

Evaluation of the Performance of ChatGPT 4.5 in LI-RADS Categorization and Management Suggestion: Zero-shot versus Few-shot Prompting Method

Eren Çamur¹ , Yasin Celal Güneş² ¹ Department of Radiology, Ministry of Health Ankara 29 Mayıs State Hospital, Ankara, Türkiye² Department of Radiology, Kırıkkale Yüksek İhtisas Hospital, Kırıkkale, Türkiye

Received: 2025-05-08

Accepted: 2025-08-08

Published Online: 2025-09-05

Corresponding Author

Eren Çamur, M.D.

Address: Department of Radiology,
Ministry of Health Ankara 29 Mayıs State
Hospital, Ankara, TürkiyeE-mail: eren.camur@outlook.com

ABSTRACT

Objective: To evaluate whether soft-prompt-based conditioning through “Few-shot” prompting improves the accuracy and clinical utility of ChatGPT 4.5 in classifying hepatic lesions and management recommendations according to the Liver Imaging Reporting and Data System (LI-RADS).

Methods: This cross-sectional observational study assessed ChatGPT 4.5 using fifty fictional radiology reports covering eight LI-RADS categories. The reports were evaluated under Zero-shot and “Few-shot” prompting conditions. Two board-certified radiologists independently scored the model’s LI-RADS categories and management suggestions using a binary correct/incorrect system. The model performance was compared to that of a radiologist, and statistical analysis was conducted using McNemar’s test, with $p < 0.05$ considered significant.

Results: With zero-shot prompting, ChatGPT 4.5 correctly classified 84% of the LI-RADS categories and 70% of the management suggestions. “Few-shot” prompting improved performance, with 92% correct LI-RADS classification and 84% accurate management recommendations. Although the improvement in categorization was not statistically significant ($p = 0.125$), the enhancement in management suggestions was significant ($p = 0.016$). The radiologist comparator achieved 82% accuracy for the LI-RADS classification and 60% for management suggestions. Notably, ChatGPT 4.5, when supported by “Few-shot” prompting, outperformed the radiologist in recommending appropriate management.

Conclusion: “Few-shot” prompting transforms ChatGPT 4.5 from a diagnostic assistant into a powerful tool for clinical decision-making, significantly enhancing its ability to generate patient-centered management recommendations. This study is among the earliest to benchmark ChatGPT 4.5 against a radiologist in LI-RADS-based diagnostic and management tasks, underscoring its potential not only to streamline reporting but also to elevate the quality of patient care. As LLMs continue to evolve, they may become supportive tools in radiology, bridging between image interpretation and clinical decision.

Keywords: ChatGPT, LI-RADS, artificial intelligence, liver, few-shot

© 2025. The copyright of this article is retained by the author(s).

OPEN  ACCESS

This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/).

This license permits the free sharing and adaptation of the work for non-commercial purposes, provided that appropriate attribution is given to the original author(s) and to its initial publication in this journal.

INTRODUCTION

Large language models (LLMs) are advanced neural networks that can interpret and generate human-like text with exceptional accuracy and coherence. Utilizing extensive datasets and sophisticated algorithms, these models have shown significant potential in radiology, particularly in improving patient triage, clinical workflow optimization, and automated structured reporting [1–3]. LLMs can efficiently automate the selection and prioritization of imaging examinations, expediting urgent case management and enhancing departmental productivity [4,5]. Given the ongoing global shortage of radiologists, these models could notably alleviate clinical workloads [2,6]. However, their utility is limited by their propensity to generate inaccurate or contextually inappropriate outputs, potentially leading to diagnostic errors and compromising patient safety [7,8].

Fine-tuning involves refining a pretrained large language model with specialized domain-specific datasets, enhancing its performance for targeted medical applications without the need for complete retraining [9]. By adjusting the model parameters, fine-tuning significantly increases the accuracy and contextual relevance, ensuring that the outputs are tailored to clinical needs. This approach is crucial because generic LLMs often lack the nuanced clinical specificity required by healthcare professionals. Fine-tuned models such as BioBERT, ClinicalBERT, and RadBERT trained on biomedical corpora consistently outperform general-purpose models in interpreting medical and radiological terminology, highlighting the value of domain-specific fine-tuning [10].

