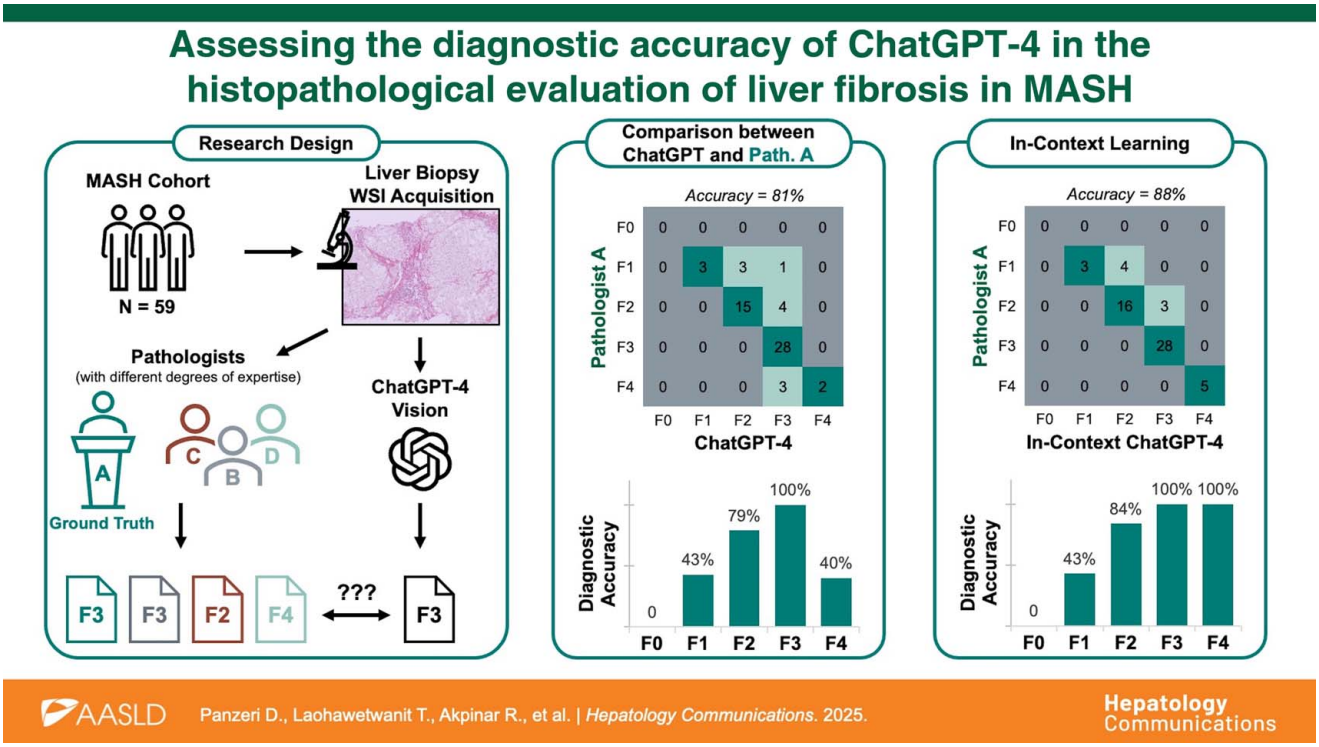
Assessing the diagnostic accuracy of ChatGPT-4 in the histopathological evaluation of liver fibrosis in MASH

VISUAL ABSTRACT



AASLD Panzeri D., Laohawetwanit T., Akpinar R., et al. | *Hepatology Communications*. 2025. Hepatology Communications

Received: 17 September 2024 | Accepted: 26 January 2025

DOI: 10.1097/HC9.0000000000000695

ORIGINAL ARTICLE

OPEN

Assessing the diagnostic accuracy of ChatGPT-4 in the histopathological evaluation of liver fibrosis in MASH

Davide Panzeri[1] | **Thiyaphat Laohawetwanit**[2] | **Reha Akpinar**[3,4] | **Camilla De Carlo**[3,4] | **Vincenzo Belsito**[4] | **Luigi Terracciano**[3,4] | **Alessio Aghemo**[3,5] | **Nicola Pugliese**[3,5] | **Giuseppe Chirico**[1] | **Donato Inverso**[6,7] | **Julien Calderaro**[8,9] | **Laura Sironi**[1] | **Luca Di Tommaso**[3,4]

[1]Department of Physics, University of Milano-Bicocca, Milan, Italy

[2]Division of Pathology, Chulabhorn International College of Medicine, Thammasat University, Pathum Thani, Thailand

[3]Department of Biomedical Sciences, Humanitas University, Milan, Italy

[4]Department of Pathology, IRCCS Humanitas Research Hospital, Milan, Italy

[5]Department of Gastroenterology, Division of Internal Medicine and Hepatology, IRCCS Humanitas Research Hospital, Milan, Italy

[6]Division of Immunology, Transplantation and Infectious Diseases IRCCS San Raffaele Scientific Institute, Milan, Italy

[7]Vita-Salute San Raffaele University, Milan, Italy

[8]Team «Viruses, Hepatology, Cancer», Institut Mondor de Recherche Biomédicale, INSERM U955, Hôpital, Henri Mondor (AP-HP), Université Paris-Est, Créteil, France

[9]Department of Pathology, AP-HP, Henri Mondor University Hospital, Créteil, France

Correspondence
Luca Di Tommaso, Department of Pathology, IRCCS Humanitas Research Hospital, Via Manzoni 56, 20089 Rozzano, MI, Italy.
Email: luca.di_tommaso@hunimed.eu

Laura Sironi, Department of Physics, University of Milano Bicocca, Piazza della Scienza 3, MI 20126, Italy.
Email: laura.sironi@unimib.it

Abstract

Background: Large language models like ChatGPT have demonstrated potential in medical image interpretation, but their efficacy in liver histopathological analysis remains largely unexplored. This study aims to assess ChatGPT-4-vision's diagnostic accuracy, compared to liver pathologists' performance, in evaluating liver fibrosis (stage) in metabolic dysfunction–associated steatohepatitis.

Methods: Digitized Sirius Red-stained images for 59 metabolic dysfunction–associated steatohepatitis tissue biopsy specimens were evaluated by ChatGPT-4 and 4 pathologists using the NASH-CRN staging system. Fields of view at increasing magnification levels, extracted by a senior pathologist or randomly selected, were shown to ChatGPT-4, asking for fibrosis staging. The diagnostic accuracy of ChatGPT-4 was compared with pathologists' evaluations

Abbreviations: AI, artificial intelligence; ChatGPT, Chat Generative Pre-trained Transformer; CNN, Convolutional Neural Networks; CPA, collagen proportionate area; FOV, field of view; GenAI, Generative Artificial Intelligence; GT, ground truth; LLM, large language models; MASH, metabolic dysfunction–associated steatohepatitis; MASLD, metabolic dysfunction-associated steatotic liver disease; WSI, whole slide image.

Davide Panzeri, Thiyaphat Laohawetwanit, and Reha Akpinar contributed equally to this work.

Laura Sironi and Luca Di Tommaso are co-last and corresponding authors.

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.

and correlated to the collagen proportionate area for additional insights. All cases were further analyzed by an in-context learning approach, where the model learns from exemplary images provided during prompting.

Results: ChatGPT-4's diagnostic accuracy was 81% when using images selected by a pathologist, while it decreased to 54% with randomly cropped fields of view. By employing an in-context learning approach, the accuracy increased to 88% and 77% for selected and random fields of view, respectively. This method enabled the model to fully and correctly identify the tissue structures characteristic of F4 stages, previously misclassified. The study also highlighted a moderate to strong correlation between ChatGPT-4's fibrosis staging and collagen proportionate area.

