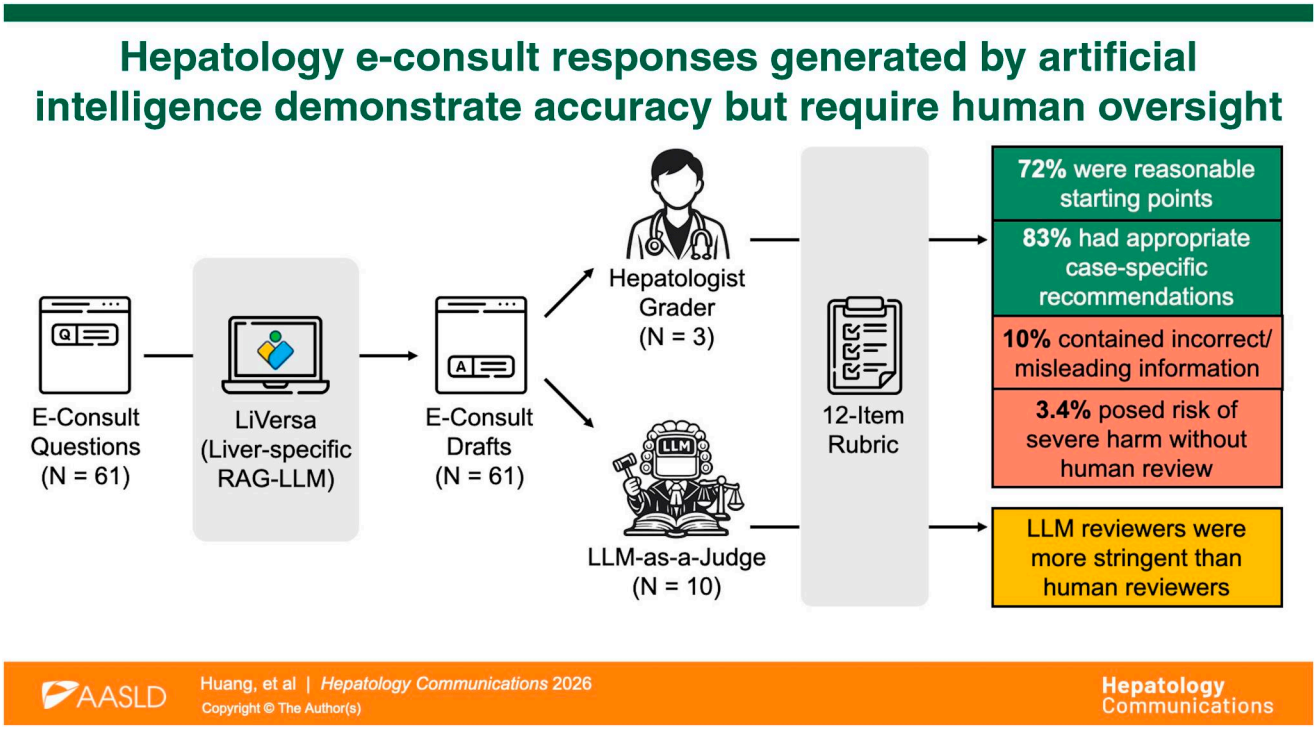
Hepatology e-consult responses generated by artificial intelligence demonstrate accuracy but require human oversight

VISUAL ABSTRACT



Received: 3 March 2026 | Accepted: 25 April 2026

DOI: 10.1097/HC9.0000000000000983

ORIGINAL ARTICLE

OPEN

Hepatology e-consult responses generated by artificial intelligence demonstrate accuracy but require human oversight

Holly K.T. Huang[1] | **Debra W. Yen**[2] | **Michelle Y. Li**[1] | **Gabrielle Jutras**[3] | **Lisa X. Deng**[1,4] | **Charles E. McCulloch**[5] | **Mark J. Pletcher**[5] | **Jennifer C. Lai**[1] | **Bilal Hameed**[1] | **Jin Ge**[1]

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

[2]Division of Hepatology, Department of Transplant, California Pacific Medical Center, San Francisco, California, USA

[3]Division of Hepatology, Department of Medicine, Centre Hospitalier de l'Université de Montréal, Montréal, Québec, Canada

[4]Department of Medicine, Veterans Affairs Medical Center, San Francisco, California, USA

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

Correspondence
Jin Ge, University of California San Francisco, 513 Parnassus Avenue, S-357, San Francisco, CA 94143, USA.
Email: jin.ge@ucsf.edu

Abstract

Background: Electronic consultations (e-consults) improve specialist access but burden providers. We developed LiVersa, a customized large language model (LLM) for liver diseases. We evaluated its performance in drafting hepatology e-consult responses and the equivalence between human and machine reviewers.

Methods: LiVersa-generated responses for hepatology e-consults answered at the University of California San Francisco (UCSF) from January to March 2025. Using a 12-item rubric, 3 independent hepatologists and "LLM-as-a-judge" (OpenAI-o1) evaluated drafts against original responses. We tested equivalence between human reviewers and "LLM-as-a-judge" using two one-sided tests (TOST).