Main Points

- This study demonstrates the performance of ChatGPT 4.5 in comparison with a radiologist for LI-RADS categorization, showing comparable diagnostic accuracy and improved management suggestions when using “Few-shot” prompting.
- The “Few-shot” prompting method significantly enhanced ChatGPT 4.5’s ability to suggest clinically appropriate management approaches for LI-RADS, establishing its potential as a valuable clinical decision support tool.
- ChatGPT 4.5 has great potential for supporting radiologists in guiding the management of cirrhotic patients with liver lesions, heralding a transformative shift in abdominal radiology.

Liver Imaging Reporting and Data System (LI-RADS) provides a structured, standardized framework for evaluating hepatic lesions in patients at risk for hepatocellular carcinoma (HCC) [11]. Designed to ensure consistency and clarity in diagnostic imaging, LI-RADS categorizes lesions based on major and ancillary imaging features, ranging from clearly benign (LR-1) to definitively diagnostic for HCC (LR-5), each associated with explicit management recommendations [12,13]. Beyond aiding clinical decision-making, LI-RADS standardizes radiological reporting, thereby facilitating reliable data collection for research and continuous quality improvement [13].

Prior research has consistently demonstrated improved accuracy and performance of LLMs following fine-tuning in radiological contexts [10,14–18]. For example, a fine-tuned BERT-based model achieved approximately 92% accuracy in selecting appropriate CT imaging protocols, significantly improving trainee performance [14]. Another study demonstrated that a smaller-scale LLM fine-tuned with synthetic labels achieved disease detection performance (micro F1 \approx 0.91) comparable to that of manually annotated datasets [15,16]. Additionally, the radiology-focused RadBERT model fine-tuned on radiological reports consistently outperformed general-purpose models in interpreting imaging data [15]. Soft prompt-based tuning is a fine-tuning method that is implemented using different prompting methods [19].

To the best of our knowledge, no study has specifically evaluated fine-tuned LLMs for the LI-RADS categorization. This study aims to investigate whether “Few-shot” prompting ChatGPT 4.5 improves its accuracy and clinical applicability in classifying hepatic lesions according to LI-RADS criteria.

MATERIALS AND METHODS

Study Design

This cross-sectional observational study assessed the performance of ChatGPT 4.5 in classifying the Liver Imaging Reporting and Data System (LI-RADS) categories and management suggestions from radiology reports using both “Zero-shot” and “Few-shot” prompting methods. This study did not perform fine-tuning of the model at the parameter level. All experiments used prompt engineering (“Few-shot” prompting) to condition the model outputs without modifying the underlying model weights.

The study exclusively utilized fictitious radiology reports without any identifiable patient information; thus, ethics committee

approval was not applicable. The study design adhered to the Standards for Reporting of Diagnostic Accuracy Studies (STARD) [20].

An overview of the study workflow is illustrated in Figure 1.

Preparation of Questions

A total of 50 distinct fictional radiology reports were generated to represent eight different LI-RADS categories: LI-RADS NC, LI-RADS 1, LI-RADS 2, LI-RADS 3, LI-RADS 4, LI-RADS 5, LI-RADS M, and LI-RADS TIV. All reports were independently created by Radiologist 1 (R1)(E.Ç.) without employing ChatGPT. This approach ensured that the generated reports were free from potential bias stemming from the internal training data of ChatGPT or context leakage.

All prepared radiology reports are provided in Supplementary Material 1.

Input-Output Procedure and Evaluation of LLM Performance

Two distinct prompting strategies were applied to evaluate the performance of ChatGPT 4.5. Ten different prompt examples were deliberately selected to span the range of LI-RADS categories and management recommendations, maximizing category diversity. They were not chosen to optimize the model’s performance on the test set. The most inclusive of these examples, following prompt, was chosen by R1 for “Few-shot” prompting: “Zero-shot” prompting: “I am solving a radiology quiz and

will provide you radiology report sentences. Please act as a radiology professor with 30 years of experience. According to LI-RADS classification, please assign an appropriate category and management suggestion for each radiology report.”

“Few-shot” prompting: The following prompt, including seven illustrative examples of LI-RADS categories, was utilized:

“I am solving a radiology quiz and will provide you radiology report sentences. Please act as a radiology professor with 30 years of experience. I will share with you the definitions of LI-RADS categories with corresponding examples:

Example 1: MRI of the liver demonstrates a 2.0 cm focal lesion in segment VII of the cirrhotic liver. The lesion is inadequately visualized due to significant motion artifacts affecting all sequences, limiting accurate assessment of enhancement patterns in arterial, portal venous, and delayed phases. The lesion’s intrinsic characteristics, including T1 and T2 signal intensities and diffusion-weighted imaging, cannot be reliably assessed due to image degradation. LI-RADS NC.