Conclusions: ChatGPT-4 showed remarkable results with a diagnostic accuracy overlapping those of expert liver pathologists. The in-context learning analysis, applied here for the first time to assess fibrosis deposition in metabolic dysfunction–associated steatohepatitis samples, was crucial in accurately identifying the key features of F4 cases, critical for early therapeutic decision-making. These findings suggest the potential for integrating large language models as supportive tools in diagnostic pathology.

Keywords: artificial intelligence, collagen, hepatology, large language models, stage

INTRODUCTION

The degree of liver fibrosis in metabolic dysfunction–associated steatohepatitis (MASH) is the primary prognostic determinant of a patient's clinical outcomes.[1,2] Treatment efficacy and clinical trial enrollment are crucially dependent on the accurate assessment of the extent and architecture of liver fibrosis.[3] This evaluation typically involves histological analysis of liver biopsies performed by expert pathologists. Recently, the advent of whole slide imaging scanners has enabled the digitization of high-resolution images of entire tissue sections, paving the way toward the development of novel artificial intelligence (AI) algorithms designed for multiple tasks, such as image segmentation, classification and integration of anatomical, clinical, and molecular information.[4,5] Notably, AI algorithms have been recently developed to assess the progression or regression of fibrosis with an enhanced dynamic range, aiming to identify potential intra-stage or intra-grade improvements, currently overlooked due to the categorical nature of the staging systems.[6–8]

A significant advancement in AI is the development of large language models (LLMs), specialized forms of generative AI (GenAI) systems that process and generate human-like text based on extensive training with textual data.[9,10] One of the most ubiquitous LLM is Chat Generative Pre-trained Transformer (ChatGPT), developed by OpenAI, an AI chatbot that mimics human conversation, offering responses across various topics.[11] ChatGPT was not designed to solve a specific problem, such as image interpretation or medical document analysis. Instead, it was engineered as a general-purpose model in order to have broad "cognitive" abilities to assist users in a wide range of tasks.[12] Its advanced understanding of context allows for relevant and engaging exchanges easily accessible through simple text prompts. Recent improvements to ChatGPT, namely the multimodal LLM model "GPT-4-vision," allow it to recognize and interpret images, extending its interactive functionality to include visual inputs alongside text prompts.[13]

ChatGPT-4, at the moment, cannot currently be retrained on specific datasets; however, in-context learning can be applied to condition the model's behavior based on the provided context. In-context learning involves presenting the model with relevant examples directly within the prompt, allowing it to use these examples as guidance for the task at hand.[14,15] This approach contrasts with zero-shot learning,[16] where the model must generalize its responses without prior examples or guidance. In-context learning is closely related to few-shot learning,[15] as it relies on providing a limited number of examples in the prompt to enhance the model's reasoning capabilities. Although in-context learning has been widely applied to general-

purpose tasks,[17] its use in cancer classification from hematoxylin and eosin digital images has only recently been reported.[18]

In the hepatology field, ChatGPT has shown promising outcomes in text interpretation.[19] Its responses to medical inquiries from patients with metabolic dysfunction–associated steatotic liver disease (MASLD) were found to be complete and comprehensible.[20] Additionally, ChatGPT was shown to generate responses ranging from very good to excellent concerning liver transplantation queries.[21] Furthermore, a comparative analysis between ChatGPT-generated cirrhosis patient education materials and human-provided materials showed similarities in readability, quality, understandability, and accuracy.[22] However, although ChatGPT demonstrated knowledge and accuracy in responding to most questions related to cirrhosis and HCC, its responses were not consistently comprehensive, particularly in diagnosis and preventive medicine.[23]

Due to the recent introduction of ChatGPT-4, there is limited data available on the application of this technology to medical images. In histopathology, this AI tool has shown relatively promising results when both text and images are provided, though its performance in image interpretation is strongly influenced by the prompt.[24,25] Specifically, regarding liver diseases, ChatGPT-4 demonstrated an accuracy of 87% for staging MASLD when analyzing digital histopathological slides.[26] Despite its high innovation and revolutionary potential, this study is limited by the number of image inputs and the reliance on internet-sourced images.[26]

In this scenario, it is evident that the medical community is increasingly interested in testing and evaluating the capability of ChatGPT as a possible educational tool or diagnostic support. To foster a more extensive and deeper debate within the scientific community, we present an evaluation of the diagnostic performance of ChatGPT-4 in assessing MASH fibrosis, comparing its accuracy to that of pathologists with varying levels of experience. In particular, we focused on ChatGPT-4 capability to interpret Sirius Red-stained liver histology images according to the NASH-CRN staging system. ChatGPT-4 performance is analyzed in dependence on different tuning parameters related to image properties, fibrosis stage and potential users, aiming to juxtapose our findings with results reported in the literature and obtained by applying other AI algorithms. Additionally, we report enhanced fibrosis stratification achieved through the application of the in-context learning strategy. Our goal is to provide comprehensive data to the ongoing debate regarding the use of ChatGPT-4 and provide initial data supporting the potential application of LLMs in diagnostic pathology.

METHODS

Case collection, slide digitization, and pathologists

Sirius Red-stained slides from a series of 59 cases of MASH biopsies (detailed in Table 1) were obtained from the Department of Pathology at the Humanitas Clinical and Research Center in Rozzano, Milan, Italy. We selected cases with a matched clinical and pathological diagnosis of MASH. Specifically, we selected only cases fulfilling the following criteria: (1) clinical diagnosis of MASLD or MetALD (MASLD predominant), (2) histopathological diagnosis of steatohepatitis, and (3) biopsy tissue of >15 mm length, showing >10 portal spaces.

These slides were digitized using a Philips Ultra Fast Scanner with an Olympus 40× air objective (NA = 0.75, Plan Apo, pixel-size = 0.25 µm). The digitized whole slide images (WSIs) underwent independent evaluation by 2 liver pathologists (pathologists A and B) and two general pathologists (pathologists C and D). Utilizing the NASH-CRN system, each pathologist assigned a fibrosis stage to the slides.[27] The stage of fibrosis determined by pathologist A, the most experienced liver pathologist, served as the reference standard or ground truth (GT).

Multiple fields of view (FOVs) at 4×, 10×, and 20× magnifications were extracted under the supervision of an expert liver pathologist who was not directly involved in the study (Supplemental Table 1 and Supplemental Figure S1 http://links.lww.com/HC9/B966). This approach was applied to the entire set of Sirius Red WSIs (N = 59). To further assess ChatGPT-4's performance, an additional dataset of randomly cropped FOVs (N = 35) was generated. From the pool of selected images, at least 6 FOVs (2

TABLE 1 Details of the dataset and strategy pursued during the interaction with ChatGPT

	Expert selection of FOVs	Randomly cropped FOVs
N	59 F = 22, M = 37	35 F = 13, M = 22
F1	7 (12%)	7 (20%)
F2	19 (32%)	9 (26%)
F3	28 (47%)	14 (40%)
F4	5 (9%)	5 (14%)
Image data	FOVs are selected by an expert pathologist	FOVs are randomly selected in the WSI
Number of images for each case	- Two 4× FOVs - Two 10× FOVs - Two 20× FOVs	
Evaluation	By 4 independent pathologists (different degrees of expertise). The senior pathologist was taken as a reference	
Model	ChatGPT-4Vision (released 09/2023)	

Abbreviations: ChatGPT, Chat Generative Pre-trained Transformer; F, female; FOVs, fields of view; M, male; WSI, whole slide image.

at 4×, 2 at 10×, and 2 at 20× magnifications) were randomly chosen and presented to ChatGPT for evaluation, ensuring that the analysis was both standardized and sufficiently randomized to reduce potential bias in FOV selection. Initially, we employed a standard few-shot learning protocol to evaluate ChatGPT-4's performance. Subsequently, all cases were reanalyzed using an in-context learning approach, where the model was primed with external examples beyond the WSIs collected in this study to improve its reasoning (as described in the Supplemental Digital Content, http://links.lww.com/HC9/B966).[14,15,18]

Ethics statement

All research was conducted in accordance with both the Declarations of Helsinki and Istanbul and with the local legislation and institutional requirements. The human samples used in this study were acquired from previously diagnosed human biopsies. Written informed consent was obtained from the individuals for the publication of any potentially identifiable images or data included in this article. Humanitas University review board exempted this study from further ethical approval.