Results: Among 61 e-consults, the most common categories were abnormal liver function tests (34%), hepatitis B (23%), and abnormal imaging (21%). LiVersa drafts demonstrated no differences from hepatologist responses in word count (284 vs. 264, $p = 0.47$) and verbosity (24 vs. 25 words per sentence, $p = 0.44$). Human reviewers rated 72% of drafts as reasonable starting points and 83% as providing appropriate case-specific recommendations; 10% contained misleading/incorrect information, and 3.4% posed a risk of severe harm. LiVersa performed better at avoiding misleading information and extraneous suggestions but scored lower on clinical equivalence, immediate usability, and comprehensiveness. LLM-

Abbreviations: AASLD, American Association for the Study of Liver Diseases; AI, artificial intelligence; CAD, coronary artery disease; CI, confidence interval; CT, computed tomography; DILI, drug-induced liver injury; e-consult, electronic consultation; EHR, electronic health record; FDG, fluorodeoxyglucose; GenAI, generative artificial intelligence; HBcAb, hepatitis B core antibody; HBV, hepatitis B virus; ICC, intraclass correlation coefficient; INR, international normalized ratio; IVIG, intravenous immunoglobulin; LLM, large language model; MELD, Model for End-stage Liver Disease; PHI, protected health information; PT, prothrombin time; RAG, retrieval augmented generation; SD, standard deviation; TOST, two one-sided tests; 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.hepcommjournal.com.

based reviewers were more stringent than human reviewers, rating fewer drafts as clinically equivalent (27% vs. 48%) and more as potentially harmful (67% vs. 20%), with agreement on accuracy, precision, and comprehensiveness (mean difference 0.026–0.029; TOST $p < 0.05$).

Conclusions: Customized LLMs like LiVersa show promise for e-consult drafting but require human oversight. LLM-as-a-judge was more conservative than humans, supporting its role in rapid quality assurance during model updates.

Keywords: cirrhosis, clinical decision support systems, liver diseases, referral and consultation, telemedicine

INTRODUCTION

Electronic consultations (e-consults) are asynchronous generalist–specialist communication via the electronic health record (EHR) that may eliminate the need for face-to-face specialist visits for low-acuity and low-complexity clinical questions. Prior studies have consistently demonstrated that e-consults significantly improve access to specialist care, reduce wait times, increase cost savings, and enhance provider communication.[1–3] These outcomes have been demonstrated in hepatology, with one program achieving a 22-hour average response time compared with 68 days for traditional in-person visits and a reduction of patient travel burden by over 10,000 cumulative miles over 38 months; moreover, primary care providers were able to successfully implement specialist recommendations in the e-consults in 97% of cases.[4]

Despite these promising outcomes in enhancing access to specialist care, the results were mixed for the specialist experience. In prior studies of e-consults, many specialist respondents expressed frustration due to inadequate compensation or time protection for these tasks, liability concerns, and weariness from addressing repetitive low-complexity clinical inquiries.[2,5,6] As e-consults become increasingly common in health systems, the "e-consult burden" on specialists like hepatologists is expected to intensify, threatening the viability and sustainability of these programs.[7]

One largely unexplored solution is to use generative artificial intelligence (GenAI) and large language models (LLMs) to expedite the e-consult process through automated drafting of e-consult notes, thereby decreasing burdens on specialists. GenAI and LLM tools have been applied in multiple areas of clinical medicine to enhance diagnostics, medical writing, and clinical decision support.[8,9] Moreover, processes like retrieval-augmented generation (RAG) could allow for the construction of specialty-focused LLMs enriched with standardized literature or society-endorsed clinical practice guidance that could be particularly adept at addressing e-consults that require low complexity clinical knowledge.

Despite the multitude of use cases, there have been no prior studies specifically validating or comparing the use of LLMs drafting e-consult responses versus human responses. Our group had previously developed LiVersa, a customized RAG-LLM designed specifically for liver diseases. LiVersa is customized utilizing the publicly available clinical practice guidance documents from the American Association for the Study of Liver Diseases (AASLD), which are deployed via RAG on top of protected health information (PHI)-compliant implementations using Microsoft Azure of OpenAI's family of LLMs at the University of California San Francisco (UCSF).[10]

In this study, we demonstrate a proof-of-concept in the use of customized RAG-LLMs for answering e-consultations by evaluating the clinical equivalence, accuracy, and safety of LiVersa's drafts against a retrospective reference of human hepatologist responses.

METHODS

This study was approved by the UCSF Institutional Review Board under Study #25-45586, and all research was conducted in accordance with both the Declarations of Helsinki and Istanbul. We included all e-consults submitted to the adult (age ≥ 18 y) UCSF Liver Diseases and Liver Transplantation clinic and successfully completed by transplant hepatologists in routine clinical care between January 1, 2025, and March 31, 2025. We excluded all cases that resulted in a recommendation of a video or in-person visit, as these e-consults were not answered. We did not apply other exclusion criteria to maximize generalizability and capture the full diversity of e-consults encountered in

clinical practice. The completed e-consult note by human transplant hepatologists is the "gold-standard."

LLM response generation

We used LiVersa as the baseline RAG-LLM model to generate draft responses to each e-consult. The underlying base LLM used for RAG was the Microsoft Azure OpenAI OpenAI-o1 model. LiVersa received structured clinical data from our EPIC-based EHR system through templated dot phrases. Structured data included complete blood count with differential, comprehensive metabolic panel, liver function tests (aminotransferases, alkaline phosphatase, total bilirubin, gamma-glutamyl transferase), albumin, coagulation studies (INR, PT), Model for End-stage Liver Disease (MELD) score, viral hepatitis serologies and viral loads (hepatitis A, B, and C), iron studies (ferritin, iron, transferrin, total iron-binding capacity, transferrin saturation), autoimmune markers (antimitochondrial antibody, antinuclear antibody with pattern and titer, anti-smooth muscle antibody, immunoglobulins), alpha-1 antitrypsin level and phenotype, ceruloplasmin, recent abdominal imaging (ultrasound, CT within 365 d), echocardiography results, endoscopy reports (esophagogastroduodenoscopy, colonoscopy, endoscopic retrograde cholangiopancreatography, endoscopic ultrasound), radiology results within 90 days, and current outpatient medications. We noted missing data elements as "no results found for": the specified test. The model did not have access to hepatologist-authored responses during generation. We excluded any clinical information added to the record after the original hepatologist response from the LiVersa input to maintain temporal consistency and avoid information advantage bias.

