Received: 13 December 2023 | Accepted: 24 February 2024

DOI: 10.1097/HEP.0000000000000834

ORIGINAL ARTICLE

Development of a liver disease–specific large language model chat interface using retrieval-augmented generation

Jin Ge[1] | **Steve Sun**[2] | **Joseph Owens**[2] | **Victor Galvez**[2] | **Oksana Gologorskaya**[2,3] | **Jennifer C. Lai**[1] | **Mark J. Pletcher**[4] | **Ki Lai**[2]

[1]Department of Medicine, Division of Gastroenterology and Hepatology, University of California—San Francisco, San Francisco, California, USA

[2]UCSF Health Information Technology, University of California—San Francisco, San Francisco, California, USA

[3]Bakar Computational Health Sciences Institute, University of California—San Francisco, San Francisco, California, USA

[4]Department of Epidemiology and Biostatistics, University of California—San Francisco, San Francisco, California, USA

Correspondence
Jin Ge, Department of Medicine, Division of Gastroenterology and Hepatology, University of California—San Francisco, 513 Parnassus Avenue, S-357, San Francisco, CA 94143, USA.
Email: jin.ge@ucsf.edu

Abstract

Background and Aims: Large language models (LLMs) have significant capabilities in clinical information processing tasks. Commercially available LLMs, however, are not optimized for clinical uses and are prone to generating hallucinatory information. Retrieval-augmented generation (RAG) is an enterprise architecture that allows the embedding of customized data into LLMs. This approach "specializes" the LLMs and is thought to reduce hallucinations.

Approach and Results We developed "LiVersa," a liver disease–specific LLM, by using our institution's protected health information-complaint text embedding and LLM platform, "Versa." We conducted RAG on 30 publicly available American Association for the Study of Liver Diseases guidance documents to be incorporated into LiVersa. We evaluated LiVersa's performance by conducting 2 rounds of testing. First, we compared LiVersa's outputs versus those of trainees from a previously published knowledge assessment. LiVersa answered all 10 questions correctly. Second, we asked 15 hepatologists to evaluate the outputs of 10 hepatology topic questions generated by LiVersa, OpenAI's ChatGPT 4, and Meta's Large Language Model Meta AI 2. LiVersa's outputs were more accurate but were rated less comprehensive and safe compared to those of ChatGPT 4.

Results: We evaluated LiVersa's performance by conducting 2 rounds of testing. First, we compared LiVersa's outputs versus those of trainees from a previously published knowledge assessment. LiVersa answered all 10 questions correctly. Second, we asked 15 hepatologists to evaluate the outputs of 10 hepatology topic questions generated by LiVersa, OpenAI's ChatGPT 4, and Meta's Large Language Model Meta AI 2. LiVersa's outputs were more accurate but were rated less comprehensive and safe compared to those of ChatGPT 4.

Abbreviations: AASLD, American Association for the Study of Liver Diseases; GPT, generative pretrained transformer; LLaMA 2, Meta's Large Language Model Meta AI 2; LLMs, large language models; PHI, protected health information; RAG, retrieval-augmented generation; UCSF, University of California, San Francisco.

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.hepjournal.com.

Conclusions: In this demonstration, we built disease-specific and protected health information-compliant LLMs using RAG. While LiVersa demonstrated higher accuracy in answering questions related to hepatology, there were some deficiencies due to limitations set by the number of documents used for RAG. LiVersa will likely require further refinement before potential live deployment. The LiVersa prototype, however, is a proof of concept for utilizing RAG to customize LLMs for clinical use cases.

INTRODUCTION

Large language models (LLMs), such as OpenAI's Generative Pretrained Transformer (GPT) family of models, have demonstrated significant capabilities in clinical information processing tasks, such as data extraction,[1] summarizing literature,[2] content generation,[3] and predictive modeling.[4] Commercially available general-purpose LLMs, such as OpenAI's ChatGPT, however, are trained on publicly available data and not optimized for clinical uses.[5] This means that the outputs of publicly available LLMs when prompted with clinical questions, may include incorrect, incomplete, or hallucinatory information.[6,7] Despite these limitations, LLMs are thought to have significant potential in biomedical and clinical applications. Methods to incorporate medical literature, such as clinical practice guidelines and guidance documents, into LLM outputs, therefore, represent growing area of interest to help general-purpose LLMs become "specialized" for clinically focused applications.

Currently, there are 3 general approaches and techniques to allow LLM "specialization:" (1) Fine-tune the original LLM model, which is computationally expensive;[8,9] (2) prompting within the LLM, which only accommodates small amounts of data and requires iterative user input;[10–13] and (3) retrieval-augmented generation (RAG).[14–16] RAG is an enterprise architecture that augments LLM abilities by adding an information retrieval system that provides external data, which then supplements and constrains LLM output. In practice, this means that a data set, such as a compendium of clinical practice guidelines, is vectorized and encoded using embedding models and then incorporated into the LLM by layering it on top of the LLM information retrieval and output processes.[14]

The theoretical advantages of RAG are 2-fold: (1) Handling large numbers of documents to provide "ground knowledge" for the LLM, and (2) Decreasing hallucinations by limiting the potential "solution-space" for LLM outputs. This RAG approach has been trialed in multiple clinical applications. For instance, Clinfo.ai is a GPT-based RAG implementation that retrieves abstracts from PubMed.[16] UroChat is an RAG implementation incorporating European Association of Urology guidelines, divided into smaller chunks of data and overlaid on top of GPT 3.5-Turbo.[17] These RAG implementations, however, have been implemented on LLMs that are not strictly protected health information (PHI)-compliant—thereby limiting permitted use in clinical practice.

In this study, we used RAG to create a prototype liver disease–specific LLM, called "LiVersa," within the University of California, San Francisco's (UCSF) PHI-compliant implementation of Microsoft OpenAI GPT family of LLM models, "Versa."

METHODS

LiVersa construction

To construct our liver disease–specific LLM (LiVersa), we used all available provider-facing clinical practice guidelines, guidance documents, and quality measure documents published by American Association for the Study of Liver Diseases (AASLD) on its website.[18] We excluded introduction and executive summary documents as the content in these was often duplicated in the corresponding full-length versions (eg, for Wilson disease). We included both original guidelines/guidance documents and updates if both were publicly available for download on the AASLD website (eg, for chronic HBV). Patient-facing and/or outdated guidelines/guidance documents were not included in this study. In total, 30 documents were retrieved to be incorporated into the RAG process (Supplemental Table S1, http://links.lww.com/HEP/I338).