Collagen proportionate area quantification

Collagen proportionate area (CPA) was estimated as a fraction of collagen pixels (Sirius Red positive) over the total number of pixels representing the tissue section.[28] CPA was reported as a percentage (0%–100%) referring to the whole tissue section.

Statistical analysis

A confusion matrix was employed to evaluate and quantify the agreement between the predictions made by ChatGPT and the assessments provided by the pathologists. In the context of our multi-class classification approach, this matrix recorded the instances of true positives, true negatives, false positives, and false negatives for each category within the 5-stage classification system (F0–F4). This matrix served as an overarching indicator of classification precision across all categories. Accuracy was computed as the ratio of correctly classified cases (sum of the diagonal elements) over the total number of samples, presented as a percentage (0%–100%).

The diagnostic accuracy of ChatGPT is here expressed as the recall metric (also known as sensitivity or true positive rate) which measures the ability of a model to correctly identify all relevant instances of a particular class (here, fibrosis stage). Specifically, it is the ratio between the number of correctly identified cases with a specific stage among all the samples with the same diagnosis. This will be hereby presented as a percentage (0%–100%).

Cohen Kappa with quadratic weighting was selected to evaluate the interobserver agreement. Quadratic weighting assigns a higher degree of significance to disagreements between categories that are further apart than those closer together. Kappa values are interpreted with the following classifications: values from 0.01 to 0.20 indicate slight agreement; from 0.21 to 0.40, fair agreement; from 0.41 to 0.60, moderate agreement; from 0.61 to 0.80, substantial agreement; and values >0.81 indicate almost perfect agreement.[29]

The diagnostic accuracy of ChatGPT was evaluated compared to the reference standard (pathologist A, GT). The diagnostic accuracy of ChatGPT and that of other human pathologists were statistically compared using Fisher exact test. Pearson correlation was used to evaluate the correlation between ChatGPT's fibrosis stage and CPA.

Detailed descriptions of methods used in this study are provided in the Supplemental Digital Content, http://links.lww.com/HC9/B966 section.

RESULTS

Cases under investigation

Table 1 shows the details of the dataset and the distribution of liver biopsies according to fibrosis stage in 2 study series submitted to ChatGPT: the first containing FOVs selected by an expert pathologist ($N=59$), while the second related to the randomly collected FOVs ($N=35$).

In order to assess ChatGPT's capability to interpret liver histology images and suggest a fibrosis stage, FOVs were uploaded onto the ChatGPT platform, providing prompts including a brief description of the provided images. To verify the repeatability of the answers and to promote the visibility of multiple FOVs of the same biopsy, avoiding influences from previous conversations and interpretations, we replicated our inquiry in 3 different conversation sessions. Further information related to the followed protocols and the strategies for interaction with ChatGPT are extensively reported in the Supplemental Digital Content, http://links.lww.com/HC9/B966.

Diagnostic accuracy of ChatGPT-4

ChatGPT achieved an overall accuracy of 81% compared to the most senior pathologist (pathologist A, GT). A comparison of the diagnoses provided by ChatGPT and pathologist A is shown in Figure 1A. In detail, for stage F1, the diagnostic accuracy (recall) is 43% (3 out

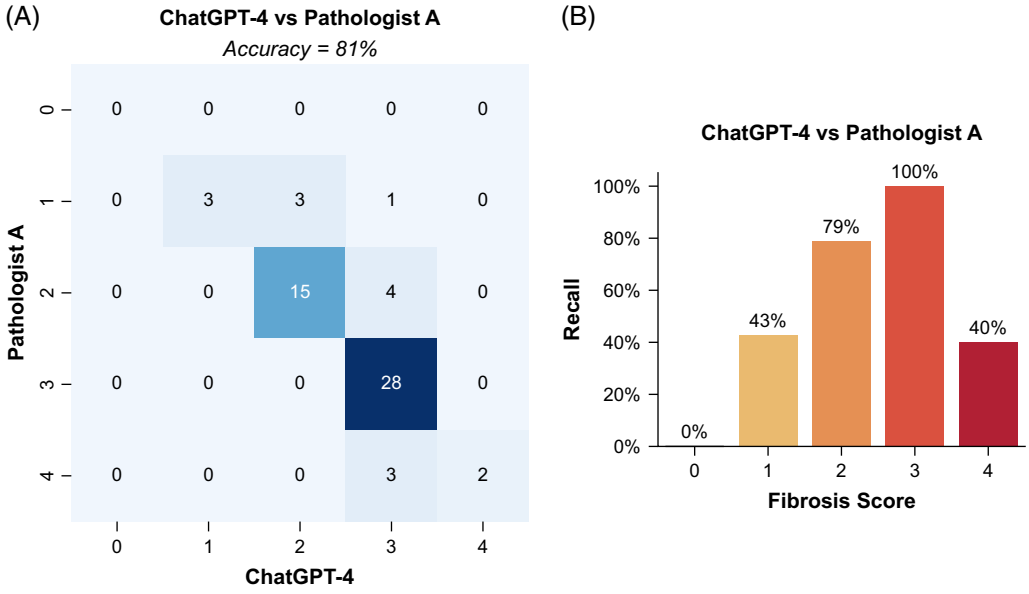


FIGURE 1 Comparison of ChatGPT and pathologist A (ground truth) diagnoses. (A) Confusion matrix comparing the performance of ChatGPT-4 and pathologist A, with the *x*-axis representing ChatGPT-4's diagnoses and the *y*-axis representing pathologist A's diagnoses. The shaded squares indicate the number of scoring comparisons within each category, with darker shades representing higher counts. (B) Bar chart showing ChatGPT-4's diagnostic accuracy (recall) for each fibrosis stage.

of 7 cases); for stage F2, it is 79% (15 out of 19 cases); for stage F3, ChatGPT achieves a perfect diagnostic accuracy rate of 100% (28 out of 28 cases); and for stage F4, the accuracy drops to 40% (2 out of 5 cases) (Figure 1B). Although F4 fibrosis samples are represented by few biopsies in our dataset, in more than half cases, ChatGPT was not able to recognize the presence of nodules, reporting the following sentences: "*without evidence of nodular regeneration and complete disruption of the liver architecture, it is not indicative of cirrhosis (F4)*"; "*However, the lack of complete nodule formation and the preservation of some normal architecture might still be more consistent with NASH-CRN stage F3, indicating numerous septa without cirrhosis*". A further comparison between pathologist A and ChatGPT is reported in Supplemental Figure S2, http://links.lww.com/HC9/B966, showing 2 F4 cases: 1 correctly diagnosed (panel A) and 1 incorrectly (panel B) classified as F3 stage by ChatGPT.

Comparison of ChatGPT-4's and pathologists' performance

Table 2 shows the diagnostic accuracy and interobserver agreement among the participants (including ChatGPT). ChatGPT shows a substantial agreement with the 2 liver pathologists (A and B), while a moderate agreement has been obtained with the general pathologists (C and D). In all cases, the agreement shows an increasing trend from F1 to F3, while for F4 the highest disagreement has been obtained due to the inability of ChatGPT to recognize the features related to this stage, as previously illustrated.

The table also reports the overall accuracy for pathologists B (80%), C (64%), and D (59%) with respect to pathologist A (GT). The quadratic weighted Cohen Kappa revealed almost perfect agreement between pathologists B and A, while a substantial agreement was obtained for all other comparisons. Notably, ChatGPT provided the lowest diagnostic accuracy of F4 among all participants.