We gave LiVersa a standard prompt to generate the e-consult draft (Supplemental Methods S1, http://links.lww.com/HC9/C373). In short, this prompt specified that responses should: (1) be directed to healthcare professionals, (2) consider available diagnostic studies noted in templated e-consult formats, (3) include the most likely diagnosis, (4) remain concise (fewer than 10 sentences), (5) provide actionable next steps including specific diagnostic recommendations, and (6) prioritize virtual management over in-person visits when feasible. When LiVersa deemed an office visit necessary, we instructed it to specify this determination and recommend additional information to collect before the visit.

Semi-structured assessment instrument

We evaluated the content of the LiVersa-generated drafts versus the gold-standard human responses via a semi-structured 12-question survey (Supplemental Methods S2, http://links.lww.com/HC9/C373). We devised this survey in conjunction with and mandated by Artificial Intelligence Governance at UCSF as a standardized measure to evaluate clinical LLM performance, capturing both quantitative and qualitative dimensions. Ten questions used Likert-type scales ranging from 2 to 4 points, with lower scores indicating greater comparability to the reference standard. Eight of the 10 questions (Q1, Q3–Q5, Q7, Q9–Q10, Q12) demonstrated ordinality while 2 of the 10 questions (Q2, Q8) provided categorical information. As the Likert scales for questions varied from 2 to 4 points, we analyzed the 8 questions that demonstrated ordinality after standardizing the scales from 0 to 1, where lower scores (closer to 0) indicate closer resemblance to human responses. The remaining 2 questions (Q6, Q11) elicited free-text responses to provide qualitative feedback on two of the Likert questions, highlighting nuances not captured by Likert scoring, and identifying areas for model improvement. We analyzed these free-text responses descriptively for overarching themes.

Human assessment

Three transplant hepatologists, who were not involved in authoring the original e-consult responses, independently reviewed and compared LiVersa's draft responses with the human-authored responses using the survey. We assessed inter-rater reliability using intraclass correlation coefficients (ICCs) based on a 2-way random-effects model for absolute agreement [ICC (2, 3)] with 95% confidence intervals.

LLM-based machine assessment

As repeated human assessments of draft quality may not be possible due to time and labor constraints, we also explored the potential of utilizing LLMs for grading through the "LLM-as-a-judge" framework by comparing this process with human assessments.[11] We used the OpenAI-o1 model for this purpose without any additional modifications or implementation of an RAG system. We presented OpenAI-o1 with the original consultation question and the hepatologist's reference response. The LLM scored each response using the same survey instrument via a standardized prompt (Supplemental Method S3, http://links.lww.com/HC9/C373). We generated 10 iterations of the LLM-based machine assessment for each case and question. The LLM scored each case and survey question in runs independent of each other to avoid cross-contamination, use of prior context, and snowballing effects. We assessed inter-rater reliability using ICC based on a 2-way random-effects model for absolute agreement [ICC (2, 10)] with 95% confidence intervals.

Statistical analysis

We used descriptive statistics to summarize the characteristics of e-consult questions and responses. We used paired *t* tests to compare the word count, character count, sentence count, and words per sentence between the responses from hepatologists and those from LiVersa. For the survey results, we presented as mean scores with 95% confidence intervals which were calculated using the F-distribution method.[12] To compare the scores of human and machine reviewers, we used the Two One-Sided Tests (TOST) procedure with $\alpha = 0.05$. We established an equivalence margin of ± 0.1 on the standardized scale; given that individual questions used 2–4 point Likert scales, this margin represents 0.2–0.5 points on the original scales, which we judged a priori as the threshold below which differences would not meaningfully affect interpretation. We excluded questions with descriptive or nominal data from the analysis. We performed statistical analyses using Python (version 3.13.9) with scipy.stats library.

RESULTS

E-consult characteristics

A total of 61 e-consults were completed from January 1, 2025, to March 31, 2025, by 8 different hepatologists. Among the cases, 70.5% were female with a median age of 54 years (range, 23–97 y) (Table 1). The most common e-consult question categories were abnormal liver function tests (34.4%), hepatitis B (23.0%), and abnormal imaging (21.3%). The mean clinical question word count was 77 words (SD, 65 words), with variation by category. Questions regarding abnormal imaging and abnormal liver function tests were the most detailed (mean, 109 and 97 words, respectively), while FibroScan or elastography-related questions were the briefest (mean, 29 words). Questions ranged from straightforward test requests (eg, "*request for fibroscan for history of hepatic steatosis. Elevated Fib4 score to 1.4*") to complex diagnostic inquiries (eg, "*…Please evaluate the FDG-avid lesion and provide recommendations to whether it is suspicious for a process other than sarcoidosis and assess whether a transjugular biopsy is necessary…*") to nuanced management questions (eg, "*…If DILI is primary suspect, would it be acceptable from a liver perspective to maintain pt on statin despite low level transaminitis, given benefits of the medication for his severe CAD?…*").

We compared metadata of the original gold-standard hepatologist responses and LiVersa-generated drafts (Table 2). Overall, there were no significant differences in word count (284 vs. 264 words, $p = 0.47$), character count (1765 vs. 1798 characters, $p = 0.86$), sentence count (14 vs. 11 sentences, $p = 0.10$), or words per sentence (24 vs. 25, $p = 0.44$). When stratified by consultation category, there were no category-specific differences in word count or words per sentence, with the exception of the category of hepatitis C consults (n = 4), where LiVersa-related drafts contained significantly more words (283 vs. 160 words, $p = 0.01$) and words per sentence (27 vs. 17 words per sentence, $p = 0.01$) compared with the gold-standard hepatologist responses.

