JHEP|Reports

Guideline-enhanced large language models outperform physician-test takers on EASL Campus quizzes multiple choice questions

To the Editor:

Recent advances in large language models (LLMs) have garnered widespread attention in the medical community, especially given their ability to handle multimodal data such as images or audio-recordings. Safe deployment into clinical practice remains unclear, with a recent systematic review identifying a broad range of accuracy (6.4-91.4%) when ChatGPT answered clinical questions on digestive disease-related topics.[1] This wide range of performance can be attributed to the use of baseline models without incorporating external medical knowledge from relevant medical guidelines.[2–4] A valuable resource to provide a useful benchmark for hepatology is the publicly available multiple-choice question (MCQ) database from the European Association for the Study of the Liver (EASL) Campus, designed to test knowledge of the content of current EASL guidelines. This benchmark is especially critical because, unlike other digestive disease societies, it offers a comprehensive, guideline-referenced, and publicly accessible tool that sets a new standard for evaluating LLMs' performance in the field.

We compiled a dataset of 110 MCQs from all available quizzes up to November 30th, excluding questions on basic science (n = 2) and rare liver diseases (n = 3) due to insufficient knowledge outlined in guideline documents. To establish a human performance benchmark, we used data from previous online physician test-takers. Specifically, for each question, we determined the percentage of participants who answered correctly, then averaged these percentages across all questions to obtain an overall physician accuracy of 56.9%. We evaluated state-of-the-art LLMs including Gemini-1.5-Pro (Google's AI), Claude-3-Opus (Anthropic), and GPT-4o (OpenAI). We incorporated specific domain knowledge from current EASL guidelines (published up to October 15th, 2024), using retrieval-augmented generation (RAG), supervised fine-tuning (SFT), only available for GPT-4o, or a combined approach. Full details are reported in the supplementary materials.[5–7] The accuracy of each LLM configuration was measured as the percentage of correctly answered questions and compared against human performance by physician test-takers, using Fisher's exact test. We defined statistical significance as a two-tailed p value less than 0.05.

As shown in Fig. 1, the baseline performance of each LLM showed varying degrees of improvement over human performance, with GPT-4o and Gemini-1.5-Pro achieving non-statistically significant higher accuracies of 65.7% and 67.6%, respectively, while baseline Claude-3-Opus demonstrated statistically significant higher accuracy than physicians (72.4%, $p = 0.026$). The implementation of RAG consistently enhanced model performance across all models, with RAG-enhanced versions showing significant improvements over physician performance: GPT-4o (72.4%, $p = 0.026$), Gemini-1.5-Pro (74.3%, $p = 0.017$), and Claude-3-Opus (81.9%, $p < 0.001$). SFT-GPT-4o also yielded substantial improvements (78.1%, $p = 0.002$). The combined SFT-RAG-GPT-4o configuration achieved the highest overall accuracy of 87.6% ($p < 0.001$). Performance across different topical domains is reported in the supplementary materials.

This study confirms both the comparable performance among baseline LLM configuration and physician test-takers,[7–9] and the fact that injection of domain knowledge improved

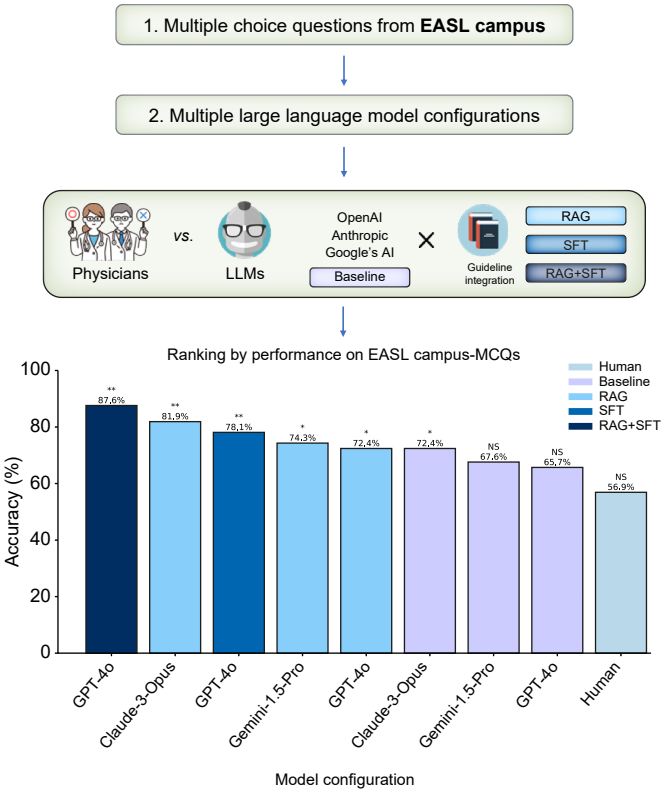

Fig. 1. Evaluation framework of liver disease knowledge using EASL Campus multiple choice questions. A total of 105 multiple choice questions were collected and used to assess both human performance and various LLM configurations. The LLMs were tested in three configurations: (1) Baseline performance using standard models, (2) RAG-enhanced models incorporating relevant EASL guideline content, and (3) models improved through SFT or combined RAG-SFT approaches. Performance is shown as percentage accuracy for each configuration compared to human baseline performance. n.s., not significant; *p <0.05; **p <0.01. LLM, large language model; RAG, retrieval-augmented generation; SFT, supervised fine-tuning.