A heat map showing the responses of all participants is reported in Figure 2. In particular, for ChatGPT, we reported the diagnosis proposed in each of the 3 single rounds of response (CG1, CG2, CG3), the most frequent response (CG-S), and the forced value (CG-F, the value that ChatGPT was forced to provide in case of uncertain fibrosis staging during conversations). Regardless of accuracy, fibrosis staging provided by ChatGPT was relatively consistent in every round. Pairwise comparisons containing a confusion matrix and percent agreement of participants are shown in Supplemental Figures S3–S8, http://links.lww.com/HC9/B966.

Effect of random cropping on ChatGPT-4's diagnostic accuracy

In order to simulate the possible interaction of a nonexpert in liver pathology, we chose to submit random selected FOVs to ChatGPT. When images subjected to random cropping were provided to ChatGPT-4, its overall accuracy dropped to 54% with

TABLE 2 Percent agreement and weighted Cohen Kappa between participants

Comparison	Overall accuracy (%)	Diagnostic accuracy (recall)					Cohen Kappa (quadratic weighting)	
		F0 (%)	F1 (%)	F2 (%)	F3 (%)	F4 (%)	Value	Interpretation
ChatGPT vs. pathologist A	81	NA	43	79	100	40	0.78	Substantial agreement
ChatGPT vs. pathologist B	76	NA	67	68	93	33	0.75	Substantial agreement
ChatGPT vs. pathologist C	53	NA	25	48	85	17	0.53	Moderate agreement
ChatGPT vs. pathologist D	47	0	7	59	89	14	0.48	Moderate agreement
Pathologist B vs. pathologist A	80	NA	29	79	89	100	0.83	Almost perfect agreement
Pathologist C vs. pathologist A	64	NA	71	63	61	80	0.70	Substantial agreement
Pathologist D vs. pathologist A	59	0	57	58	57	80	0.67	Substantial agreement
ChatGPT random crop vs. pathologist A	54	NA	29	44	86	20	0.52	Moderate agreement
In-context ChatGPT vs. pathologist A	88	NA	43	84	100	100	0.90	Almost perfect agreement
In-context ChatGPT random crop vs. pathologist A	77	NA	29	67	100	100	0.75	Substantial agreement

single-stage outcomes outlined in Table 2 and Figure 3. Considering these results, it is clear that ChatGPT-4 accuracy is strongly dependent on the user and on the selected FOVs. Indeed, we want to stress that in this case, we exploited OpenAI instructions regarding step-by-step prompting and prompt engineering.

Correlation between ChatGPT-4's fibrosis staging and CPA

A total of 12 (20%) ChatGPT's responses were characterized by an ambiguity between 2 fibrosis stages. These indeterminate responses were observed in 3, 7, and 2 cases for F1/2, F2/3, and F3/4, respectively (see Supplemental Figure S9, http://links.lww.com/HC9/B966 for some examples). These ambiguous results prompted the investigation of a possible correlation between ChatGPT and the proportion of biopsy occupied by collagen CPA. The analysis revealed a moderate to strong correlation between ChatGPT's staging and CPA (Pearson correlation coefficient: 0.69; $p < 0.01$, Figure 4).

Indeed, recently, different authors[6–8] reported the possibility of highlighting intra-score fibrosis stages by exploiting also the information provided by the percentage of collagen in the liver biopsy as a quantitative and understandable feature. Moreover, in literature, a correlation of the fibrosis stage with the CPA has been suggested,[30–34] and ChatGPT is trained about this topic since it reports during conversations that: "*The stage of fibrosis in a liver biopsy is indeed correlated with the percentage of collagen in the tissue. In liver biopsies, the fibrosis stage is typically assessed using various scoring systems, which categorize the extent and pattern of fibrosis. These scoring systems often indirectly reflect the amount of collagen deposition in the liver. As fibrosis progresses, the percentage of collagen typically increases, leading to more extensive scarring. This can be quantified using histological techniques that specifically stain collagen, such as Trichrome or Sirius Red staining, and can be visually or digitally quantified to estimate the collagen content. Thus, a higher percentage of collagen generally indicates a more advanced stage of fibrosis. This correlation is crucial for diagnosing the severity of liver disease and guiding treatment decisions*."

In light of these observations, it is therefore possible that ChatGPT attention-based mechanism might also take into account the amount/extension of collagen fibers during its assessment of the fibrosis stage.

Effect of in-context learning on ChatGPT-4's diagnostic accuracy

Drawing inspiration from general-purpose models[17] that use exemplary images and corresponding descriptions

A(GT)	B	C	D	CG1	CG2	CG3	CG-S	CG-F
1	1	1	0	2.5	1.5	1.5	1.5	1
	2	1	2	3.5	2	2	2	2
	2	2	1	3	2	2	2	2
	2	2	1	3	2	2	2	2
	2	1	1	3	3	3	3	3
	2	1	1	2.5	1.5	1.5	1.5	1
	1	1	0	1.5	1.5	1	1.5	1
2	2	2	1	2.5	2.5	2.5	2.5	3
	2	2	2	2	2.5	2	2	2
	3	2	1	3	3	3	3	3
	3	2	2	2.5	1.5	2.5	2.5	2
	2	2	2	1.5	1.5	2.5	2	2
	2	2	2	2.5	2	2.5	2.5	2
	2	2	2	2.5	2	2	2	2
	2	2	2	2.5	3	2.5	2.5	2
	2	2	2	2	3	3	3	3
	2	2	2	2.5	2.5	3	2.5	3
	2	2	2	1.5	2	2.5	2	2
	2	1	1	1.5	2	2	2	2
	2	1	1	1.5	2	2	2	2
	2	2	1	3	2.5	2.5	2.5	2
	3	3	3	2.5	3	2	2	2
	1	1	1	3	1.5	2.5	2	2
	2	1	2	2	2	2.5	2	2
	2	1	1	2.5	2.5	3	2.5	2
	2	3	2	3	2	2.5	2	2
3	3	3	3	3	3	3	3	3
	3	2	3	3	3	3	3	3
	3	3	2	4	3	3	3	3
	3	3	3	3	3	3	3	3
	3	1	1	3	3	3	3	3
	2	2	2	3	3.5	3	3	3
	3	3	3	3	4	3	3	3
	3	3	3	3	3	3	3	3
	2	2	1	3.5	3	3.5	3.5	3
	3	4	3	3	3	3	3	3
	3	2	3	3	3	3	3	3
	3	3	3	3	3	3	3	3
	4	3	3	3.5	3	3.5	3.5	3
	3	3	3	3	3	3	3	3
	3	3	4	3	3	3	3	3
	3	3	3	3	3	2.5	3	3
	3	3	3	3	3	3	3	3
	3	1	2	3	3	3	3	3
	3	3	2	3	3	3	3	3
	3	3	4	3	2.5	3	3	3
	3	3	1	3	3	2.5	3	3
	3	2	3	3	3	3	3	3
	3	3	3	3	3	3.5	3	3
	3	2	2	3	3.5	3	3	3
	3	3	3	3	3	3	3	3
	3	2	1	3	3	3	3	3
	3	3	3	3	3	3	3	3
	3	4	4	3	3	4	3	3
4	4	4	4	4	4	4	4	4
	4	4	4	3	3	4	3	3
	4	4	4	3	3	3	3	3
	4	3	3	4	3.5	4	4	4
	4	4	4	3.5	3	3	3	3

FIGURE 2 Heat map showing responses of pathologists (A, B, C, D) and ChatGPT-4. Abbreviations: A (GT) = pathologist A (ground truth); B = pathologist B; C = pathologist C; D = pathologist D; CG1 = ChatGPT-4 attempt 1; CG2 = ChatGPT-4 attempt 2; CG3 = ChatGPT-4 attempt 3; CG-S = ChatGPT-4 summary; CG-F = ChatGPT-4 forced fibrosis stage. Numerical values refer to the pathological stage according to the NASH-CRN staging system. In some cases, ChatGPT-4 generated a response staying in the middle of categories of NASH-CRN (1.5 for indecision between F1 and F2; 2.5 for indecision between F2 and F3; 3.5 for indecision between F3 and F4).