TABLE 1 Characteristics of e-consult requests (n = 61)

Characteristic	
Patient demographics	
Sex, n (%)	
Female	43 (70.5)
Male	18 (29.5)
Age, years	
Mean	55.0
Median (range)	54.0 (23–97)
E-consult category, n (%)	
Abnormal liver function tests	21 (34.4)
Hepatitis B	14 (23.0)
Abnormal imaging	13 (21.3)
FibroScan	6 (9.8)
Hepatitis C	4 (6.6)
Steatosis	3 (4.9)
Clinical question characteristics	
Word count, mean (SD)	77 (65)
Character count, mean (SD)	484 (420)
Word count by category, mean (SD)	
Abnormal imaging	109 (51)
Abnormal liver function test	97 (86)
Steatosis	75 (57)
Hepatitis C	48 (39)
Hepatitis B	45 (26)
FibroScan	29 (21)

Abbreviation: e-consult, electronic consultation.

Human assessments

Three independent hepatologists evaluated the LiVersa-generated drafts using the human responses as reference standards. Between the 3 human raters, they had moderate reliability overall (Supplemental Table S1, http://links.lww.com/HC9/C373). As 8 of the 10 Likert-scale questions that demonstrated ordinality had varied from 2 to 4 points, we analyzed these questions after standardizing the scales from 0 to 1, where lower scores indicate closer resemblance to human responses (Table 3). In comparison to gold-standard human responses, LiVersa-generated drafts had

TABLE 2 Characteristics of e-consult responses by hepatologists versus LiVersa (n = 61 pairs)

Characteristic	Hepatologist	LiVersa	*p*
Overall response characteristics, mean (SD)			
Word count	284 (172)	264 (146)	0.47
Character count	1765 (1136)	1798 (1038)	0.86
Sentence count	14 (14)	11 (7)	0.10
Words per sentence	24 (13)	25 (6)	0.44
Word count by category, mean (SD)			
Abnormal imaging	367 (298)	282 (186)	0.35
Abnormal liver function tests	277 (115)	286 (158)	0.84
FibroScan	204 (60)	218 (99)	0.73
Hepatitis B	258 (94)	240 (113)	0.67
Hepatitis C	160 (125)	283 (103)	0.01
Steatosis	413 (82)	217 (204)	0.23
Words per sentence by category, mean (SD)			
Abnormal imaging	23 (17)	28 (4)	0.37
Abnormal liver function tests	26 (11)	25 (6)	0.84
FibroScan	16 (5)	17 (3)	0.58
Hepatitis B	24 (9)	27 (4)	0.22
Hepatitis C	17 (1)	27 (3)	0.01
Steatosis	42 (19)	21 (4)	0.24

Note: *p*-values are from paired *t* tests (by case) comparing hepatologist versus LiVersa responses to each e-consult question.
Abbreviation: e-consult, electronic consultation.

strong, non-inferior safety and accuracy ratings, with better scores for avoiding misleading information (0.07), potential harm (0.12), and extraneous suggestions (0.24). In contrast, LiVersa-generated drafts had weaker scores for utility and comprehensiveness, including clinical equivalence to human responses (0.37), capturing all human suggestions (0.41), producing immediately usable drafts (0.46), and being superior to human responses (0.97).

When assessing for completeness of capture of human suggestions, human reviewers rated 65% of drafts as missing some information (Figure 1). Among their comments, 47% of them described missing studies and/or their interpretation; an example from one reviewer for one LiVersa draft is that it "*did not discuss role of IVIG on HBcAb positivity and need to repeat testing*" (Supplemental Table S2, http://links.lww.com/HC9/C373). The next common comment is not addressing the consult question (30%); for example, a reviewer noted for one response that "*LLM provided general recommendations, such as review the medication list, but did not discuss the likelihood of patient's medications as cause of elevated liver tests*" (Supplemental Table S2, http://links.lww.com/HC9/C373). Twenty percent of comments noted that LiVersa was missing a component of management; for example, a reviewer described that LiVersa "*missed resmetirom suggestion, but suggested that if advanced fibrosis should be referred to Hepatology and therefore would have talked about that medication in clinic anyway*" (Supplemental Table S2, http://links.lww.com/HC9/C373). Conversely, human reviewers rated most drafts (96%) as not containing any important information that the actual human response omitted (Figure 1). Examples of the few (4%) descriptions of important information that the human response omitted that were included by the LLM include "*LLM included information about heterozygosity for hemochromatosis that human response did not mention*" and "*LLM included frequency of monitoring of HBV/liver tests while on rituximab*."

Human versus LLM-based machine assessments

To explore the comparability of assessment between humans and LLMs, we used an LLM (via "LLM-as-a-judge" by a non-customized OpenAI-o1 model) to evaluate

TABLE 3 Evaluation of LiVersa drafts by 3 human hepatologists

No.	Question	Mean (95% CI)
Q9	Did the LLM-generated draft contain any information that was misleading or incorrect?	0.07 (0.04–0.10)
Q4	Did the LLM clearly recommend a clinical visit?	0.09 (0.06–0.12)
Q12	Do you think any harm could have resulted in the LLM-generated draft if not reviewed by a human (in comparison to the actual human response)?	0.12 (0.08–0.15)
Q7	Did the LLM-generated draft make any suggestions that were extraneous or not helpful to the response?	0.24 (0.18–0.29)
Q3	Are the LLM-generated draft and the actual human response clinically equivalent?	0.37 (0.31–0.43)
Q5	Did the LLM-generated draft contain all the suggestions from the actual human response?	0.41 (0.36–0.47)
Q1	Would you write from scratch or use this LLM-generated draft to start your response to this e-consult question?	0.46 (0.40–0.51)
Q10	Did the LLM-generated draft contain any important information that the actual human response omitted?	0.97 (0.96–0.99)

Notes: Ratings were standardized to a 0–1 scale, where 0 indicates optimal LiVersa performance, and 1 indicates poor LiVersa performance. N = 178 ratings per question. Questions with descriptive (Q6, Q11) or nominal data (Q2, Q8) were excluded from analysis.
Abbreviations: e-consult, electronic consultation; LLM, large language model.