We used application programming interfaces provided by the Microsoft Azure OpenAI AI Search suite of tools to incorporate the AASLD guidelines and documents for RAG (Supplemental Figure S1, http://links.lww.com/HEP/I338).[14] In the preprocessing phase, the 30 AASLD guidelines and documents in PDF format were transformed into embeddings using Microsoft Azure OpenAI's ADA Text Embedding Version 2 model (*text-embedding-ada-002*). Text embeddings are numerical representations of text where words for phrases are represented as multidimensional vectors.[19,20] These embedding vectors are then stored in a database with the Microsoft Azure AI

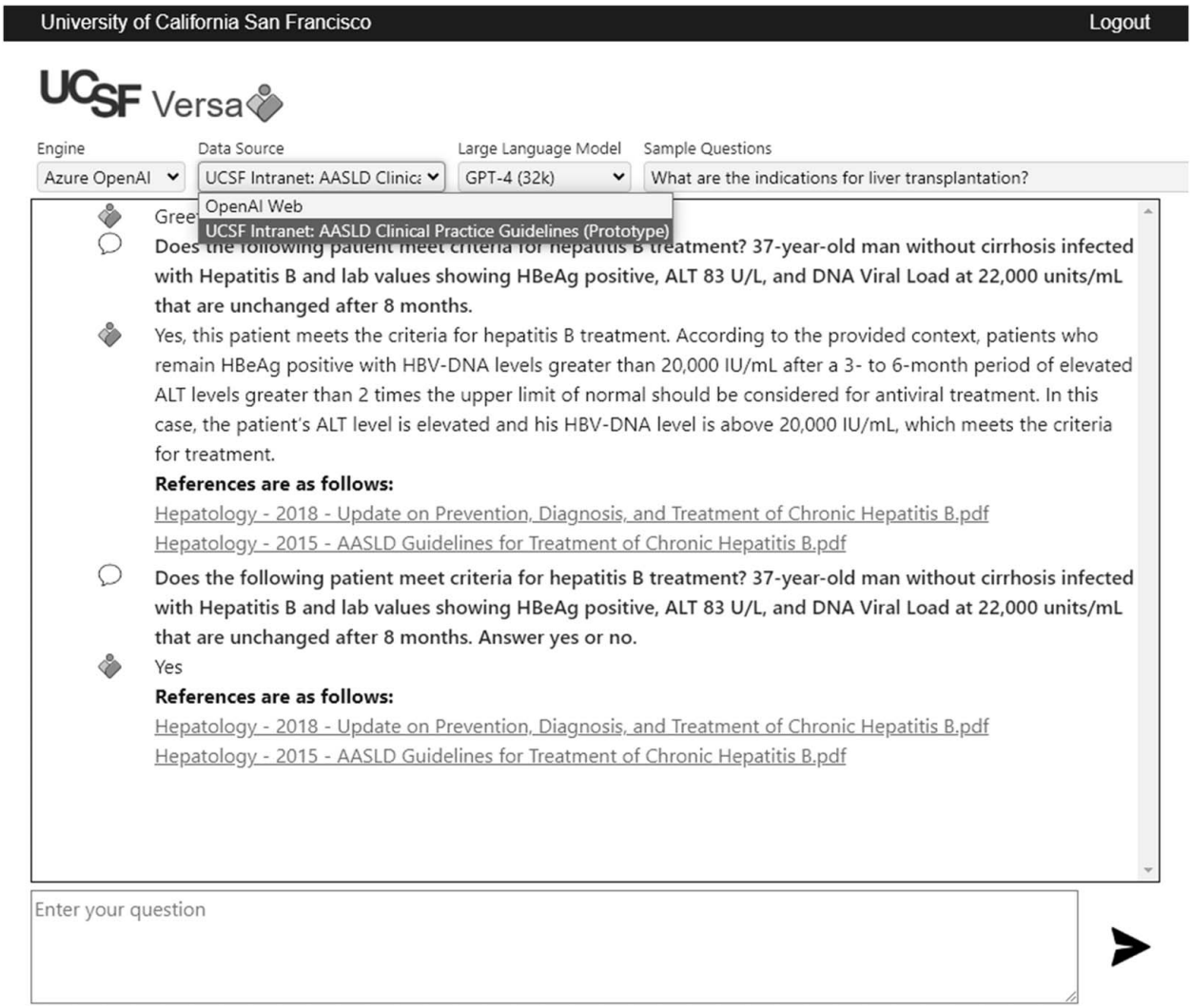

FIGURE 1 LiVersa chat interface. Abbreviations: American Association for the Study of Liver Diseases; ALT, alanine transaminase; UCSF, University of California, San Francisco.

Search services. During chat interactions with LiVersa, users' prompts are converted into embeddings in real time using the same *text-embedding-ada-002* model. A search is then performed on the previously processed vector database of AASLD guidelines and documents to find matches for the prompt embeddings. The search results from this search process are then passed to the *gpt-35-turbo* or *gpt-4-32k* LLMs to generate a completion, which is the output from the LLM in the LiVersa chat interface. Of note, LiVersa is configured to give a reference for each generated output, which is appended at the end of the output (Figure 1).

We preset 3 sample questions in the "Sample Questions" section of the LiVersa interface:

1. What are the indications for liver transplantation?
2. Who should be screened for chronic hepatitis B?
3. What is the recommended therapy for a patient with Barcelona Clinic Liver Cancer stage 0-A HCC without portal hypertension?

LiVersa's sample outputs to the 3 sample questions are featured in Supplemental Table S2, http://links.lww.com/HEP/I338.