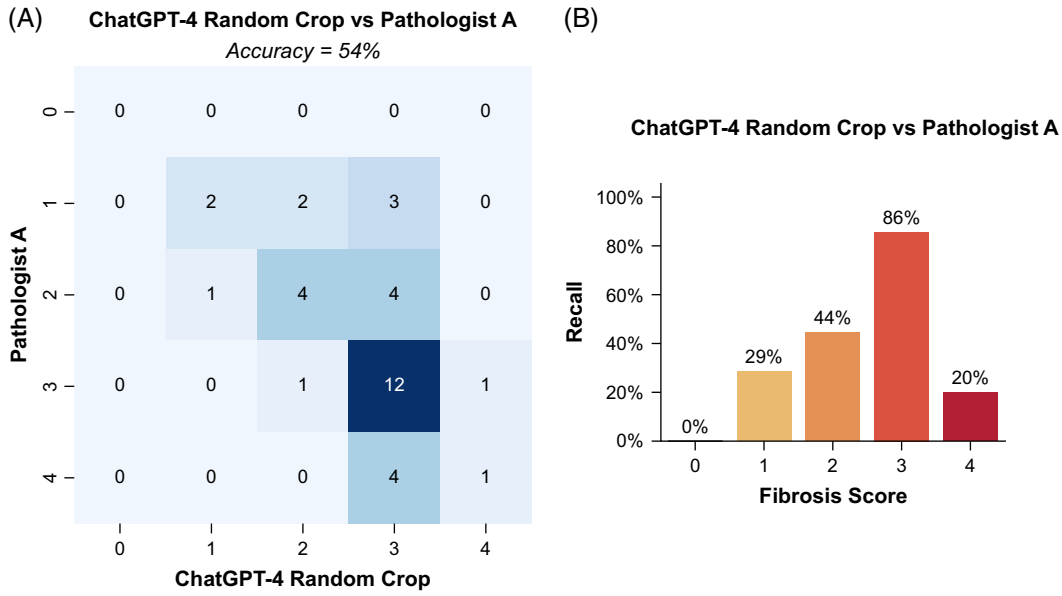


FIGURE 3 Effect of random cropping on ChatGPT-4's diagnostic accuracy. (A) Confusion matrix comparing the performance of ChatGPT-4 and pathologist A (ground truth). (B) Bar chart showing ChatGPT-4's diagnostic accuracy (recall) of each fibrosis stage.

to enhance LLM performance, we applied in-context learning to all selected and random FOVs for fibrosis stratification.[14,15,18]

In this light, the misclassification of F4 cases was particularly unexpected, given that these cases are defined by distinctive nodular structures and collagen deposition patterns that ChatGPT-4 initially failed to recognize. To address this issue, we presented images of nodular structures at different magnifications, prompting ChatGPT to learn the key features of nodular formations for following stratification task (see also Supplemental Digital Content, http://links.lww.com/HC9/B966). This approach aimed to improve the model's ability to recognize and classify nodular structures.

Furthermore, a second in-context learning approach was applied to all F1–F4 cases, by providing examples of FOVs at different magnifications and across various fibrosis stages extracted from images external to the study cohort. This procedure has been applied to both selected and random FOVs datasets (see also Supplemental Digital Content, http://links.lww.com/HC9/B966). This approach offered the model a comparative framework, emphasizing the structural transitions characteristic of each stage.

Both in-context learning strategies resulted in a consistent improvement in staging performance, raising the overall accuracy to 88% and 77% for selected and random FOVs, respectively (Figures 5A, C). Interestingly, F1 cases did not show significant improvement with this approach (see Figures 5B, D), whereas F4 cases were consistently staged correctly (100%, 5 out of 5 cases). This outcome may also reflect the variability among pathologists in staging early fibrosis cases (F1/F2), while more advanced stages tend to be more consistently recognized by pathologists.[28]

The final rows in Table 2 show the improvement in overall accuracy and interobserver agreement between ChatGPT versus pathologist A. The quadratically weighted Cohen Kappa, which measures the overall agreement between ChatGPT's scoring and the GT scoring by pathologist A, increased from 0.78 in the previous protocol to 0.90 with the implementation of the in-context strategy. This result indicates almost perfect agreement, comparable to the level of concordance observed between the 2 most senior pathologists (0.83, pathologists A vs. B).

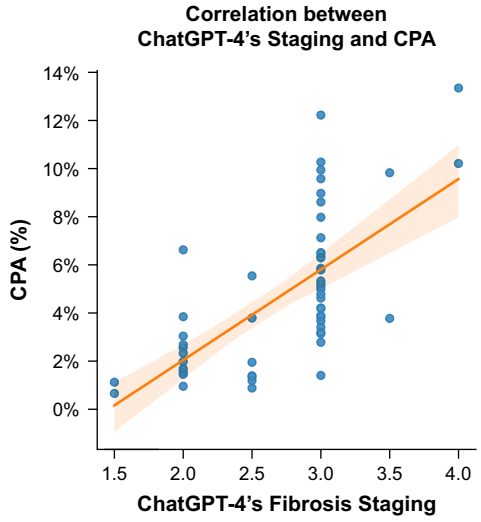


FIGURE 4 ChatGPT-4's indecisive fibrosis staging and CPA. Correlation between ChatGPT-4's staging and CPA. Data is displayed in blue, and a linear regression is overlaid in orange. Pearson correlation coefficient is 0.69 ($p < 0.01$). Abbreviation: CPA, collagen proportionate area.