Management: Repeat or alternative imaging in < 3 months.

Example 2: A 1.5 cm lesion in a non-cirrhotic liver appears hypointense on T1-weighted imaging, hyperintense on T2-weighted imaging, and does not show arterial phase hyperenhancement. LI-RADS 1 (Definitely Benign).

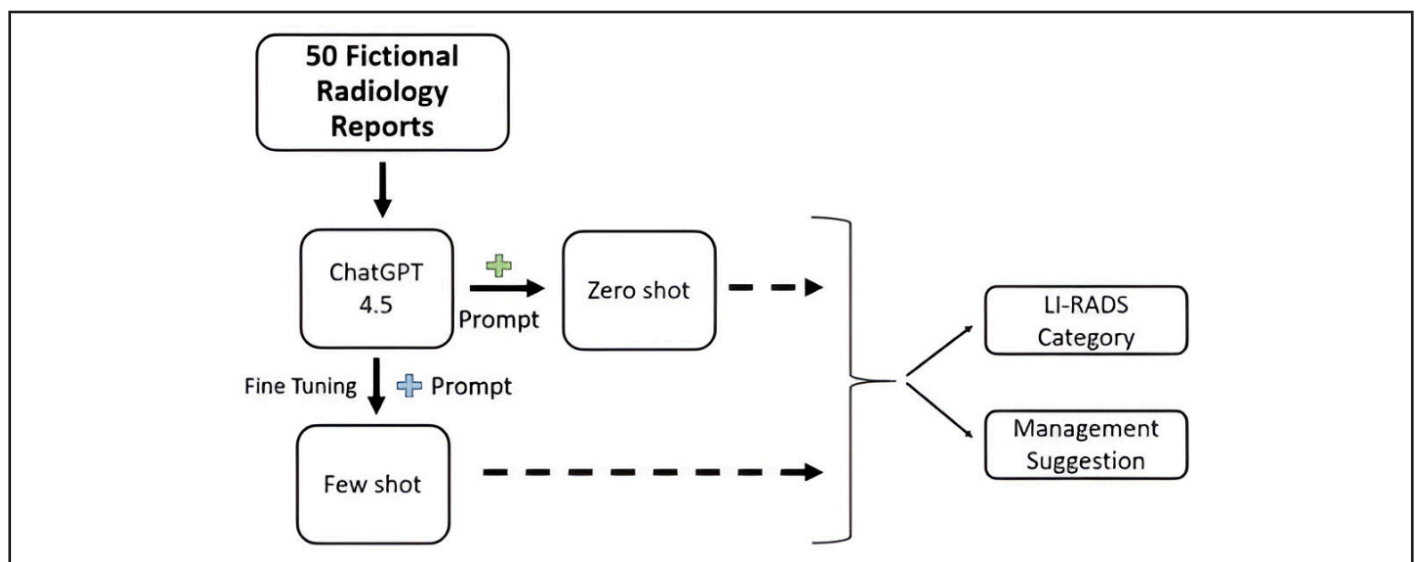


Figure 1. The Workflow of the Study

Management: Return to surveillance in 6 months.

Example 3: A 3 cm lesion in a cirrhotic liver demonstrates T2 hyperintensity with peripheral discontinuous nodular enhancement in the arterial phase and persistent enhancement on delayed phases. LI-RADS 2 (Probably Benign).

Management: Return to surveillance in 6 months or consider repeat diagnostic imaging in 6 months.

Example 4: A 2 cm lesion in a cirrhotic liver shows arterial phase hyperenhancement but lacks washout in the portal venous or delayed phases. There is no capsule or threshold growth. LI-RADS 3 (Intermediate Probability of HCC).

Management: Repeat or alternative diagnostic imaging in 3-6 months.

Example 5: A 2.3 cm lesion in a cirrhotic liver shows arterial phase hyperenhancement, mild washout in the portal venous phase, and an enhancing capsule. LI-RADS 4 (Probably HCC).

Management: Multidisciplinary discussion for tailored workup.

Example 6: A 3 cm lesion in a cirrhotic liver shows arterial phase hyperenhancement, washout in the portal venous phase, an enhancing capsule, and threshold growth. LI-RADS 5 (Definite HCC).