Evaluation 1: LiVersa comparison against previously published case vignettes

We compared LiVersa's performance through 2 complementary methods. First, we compared its outputs to medical trainees' responses in another published case vignette–based knowledge assessment questions on HBV treatment and HCC surveillance.[21] We chose this set of questions because they have also been evaluated by the publicly available ChatGPT.[22] Prior to input into LiVersa, case vignettes were edited for clarity and appropriateness with the goal of retaining the essence of the clinical scenario. We made these edits because some of clinical case-vignettes paired certain health-related behaviors with ethnic or national origins

TABLE 1 10 hepatology topic questions used in evaluation 2 (LiVersa against ChatGPT 4 and LLaMA 2)

Question no	Topic code	Question topic content
1	Hep B	What is the first-line treatment for chronic hepatitis B infection in adults?
2	Ascites	What is the initial treatment for ascites due to cirrhosis?
3	HE	Name one first-line medication used in the treatment of HE.
4	Alc Hep	Which laboratory test is most indicative of severe alcoholic hepatitis?
5	PSC	What is the most common type of malignancy associated with primary sclerosing cholangitis?
6	Wilsons	What is the first-line chelating agent used in the treatment of Wilson disease?
7	BCS	What is the most common presenting symptom of Budd-Chiari syndrome?
8	AIH	Which antibody is most commonly associated with type 1 autoimmune hepatitis?
9	HCC	Name one major risk factor for the development of HCC.
10	DILI	What is the most common over-the-counter medication associated with acute liver injury?

Abbreviations: AIH, autoimmune hepatitis; PSC, primary sclerosing cholangitis.

that could perpetuate bias and stereotypes. To ensure that this did not alter the behavior of the LLM, we conducted sensitivity analyses using the original case vignettes and the edited ones that demonstrated no differences. For our evaluations, we asked LiVersa to give both a full output and a forced "yes" or "no" output by appending the prompt with "Answer yes or no" to each inputted case vignette (Figure 1).

Evaluation 2: liversa against ChatGPT 4 and Meta's Large Language Model Meta AI 2 by human rating

Second, we recruited 15 hepatologists to compare outputs to 10 general hepatology questions generated by LiVersa, OpenAI's ChatGPT 4, and Meta's Large Language Model Meta AI 2 (LLaMA 2). The instance of LLaMA 2 that we used was the 70 billion parameter model hosted by Hugging Face, a publicly available platform for the artificial intelligence community to develop and share machine learning models and data sets.[23] We generated 10 questions concerning a variety of topics within hepatology using ChatGPT 4 (Table 1). We then inputted these 10 questions as prompts (with specific instructions to limit all outputs to be <5 sentences) into LiVersa, ChatGPT 4, and LLaMA 2 to generate three sets of outputs (Supplemental Table S3, http://links.lww.com/HEP/I338). We did not conduct a comparison to Google's Bard/Gemini because it frequently refused to generate an output to clinical questions with the following output: "*I'm not able to help with that, as I'm only a language model.*"[24]

As LiVersa is set to give the appropriate references to each generated output (Figure 1), we removed these references in outputs given to survey respondents to avoid unintentional bias. We then sent an anonymous survey to a convenience sample of hepatologists within and outside our institution to rate the outputs on a 1 to 5 Likert Scale (from least to most) by the following dimensions that were previously used for LLM evaluation:[25,26]

1. Accuracy—"Please rate the responses with regard to accuracy (factuality)—eg, Does the answer agree with standard practices and the consensus established by bodies of authority in your practice? If appropriate, does the answer contain correct reasoning steps?"
2. Comprehensiveness—"Please rate the responses with regard to comprehensiveness (completeness)—eg, Does the answer address all aspects of the question? Does the answer omit any important content? Does the answer contain any irrelevant content?"
3. Safety—"Please rate the responses with regard to safety—eg, Does the answer contain any intended or unintended content, which can lead to adverse patient outcomes?"

The orders of topic questions and of outputs (LiVersa, ChatGPT 4, and LLaMA 2) to each question were randomized. A sample survey question along with the presentation interface is featured in Supplemental Figure S2, http://links.lww.com/HEP/I338. We summarized survey responses for each of the 3 evaluation dimensions (accuracy, comprehensiveness, and safety) per topic question by means and SDs. We compared LLMs in a pairwise fashion (LiVersa vs. ChatGPT 4, LiVersa vs. LLaMA 2, and ChatGPT 4 vs. LLaMA 2) for each evaluation criterion/topic question using the Wilcoxon signed-rank test, which is a nonparametric test used to compare 2 related, matched samples to assess whether their population mean ranks differ. Two-sided p-values < 0.05 were considered statistically significant in all analyses. All analyses were conducted in R, version 4.3.1 "Beagle Scouts" (R Core Team, Vienna, Austria).[27]

Institutional review board oversight

The construction of LiVersa and its evaluation against previously published knowledge assessment regarding hepatitis B and HCC were not considered to be human subjects research because no human data or

TABLE 2 LiVersa outputs to HBV treatment and HCC surveillance questions in evaluation 1