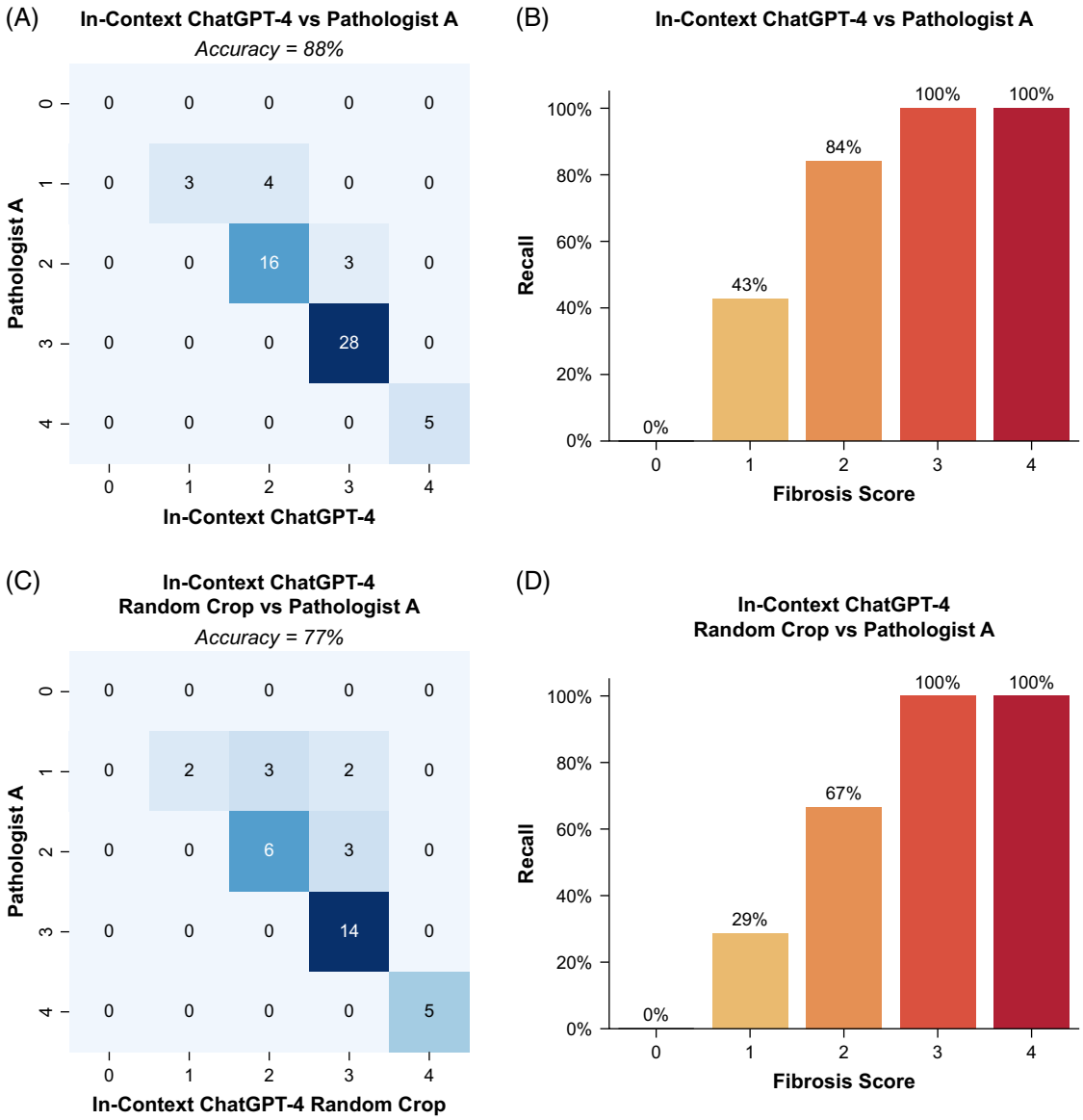


FIGURE 5 Comparison of in-context ChatGPT-4 and pathologist A (ground truth) diagnoses. (A) Confusion matrix comparing the performance of in-context ChatGPT-4 and pathologist A for selected FOVs, with the *x*-axis representing ChatGPT-4's diagnoses and the *y*-axis representing pathologist A's diagnoses. The shaded squares indicate the number of instances in each category, with darker shades representing higher counts. (B) Bar chart showing the recall (diagnostic accuracy) of in-context ChatGPT-4 for each fibrosis stage using selected FOVs. Similarly, (C) and (D) present the confusion matrix and recall for the in-context analysis performed on the dataset comprising randomly cropped FOVs. Abbreviation: FOVs, fields of view.

Supplemental Figure S10, http://links.lww.com/HC9/B966, illustrates the Fleiss Kappa agreement values across fibrosis stages (F0–F4), assessing agreement for each stage individually—unlike Cohen Kappa, which provides an overall agreement between 2 raters. The figure shows values for datasets without (Supplemental Figures 10A, C, http://links.lww.com/HC9/B966) and with in-context learning (Supplemental Figures 10B, D, http://links.lww.com/HC9/B966), using selected and random FOVs, respectively. The implementation of in-context learning significantly improved the agreement for F4 cases, as demonstrated by the increase in Fleiss Kappa from 0.54 to 1.00 for the selected dataset (Supplemental Figures 10A, C, http://links.lww.com/HC9/B966) and from 0.21 to 1.00 for random FOVs (Supplemental Figures 10B, D, http://links.lww.com/HC9/B966). Additionally, the diagnostic accuracy (recall) rises from 40% to 100% with the in-context approach for selected FOVs of F4 cases (Figure 5B). Furthermore, even when random FOVs are used, the diagnostic accuracy (recall) improves significantly from 20% to 100% (Figure 5D). Overall, the in-context learning strategies enhance ChatGPT's staging performance, increasing the accuracy with random FOVs from 54% to 77% (Figure 5C). However, this value remains lower than the accuracy obtained with FOVs selected by an expert liver pathologist (88%, Figure 5A).

DISCUSSION

In this study, we provided comprehensive data on the diagnostic accuracy of ChatGPT-4 in staging MASH. We employed a large series of MASH WSIs, each accompanied by a histological diagnosis and collagen area quantification. Selected and random FOVs from this series were submitted to ChatGPT-4 with an interaction mimicking the pathologist's approach while interpreting liver fibrosis in histologic sections. Typically, pathologists begin by evaluating tissue samples at lower magnifications (4×) and then progress to higher magnifications (10× and 20×). Following this approach, ChatGPT-4 was provided with a series of images, starting from lower to higher magnification. This strategy, which involves applying a series of intermediate reasoning steps while interacting with LLMs, is known as chain-of-thought prompting, and it has been reported to enhance the model, performance on tasks involving arithmetic, commonsense, and symbolic reasoning.[35] As it happens during pathologists' assessment, the lower magnification FOVs are already endowed with a sufficient amount of relevant information to collocate the biopsy in between 2 fibrosis stages or to exclude the opposite cases. For example, for low fibrosis extent (F1), ChatGPT immediately excludes the possibility of F3–F4 stages. The higher magnification images are necessary to refine the description and the final suggestion of the fibrosis stage, particularly for lower fibrosis levels.

Using this step-by-step protocol and submitting properly selected images, ChatGPT-4 demonstrated an overall accuracy of 81% in staging MASH, comparable to that of 2 expert liver pathologists (pathologists A and B). The performance dropped to 54% with randomly cropped images. This decline was expected to be similar to the situation observed with junior pathology residents who lack the skill to select the appropriate region of interest for diagnostic purposes.[36] This reflects the strong dependence of ChatGPT's responses on the user and the presented FOVs. Although our workflow relied on expert pathologist-selected FOVs to ensure human-in-the-loop validation, future adaptations could employ self-supervised or weakly supervised algorithms to automate FOV selection. Such approaches can process larger image datasets at once, potentially filtering regions of diagnostic interest with performance approximating that of a senior pathologist.

Further observations emerged while dissecting results according to the fibrosis stage. ChatGPT-4 performed best on F3 with selected (100%) and random images (86%). On the other hand, the worst results were observed for F4 (40% and 20% for selected and random FOVs). This is in striking contrast to pathologists' performance. Indeed, while the overall agreement on F1, F2, and F3 strongly reflects the pathologist's experience, the concordance on F4 was always the highest observed. This unexpected result may suggest that ChatGPT-4 is capable of recognizing fibrosis based on the color of histopathological staining (red for Sirius) and its extent within the tissue. However, its higher error rate in F4 cases might be related to difficulties in accurately identifying more complex structures, such as nodular formations and the disruption of liver architecture, although present in the shown images.

To improve upon this few-shot strategy, which relied on different magnifications of the same biopsy, we implemented a proof-of-concept in-context learning approach. In this strategy, we provided ChatGPT with examples of key defining features of fibrosis stages F1–F4 before presenting the case under investigation. This approach increased the overall accuracy of ChatGPT's staging to almost 90% and 80% for selected and random FOVs, respectively. Furthermore, the diagnostic accuracy trend across fibrosis stages closely matched the assessments performed by pathologists. Specifically, the staging accuracy for more advanced cases (F3–F4), which exhibit the most accordance between pathologists[28,37,38] and are most critical for therapeutic and clinical decisions, reached 100%. By contrast, the accuracy for early-stage cases (F1) was lower at 43%, and the in-context strategy did not have any improvement, which probably also reflects the variability among pathologists in staging early-onset fibrosis.