Management: Multidisciplinary discussion for consensus management.

Example 7: A 4 cm lesion in a cirrhotic liver shows irregular rim enhancement in the arterial phase, progressive enhancement in later phases, and restricted diffusion. LI-RADS M (Probably Malignant but Not HCC-Specific).

Management: Multidisciplinary discussion for tailored workup
Example 8: A cirrhotic patient has a 5 cm liver lesion with arterial phase hyperenhancement and an adjacent portal vein thrombus that enhances in the arterial phase. LI-RADS TIV (Tumor in Vein).

Management: Multidisciplinary discussion for tailored workup

Now, please answer the following question and assign the

appropriate LI-RADS category management suggestion for each case that will be given to you”.

The prompts used in this study were designed using a structured approach without iterative adjustments during the study. They utilized role-based contextualization to replicate the thought process and clinical reasoning of an experienced senior radiologist, aiming to enhance clinical applicability and encourage comprehensive differential diagnosis. To minimize bias from inconsistent prompt structures throughout the sessions, a uniform prompt format was consistently employed. In addition, it is important to emphasize that longer few-shot prompts can approach context window limits, potentially affecting model performance, especially in production use cases. However, because this study only aims to reveal the performance difference between “Zero-shot” and “Few-shot” prompting, the effect of the window limit was not evaluated.

These prompts were presented by R1 in March 2025 using OpenAI’s ChatGPT 4.5 (<https://chat.openai.com>) with the default model hyperparameters. Notably, the model was not pretrained on the specific prompts, datasets, or question sets used in this study by the authors before the study.

R1 submitted radiology reports to ChatGPT 4.5 and documented the responses generated. Additionally, R1 evaluated the performance of both “Zero-shot” and “Few-shot” prompting approaches by categorizing “category” and “management suggestion” as either correct (1) or incorrect (0). This binary scoring system was adopted because each case had a clearly defined, single, correct diagnosis from the dataset, facilitating objective assessment. When two different management suggestions were clinically appropriate, both were scored as correct.

The Background of Radiologists and Evaluation of the Performance of Radiologist

Two board-certified radiologists-(EDiR)-(R1 and Radiologist 2), each with seven years of general radiology experience, participated in this study.

Radiologist 2 (R2)(Y.C.G.) independently evaluated the same radiology reports and gave LI-RADS category and management suggestion for each case. R2 evaluated these reports using R1’s computer without internet access to minimize potential evaluation bias.

Statistical Analysis

Descriptive statistics were calculated, including the mean, median, interquartile range (IQR), frequency, and percentage values. The normality of the variable distribution was assessed using the Kolmogorov–Smirnov test. Given the characteristics of our data distribution, non-parametric tests were used to compare the quantitative data. McNemar’s test was applied to compare the proportion of correct responses across different questions. The consistency assessment between “Zero-shot” and “Few-shot” prompting methods was assessed using Cohen’s kappa. Statistical analyses were conducted using SPSS 26.0 (IBM, USA), with statistical significance set at $p < 0.05$.

RESULTS

A total of 50 radiology reports were evaluated using two different prompting methods (“Zero-shot” and “Few-shot”) to classify LI-RADS categories and management suggestions with ChatGPT 4.5.

ChatGPT 4.5 correctly classified the LI-RADS category in 42 out of 50 reports (84.0%), with 8 reports incorrectly classified (16.0%) using the “Zero-shot” prompting method. When employing the “Few-shot” prompting method, the model’s performance improved, achieving correct classification for 46 out of 50 reports (92.0%) and incorrectly classifying four reports (8.0%) (Table 1).

Table 1. Accuracy Comparison of ChatGPT and Radiologist for only Category Selection: “Zero-shot” vs.” Few-shot” Prompting

	Zero-shot (n=50)	Few-shot (n=50)	R2 (n=50)
	Category		
True (%)	42 (84%)	46 (92%)	41 (82%)
False (%)	8 (16%)	4 (8%)	9 (18%)

Information: p- values were derived from McNemar test,
R2: Radiologist 2

With the “Zero-shot” prompting method, the model accurately suggested management strategies in 35 out of 50 reports (70.0%), while incorrectly suggested 15 reports (30.0%). Conversely, using “Few-shot” prompting method, the accuracy increased, correctly recommended management suggestions 42 out of 50 reports (84.0%) and incorrectly suggested 8 reports (16.0%) (Table 2,3).

