İç Hastalıkları Dergisi
Turkish Archives of Internal Medicine

RESEARCH ARTICLE/ARAŞTIRMA MAKALESİ

İç Hastalıkları Dergisi 2025;26(2):29-37 • **Doi:** 10.5578/ichastaliklari.2025027

Performance Evaluation of Gastroenterology Specific Language Model GastroGPT on Hepatocellular Carcinoma Management

Gastroenteroloji Özel Dil Modeli GastroGPT'nin Hepatoselüler Karsinom Yönetimindeki Performans Değerlendirmesi

Cem Şimşek[1](İD), Mete Üçdal[2](İD), Yiğit Yazarkan[3](İD), Muhammed Bahaddin Durak[1](İD), Hatice Yasemin Balaban[1](İD)

[1] Department of Gastroenterology, Hacettepe University Faculty of Medicine, Ankara, Türkiye

[2] Department of Internal Medicine, Hacettepe University Faculty of Medicine, Ankara, Türkiye

[3] Hacettepe University Faculty of Medicine, Ankara, Türkiye

Cite this article as: Şimşek C, Üçdal M, Yazarkan Y, Durak MB, Balaban HY. Performance evaluation of gastroenterology specific language model GastroGPT on hepatocellular carcinoma management. İç Hastalıkları Dergisi 2025;26(2):29-37.

ABSTRACT

Introduction: *This study aimed to evaluate the performance of GastroGPT, a gastroenterology-specific large language model (LLM), compared to ChatGPT and expert physician notes using real-life hepatocellular carcinoma (HCC) cases.*

Materials and Methods: *Ten cases similar to HCC patients in the MIMIC-III dataset were constructed and evaluated using GastroGPT, ChatGPT, and expert notes. Model outputs were blindly scored using an evaluation system covering seven clinical domains.*

Results: *GastroGPT scored higher than ChatGPT across seven clinical domains: Assessment and Summary Evaluation (8.80 ± 0.84 vs. 6.31 ± 1.69), Additional Questions and History Taking (8.56 ± 0.82 vs. 3.54 ± 2.44, p< 0.05), Diagnostic Evaluation (9.09 ± 0.97 vs. 5.11 ± 3.04, p< 0.05), Treatment Plan (7.36 ± 2.75 vs. 3.90 ± 3.60, p< 0.05), Follow-Up Plan (8.31 ± 1.70 vs. 4.75 ± 1.36, p< 0.05), Referral and Consultation (8.80 ± 0.85 vs. 5.40 ± 1.69, p< 0.05), and Patient Education (7.76 ± 1.85 vs. 2.89 ± 1.81, p< 0.05). GastroGPT also demonstrated higher scores compared to the attending physician in all evaluated domains: Assessment and Summary Evaluation (8.80 ± 0.84 vs. 7.50 ± 0.81, p< 0.05), Diagnostic Evaluation (9.09 ± 0.97 vs. 7.00 ± 2.01, p< 0.05), Treatment Plan (7.36 ± 2.75 vs. 7.15 ± 2.08, p< 0.05), Follow-Up Plan (8.31 ± 1.70 vs. 6.83 ± 0.92), and Referral and Consultation (8.80 ± 0.85 vs. 4.58 ± 1.25, p< 0.05). In the overall rubric, GastroGPT scored 8.73, while ChatGPT scored 4.69.*

Address for Correspondence/Yazışma Adresi

Cem Şimşek

Department of Gastroenterology, Hacettepe University Faculty of Medicine, Ankara, Türkiye

e-mail: cemgsimsek@gmail.com

Received: 16.04.2025
Accepted: 18.06.2025
Available Online Date: 26.06.2025

Data Sharing Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conclusion: *This study demonstrates the feasibility and potential of specialized clinical LLMs in managing HCC. GastroGPT exhibited superior scores compared to ChatGPT and expert notes. Future research should focus on validating the model with real clinical data and assessing its performance across various medical fields.*

Key Words: *Hepatocellular cancer, artificial intelligence, large language models*

ÖZ

Giriş: *Bu çalışmada, gerçek hepatosellüler karsinom (HCC) vakaları kullanılarak, gastroenterolojiye özgü bir büyük dil modeli (LLM) olan GastroGPT'nin performansının, ChatGPT ve uzman hekim notlarıyla karşılaştırılarak değerlendirilmesi amaçlanmıştır.*

Materyal ve Metod: *MIMIC-III veri setindeki HCC hastalarına benzer on vaka oluşturulmuş ve GastroGPT, ChatGPT ve uzman notları kullanılarak değerlendirilmiştir. Model çıktıları, yedi klinik alanı kapsayan bir değerlendirme sistemi kullanılıp kör olarak puanlanmıştır.*