Our study may also suggest that fibrosis is interpreted by ChatGPT as a continuous process rather than distinct categories, explaining its tendency to provide indecisive diagnosis. Indeed, ChatGPT may respond with an indecisive stage highlighting how some FOVs are characterized by features related to contiguous stages. To explore this aspect, we compared ChatGPT-4's responses with CPA, a parameter representing the total collagen area in liver biopsies and a crucial predictor of clinical outcomes in liver diseases.[28,30] We found a moderate to strong correlation between ChatGPT-4's assessments and CPA, indicating that ChatGPT-4's evaluations are not arbitrary but reflect a nuanced understanding of the fibrosis process, as also reported in the literature.[28,38]

Findings from traditional Convolutional Neural Networks (CNNs),[7,8,34,37,39,40] trained with annotated WSIs, are comparable to the results obtained in our study using a general LLM like ChatGPT, which was not specifically trained on medical images. However, CNN-based tools are costly to develop and often remain confined to their creating institutions, restricting broader access. LLMs, such as ChatGPT-4, even though they still would require extensive trials, might be considered as an alternative method given their ease of use and interactive approach, while also outsourcing the training of the underlying model, which not all hospitals are equipped to perform. While LLMs have already

succeeded in text-based applications, offering potential benefits in clinical decision support, patient engagement, and public health efforts, their application on multimodal data, such as medical images, is recently emerging.

Additionally, a recent study by Ferber et al[18] revealed that by employing few-shot in-context prompting strategies, GPT-4V achieved classification accuracy comparable to dedicated image classifiers, highlighting the potential of multimodal LLMs to bridge the performance gap between generalist and specialized AI systems in pathology.

Notably, our dataset has never been published or made publicly available prior to this work, ensuring that ChatGPT had no prior exposure to the images employed here during its training process. This minimizes bias and eliminates the possibility that the model's performance was influenced by preexisting familiarity with our dataset.

More generally, the development of foundation models for medical AI demonstrated significant improvements in interpreting diverse medical images, contributing to generalist AI capabilities in healthcare.[41–45] Among them, in May 2024, Google has announced Med-Gemini a family of LLM capable of interpreting multimodal inputs coming from different sources (health records, diagnosis reports, lab analysis) and imaging modalities (radiology, histology, and microscopy).[46] Their performance is reported to be superior to ChatGPT-4Vision by an average margin of 44.5%. We could not validate these findings on our dataset since, at the moment, this model is not widely available.

While LLM-based applications in diagnostic medicine offer potential benefits, several limitations and ethical concerns may affect their effective integration into clinical practice. In general, these models are characterized by numerous issues, such as data bias, limited interpretability, accuracy, and reliability of the generated information.[47,48] Moreover, the performance of these models is heavily related to the representativeness and quality of the training data, which is actually undisclosed for closed-source models such as ChatGPT or Gemini.[49] Also, the lack of consistency in answers of LLMs over time is a concern for large-scale applications in the medical field.[50] However, the interaction strategy between pathologists and LLMs through in-context examples, which we tested in our study, could be beneficial for improving explainability and guiding the model's reasoning processes, ultimately fostering more transparent and reliable AI-assisted diagnostics.

In summary, this study has been designed to provide scientific data to enrich the discussion on the ongoing application of generative AI technology to medicine and, in particular, hepatology. Our objective was not to debate the ethical aspects of this development or to promote its use. Larger studies, prospective validations, and additional training strategies will be required to fully delineate the safe and effective integration of LLMs into daily clinical practice.[42,48] The obtained results are remarkable and, in agreement with other sources,[26] show that ChatGPT can match the diagnostic accuracy of expert liver pathologists in staging MASH if adequate images are provided. Importantly, the application of an in-context learning approach effectively resolved the previously observed challenges with accurately classifying F4 fibrosis cases. This enhancement not only improved the model's ability to identify advanced fibrosis stages but also reinforced its potential as a reliable tool in diagnostic pathology. These findings suggest that with appropriate guidance and structured input, LLMs like ChatGPT can provide valuable assistance to human expertise in clinical and research settings.

AUTHOR CONTRIBUTIONS

Davide Panzeri: conception and design, data acquisition, data analysis, and writing—original, review, and editing. Thiyaphat Laohawetwanit: data analysis, writing—original, review, and editing. Reha Akpinar: conception and design, data acquisition, data analysis, and writing—review and editing. Camilla De Carlo: data analysis and writing—review and editing. Vincenzo Belsito: data acquisition, data analysis, and writing—review and editing. Luigi Terracciano: data acquisition, data analysis, and writing—review and editing. Alessio Aghemo: data analysis and writing—review and editing. Nicola Pugliese: data analysis and writing—review and editing. Giuseppe Chirico: data analysis and writing—review and editing. Donato Inverso: data analysis and writing—review and editing. Julien Calderaro: data analysis and writing—review and editing. Laura Sironi: conception and design, data acquisition, data analysis, and writing—review and editing. Luca Di Tommaso: conception and design, data acquisition, data analysis, and writing—review and editing.

ACKNOWLEDGMENTS

The publication fee for this work was covered by the Italian Ministry of Health's “Ricerca Corrente” funding to IRCCS Humanitas Research Hospital.

CONFLICTS OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORCID

Davide Panzeri https://orcid.org/0009–0000–8773–5883
Thiyaphat Laohawetwanit https://orcid.org/0000–0003–3805–7291
Reha Akpinar https://orcid.org/0000–0002–6444–0932

Camilla De Carlo https://orcid.org/0000–0001–5236–8307
Luigi Terracciano https://orcid.org/0000–0002–9393–9660
Alessio Aghemo https://orcid.org/0000–0003–0941–3226
Nicola Pugliese https://orcid.org/0000–0001–6466–1412
Giuseppe Chirico https://orcid.org/0000–0001–6578–6460
Donato Inverso https://orcid.org/0000–0003–0987–3345
Laura Sironi https://orcid.org/0000–0002–6638–7709
Luca Di Tommaso https://orcid.org/0000–0002–9013–4728