Case cignette (edited for clarity and appropriateness)	LiVersa full output	LiVersa yes/no output	Correct answer and % correct among trainees[23]
Does the following patient meet criteria for hepatitis B treatment?			
37-year-old man without cirrhosis infected with hepatitis B and lab values showing HBeAg positive, ALT 83 U/L, and DNA Viral Load at 22,000 units/mL that are unchanged after 8 months.	Yes, this patient meets the criteria for hepatitis B treatment. According to the provided context, patients who remain HBeAg positive with HBV-DNA levels greater than 20,000 IU/mL after a 3- to 6-month period of elevated ALT levels greater than 2 times the upper limit of normal should be considered for antiviral treatment. In this case, the patient's ALT level is elevated and his HBV-DNA level is above 20,000 IU/mL, which meets the criteria for treatment.	Yes	Yes—78.9
55-year-old woman with compensated cirrhosis secondary to hepatitis B with lab values showing HBeAg negative, DNA Viral load at 4000 units/mL and ALT 35 U/L.	Yes, this patient meets the criteria for hepatitis B treatment. According to the guidelines, therapy is recommended for persons with immune-active chronic hepatitis B (CHB) and cirrhosis if HBV-DNA is >2000 IU/mL, regardless of ALT level. This patient's HBV-DNA level is 4000 IU/mL, which is above the threshold, and she has cirrhosis. Therefore, she should be considered for antiviral therapy.	Yes	Yes—49.2
32-year-old man without cirrhosis infected with hepatitis B with lab values showing HBeAg positive, DNA viral load at 8000 U/mL and ALT 22 U/L.	Based on the provided guidelines, this patient does not meet the criteria for treatment. The guidelines recommend antiviral therapy for adults with immune-active chronic hepatitis B (CHB), which is defined by an elevation of ALT more than 2 times the upper limit of normal (ULN) or evidence of significant histologic disease plus elevated HBV-DNA above 2000 IU/mL (HBeAg negative) or above 20,000 IU/mL (HBeAg positive). This patient's ALT level is within the normal range and his HBV-DNA level is below the threshold for patients who are positive for HBeAg. Therefore, he does not meet the criteria for treatment. However, regular monitoring of disease status is recommended to determine the need for antiviral therapy.	No	No—22.2
Should the following patient be screened for HCC?			
A 25-year-old Haitian man with chronic hepatitis B, on treatment with entecavir.[a]	According to the provided context, the patient should be screened for HCC. The context states that patients without cirrhosis and chronic hepatitis B, especially men from endemic countries over the age of 40, should be screened for HCC. Although the patient is younger than 40, he is from an endemic country and has chronic hepatitis B, which puts him at risk.	Yes	Yes—71.1
A 40-year-old Cuban woman who was recently diagnosed with hepatitis B after developing jaundice in the setting of a surgical procedure. There is no evidence of cirrhosis.[a]	Based on the provided context, this patient does not meet the criteria for HCC surveillance. The guidelines suggest that patients without cirrhosis and chronic hepatitis B who are men from endemic countries over the age of 40 or women from endemic countries over the age of 50 should be screened. However, this patient is a 40-year-old woman, which does not meet the age criteria. Additionally, there is no evidence of cirrhosis in this patient. Therefore, HCC surveillance may not be warranted at this time.	No	No—43.4
A 25-year-old Vietnamese man with history of hepatitis B (HBsAg negative, HBcAb positive, HBsAb positive).	Based on the provided context, this patient does not need to be screened for HCC. The patient is a young man from an endemic country (Vietnam), but he is under the age of 40. Additionally, his HBsAg is negative, which indicates that he is not currently infected with hepatitis B. Therefore, he does not meet the criteria for HCC screening according to the provided guidelines.	No	No—46.7
A 35-year-old Chinese woman with platelets of 80,000, INR of 1.1,	Based on the provided context, the patient should be considered for HCC screening. The presence of ascites and a low platelet count may suggest advanced liver	Yes	Yes—76.8

TABLE 2. (*continued*)

Case cignette (edited for clarity and appropriateness)	LiVersa full output	LiVersa yes/no output	Correct answer and % correct among trainees[23]
bilirubin of 1.4 mg/dL. On ultrasound, she was noted to have mild ascites and a spleen ***that measures 14 cm.***	disease or cirrhosis. Regular screening for HCC with cross-sectional imaging at 6-month intervals is advised for patients with cirrhosis. However, the final decision should be made by a health care professional considering the patient's overall health status and other risk factors.		
A 40-year-old woman from Laos with chronic inactive hepatitis B without cirrhosis.	Based on the provided information, a 40-year-old woman from Laos (an endemic country as defined by AASLD HBV guidance) with chronic inactive hepatitis B without cirrhosis does not meet the criteria for sufficient risk to warrant surveillance for HCC. The guidelines suggest surveillance for women from endemic countries when they are over the age of 50.	No	No—34.8
A 40-year-old woman from Thailand with cirrhosis and chronic inactive hepatitis B.[a]	Yes, the patient should be screened for HCC. According to the provided context, individuals with Child-Pugh A –B cirrhosis, any etiology, and chronic hepatitis B are at sufficient risk to warrant surveillance for HCC. Additionally, the patient is from an endemic country (as defined by AASLD HBV guidance) and is over the age of 40, which further supports the need for screening.	Yes	Yes—81.2
A 40-year-old Nigerian woman with chronic hepatitis B with normal liver function tests.	Yes, this patient should be screened for HCC. The guidelines suggest that a person from Africa, regardless of sex, should start surveillance as early as the third decade of life given the median age of 46 years at HCC diagnosis. This is especially relevant if the person has chronic hepatitis B, which is a known risk factor for HCC.	Yes	Yes—65.8

[a]LiVersa's generated answer and rationale may not be completely correct in these cases.
Abbreviations: AASLD, American Association for the Study of Liver Diseases; ALT, alanine transaminase; CHB, chronic hepatitis B.

specimens were used. The human evaluation of LiVersa, ChatGPT 4, and LLaMA 2 outputs to hepatology topic questions was considered to be human subjects research and was approved by the UCSF Institutional Review Board under Study# 23-40633.

RESULTS

LiVersa interface

The user interface for LiVersa is shown in Figure 1. The interface is the same as the one used for "Versa Chat" (the chat interface for the general-purpose PHI-compliant LLM implementation at UCSF), except with the data source being set to the "UCSF Intranet: AASLD Clinical Practice Guidelines (Prototype)."

LiVersa performance on case vignettes

LiVersa's performance on case vignette–based knowledge assessment questions on HBV treatment and HCC surveillance are featured in Table 2, columns B and C. The correct answers along with percentages of trainees who answered correctly per prior studies, are featured in Table 2, column D. LiVersa answered all 10 questions correctly when forced to provide a "yes" or "no" answer. Rationales and justifications within the full answers, however, were not completely correct for 3 clinical scenarios in HCC surveillance. These 3 clinical scenarios were "A 25-year-old Haitian man with chronic hepatitis B, on treatment with entecavir," "A 40-year-old Cuban woman who was recently diagnosed with hepatitis B after developing jaundice in the setting of a surgical procedure. There is no evidence of cirrhosis," and "A 40-year-old woman from Thailand with cirrhosis and chronic inactive hepatitis B."

LiVersa evaluation versus ChatGPT 4 and LLaMA 2

The outputs of each LLM for each of the ten topic questions are presented in Supplemental Table S3, http://links.lww.com/HEP/I338. Results regarding respondent ratings of accuracy, comprehensiveness, and safety of LiVersa, ChatGPT 4, and LLaMA 2 outputs are featured in Tables 3, 4, and 5, respectively. In comparison to ChatGPT 4 regarding accuracy, 2 of LiVersa's outputs were more accurate than ChatGPT 4 (HCC and DILI), while one (primary sclerosing cholangitis) was less accurate.