Bulgular: *GastroGPT, yedi klinik alanda ChatGPT'den daha yüksek puan aldı: Değerlendirme ve Özet Değerlendirme (8.80 ± 0.84'e karşı 6.31 ± 1.69), Ek Sorular ve Anamnez Alma (8.56 ± 0.82'ye karşı 3.54 ± 2.44, p< 0.05), Tanı Değerlendirmesi (9.09 ± 0.97'ye karşı 5.11 ± 3.04, p< 0.05), Tedavi Planı (7.36 ± 2.75'e karşı 3.90 ± 3.60, p< 0.05), Takip Planı (8.31 ± 1.70'e karşı 4.75 ± 1.36, p< 0.05), Sevk ve Konsültasyon (8.80 ± 0.85'e karşı 5.40 ± 1.69, p< 0.05) ve Hasta Eğitimi (7.76 ± 1.85'e karşı 2.89 ± 1.81, p< 0.05). GastroGPT, değerlendirilen tüm alanlarda da ilgili hekime kıyasla daha yüksek puanlar almıştır: Değerlendirme ve Özet Değerlendirme (8.80 ± 0.84'e karşı 7.50 ± 0.81, p< 0.05), Tanı Değerlendirme (9.09 ± 0.97'ye karşı 7.00 ± 2.01, p< 0.05), Tedavi Planı (7.36 ± 2.75'e karşı 7.15 ± 2.08, p< 0.05), Takip Planı (8.31 ± 1.70'e karşı 6.83 ± 0.92) ve Sevk ve Konsültasyon (8.80 ± 0.85'e karşı 4.58 ± 1.25, p< 0.05). Genel değerlendirmede GastroGPT 8.73 puan alırken, ChatGPT 4.69 puan almıştır.*

Sonuç: *Bu çalışma, HCC yönetiminde özel klinik LLM'lerin uygulanabilirliğini ve potansiyelini göstermektedir. GastroGPT, ChatGPT ve uzman notlarına kıyasla üstün puanlar almıştır. Gelecekteki araştırmalar, modeli gerçek klinik verilerle doğrulamaya ve çeşitli tıbbi alanlarda performansını değerlendirmeye odaklanmalıdır.*

Anahtar Kelimeler: *Hepatoselüler karsinom, yapay zeka, büyük dil modelleri*

INTRODUCTION

Hepatocellular carcinoma (HCC) ranks as the sixth most prevalent malignancy and the third principal contributor to cancer-related mortality on a global scale (1). Furthermore, prognostic trends indicate a projected rise in both mortality and incidence rates over the forthcoming decade (2). The treatment landscape for HCC is evolving and requires a multidisciplinary approach that includes surgery, transplantation, locoregional therapies, and systemic treatments, and a personalized plan should be devised for every patient by a multidisciplinary team, considering liver function, patients functional status, comorbidities, cancer characteristics, and patient preferences (3,4). However, this necessary multidisciplinary expertise is limited to centers with high volume, at the end restricting access to care for HCC patients, and solutions to standardize and democratize care is required.

In the recent years, artificial intelligence (AI) is transforming various aspects of medicine and providing solutions to long-standing problems in diagnostics, drug discovery, risk analyses and other domains of healthcare (5). Among AI models, Large Language Models (LLMs) are being used more frequently with new advancements in the field (6). Trained on extensive and various datasets, LLMs can retrieve information and produce human-like text, making them potentially valuable for clinical applications. However, current general-purpose LLMs, such as ChatGPT face limitations, including inconsistency, unreliability, and the risk of generating misleading or incorrect information (7,8). If used, LLMs can be useful in virtually every setting of patient care, but these limitations impede their direct use in clinical settings. To effectively integrate LLMs into patient care, one solution may be to develop specialized models tailored to specific clinical domains, ensuring enhanced reliability, consistency, and accuracy (9).

One effort of this specialized clinical LLMs is GastroGPT. GastroGPT is a deployable, gastroenterology specific AI model trained and tuned for clinical tasks using up-to-date evidence-based documents. GastroGPT showed potential benefit when used in more specific clinical tasks. As such in this study, our objective was to compare the performance of GastroGPT with that of general knowledge LLMs and human experts using real-life patient cases. To ensure patient privacy and adhere to ethical standards, we employed an open, publicly available, anonymized dataset. By evaluating GastroGPT's performance in a simulated real-world environment, we aimed to assess its feasibility and

utility in clinical practice. We also aimed to identify the level of expertise needed to create effective clinical task models for improving patient care. We believed that GastroGPT would perform better than general knowledge language models and produce text similar in quality to that of human experts. The results of this study could help in the development and use of specialized clinical language models, ultimately improving patient care and outcomes in managing HCC and other medical conditions.