REFERENCES

1. Dulai PS, Singh S, Patel J, Soni M, Prokop LJ, Younossi Z, et al. Increased risk of mortality by fibrosis stage in nonalcoholic fatty liver disease: Systematic review and meta-analysis. Hepatology. 2017;65:1557–65.
2. Ekstedt M, Hagström H, Nasr P, Fredrikson M, Stål P, Kechagias S, et al. Fibrosis stage is the strongest predictor for disease-specific mortality in NAFLD after up to 33 years of follow-up. Hepatology. 2015;61:1547–54.
3. Gawrieh S, Sethunath D, Cummings OW, Kleiner DE, Vuppalanchi R, Chalasani N, et al. Automated quantification and architectural pattern detection of hepatic fibrosis in NAFLD. Ann Diagn Pathol. 2020;47:151518.
4. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology—New tools for diagnosis and precision oncology. Nat Rev Clin Oncol. 2019;16:703–15.
5. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28:31–8.
6. Soon G, Wee A. Updates in the quantitative assessment of liver fibrosis for nonalcoholic fatty liver disease: Histological perspective. Clin Mol Hepatol. 2021;27:44–57.
7. Naoumov NV, Brees D, Loeffler J, Chng E, Ren Y, Lopez P, et al. Digital pathology with artificial intelligence analyses provides greater insights into treatment-induced fibrosis regression in NASH. J Hepatol. 2022;77:1399–409.
8. Heinemann F, Gross P, Zeveleva S, Qian HS, Hill J, Höfer A, et al. Deep learning-based quantification of NAFLD/NASH progression in human liver biopsies. Sci Rep. 2022;12:19236.
9. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 2024;30:1134–42.
10. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Publisher correction: Large language models encode clinical knowledge. Nature. 2023;620:E19.
11. OpenAI. Introducing ChatGPT. *OpenAI.com*. Published online 2022. https://openai.com/blog/chatgpt
12. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388:1233–9.
13. OpenAI. ChatGPT can now see, hear, and speak. *OpenAI.com*. 2023. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak
14. Dong Q, Li L, Dai D, Zheng C, Ma J, Li R, et al. A Survey on In-context Learning, arXiv. 2023. doi:10.48550/ARXIV.2301.00234.
15. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. 2020. doi:10.48550/ARXIV.2005.14165.
16. Romera-Paredes B, Torr PHS. An embarrassingly simple approach to zero-shot learning. In: Feris RS, Lampert C, Parikh D, eds. *Visual Attributes* Advances in Computer Vision and Pattern Recognition. Springer International Publishing; 2017:11–30. doi:10.1007/978-3-319-50077-5_2
17. Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: A Visual Language Model for Few-Shot Learning, Springer, Cham; 2022. Published online. doi:10.48550/ARXIV.2204.14198.
18. Ferber D, Wölflein G, Wiest IC, Ligero M, Sainath S, Ghaffari Laleh N, et al. In-context learning enables multimodal large language models to classify cancer pathology images. Nat Commun. 2024;15:10104.
19. Ge J, Lai JC. Artificial intelligence-based text generators in hepatology: ChatGPT is just the beginning. Hepatol Commun. 2023;7:e0097.
20. Pugliese N, Wai-Sun wong V, Schattenberg JM, Romero-Gomez M, Sebastiani G, Aghemo A, et al. Accuracy, reliability, and comprehensibility of ChatGPT-generated medical responses for patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol. 2024;22:886–89.e5.
21. Endo Y, Sasaki K, Moazzam Z, Lima HA, Schenk A, Limkemann A, et al. Quality of ChatGPT responses to questions related to liver transplantation. J Gastrointest Surg. 2023;27:1716–9.
22. Pradhan F, Fiedler A, Samson K, Olivera-Martinez M, Manatsathit W, Peeraphatdit T. Artificial intelligence compared with human-derived patient educational materials on cirrhosis. Hepatol Commun. 2024;8:e0367.
23. 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;29:721–32.
24. Apornvirat S, Namboonlue C, Laohawetwanit T. Comparative analysis of ChatGPT and Bard in answering pathology examination questions requiring image interpretation. Am J Clin Pathol. 2024;162:252–60.
25. Oon ML, Syn NL, Tan CL, Tan KB, Ng SB. Bridging bytes and biopsies: A comparative analysis of ChatGPT and histopathologists in pathology diagnosis and collaborative potential. Histopathology. 2024;84:601–13.
26. Zhang Y, Liu H, Sheng B, Tham YC, Ji H. Preliminary fatty liver disease grading using general-purpose online large language models: ChatGPT-4 or Bard? J Hepatol. 2024;80:e279–81.
27. Kleiner DE, Brunt EM, Van Natta M, Behling C, Contos MJ, Cummings OW, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology. 2005;41:1313–21.
28. Akpinar R, Panzeri D, De Carlo C, Belsito V, Durante B, Chirico G, et al. Role of artificial intelligence in staging and assessing of treatment response in MASH patients. Front Med. 2024;11:1480866.
29. Fleiss JL, Levin B, Paik MC. The measurement of interrater agreement. Ch. 18. In: Shewart WA, Wilks SS, eds. Statistical Methods for Rates and Proportions, Wiley; 2003.
30. Buzzetti E, Hall A, Ekstedt M, Manuguerra R, Guerrero Misas M, Covelli C, et al. Collagen proportionate area is an independent predictor of long-term outcome in patients with non-alcoholic fatty liver disease. Aliment Pharmacol Ther. 2019;49:1214–22.
31. Stasi C, Tsochatzis EA, Hall A, Rosenberg W, Milani S, Dhillon AP, et al. Comparison and correlation of fibrosis stage assessment by collagen proportionate area (CPA) and the ELF panel in patients with chronic liver disease. Dig Liver Dis. 2019;51:1001–7.
32. Serdjebi C, Bertotti K, Huang P, Wei G, Skelton-Badlani D, Leclercq IA, et al. Automated whole slide image analysis for a translational quantification of liver fibrosis. Sci Rep. 2022;12:17935.

33. Israelsen M, Guerrero Misas M, Koutsoumourakis A, Huang Y, Thiele M, Hall A, et al. Collagen proportionate area predicts clinical outcomes in patients with alcohol-related liver disease. Aliment Pharmacol Ther. 2020;52:1728–39.
34. Wang Z, Jeffrey GP, Huang Y, De Boer B, Garas G, Wallace M, et al. Liver fibrosis quantified by image morphometry predicts clinical outcomes in patients with non-alcoholic fatty liver disease. Hepatol Int. 2023;17:1162–9.
35. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. doi: 10.48550/arXiv.2201.11903.
36. Brunyé TT, Balla A, Drew T, Elmore JG, Kerr KF, Shucard H, et al. From image to diagnosis: Characterizing sources of error in histopathologic interpretation. Mod Pathol. 2023;36:100162.
37. Taylor-Weiner A, Pokkalla H, Han L, Jia C, Huss R, Chung C, et al. A machine learning approach enables quantitative measurement of liver histology and disease monitoring in NASH. Hepatology. 2021;74:133–47.
38. Schuppan D, Surabattula R, Wang XY. Determinants of fibrosis progression and regression in NASH. J Hepatol. 2018;68:238–50.
39. Naik SN, Forlano R, Manousou P, Goldin R, Angelini ED. Fibrosis severity scoring on Sirius red histology with multiple-instance deep learning. Biol Imaging. 2023;3:e17.
40. Ratziu V, Hompesch M, Petitjean M, Serdjebi C, Iyer JS, Parwani AV, et al. Artificial intelligence-assisted digital pathology for non-alcoholic steatohepatitis: Current status and future directions. J Hepatol. 2024;80:335–51.
41. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616:259–65.
42. Han T, Adams LC, Nebelung S, Kather JN, Bressem KK, Truhn D. Multimodal Large Language Models are Generalist Medical Image Interpreters. medRxiv. 2023. doi:10.1101/2023.12.21.23300146
43. Lu MY, Chen B, Williamson DFK, Chen RJ, Liang I, Ding T, et al. A visual-language foundation model for computational pathology. Nat Med. 2024;30:863–74.
44. Lu MY, Chen B, Williamson DFK, Chen RJ, Zhao M, Chow AK, et al. A multimodal generative AI copilot for human pathology. Nature. 2024;634:466–73.
45. Vorontsov E, Bozkurt A, Casson A, Shaikovski G, Zelechowski M, Liu S, et al. Virchow: A Million-Slide Digital Pathology Foundation Model. arXiv preprint arXiv. 2023. Published online. doi: 10.48550/arXiv.2309.07778.
46. Yang L, Xu S, Sellergren A, Kohlberger T, Zhou Y, Ktena I, et al. Advancing multimodal medical capabilities of Gemini. arXiv preprint arXiv. 2024. doi: 10.48550/arXiv.2405.03162.
47. Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology—A recent scoping review. Diagn Pathol. 2024;19:43.
48. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30:2613–22.
49. Balloccu S, Schmidtová P, Lango M, Dušek O. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. arXiv preprint arXiv. 2024. doi:10.48550/arXiv.2402.03927
50. Reese JT, Danis D, Caufield JH, Groza T, Casiraghi E, Valentini G, et al. On the limitations of large language models in clinical diagnosis. medRxiv. 2024. doi:10.1101/2023.07.13.23292613.

How to cite this article: Panzeri D, Laohawetwanit T, Akpinar R, De Carlo C, Belsito V, Terracciano L, et al. Assessing the diagnostic accuracy of ChatGPT-4 in the histopathological evaluation of liver fibrosis in MASH. Hepatol Commun. 2025;9:e0695. https://doi.org/10.1097/HC9.0000000000000695