The accuracies of ChatGPT 4.5 and radiologists’ LI-RADS category selection and management suggestions are presented in

Tables 1 and 2.

Table 2. Accuracy Comparison of ChatGPT and Radiologist for only Management Suggestion: “Zero-shot” vs. “Few-shot” Prompting

	Zero-shot (n=50)	Few-shot (n=50)	R2 (n=50)
	Management Suggestion		
True (%)	35 (70%)	42 (84%)	30 (60%)
False (%)	15 (30%)	8 (16%)	20 (40%)

Information: p- values were derived from McNemar test,
R2: Radiologist 2

Table 3. Accuracy Comparison of ChatGPT for Management Suggestion Across LI-RADS Categories: “Zero-shot” vs. “Few-shot” Prompting

LI-RADS Category	Accuracy with Zero-shot Prompting (%)	Accuracy with Few-shot Prompting (%)
1	25.0%	75.0%
2	75.0%	100.0%
3	66.7%	66.7%
4	75.0%	75.0%
5	100.0%	100.0%
M	100.0%	100.0%
NC	20.0%	80.0%
TIV	57.1%	71.4%

The consistency of responses between “Zero-shot” and “Few-shot” prompting for LI-RADS category classification is $\kappa = 0.627$. Similarly, for management suggestions, $\kappa = 0.615$, suggesting moderate-substantial consistency.

There was no statistically significant difference between “Zero-shot” and “Few-shot” prompting methods for classifying LI-RADS category ($p = 0.125$). However, management suggestions were significantly improved with “Few-shot” prompting method ($p = 0.016$).

R2 demonstrated an accuracy of 82.0% in correctly assigning the LI-RADS categories, correctly classifying 41 out of 50 reports, while misclassifying the remaining 9 reports (18.0%). In contrast, the accuracy for appropriate management suggestions were lower, with R2 correctly recommending management suggestions in 30 of the 50 cases (60.0%) and providing incorrect suggestions in 20 cases (40.0%)

A comparison of the performance and accuracy of ChatGPT 4.5 suggestions and LI-RADS category classification are shown in Figure 2 and Table 4. The confusion matrices for management