MATERIALS and METHODS

Study Design and Patient Data

This study employed a blinded, controlled experimental design to evaluate the feasibility, performance and potential utility of a domain-specific AI system, GastroGPT, in the context of managing patients with HCC. The study aimed to compare the outputs of GastroGPT, with general-purpose language model (ChatGPT), and attending physician notes to assess the feasibility and benefits of using a specialized clinical AI system in gastroenterology.

ChatGPT, a widely adopted and state-of-the-art general-purpose language model, was selected to compare the performance differences between domain-specific and general AI systems in managing HCC patients. Although ChatGPT has not been specifically designed for clinical applications, it has demonstrated capabilities in various clinical tasks, including question answering, text summarization, and content generation (10). The ChatGPT model used in this study was GPT-4.0.

This study was based on the Medical Information Mart for Intensive Care III (MIMIC-III) database, which contains de-identified and comprehensive clinical patient data from hospital admissions at Beth Israel Deaconess Medical Center in Boston, Massachusetts (11). To identify relevant cases, a screening process was conducted on over 18.000 patient records using Medical Subject Headings keywords. This screening process categorized patients based on the presence of HCC and various gastroenterology subspecialties, including hepatology, oncology, and emergency admissions. Based on this categorized dataset, ten HCC-like cases were constructed rather than directly selecting patients, ensuring alignment with real-world clinical scenarios. A representative example of GastroGPT's structured recommendation alongside the original physician note is provided in Supplementary Table S1 to illustrate key differences in clinical phrasing and rubric grading.

Gastroenterology Specific Large Language Model

GastroGPT is an AI system designed specifically for clinical tasks within the field of gastroenterology. The model's architecture is modular, which allows for potential future enhancements such as incorporating imaging data and specializing in gastrointestinal subspecialties. GastroGPT is built upon a transformer-based neural network architecture with attention mechanisms, aiming to generate clinically relevant language related to gastroenterology. The model employs an encoder-decoder structure and multi-headed self-attention to handle complex clinical notes. During development, GastroGPT underwent an iterative process with expert-curated data with frequent performance evaluations to align with model with clinical needs and optimize content. Notably, the model was trained solely on expert-curated data, omitting patient cases or clinical data to isolate the effects of its design and enhance transparency. Current evidence suggests GastroGPT can be more successful than the available general LLMs and can be comparable to human notes in real life patient cases (12-14).

GastroGPT was developed by fine-tuning OpenAI's GPT-3.5-turbo base model using a corpus of approximately 500.000 de-identified, expert-curated gastroenterology resources, including clinical guidelines, academic literature, radiology/pathology reports, and synthetic cases. No identifiable patient data were used during training

Evaluation

The study involved providing LLMs with case presentation notes and comparing their outputs to the corresponding attending notes for each case. To facilitate a comprehensive evaluation, an extensive rubric was developed by experts, covering tasks in seven domains that mirror the clinical workflow: Assessment and Summary, Additional History Gathering, Diagnostic Studies, Proposed Management, Follow-up Recommendations, Referral Guidance, and Patient Counseling and Communication. The model outputs were evaluated in a blinded manner. A systematic scoring system was developed to evaluate the quality of model outputs across seven clinical task domains, using a scale ranging from 1 to 10. This rubric considered aspects such as comprehensiveness, relevance, evidence support, safety, feasibility, and patient focus for each task. Each domain was assigned a weight factor to compute the comprehensive score based on its relevance in patient care (20%, 15%,

15%, 20%, 5%, 10%, and 20%, respectively). Each reviewer individually assessed all model outputs for each case and provided qualitative feedback when necessary (15).

Statistical Analysis

All statistical analyses were performed using R (version 4.1.0) and Python (version 3.9.5) software. The normality of score distributions was assessed using the Shapiro-Wilk test. Descriptive statistics, including means, standard deviations, and percentages, were used to summarize reviewer ratings. Interrater reliability was evaluated using the intraclass correlation coefficient (ICC) with a two-way random-effects model and absolute agreement.

The primary outcome was the overall weighted performance, calculated as the average of scores across all clinical tasks. Secondary outcomes included individual task performances and consistency across case complexity. Due to the lack of relevant responses in the attending notes for Additional History Gathering and Patient Counseling and Communication, these sections were excluded from the analyses.