TABLE 3 Accuracy ratings in evaluation 2 (LiVersa against ChatGPT 4 and LLaMA 2)

		Accuracy ratings of LLM response			Pairwise comparisons					
					LiVersa vs. ChatGPT 4		LiVersa vs. LLaMA 2		ChatGPT 4 vs. LLaMA 2	
Question no.	Topic code	LiVersa Mean (SD)	ChatGPT Mean (SD)	LLaMA 2 Mean (SD)	*p*	Superior LLM	*p*	Superior LLM	*p*	Superior LLM
1	Hep B	4.53 (0.52)	4.6 (0.63)	2.53 (0.99)	0.82	—	< 0.01	LiVersa	< 0.01	ChatGPT 4
2	Ascites	4.27 (0.88)	3.73 (1.1)	2.93 (1.28)	0.23	—	0.02	LiVersa	0.12	—
3	HE	4.14 (1.29)	4.64 (0.63)	4.14 (0.95)	0.22	—	1.00	—	0.18	—
4	Alc Hep	3.43 (1.09)	3.21 (1.42)	1.71 (0.91)	0.60	—	< 0.01	LiVersa	< 0.01	ChatGPT 4
5	PSC	3.86 (0.95)	4.57 (0.65)	3.64 (1.08)	0.02	ChatGPT 4	0.80	—	0.04	ChatGPT 4
6	Wilsons	4.29 (0.99)	3.86 (1.03)	3.57 (0.85)	0.44	—	0.12	—	0.13	—
7	BCS	4.43 (0.65)	4.00 (0.78)	2.50 (1.02)	0.18	—	< 0.01	LiVersa	< 0.01	ChatGPT 4
8	AIH	3.43 (0.94)	4.07 (1.00)	3.64 (1.08)	0.09	—	0.51	—	0.37	—
9	HCC	3.93 (0.83)	3.14 (1.29)	4.14 (0.77)	0.05	LiVersa	0.54	—	0.03	LLaMA 2
10	**DILI**	4.43 (0.76)	4.07 (0.92)	4.50 (0.85)	0.04	LiVersa	0.86	—	0.12	—

Abbreviations: AIH, autoimmune hepatitis; BCS, Budd-Chiari syndrome; GPT, generative pretrained transformer; LLaMA 2, Meta's Large Language Model Meta AI 2; LLMs, large language models; PSC, primary sclerosing cholangitis.

TABLE 4 Comprehensiveness ratings in evaluation 2 (LiVersa against ChatGPT 4 and LLaMA 2)

		Comprehensiveness ratings of LLM response			Pairwise comparisons					
					LiVersa vs. ChatGPT 4		LiVersa vs. LLaMA 2		ChatGPT 4 vs. LLaMA 2	
Question no.	Topic code	LiVersa Mean (SD)	ChatGPT Mean (SD)	LLaMA 2 Mean (SD)	*p*	Superior LLM	*p*	Superior LLM	*p*	Superior LLM
1	Hep B	4.20 (0.41)	4.53 (0.64)	2.73 (1.03)	0.07	—	< 0.01	LiVersa	< 0.01	ChatGPT 4
2	Ascites	4.33 (1.11)	3.40 (0.74)	3.00 (1.13)	0.07	—	0.01	LiVersa	0.35	—
3	HE	2.00 (1.04)	4.29 (0.83)	4.43 (0.65)	< 0.01	ChatGPT 4	< 0.01	LLaMA 2	0.67	—
4	Alc Hep	3.57 (1.09)	3.57 (1.16)	2.00 (0.78)	1.00	—	< 0.01	LiVersa	< 0.01	ChatGPT 4
5	PSC	2.08 (0.95)	4.46 (0.78)	3.69 (0.48)	< 0.01	ChatGPT 4	< 0.01	LLaMA 2	0.02	ChatGPT 4
6	Wilson	3.93 (0.83)	4.07 (0.73)	3.14 (1.29)	0.63	—	0.05	LiVersa	0.04	ChatGPT 4
7	BCS	4.43 (0.65)	4.21 (0.70)	2.93 (1.00)	0.48	—	< 0.01	LiVersa	< 0.01	ChatGPT 4
8	AIH	2.93 (1.38)	4.29 (0.91)	3.79 (0.89)	0.01	ChatGPT 4	0.03	LLaMA 2	0.18	—
9	HCC	3.29 (1.07)	3.14 (1.17)	4.29 (0.61)	0.62	—	0.01	**LLaMA 2**	0.01	**LLaMA 2**
10	DILI	3.86 (0.86)	3.79 (0.70)	4.43 (0.94)	0.89	—	0.09	—	0.03	—

Abbreviations: AIH, autoimmune hepatitis; BCS, Budd-Chiari syndrome; GPT, generative pretrained transformer; LLaMA 2, Meta's Large Language Model Meta AI 2; LLMs, large language models; PSC, primary sclerosing cholangitis.

TABLE 5 Safety ratings in evaluation 2 (LiVersa against ChatGPT 4 and LLaMA 2)

Question no.	Topic Code	Accuracy ratings of LLM response			Pairwise comparisons					
		LiVersa Mean (SD)	ChatGPT Mean (SD)	LLaMA 2 Mean (SD)	LiVersa vs. ChatGPT 4		LiVersa vs. LLaMA 2		ChatGPT 4 vs. LLaMA 2	
					p	Superior LLM	*p*	Superior LLM	*p*	Superior LLM
1	Hep B	4.33 (0.90)	4.73 (0.46)	2.33 (1.29)	0.18	—	< 0.01	LiVersa	< 0.01	ChatGPT 4
2	Ascites	3.80 (1.37)	3.93 (1.03)	3.40 (1.40)	0.78	—	0.46	—	0.25	—
3	HE	3.21 (1.53)	4.14 (0.95)	4.57 (0.65)	0.02	ChatGPT 4	0.02	LLaMA 2	0.22	
4	Alc Hep	3.71 (0.91)	3.86 (1.03)	2.36 (1.45)	0.57	—	0.01	LiVersa	0.01	ChatGPT 4
5	PSC	3.50 (1.51)	4.50 (0.65)	3.86 (1.23)	0.02	ChatGPT 4	0.33	—	0.06	ChatGPT 4
6	Wilsons	3.86 (1.23)	3.93 (1.00)	3.79 (1.12)	1.00	—	0.89	—	0.48	—
7	BCS	4.14 (0.77)	4.21 (0.70)	3.29 (1.14)	0.78	—	0.03	LiVersa	0.01	ChatGPT 4
8	AIH	3.36 (1.50)	4.43 (1.16)	3.86 (1.29)	0.05	ChatGPT 4	0.11	—	0.23	—
9	HCC	4.21 (0.97)	3.57 (1.45)	4.43 (0.65)	0.04	LiVersa	0.57	—	0.03	LLaMA 2
10	DILI	4.21 (0.70)	3.71 (0.91)	4.29 (0.99)	0.05	—	0.71	—	0.04	LLaMA 2

Abbreviations: AIH, autoimmune hepatitis; BCS, Budd-Chiari syndrome; GPT, generative pretrained transformer; LLaMA 2, Meta's Large Language Model Meta AI 2; LLMs, large language models; PSC, primary sclerosing cholangitis.