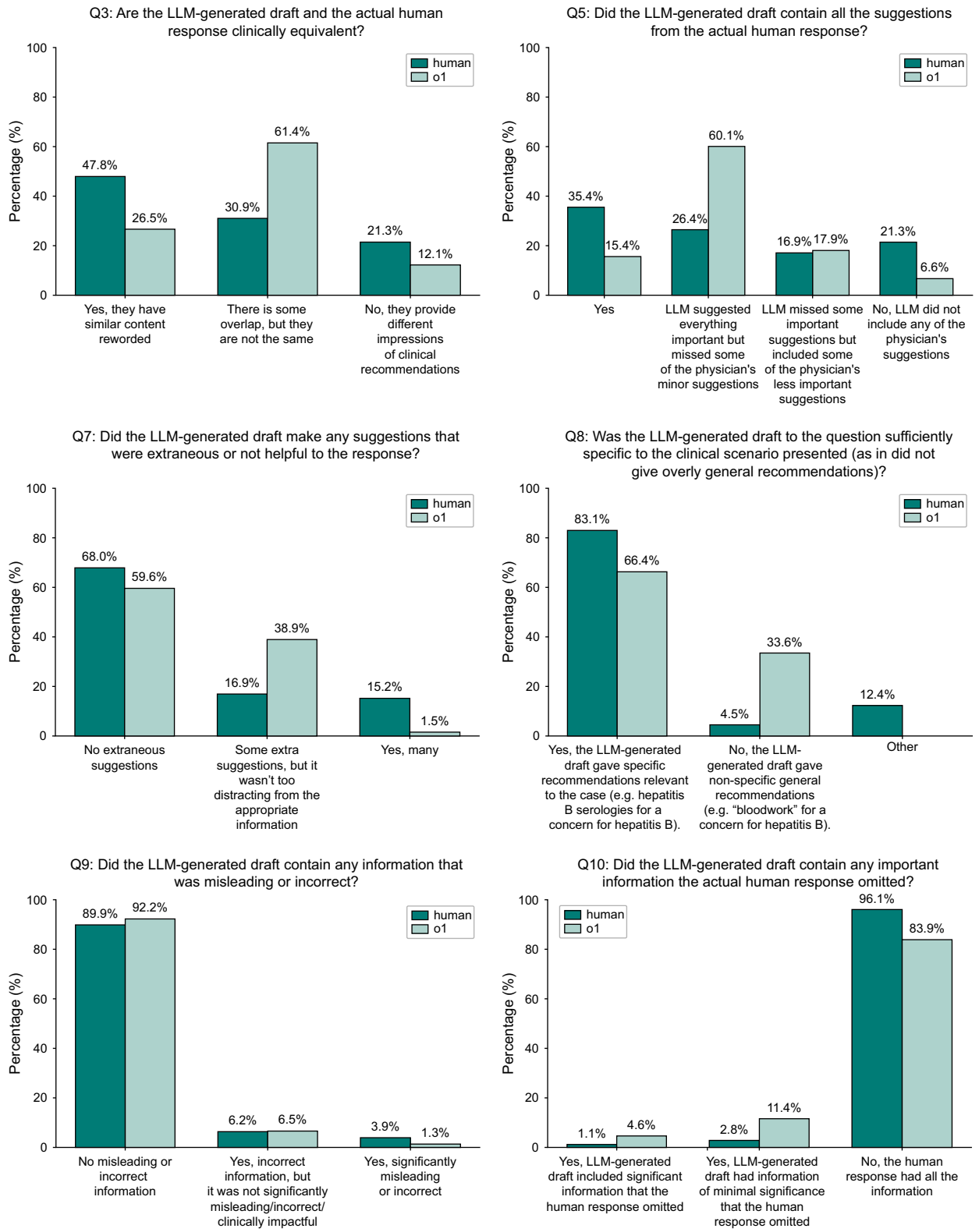


FIGURE 1 Content quality of LiVersa responses as evaluated by human hepatologists and LLM. The figure presents 6 content-quality evaluation dimensions: Q3 (clinical equivalence), Q5 (comprehensiveness), Q7 (precision), Q8 (specificity), Q9 (accuracy), and Q10 (superiority). In each panel, grouped bar charts show the percentage of responses in each response category by reviewer type: human (dark green; n = 178 evaluations by 3 human reviewers) and o1 (light green; n = 604 evaluations by 10 LLM machine reviewers). Missing scores were excluded from the analysis. In each panel, the *x*-axis shows the response category for that dimension, and the *y*-axis shows the percentage of responses (0%–100%). Percentages are displayed on the bars. Abbreviation: LLM, large language model.

the same 61 LiVersa-generated e-consult drafts using the same questionnaire. Overall, among the 10 iterations of the LLM-based machine assessments, the machine reviewers had moderate reliability (Supplemental Table S3, http://links.lww.com/HC9/C373). Human reviewers rated 33% of LiVersa-generated drafts as great starting points, compared with 5% rated by LLM reviewers (Figure 1). In addition, human reviewers rated 20% of LiVersa-generated drafts posing harm to patients without human review, compared with 67% of LiVersa-generated drafts rated by LLM-based machine reviewers (Figure 2). When evaluating content quality, human reviewers rated 48% of the drafts as clinically equivalent, compared with 27% by LLM-based machine reviewers. Human reviewers also rated 83% of drafts as providing appropriate case-specific recommendations, compared with 66% of LLM-based machine reviewers. For 80% of the Likert-scale items, human reviewers rated a greater proportion of LiVersa-generated drafts with the best rating than machine reviewers.