Paired t-tests were employed to compare the performance of GastroGPT with general language models for each clinical task. Analysis of variance (ANOVA) was used to assess overall performance differences between the models. Subgroup analyses were conducted to investigate potential interactions between model scores and case characteristics, such as complexity, clinical presentation, and disorder frequency, using ANOVA and linear regression techniques. Levene's test was used to compare score variance between models.

Multiple comparisons were addressed using Bonferroni correction to control the familywise error rate. Statistical significance was set at a two-tailed p-value <0.05. Effect sizes were reported using Cohen's d for pairwise comparisons and partial eta-squared (η^2p) for ANOVA. Confidence intervals (95% CI) were provided for all point estimates.

Qualitative feedback from reviewers was analyzed using an inductive thematic analysis approach. Two researchers independently coded the responses and identified emerging themes. Discrepancies in coding were resolved through discussion until consensus was reached. The frequency of each theme was reported, and representative quotes were selected to illustrate key findings. Likert scale responses were compared between models using the Mann-Whitney U test, with response categories coded as numerical scores (1-5). The null hypothesis assumed no difference in the responses generated by the two models and attending notes.

RESULTS

Based on the MIMIC-III dataset, a total of 10 HCC-like cases were constructed for this proof-of-concept study instead of directly selecting patients. Detailed information regarding the ages, genders, case complexity, chief complaints, and chronic diseases of the constructed cases is provided in Table 1. The word counts of the patient document as well as the word counts of the LLM output are displayed in Table 2.

Across seven clinical domains, GastroGPT scored higher than ChatGPT: Assessment and Summary Evaluation (8.80 ± 0.84 vs. 6.31 ± 1.69), Additional Questions and History Taking (8.56 ± 0.82 vs. 3.54 ± 2.44 <0.05), Diagnostic Evaluation (9.09 ± 0.97 vs. 5.11 ± 3.04 <0.05), Treatment Plan (7.36 ± 2.75 vs. 3.90 ± 3.60 <0.05), Follow-Up Plan (8.31 ± 1.70 vs. 4.75 ± 1.36 <0.05), Referral and Consultation (8.80 ± 0.85 vs. 5.40 ± 1.69 <0.05), and Patient Education (7.76 ± 1.85 vs. 2.89 ± 1.81 <0.05). Likewise, GastroGPT demonstrated higher evaluation scores across all evaluated domains compared to the attending physician: Assessment and Summary Evaluation (8.80 ± 0.84 vs. 7.50 ± 0.81 <0.05), Diagnostic Evaluation (9.09 ± 0.97 vs. 7.00 ± 2.01 <0.05), Treatment Plan (7.36 ± 2.75 vs. 7.15 ± 2.08 <0.05), Follow-Up Plan (8.31 ± 1.70 vs. 6.83 ± 0.92), and Referral and Consultation (8.80 ± 0.85 vs. 4.58 ± 1.25 <0.05). When comparing ChatGPT with

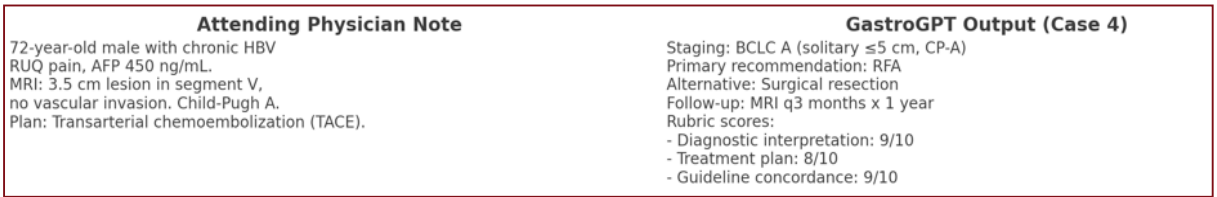

Figure 1. Comparison between the attending physician note and GastroGPT output for case 4.

Table 1. Age, sex, complexity, chief complaint, and chronic diseases of the cases