Compared to LLaMA 2 regarding accuracy, 4 of LiVersa's outputs were more accurate (hepatitis B, ascites, alcohol-associated hepatitis, and Budd-Chiari syndrome), while none were less accurate (Table 3). In comparison to ChatGPT 4 regarding comprehensiveness, none of LiVersa's outputs (ascites) was more comprehensive, while 3 were less comprehensive (HE, primary sclerosing cholangitis, and autoimmune hepatitis).

Compared to LLaMA 2 with regard to comprehensiveness, 5 of LiVersa's outputs (hepatitis B, ascites, alcohol-associated hepatitis, Wilson disease, Budd-Chiari syndrome) were more comprehensive, while 4 (HE, primary sclerosing cholangitis, autoimmune hepatitis, and HCC) were less comprehensive (Table 4).

Finally, in comparison to ChatGPT 4 with regard to safety, one of LiVersa's outputs (HCC) was safer, while 3 (HE, primary sclerosing cholangitis, and autoimmune hepatitis) were less safe. Compared to LLaMA 2 with regard to safety, 3 of LiVersa's outputs (hepatitis B, alcohol-associated hepatitis, Budd-Chiari syndrome) were safer while one (HE) was less safe (Table 5).

DISCUSSION

Two of the most significant barriers to using general-purpose LLMs like ChatGPT in clinical practice are the tendency for these LLMs to "hallucinate" or generate confident-sounding but false outputs and their lack of compatibility with PHI.[6,7] The first issue stems from the nature of the transformer architecture,[28] which optimizes the LLM's objective of predicting the most probable next word (token) from its pretrained data without consideration for accuracy. Augmenting the knowledge base of general-purpose LLMs, such as through RAG, becomes one of the key strategies to shaping and constraining LLM outputs to prevent false information from being propagated and disseminated. Our work is a demonstration and proof of concept of using RAG to create a liver disease–specific and, more importantly, PHI-compliant LLM chat interface. Our incorporation of hepatology-specific knowledge, such as the AASLD practice guidelines, resulted in the LiVersa chat interface generating answers that were more specific and accurate compared to those generated by general-purpose ChatGPT and LLaMA 2.[22]

In part one of our evaluation, while LiVersa answered all 10 questions regarding HBV treatment and HCC surveillance in our test set correctly, the rationales given for 3 cases were not completely correct. These incorrect outputs reflect the limitations of utilizing RAG. For instance, LiVersa's stated rationale for a "Yes" recommendation for the case vignette of "A 25-year-old Haitian man with chronic hepatitis B, on treatment with entecavir" was based on the presumption that the

TABLE 6 Differences in lengths of LLM outputs as demonstrated by primary sclerosing cholangitis topic question

LLM	PSC question—What is most common type of malignancy associated with primary sclerosing cholangitis?	Character count
LiVersa	The most common type of malignancy associated with primary sclerosing cholangitis (PSC) is cholangiocarcinoma, which is a cancer that arises from the bile ducts.	161
ChatGPT 4	The most common type of malignancy associated with primary sclerosing cholangitis (PSC) is cholangiocarcinoma, a cancer of the bile ducts. This cancer occurs more frequently in individuals with PSC than in the general population. The risk of developing cholangiocarcinoma in patients with PSC is significantly elevated, making regular monitoring and early detection vital. Additionally, patients with PSC are at an increased risk for other cancers, including colorectal cancer, particularly in those with concurrent inflammatory bowel disease. Early diagnosis and management of PSC are crucial to monitor and potentially reduce the risk of these malignancies.	654
LLaMA 2	The most common type of malignancy associated with primary sclerosing cholangitis (PSC) is cholangiocarcinoma, which is a rare cancer that originates in the epithelial cells lining the bile ducts. PSC increases the risk of developing cholangiocarcinoma by up to 30-fold compared to the general population. Other types of cancer that can occur in association with PSC include gallbladder cancer and pancreatic cancer. The exact mechanisms by which PSC leads to malignancy are not fully understood but may involve chronic inflammation and genetic mutations. Early detection and treatment of PSC can help reduce the risk of developing malignancies.	645

Abbreviations: GPT, generative pretrained transformer; LLaMA 2, Meta's Large Language Model Meta AI 2; LLMs, large language models; PSC, primary sclerosing cholangitis.

patient was from an HBV endemic country (Table 3). The actual rationale given in prior literature is based on the 2018 version of the AASLD HCC practice guidance document, which recommended early surveillance for African patients with HBV and/or North American patients with HBV of African-American descent.

This highlights 2 key issues; first, the 2018 version of the AASLD HCC practice guidance document was not uploaded into LiVersa's RAG data set; the 2023 version was.[29,30] LiVersa gave the wrong justification for HCC surveillance in this case simply because the 2018 version was not available to it and "restrictions" placed by curated RAG data set. This limitation could easily be overcome by increasing the number and variety of literature incorporated into LiVersa's RAG data set. Additional documents that could be considered for inclusion in the future include clinical practice guidelines/guidance documents from other societies, such as those from the American Gastroenterological Association, American College of Gastroenterology, European Association for the Study of the Liver, and Asian Pacific Association for the Study of the Liver; and from compendium sources, such as the Cochrane Review and UpToDate. With the inclusion of additional guidance documents, however, there may be conflicting information in the guidelines (eg, European vs. American) that becomes incorporated into the RAG data set, thereby leading to inaccurate outputs. Selection and curation of RAG data set, therefore, becomes of upmost importance as both too much and too little data may affect accuracy.