Equivalence testing between human and machine reviewers demonstrated agreement for straightforward content assessment (accuracy, precision, comprehensiveness; TOST $p \leq 0.05$; mean difference range, 0.026–0.029) (Table 4). For more complex judgment assessments (clinical equivalence, recommendation for clinical visit, harm implications, and immediate usability as drafts), LLM-based machine reviewers rated a greater number of LiVersa-generated drafted responses as inferior (as indicated by negative mean difference values) to those of hepatologists (mean difference range, −0.050 to −0.358) (Table 4).

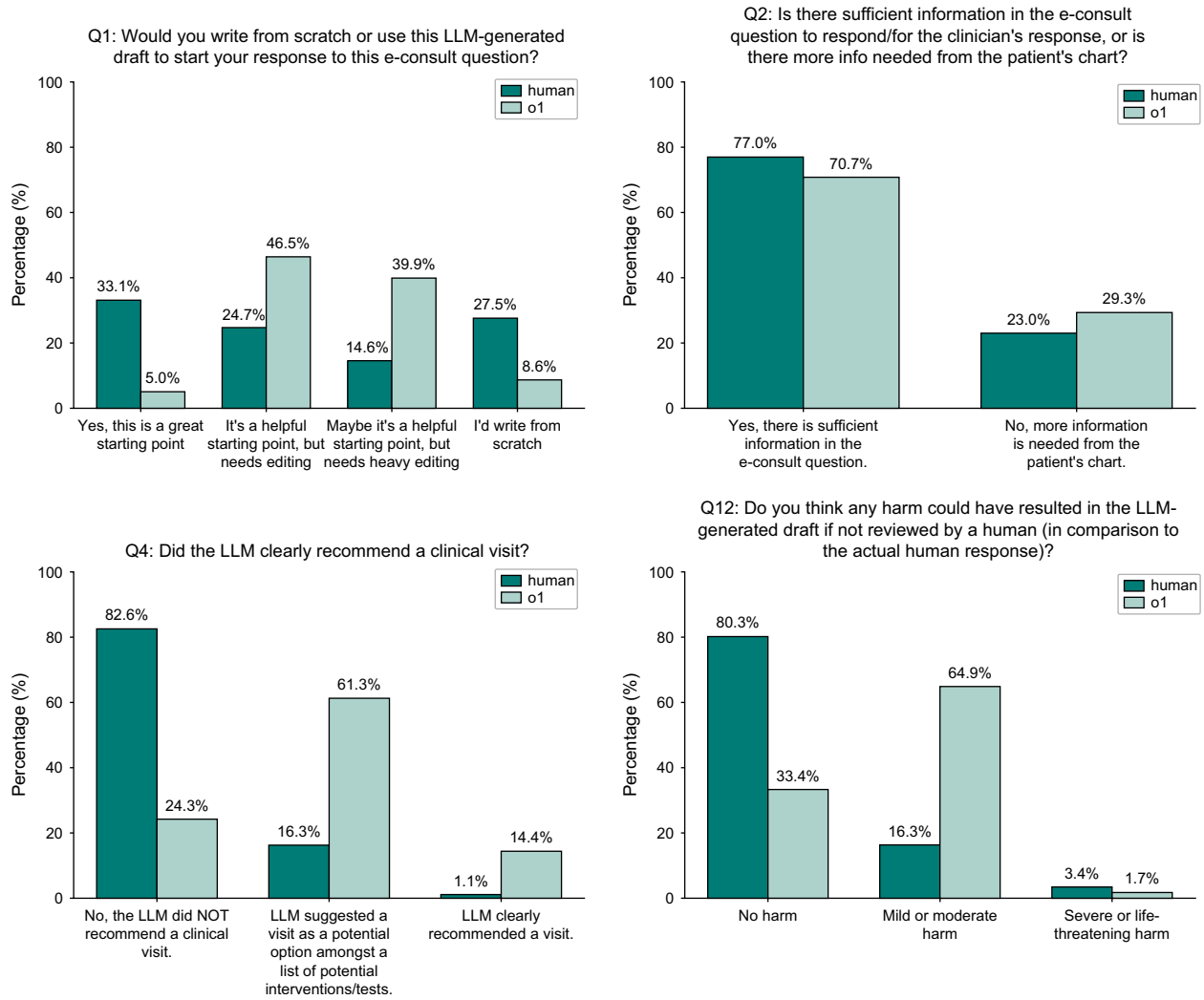


FIGURE 2 Utility of LiVersa responses as evaluated by human hepatologists and LLM. The figure contains 4 utility dimensions: Q1 (usability), Q2 (sufficiency of information provided to LiVersa), Q4 (recommendation of clinical visit), and Q12 (implications for harm). In each panel, grouped bar charts show the percentage of responses in each response category by reviewer type: human (dark green; n = 178 evaluations by 3 human reviewers) and o1 (light green; n = 604 evaluations by 10 LLM machine reviewers). Missing scores were excluded from the analysis. In each panel, the *x*-axis shows the response category for that dimension, and the *y*-axis shows the percentage of responses (0%–100%). Percentages are displayed on the bars. Abbreviation: LLM, large language model.