accuracy over the pooled human baseline. Previous studies have highlighted concerning limitations in LLM performance on standardized medical assessment, with Suchman *et al.*[9] demonstrating that ChatGPT-3 and ChatGPT-4 failed to pass the American College of Gastroenterology (ACG) self-assessment tests, achieving only 65.1% and 62.4% accuracy, respectively, below the required 70% passing thresholds. Similarly, in a previous study we focused on questions related to upper gastrointestinal bleeding from the ACG self-assessment tests, confirming that baseline models did not surpass average physician-test takers' scores (average score: 75%), with SFT-GPT-4o outperforming test takers with accuracy approaching 90%.[7] With this study we confirm the potential of domain-knowledge injection to increase LLMs' accuracy on a publicly available dataset released by one of the leading hepatological societies worldwide. The improvement in performance with SFT-GPT-4o is primarily due to its ability to internalize domain-specific knowledge from the fine-tuning process, enhancing the model's understanding of clinical concepts, while RAG strengthens accuracy by dynamically incorporating relevant guideline-based information during inference, thus suggesting that integrating domain-specific knowledge is crucial for enhancing LLM performance in specialized medical fields.

However, we acknowledge that MCQ databases, while valuable for standardized assessment, have inherent limitations for evaluating LLM-based systems in clinical practice. Real-world clinical scenarios differ substantially from this controlled environment, requiring physicians to gather patient information through open-ended questioning, integrate multiple data sources, and employ multistep reasoning for diagnosis and treatment planning.[10] Future work should focus on creating comprehensive benchmarks that incorporate these aspects of clinical practice, capturing not only guideline recommended practices but pooled expertise to answer more complex clinical scenarios. Clinical questions posed by patients and providers rarely exist within a vacuum, and benchmarks should be designed to handle uncertainty quantification, sequential decision-making across different providers and systems, and multimodal data (laboratory results, imaging, clinical notes) irregularly sampled over time.

Mauro Giuffrè[1,2,*]
Alessia Distefano[2]
Simone Kresevic[3]
Lory Saveria Crocè[2]
Dennis Legen Shung[1], Collaborators

[1]*Section of Digestive Diseases, Department of Internal Medicine, Yale School of Medicine, New Haven, USA*
[2]*Department of Medical, Surgical, and Health Sciences, University of Trieste, Trieste, Italy*

[3]*Department of Engineering and Architecture, University of Trieste, Trieste, Italy*

[*]Corresponding author. Address: Mauro Giuffrè, Department of Internal Medicine (Digestive Diseases), PO Box 208019, New Haven, CT 05520-8019, USA.
E-mail address: mauro.giuffre@yale.edu (M. Giuffrè)

Received 28 December 2024; Received in revised form 3 July 2025; Accepted 7 July 2025; Available online 14 July 2025
https://doi.org/10.1016/j.jhepr.2025.101523

Financial support

DLS is supported by NIH NIDDK grant DK125718. MG is supported by the American-Italian Cancer Foundation Post-Doctoral Research Fellowship (Year 2024/2025).

Conflict of interest

Please refer to the accompanying ICMJE disclosure forms for further details.

Authors' contributions

M.G., S.K., L.S.C., and D.L.S. designed the study; A.D. processed EASL guidelines into an LLM-friendly version; M.G. and A.D. collected EASL Campus Quizzes; M.G. and S.K. performed data analysis and reviewed LLM performance. M.G., A.D., S.K., L.S.C., and D.L.S drafted and reviewed the manuscript.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhepr.2025.101523.

Collaborators

Milos Ajcevic, Marco Gulotta, Francesca Orbosuè, Lisa Fusaro, Simone Giovanni Ermacora.

References

[1] Giuffrè M, Kresevic S, You K, et al. Systematic review: the use of large language models as medical chatbots in digestive diseases. Aliment Pharmacol Ther 2024 Jul;60(2):144–166.
[2] Giuffrè M, Shung DL. Scrutinizing ChatGPT applications in Gastroenterology: a call for methodological rigor to define accuracy and preserve privacy. Clin Gastroenterol Hepatol 2024 Oct;22(10):2156–2157.
[3] Giuffrè M, You K, Shung DL. Evaluating ChatGPT in medical contexts: the imperative to guard against hallucinations and partial accuracies. Clin Gastroenterol Hepatol 2024 May;22(5):1145–1146.
[4] Alonso I, Oronoz M, Agerri R. MedExpQA: multilingual benchmarking of Large Language Models for medical question answering. Artif Intell Med 2024 Sep;155:102938.
[5] Giuffrè M, Kresevic S, Pugliese N, et al. Optimizing large language models in digestive disease: strategies and challenges to improve clinical outcomes. Liver Int 2024 Sep;44(9):2114–2124.
[6] Kresevic S, Giuffrè M, Ajcevic M, et al. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. NPJ Digit Med 2024 Apr 23;7(1):102.
[7] Giuffrè M, You K, Pang Z, et al. Expert of experts verification and alignment (EVAL) framework for Large Language Models safety in Gastroenterology. Npj Digit Med 2025;8:242.
[8] Guo Y, Li T, Xie J, et al. Evaluating the accuracy, time and cost of GPT-4 and GPT-4o in liver disease diagnoses using cases from "What is Your Diagnosis". J Hepatol 2024 Sep 20:S0168–S8278 (24) 02555-8.
[9] 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.
[10] Soroush A, Giuffrè M, Chung S, et al. Generative artificial intelligence in clinical medicine and impact on Gastroenterology. Gastroenterology 2025 Apr;15:S0016-5085(25)00634-1.