	Age	Sex	Complexity	Chief Complaint	Chronic Diseases
1.	61	M	Middle	Fevers/cholangitis	Alcoholic cirrhosis, hepatocellular carcinoma (HCC), delayed hepatic arterial thrombosis, ischemic cholangiopathy, bile lakes
2.	84	M	High	Falls and hip pain	NASH cirrhosis, hepatocellular carcinoma (HCC), portal vein thrombosis, grade II varices, hepatic encephalopathy, refractory ascites, spontaneous bacterial peritonitis (SBP), Clostridium difficile infection (CDI), arteriovenous malformations (AVMs) in the cecum, coronary artery disease (CAD), benign prostatic hyperplasia (BPH), type 2 diabetes mellitus (DM2), hypothyroidism, anemia, gastric cancer
3.	75	M	Low	Coffee ground emesis	Hepatocellular carcinoma (HCC), hepatitis B cirrhosis, varices
4.	64	M	Middle	IVC thrombus extending into the right atrium	Hepatitis C (hep C), hepatocellular carcinoma (HCC), hypertension (HTN)
5.	66	F	Middle	Chest pain	Cirrhosis secondary to hepatitis C virus (HCV), hepatocellular carcinoma (HCC), history of TIPS procedure
6.	71	M	Low	Worsening weakness and fatigue	Multifocal hepatocellular carcinoma (HCC), cirrhosis of unclear etiology
7.	66	M	Low	Gastrointestinal bleeding	Hepatitis B cirrhosis, hepatocellular carcinoma (HCC)
8.	58	M	High	Mild abdominal pain and spontaneous bacterial peritonitis (SBP)	End-stage liver disease (ESLD), hepatitis C (HepC), hepatocellular carcinoma (HCC), encephalopathy, esophageal varices, ascites refractory to paracentesis, failure to thrive (FTT), hepatorenal syndrome
9.	69	F	High	Slowly progressive weakness, acute renal failure, hyponatremia, hypotension	Metastatic hepatocellular carcinoma (HCC), nonalcoholic steatohepatitis (NASH) with cirrhosis, type 2 diabetes, hypertension, hidradenitis, arthritis, history of hip and knee replacements, history of D&C, smoking history
10.	77	M	Middle	Shortness of breath (SOB)	Alcoholic cirrhosis, hepatocellular carcinoma (HCC) status post-ablation, congestive heart failure (CHF), chronic atrial fibrillation (afib), portal vein thrombus, chronic obstructive pulmonary disease (COPD)

attending notes, the attending notes achieved higher scores in Assessment and Summary Evaluation (7.50 ± 0.81 vs. 6.31 ± 1.69 <0.05), Diagnostic Evaluation (7.00 ± 2.01 vs. 5.11 ± 3.04 <0.05), Treatment Plan (7.15 ± 2.08 vs. 3.90 ± 3.60 <0.05), and Follow-Up Plan domains (6.83 ± 0.92 vs. 4.75 ± 1.36 <0.05). However, ChatGPT attained a higher score in the Referral and Consultation domain (5.40 ± 1.69 vs. 4.58 ± 1.25 <0.05). In the overall rubric, GastroGPT outperformed ChatGPT with a score of 8.73, whereas ChatGPT obtained a lower score of 4.69. The mean scores and overall scores of the models and attending notes are presented in Table 3.

Statistical analysis results indicate significant differences in scores among GastroGPT, ChatGPT, and Attending Notes. GastroGPT and ChatGPT resulted in a statistic of 145 and a p-value of 0.01, indicating a significant difference in scores favoring GastroGPT. Similarly, comparing GastroGPT and Attending Notes showed a statistic of 130 and a p-value of 0.02, again indicating a significant difference favoring GastroGPT. However, the comparison between ChatGPT and Attending Notes yielded a statistic of 170 and a p-value of 0.15, which is not statistically significant.

Table 2. Word counts of model outputs

Input	GastroGPT	ChatGPT	Attending Notes
1623	222	477	435
1199	150	395	475
636	91	365	395
618	111	406	143
895	261	426	385
851	309	424	631
1154	241	376	477
833	237	378	210
989	180	385	534
873	261	403	254

Table 3. Gastrogpt, Chatgpt, and attending notes scores across various clinical domains, including assessment and summary, additional questions and history taking, diagnostic evaluation, treatment plan, follow-up plan, referral and consultation, and patient education

	Assessment and Summary Evaluation (20% weight)	Additional Questions and History Taking (15% weight)	Diagnostic Evaluation (15% weight)	Treatment Plan (20% weight)	Follow Up Plan (5% weight)	Referral and Consultation (10% weight)	Patient Education (20% weight)	Total Score
GastroGPT	8.80 ± 0.84	8.56 ± 0.82	9.09 ± 0.97	7.36 ± 2.75	8.31 ± 1.70	8.80 ± 0.85	7.76 ± 1.85	8.73
ChatGPT	6.31 ± 1.69	3.54 ± 2.44	5.11 ± 3.04	3.90 ± 3.60	4.75 ± 1.36	5.40 ± 1.69	2.89 ± 1.81	4.69
Attending Notes	7.50 ± 0.81	6.00 ± 1.50	7.00 ± 2.01	7.15 ± 2.08	6.83 ± 0.92	4.58 ± 1.25	5.00 ± 1.75	6.50

Pearson's correlation coefficient analysis was conducted to assess the agreement between the ratings provided by Rater 1 and Rater 2. The analysis revealed a strong positive correlation between the ratings ($r= 0.75$, $p< 0.05$), indicating statistically significant agreement. Levene's test confirmed significant differences in variances among the scores (statistic= 5.997, p= 0.007).