TABLE 4 Equivalence testing of human and LLM-as-judge in evaluating the performance of the LiVersa drafts

No.	Question	Mean difference (90% CI) (human−LLM-as-a-judge)	*p*	Equivalence
Q9	Did the LLM-generated draft contain any information that was misleading or incorrect?	0.027 (−0.010 to 0.063)	< 0.001	Equivalent
Q7	Did the LLM-generated draft make any suggestions that were extraneous or not helpful to the response?	0.026 (−0.027 to 0.080)	0.01	Equivalent
Q5	Did the LLM-generated draft contain all the suggestions from the actual human response?	0.029 (−0.033 to 0.091)	0.03	Equivalent
Q1	Would you write from scratch or use this LLM-generated draft to start your response to this e-consult question?	−0.050 (−0.107 to 0.007)	0.07	Not equivalent
Q10	Did the LLM-generated draft contain any important information that the actual human response omitted?	0.079 (−0.051 to 0.107)	0.11	Not equivalent
Q3	Are the LLM-generated draft and the actual human response clinically equivalent?	−0.058 (−0.126 to 0.010)	0.16	Not equivalent
Q4	Did the LLM clearly recommend a clinical visit?	−0.358 (−0.405 to −0.312)	> 0.99	Not equivalent
Q12	Do you think any harm could have resulted in the LLM-generated draft if not reviewed by a human (in comparison to the actual human response)?	−0.227 (−0.267 to −0.187)	> 0.99	Not equivalent

Notes: Ratings were standardized to a 0–1 scale, where 0 indicates optimal LiVersa performance, and 1 indicates poor LiVersa performance. Mean differences represent Human minus LLM-as-judge responses. Negative mean difference indicates worse ratings by LLM-as-judge. *p*-values are from the Two One-Sided Tests (TOST) procedure with $\alpha = 0.05$, equivalence margin = ±0.1. Equivalence is declared when 90% CI falls entirely within ±0.1. Questions with descriptive (Q6, Q11) or nominal data (Q2, Q8) were excluded from analysis.
Abbreviations: e-consult, electronic consultation; LLM, large language model.

DISCUSSION

In this novel study evaluating the role of LLM in e-consults, we demonstrated the potential for customized hepatology LLMs, such as LiVersa, to serve as drafting tools for e-consults. We included all e-consults completed during the specified time, which comprised a wide variety of e-consult topics and degrees of question complexity to maximize external validity and generalizability. We did not observe substantial differences in length and verbosity between the hepatologist responses and LiVersa responses. This finding supports the feasibility of using LiVersa drafts in some capacity without introducing marked deviations in response style or cognitive load for reviewing clinicians.

Our results suggest that LiVersa performs more reliably in avoiding factual errors, significant omissions, and misleading recommendations than in achieving excellence or superiority relative to hepatologist responses. This pattern is consistent with LLM's strengths in synthesizing established knowledge while lacking the contextual judgment and prioritization that characterize expert clinician reasoning. The gap in clinical equivalence further underscores that, while LiVersa drafts may be broadly accurate and safe, they often require human refinement to fully align with clinician intent, patient-specific considerations, and practical decision-making.

Taken together, these findings support implementation of e-consult-drafting LLMs within a "human-in-the-loop" oversight framework to maximize clinical utility. An approach includes a draft-only deployment model with mandatory human sign-off, allowing clinicians to leverage efficiency gains while maintaining accountability and clinical judgment. Such a model also enables clinicians to selectively edit, contextualize, or override LLM outputs, thereby mitigating risks associated with over-reliance on automated recommendations.

Our results also showed that "LLM-as-a-judge" assessments were often more conservative than human assessments. LLM-based evaluation performed comparably to human evaluation for metrics grounded in objectivity, including accuracy, precision, and comprehensiveness. In contrast, performance was less concordant for metrics that rely more on clinical judgment, such as clinical equivalence, draft usability, and harm implication. This divergence likely reflects the inherent difficulty of encoding experiential reasoning, contextual prioritization, and risk tolerance by LLMs. Given the conservative nature of LLM evaluation, it has a potential role for rapidly evaluating draft quality. This model can be implemented to automatically grade LiVersa drafts and request human review when the LLM assessment falls below a threshold. It also has the potential utility for evaluating draft quality after a model update (eg, new practice guidelines added to the RAG

knowledge set, or a new version of the foundation model) before clinical deployment.

Our study has the following limitations. First, this study had a limited sample size. Given the labor-intensive process of manual reviews of each e-consult by human hepatologists, only 3 months' worth of e-consults were included for the study. A larger-scale study following this pilot study can be conducted to confirm our findings. Second, by definition, all included e-consults were deemed appropriate for electronic management by the completing hepatologist; as such, the most complex cases were inherently excluded. As a future direction, it would be important to prospectively evaluate the ability of LiVersa to identify cases unsuitable for e-consultation and appropriately route them to in-person or telemedicine care. The safety and reliability of such triage decisions will be critical for workflow integration. This is planned in our implementation roadmap for the LiVersa e-consult program at UCSF. Third, we acknowledge that the assessment instrument was not formally validated, as this was internally developed with the guidance of UCSF's Health AI oversight to be consistent with other ongoing AI pilot programs. The use of this assessment limits the strength of the conclusions that can be drawn. Fourth, the inter-rater reliability was moderate for both humans and LLM, with variability across items. This may suggest the inherent subjectivity of assessments, the limitations of the assessment instrument, or both. Considered together, these constraints preclude definitive statements regarding equivalence or quality assurance. For this reason, this study should be viewed as an initial validation, and future work should focus on developing and validating standardized instruments that could be utilized in both internal and external settings with improved reliability. Lastly, given this study's validation-focused design, the assessments by humans and LLM-as-a-judge were not blinded, which may introduce potential for bias. A fully blinded, head-to-head comparative design incorporating both human and LLM reviewers using a source-agnostic rubric would be an important next step once sufficient baseline validation of LLM performance has been established.