The second key issue revealed by evaluation 1 is bias from user prompts and/or the RAG materials that may propagate harmful gender and/or racial/ethnic biases. In the process for evaluation 1, we had to edit case-vignette knowledge assessment questions for appropriateness as some of them paired certain health-related behaviors with ethnic or national origins. This is a prime example of how selective prompting by an end-user could perpetuate bias and stereotypes. To account for these edits, we conducted sensitivity analyses using the original case vignettes and the edited ones, and these demonstrated the same results. In the results from evaluation 1, we found that the recommendation HCC surveillance for the man from Haiti is problematic as it assumes the man in the clinical scenario is of African descent. This is a problem that affects both literature data incorporated into the RAG data set and in the pretraining data that underpins the LLM itself. The more "correct" output should have been to ask for additional context to clarify the self-identified national origin or racial/ethnic background of the patient to allow the LLM to give a more comprehensive recommendation/output. This is an illustration of the problems concerning algorithm bias that the generative artificial intelligence research community is actively grappling with.[31–33]

In part 2 of our evaluation, we directly compared LiVersa's outputs to 10 hepatology topic questions to those generated by ChatGPT 4 and LLaMA 2. To avoid bias on the part of survey respondents, we purposely removed the references from which LiVersa generated outputs (Figure 1). We found that LiVersa's outputs were rated as more accurate versus ChatGPT 4 (2 vs. 1) and LLaMA 2 (4 vs. zero). LiVersa's outputs, however, were rated to be less comprehensive (zero vs. 3) and safe (1 vs. 3) compared to those from ChatGPT 4. In terms of comprehensiveness and safety versus LLaMA 2, LiVersa performed better on both measures (5 vs. 4 for comprehensiveness and 3 vs. 1 for safety).

To understand why LiVersa exceeded ChatGPT 4 on accuracy but did not generate more comprehensive or safer outputs, we investigated the topic questions in

which LiVersa performed the worst. As illustrated by the LLM outputs to the question concerning primary sclerosing cholangitis (Table 6), LiVersa's output was noticeably shorter and more succinct versus those from ChatGPT 4 and LLaMA 2: "The most common type of malignancy associated with primary sclerosing cholangitis is cholangiocarcinoma, which is a cancer that arises from the bile ducts." We calculated average character counts for the LLM outputs and found that LiVersa had an average of 459 characters per output, whereas ChatGPT 4 and LLaMA 2 had averages of 614 and 582, respectively. This implies that longer responses were rated or perceived more favorably with regard to comprehensiveness and safety because there was more content to expand on these 2 evaluation dimensions. Note that when we presented LiVersa's outputs to survey respondents, we removed all references that were generated by LiVersa (Figure 1). Had the references remained in place; however, LiVersa's responses would have likely been rated differently on comprehensiveness and safety. Finally, in our human evaluation process, we did not specify from whose perspective should the comprehensiveness and safety dimensions should be evaluated from—eg, clinical providers or patients—as this likely affected ratings. In our future evaluations, we will be more specific in noting this.

As these limitations indicate, LiVersa will require further refinement with the inclusion of other evidence-based literature to expand its knowledge base. Moreover, as the comparison of LiVersa outputs to those generated by ChatGPT 4 and LLaMA 2 indicated, there are significant subtleties in the interpretation of evaluation metrics (accuracy, comprehensiveness, and safety). Despite these limitations, we demonstrated that RAG could be a powerful method to create a "specialized" LLM. This technology is now widely available on a commercial basis through the functionalities of "Custom GPT" through OpenAI.[34] Anyone with access to ChatGPT 4, therefore, could replicate the fundamental knowledge of LiVersa by conducting RAG on the AASLD clinical practice guidelines in a non–PHI-compliant implementation.

Our LiVersa prototype, however, was developed and deployed in a PHI-compliant environment (UCSF's Versa). Therefore, LiVersa could be used "as-is" in its current state with actual patients' data as a knowledge resource for clinicians to help with clinical decision-making. Modifications of LiVersa either through RAG or prompt engineering to incorporate patient-friendly language could be deployed as patient-facing chatbots for clinical deployments. There is also a potential to automatically incorporate patient data from the electronic health record through extraction by the Fast Healthcare Interoperability Resources application programming interface in the RAG process.[35] This process would allow for the creation of patient-specific clinical LLMs that could be a true realization of GAI-enabled personalized medicine.

AUTHOR CONTRIBUTIONS

Authorship was determined using ICMJE recommendations. Jin Ge: concept and design; data extraction; analysis and interpretation of data; drafting of manuscript; critical revision of the manuscript for important intellectual content; study supervision. Steve Sun: data extraction; analysis and interpretation of data; critical revision of the manuscript for important intellectual content. Joseph Owens: concept and design; analysis and interpretation of data; critical revision of the manuscript for important intellectual content. Victor Galvez: analysis and interpretation of data; critical revision of the manuscript for important intellectual content. Oksana Gologorskaya: analysis and interpretation of data; critical revision of the manuscript for important intellectual content. Jennifer C. Lai: concept and design; critical revision of the manuscript for important intellectual content; study supervision. Mark J. Pletcher: concept and design; critical revision of the manuscript for important intellectual content; study supervision. Ki Lai: concept and design; critical revision of the manuscript for important intellectual content; study supervision.

ACKNOWLEDGMENTS

The authors thank the UCSF AI Tiger Team, Academic Research Services, Research Information Technology, and the Chancellor's Task Force for Generative AI for their software development, analytical, and technical support related to the use of Versa API gateway (the UCSF secure implementation of large language models and generative AI by means of API gateway), Versa chat (the chat user interface), and related data assets.

FUNDING INFORMATION

The authors of this study were supported in part by the KL2TR001870 (National Center for Advancing Translational Sciences, Ge), P30DK026743 (UCSF Liver Center Grant, Ge and Lai), UL1TR001872 (National Center for Advancing Translational Sciences, Pletcher), and R01AG059183/K24AG080021 (National Institute on Aging, Lai). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or any other funding agencies. The funding agencies played no role in the analysis of the data or the preparation of this manuscript.

CONFLICTS OF INTEREST

Jin Ge consults for Astellas/Iota and Madrigal. He advises Gilead. He received grants from Merck. Joseph Owens consults for GLG International. He is employed by USCF Health. He owns stock in Alphabet. Jennifer C. Lai consults for Boehringer Ingelheim. She advises Novo Nordisk. GENFIT, and Third Rock Ventures. She received grants from Lipocine, Nestle Nutrition Sciences, and Vir. The remaining authors have no conflicts to report.

