoft.com/en-us/azure/ai-services/speech-service/how-to-custom-speech-evaluate-data?pivots=speech-studio).

19. Gholami S, Omar M. Do generative large language models need billions of parameters? 2023 (https://arxiv.org/pdf/2309.06589.pdf). Preprint.

20. Zhou K, Zhu Y, Chen Z, et al. Don't make your LLM an evaluation benchmark cheater. 2023 (https://arxiv.org/abs/2311.01964). Preprint.

21. Kim S, Yun S, Lee H, et al. Propile: probing privacy leakage in large language models. 2023 (https://arxiv.org/abs/2307.01881). Preprint.

22. Lv K, Yang Y, Liu T, et al. Full parameter fine-tuning for large language models with limited resources. 2023 (https://arxiv.org/abs/2306.09782). Preprint.

23. Zhao WX, Zhou K, Li J, et al. A survey of large language models. 2023 (https://arxiv.org/abs/2303.18223). Preprint.

24. Hobbhahn M, Lieberum T, Seiler D. Investigating causal understanding in LLMs. In: NeurIPS ML Safety Workshop. December 9, 2022 (https://neurips2022.mlsafety.org/).

25. Suchman K, Garg S, Trindade AJ. Chat generative pretrained transformer fails the multiple-choice American College of Gastroenterology self-assessment test. Am J Gastroenterol 2023;118:2280-2282. DOI: 10.14309/ajg.0000000000002320.

26. Passby L, Jenko N, Wernham A. Performance of ChatGPT on Specialty Certificate Examination in Dermatology multiple-choice questions. Clin Exp Dermatol 2023 June 2 (Epub ahead of print). DOI: 10.1093/ced/llad197.

27. King RC, Samaan JS, Yeo YH, et al. Appropriateness of ChatGPT in answering heart failure related questions. July 10, 2023 (https://www.medrxiv.org/content/10.1101/2023.07.07.23292385v1). Preprint.

28. Miao J, Thongprayoon C, Cheungpasitporn W. Assessing the accuracy of ChatGPT on core questions in glomerular disease. Kidney Int Rep 2023;8:1657-1659. DOI: 10.1016/j.ekir.2023.05.014.

29. Lewkowycz A, Andreassen A, Dohan D, et al. Solving quantitative reasoning problems with language models. 2002 (https://arxiv.org/abs/2206.14858). Preprint.