DISCUSSION

In this study, we evaluated the feasibility of a specialized clinical language model in managing HCC cases of varying complexities by comparing it with general-purpose LLMs and attending notes. We focused on criteria that simulated the comprehensive clinical care of these patients. Each criterion was assigned a specific weight to comprehensively assess the capabilities of the LLMs and attending notes. GastroGPT scored higher than ChatGPT in seven clinical domains: assessment and summary, additional questions and history, diagnostic evaluation, treatment plan, follow-up plan, referral and consultation, and patient education. GastroGPT also scored higher the attending physician in all evaluated domains. ChatGPT scored higher than the attending notes in the referral and consultation domain, while the attending notes' scores were higher in all other domains. Nevertheless, these findings underscore the potential of AI-driven tools to standardize HCC therapeutic decisions across varied clinical settings. We acknowledge that constructing ten HCC-like scenarios from the MIMIC-III intensive-care dataset-rather than enrolling consecutive, real-world patients with intermediate (BCLC B) and advanced (BCLC C) disease-inevitably limits the external validity of our proof-of-concept model. Management of BCLC B and C HCC exhibits considerable heterogeneity, encompassing diverse downstaging protocols (e.g., transarterial chemoembolization, radioembolization), expanded transplant eligibility frameworks (UCSF, up-to-seven criteria), and varying surgical resection thresholds

across Eastern and Western high-volume centers. Moreover, region-specific use of systemic agents (sorafenib, lenvatinib, atezolizumab-bevacizumab) and absence of dynamic tumor biomarkers (AFP kinetics, molecular subtypes) in our simulated cohort further constrain generalizability. Prospective, multicenter investigations that enroll unselected BCLC B/C populations, employ standardized eCRF-based data capture and align with Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) reporting guidelines, are therefore essential to validate and extend these preliminary findings (16).

There has been previous efforts on utilizing AI technology in care of HCC patients, with most models focusing on liver metastasis detection, tumor classification, differentiation of liver conditions, HCC nuclei grading, nucleus segmentation, diagnosis and prognosis after treatment (17-20). The comprehensive review can be found in other resources (21,22). Yasaka et al. utilized a convolutional neural network (CNN) to classify liver masses into five categories using dynamic contrast-enhanced CT images (23). From the prognosis side, the recently developed and validated a machine learning model demonstrated the model's superior accuracy in predicting the risk of post-liver transplant hepatocellular carcinoma recurrence compared to other conventional scores (24). Another study constructed a traditional AI model, a gradient boosting machine learning algorithm, predicting patient survival at six time points and included 100 HCC patients (80% male) with a median overall survival of 43 months (range: 0.7-256 months). Survival rates at six, 12, 24, 36, 60, and 120 months were 88%, 81%, 67%, 60%, 40%, and 11%, respectively. The model's mean area under the curve (AUC) was 0.92 for survival beyond six months, 0.81 for beyond one year, 0.78 for beyond two years, 0.81 for beyond three years, 0.82 for beyond five years, 0.81 for beyond eight years, and 0.66 for beyond 10 years. The machine learning model effectively predicted both short- and long-term survival for HCC patients (25).

More recently, LLMs have been studied in HCC, as in many fields of medicine. However, the lack of standardized methods for evaluating LLM performance has led most research to rely on multiple-choice questions and pre-made standardized tests typically used in board examinations (26-28). In gastroenterology, a study found that ChatGPT achieved passing scores on the 2021 and 2022 American College of Gastroenterology Self-Assessment (ACG-SA) exams when provided with an engineered prompt to optimize its reasoning approach. ChatGPT scored an average performance comparable to that of human test-takers (10). However, an earlier study using minimal prompt instructions found that ChatGPT failed to pass these exams, which indicatesthat prompt optimization is crucial for maximizing the clinical reasoning capabilities of general LLMs (29).