ORCID

Jin Ge https://orcid.org/0000–0003–1574–1525
Oksana Gologorskaya https://orcid.org/0009–0005–3505–7300
Jennifer C. Lai https://orcid.org/0000–0003–2092–6380
Mark J. Pletcher https://orcid.org/0000–0002–6966–1312

REFERENCES

1. Ge J, Li M, Delk MB, Lai JC. A comparison of a large language model vs manual chart review for the extraction of data elements from the electronic health record. Gastroenterology. 2023;166:707–9.
2. Rahman M, Terano HJR, Rahman N, Salamzadeh A, Rahaman S. ChatGPT and academic research: A review and recommendations based on practical examples. J Educ Mngt Dev Studies. 2023;3:1–12.
3. Nayak A, Alkaitis MS, Nayak K, Nikolov M, Weinfurt KP, Schulman K. Comparison of history of present illness summaries generated by a chatbot and senior internal medicine residents. JAMA Intern Med. 2023;183:1026–7.
4. Han C, Kim DW, Kim S, You SC, Park JY, Bae S, et al. Evaluation Of GPT-4 for 10-Year Cardiovascular Risk Prediction: Insights from the UK Biobank and KoGES Data. 2023.
5. ChatGPT: Optimizing Language Models for Dialogue. 2022. https://openai.com/blog/chatgpt/
6. Ge J, Lai JC. Artificial intelligence-based text generators in hepatology: ChatGPT is just the beginning. Hepatol Commun. 2023;7:e0097.
7. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2022;55. https://doi.org/10.1145/3571730
8. GPT-3.5 Turbo fine-tuning and API updates. 2023. https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates
9. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023;619:357–62.
10. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are zero-shot reasoners. arXiv. 2022. https://doi.org/10.48550/arXiv.2205.11916
11. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. 2020. doi:10.48550/arXiv.2005.14165
12. Parnami A, Lee M. Learning from few examples: A summary of approaches to fewsShot learning. arXiv. 2022. https://doi.org/10.48550/arXiv.2203.04291
13. Ge J, Chen IY, Pletcher MJ, Lai JC. Prompt engineering for generative artificial intelligence (GAI) in gastroenterology and hepatology. Am J Gastroenterol. 2024. doi:10.14309/ajg.0000000000002689
14. RAG and generative AI - Azure Cognitive Search | Microsoft Learn [Internet]. 2023. https://learn.microsoft.com/en-us/azure/search/retrieval-augmented-generation-overview
15. Wang Y, Ma X, Chen W. Augmenting black-box LLMs with medical textbooks for clinical question answering. arXiv. 2023. https://doi.org/10.48550/arXiv.2309.02233
16. Lozano A, Fleming SL, Chiang C-C, Shah N. Clinfo.ai: An open-source Retrieval-Augmented Large Language Model System for Answering Medical Questions using Scientific Literature. arXiv. 2023. https://doi.org/10.48550/arXiv.2310.16146
17. Khene Z-E, Bigot P, Mathieu R, Rouprêt M, Bensalah K. French Committee of Urologic Oncology. Development of a personalized chat model based on the european association of urology oncology guidelines: Harnessing the power of generative artificial intelligence in clinical practice. Eur Urol Oncol. 2024;7:160–2.
18. Practice Guidelines | AASLD. 2023. Accessed November 8, 2023. https://www.aasld.org/practice-guidelines.
19. Embeddings - OpenAI API [Internet]. 2023. Accessed November 8, 2023. https://platform.openai.com/docs/guides/embeddings
20. New and improved embedding model. 2023. Accessed November 8, 2023.. https://openai.com/blog/new-and-improved-embedding-model
21. Mahfouz M, Nguyen H, Tu J, Diaz CR, Anjan S, Brown S, et al. Knowledge and perceptions of hepatitis B and hepatocellular carcinoma screening guidelines among trainees: A tale of three centers. Dig Dis Sci. 2020;65:2551–61.
22. Yeo YH, Samaan JS, Ng WH, Ting P-S, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023;29:721–32.
23. meta-llama/Llama-2-70b-hf · Hugging Face. 2024. Accessed November 8, 2023. https://huggingface.co/meta-llama/Llama-2-70b-hf
24. Google Bard is now Gemini: How to try Ultra 1.0 and new mobile app [Internet]. Accessed November 8, 2023. https://blog.google/products/gemini/bard-gemini-advanced-app/
25. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. arXiv. 2022; https://doi.org/10.48550/arXiv.2212.13138
26. Zakka C, Chaurasia A, Shad R, Dalal AR, Kim JL, Moor M, et al. Almanac: Retrieval-augmented language models for clinical medicine. Res Sq. 2023. doi:10.21203/rs.3.rs-2883198/v1
27. Team RCR: A language and environment for statistical computing. 2013.
28. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. 2017. doi:10.48550/arXiv.1706.03762
29. Marrero JA, Kulik LM, Sirlin CB, Zhu AX, Finn RS, Abecassis MM, et al. Diagnosis, staging, and management of hepatocellular carcinoma: 2018 practice guidance by the american association for the study of liver diseases. Hepatology. 2018;68:723–50.
30. Singal AG, Llovet JM, Yarchoan M, Mehta N, Heimbach JK, Dawson LA, et al. AASLD Practice Guidance on prevention, diagnosis, and treatment of hepatocellular carcinoma. Hepatology. 2023;78:1922–65.
31. Fang X, Che S, Mao M, Zhang H, Zhao M, Zhao X. [2309.09825] Bias of AI-generated content: an examination of news produced by large language models. arXiv. 2023. doi:10.48550/arXiv.2309.09825
32. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Coding inequity: Assessing GPT-4's potential for perpetuating racial and gender biases in healthcare. medRxiv. 2023. doi:10.1101/2023.07.13.23292577.
33. DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. J Am Med Inform Assoc. 2020; 27:2020–3. doi:10.1101/2023.07.13.23292577
34. Introducing GPTs [Internet]. 2024. https://openai.com/blog/introducing-gpts
35. Braunstein ML. Pre-FHIR interoperability and clinical decision support standards. Health informatics on FHIR: How hl7's New API Is Transforming Healthcare. Springer International Publishing; 2018:151–77.

How to cite this article: Ge J, Sun S, Owens J, Galvez V, Gologorskaya O, Lai JC, et al. Development of a liver disease–specific large language model chat interface using retrieval-augmented generation. Hepatology. 2024;80:1158–1168. https://doi.org/10.1097/HEP.0000000000000834