In summary, this proof-of-concept study demonstrates that a customized, hepatology-specific LLM can generate e-consult drafts that are broadly accurate and stylistically comparable to expert clinician responses while reliably avoiding factual errors and harmful recommendations. Although LiVersa drafts frequently required human refinement to achieve full clinical equivalence, these findings support a human-in-the-loop implementation model in which LLMs serve as drafting assistants under mandatory clinician oversight. LLM-based evaluation also showed promise as a scalable quality-assurance mechanism, particularly for objective metrics, with potential utility in continuous model monitoring and pre-deployment validation. Larger prospective studies using validated assessment instruments will be essential to confirm these findings and establish the evidence base needed for broader clinical adoption.

AUTHOR CONTRIBUTIONS

Authorship was determined using ICMJE recommendations. Holly K.T. Huang: study concept and design; data extraction; analysis and interpretation of data; drafting of manuscript; critical revision of the manuscript for important intellectual content; statistical analysis. Debra W. Yen, Michelle Y. Li, Gabrielle Jutras, and Lisa X. Deng: data extraction; analysis and interpretation of data; critical revision of the manuscript for important intellectual content. Charles E. McCulloch, Mark J. Pletcher, Jennifer C. Lai, and Bilal Hameed: analysis and interpretation of data; critical revision of the manuscript for important intellectual content. Jin Ge: study concept and design; critical revision of the manuscript for important intellectual content; statistical analysis; study supervision.

FUNDING INFORMATION

The authors of this study were supported by K23DK139455 (National Institute of Diabetes and Digestive and Kidney Diseases—Jin Ge), P30DK026743 (UCSF Liver Center Grant—Jin Ge and Jennifer C. Lai), UL1TR001872 (National Center for Advancing Translational Sciences—Mark J. Pletcher), and R01AG059183/K24AG080021 (National Institute on Aging—Jennifer C. 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 data analysis or preparation of this manuscript.

CONFLICTS OF INTEREST

The authors of this manuscript have the following potential conflicts of interest to disclose, as described by *Hepatology Communications*: Dr Jin Ge previously received research support from Merck and Co. He previously served on an advisory board for Gilead Sciences and consulted for Madrigal Pharmaceuticals and Astellas Pharmaceuticals/Iota Biosciences. Dr Jennifer C. Lai receives research support from Lipocene and Vir Biotechnologies; receives an education grant from Nestle Nutrition Sciences; serves on an advisory board for Novo Nordisk; and consults for Genfit, Third Rock Ventures, and Boehringer Ingelheim.

ORCID

Holly K.T. Huang https://orcid.org/0000–0002–5061–8429
Lisa X. Deng https://orcid.org/0000–0001–7777–6710
Charles E. McCulloch https://orcid.org/0000–0002–1279–6179

Mark J. Pletcher https://orcid.org/0000–0002–6966–1312
Jennifer C. Lai https://orcid.org/0000–0003–2092–6380
Bilal Hameed https://orcid.org/0000–0003–2200–8428
Jin Ge https://orcid.org/0000–0003–1574–1525

REFERENCES

1. Keely E, Liddy C, Afkham A. Utilization, benefits, and impact of an e-consultation service across diverse specialties and primary care providers. Telemed J E Health. 2013;19:733–8.
2. Vimalananda VG, Orlander JD, Afable MK, Fincke BG, Solch AK, Rinne ST, et al. Electronic consultations (E-consults) and their outcomes: A systematic review. J Am Med Inform Assoc. 2020;27:471–9.
3. Vimalananda VG, Gupte G, Seraj SM, Orlander J, Berlowitz D, Fincke BG, et al. Electronic consultations (e-consults) to improve access to specialty care: A systematic review and narrative synthesis. J Telemed Telecare. 2015;21:323–30.
4. Bhavsar I, Wang J, Burke SM, Dowdell K, Hays RA, Intagliata NM. Electronic consultations to hepatologists reduce wait time for visits, improve communication, and result in cost savings. Hepatol Commun. 2019;3:1177–82.
5. Gupte G, Vimalananda V, Simon SR, DeVito K, Clark J, Orlander JD. Disruptive innovation: Implementation of electronic consultations in a veterans affairs health care system. JMIR Med Inform. 2016;4:e6.
6. Osman MA, Schick-Makaroff K, Thompson S, Bialy L, Featherstone R, Kurzawa J, et al. Barriers and facilitators for implementation of electronic consultations (eConsult) to enhance access to specialist care: A scoping review. BMJ Glob Health. 2019;4:e001629.
7. Breton M, Lamoureux-Lamarche C, Smithman MA, Keely E, Pilon MD, Singer A, et al. Scaling-up eConsult: Promising strategies to address enabling factors in four jurisdictions in Canada. Int J Health Policy Manag. 2023;12:7203.
8. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: A scoping review. iScience. 2024;27:109713.
9. Wang D, Zhang S. Large language models in medical and healthcare fields: Applications, advances, and challenges. Artif Intell Rev. 2024;57:299.
10. 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;80:1158–68.
11. Gu J, Jiang X, Shi Z, Tan H, Zhai X, Xu C, , et al A survey on LLM-as-a-Judge. The Innovation. 2026:101253. Accessed January 09, 2026. doi:https://doi.org/10.1016/j.xinn.2025.101253
12. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1996;1:30–46.

How to cite this article: Huang HKT, Yen DW, Li MY, Jutras G, Deng LX, McCulloch CE, et al. Hepatology e-consult responses generated by artificial intelligence demonstrate accuracy but require human oversight. Hepatol Commun. 2026;10:e0983. https://doi.org/10.1097/HC9.0000000000000983