Yeo et al. assessed the accuracy and reproducibility of ChatGPT's responses to 164 questions regarding cirrhosis and HCC. Two transplant hepatologists independently graded the responses, with a third reviewer resolving any discrepancies. ChatGPT accurately answered 79.1% of questions about cirrhosis and 74.0% of questions about HCC; however, only 47.3% and 41.1% of the responses, respectively, were deemed comprehensive. The model successfully provided basic knowledge, lifestyle advice, and treatment information but was less effective in addressing diagnosis and preventive medicine. Despite lacking specifics on regional guidelines, ChatGPT offered practical advice on next steps and adapting to a new diagnosis, indicating its potential as a supplementary informational tool for both patients and physicians (30). In another study, Ge et al. developed a liver-specific LLM utilizing retrieval-augmented generation (RAG) alongside the University of California – San Francisco's PHI-compliant text embedding platform. It demonstrated proficiency in providing binary responses to all 10 questions posed. However, detailed justifications were inaccurate for three questions. While their model exhibited promise in addressing clinical hepatology inquiries with specificity, limitations in the diversity of documents used for RAG contributed to some knowledge deficiencies. Despite these constraints, their model serves as a proof of concept for harnessing RAG to tailor LLMs for clinical contexts, thereby paving the path towards personalized medicine in the future (31).

CONCLUSION

Our investigation demonstrated numerous advantages, which encompass a meticulously controlled and replicable direct assessment between GastroGPT, ChatGPT and Attending Notes, alongside the employment of diverse simulated scenarios that reflect actual HCC clinical care. Nevertheless, we did not utilize real clinical data for validation, which may

impact the generalizability of the model's performance to real-world settings. Additionally, the study's reliance on expert feedback introduces a degree of subjectivity, which could affect the reliability of the results. Finally, the scope of the dataset, limited to ten patient cases, restricts the comprehensiveness of the evaluation.

Future research should focus on validating the model with real clinical data, minimizing biases, and assessing its performance across various medical fields to establish broader applicability. In this regard, the focus should be on comparing the efficacy of AI models with that of human specialists, enhancing GastroGPT by augmenting its training datasets, and assessing its applicability across diverse medical and surgical disciplines to ascertain its generalizability.

ETHICS COMMITTEE APPROVAL

Ethics committee approval is not required for this study.

CONFLICT of INTEREST

No conflict of interest declared.

FUNDING SOURCES

The author(s) received no financial support for the research, authorship, and/or publication of this article.

AUTHORSHIP CONTRIBUTIONS

Concept and Design: CŞ, MÜ

Analysis/Interpretation: MÜ, CŞ

Data Collection or Processing: YY, MBD

Writing: MÜ, CŞ

Review and Correction: HYB

Final Approval: All of authors

REFERENCES

1. *Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71(3):209-49. https://doi.org/10.3322/caac.21660*
2. *Petrick JL, Kelly SP, Altekruse SF, McGlynn KA, Rosenberg PS. Future of hepatocellular carcinoma incidence in the United States Forecast Through 2030. J Clin Oncol 2016;34(15):1787-94. https://doi.org/10.1200/JCO.2015.64.7412*
3. *Kinsey E, Lee HM. Management of hepatocellular carcinoma in 2024: The multidisciplinary paradigm in an evolving treatment landscape. Cancers (Basel) 2024;16(3). https://doi.org/10.3390/cancers16030666*
4. *Brown ZJ, Tsilimigras DI, Ruff SM, Mohseni A, Kamel IR, Cloyd JM, et al. Management of hepatocellular carcinoma: A review. JAMA Surg 2023;158(4):410-20. https://doi.org/10.1001/jamasurg.2022.7989*
5. *Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31-8. https://doi.org/10.1038/s41591-021-01614-0*
6. *Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023;330(9):866-9. https://doi.org/10.1001/jama.2023.14217*
7. *Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-40. https://doi.org/10.1038/s41591-023-02448-8*
8. *Deng J, Zubair A, Park YJ, Affan E, Zuo QK. The use of large language models in medicine: Proceeding with caution. Curr Med Res Opin 2024;40(2):151-3. https://doi.org/10.1080/03007995.2023.2295411*
9. *Klang E, Sourosh A, Nadkarni GN, Sharif K, Lahat A. Evaluating the role of ChatGPT in gastroenterology: A comprehensive systematic review of applications, benefits, and limitations. Therap Adv Gastroenterol 2023;16:17562848231218618. https://doi.org/10.1177/17562848231218618*
10. *Ali S, Shahab O, Al Shabeeb R, Ladak F, Yang JO, Nadkarni G, et al. General purpose large language models match human performance on gastroenterology board exam self-assessments. medRxiv 2023:2023.09.21.23295918. https://doi.org/10.1101/2023.09.21.23295918*
11. *Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data 2016;3(1):160035. https://doi.org/10.1038/sdata.2016.35*
12. *Simsek C, Tinaz B, Erol H, Demir H, Atay B, Yazarkan Y, et al. 315 gastroenterology specific ai model outperforms attending physician clinical notes in a real-world data evaluation. Gastroenterology 2024;18(166(5):S-68. https://doi.org/10.1016/S0016-5085(24)00654-1*
13. *Simsek C, Aydinli H, Yazarkan Y, Bozdogan AB, Tinaz B, Melihcan HS, et al. Gastroenterology-specific ai model GastroGPT outperforms attending physicians' and ChatGPT in analyses of clinical notes of real-world endoscopy cases. Endoscopy 2024;56(S 02):S142. https://doi.org/10.1055/s-0044-1782999*

14. *Simsek C, Ebigbo A, Vanek P, Elshaarawy O, Voiosu AM, Antonelli G, et al. LB16 GASTROGPT: First specialtyspecific AI language model outperforms general models across key clinical tasks. United European Gastroenterology Journal 2023;11(935).*
15. *Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol 2009;62(10):1013-20. https://doi.org/10.1016/j.jclinepi.2008.10.009*
16. *Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med 2015;162(1):W1-73. doi: 10.7326/M14-0698.*
17. *Ben-Cohen A, Klang E, Kerpel A, Konen E, Amitai MM, Greenspan H. Fully convolutional network and sparsity-based dictionary learning for liver lesion detection in CT examinations. Neurocomputing 2018;275:1585-94. https://doi.org/10.1016/j.neucom.2017.10.001*
18. *Trivizakis E, Manikis GC, Nikiforaki K, Drevelegas K, Constantinides M, Drevelegas A, et al. Extending 2-D convolutional neural networks to 3-D for advancing deep learning cancer classification with application to MRI liver tumor differentiation. IEEE J Biomed Health Inform 2019;23(3):923-30. https://doi.org/10.1109/JBHI.2018.2886276*
19. *Li S, Jiang H, Pang W. Joint multiple fully connected convolutional neural network with extreme learning machine for hepatocellular carcinoma nuclei grading. Comput Biol Med 2017;84:156-67. https://doi.org/10.1016/j.compbiomed.2017.03.017*
20. *Li S, Jiang H, Yao YD, Pang W, Sun Q, Kuang L. Structure convolutional extreme learning machine and case-based shape template for HCC nucleus segmentation. Neurocomputing 2018;312:9-26. https://doi.org/10.1016/j.neucom.2018.05.013*
21. *Pellat A, Barat M, Coriat R, Soyer P, Dohan A. Artificial intelligence: A review of current applications in hepatocellular carcinoma imaging. Diagn Interv Imaging. 2023;104(1):24-36. https://doi.org/10.1016/j.diii.2022.10.001*
22. *Sato M, Tateishi R, Yatomi Y, Koike K. Artificial intelligence in the diagnosis and management of hepatocellular carcinoma. J Gastroenterol Hepatol. 2021;36(3):551-60. https://doi.org/10.1111/jgh.15413*
23. *Yasaka K, Akai H, Abe O, Kiryu S. Deep Learning with Convolutional Neural Network for Differentiation of Liver Masses at Dynamic Contrast-enhanced CT: A Preliminary Study. Radiology. 2018;286(3):887-96. https://doi.org/10.1148/radiol.2017170706*
24. *Lai Q, De Stefano C, Emond J, Bhangui P, Ikegami T, Schaefer B, et al. Development and validation of an artificial intelligence model for predicting post-transplant hepatocellular cancer recurrence. Cancer Commun (Lond). 2023;43(12):1381-5. https://doi.org/10.1002/cac2.12468*
25. *Simsek C, Can Guven D, Koray Sahin T, Emir Tekin I, Sahan O, Yasemin Balaban H, et al. Artificial intelligence method to predict overall survival of hepatocellular carcinoma. Hepatol Forum. 2021;2(2):64-8. https://doi.org/10.14744/hf.2021.2021.0017*
26. *Gravina AG, Pellegrino R, Palladino G, Imperio G, Ventura A, Federico A. Charting new AI education in gastroenterology: Cross-sectional evaluation of ChatGPT and perplexity AI in medical residency exam. Dig Liver Dis. 2024. https://doi.org/10.1016/j.dld.2024.02.019*
27. *Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS One. 2023;18(8):e0290691. https://doi.org/10.1371/journal.pone.0290691*
28. *Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198*
29. *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(12):2280-2. https://doi.org/10.14309/ajg.0000000000002320*
30. *Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023;29(3):721-32. https://doi.org/10.3350/cmh.2023.0089*
31. *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. medRxiv 2023. https://doi.org/10.1101/2023.11.10.23298364*