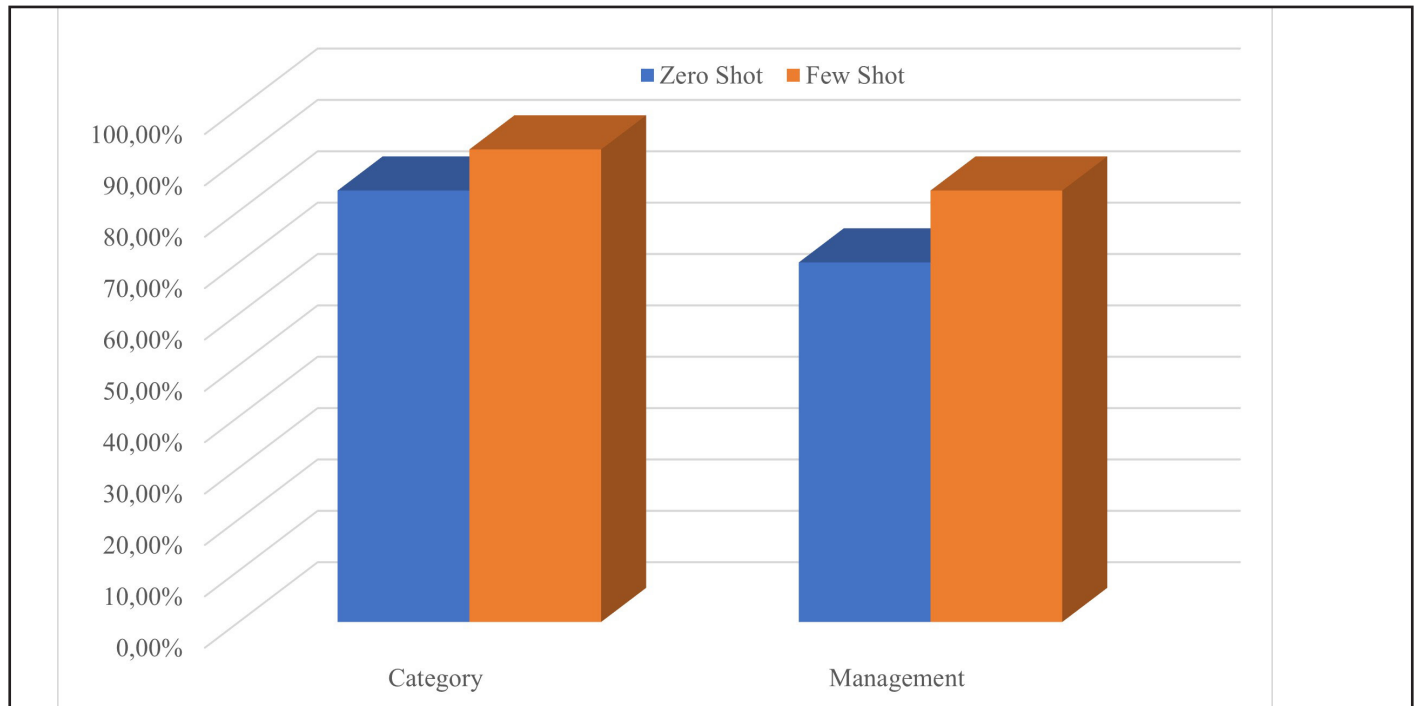


Figure 2. Comparison of “Zero-shot” and “Few-shot” Performances

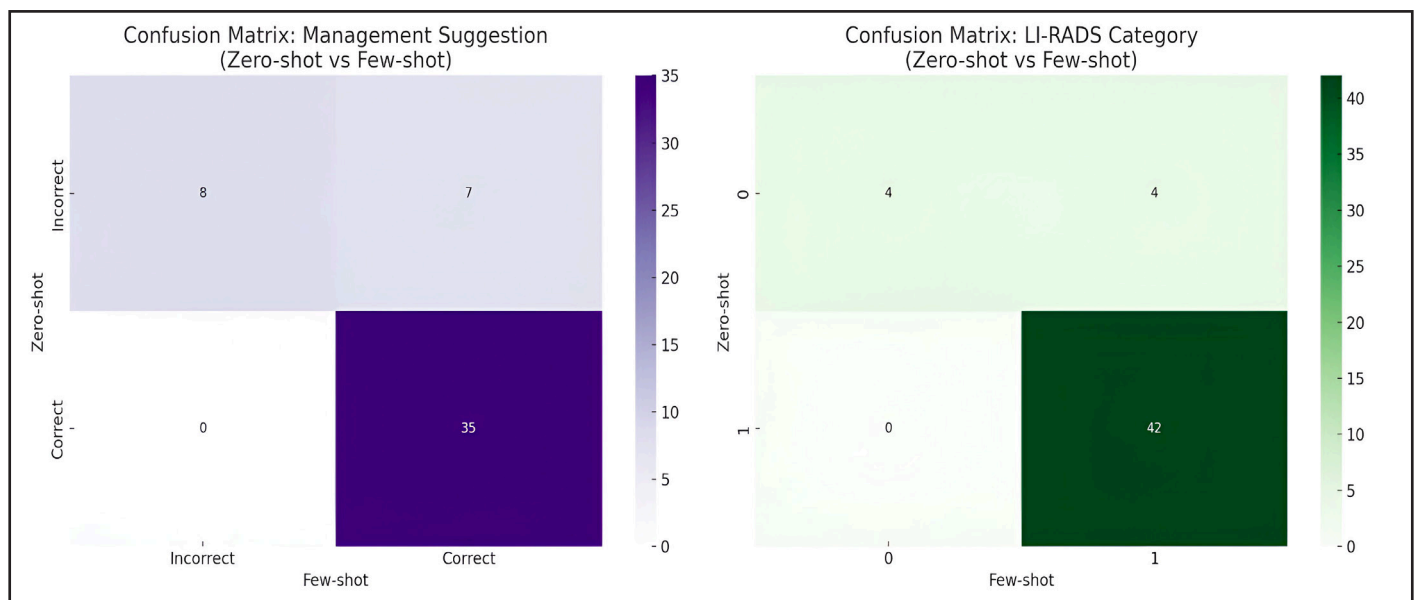


Figure 3. “Zero-shot” vs “Few-shot” Confusion Matrices for Management Suggestion and LI-RADS Category Classification

Table 4. Comparison of the Performances of ChatGPT with Different Prompting Methods and Radiologist

	Zero-shot	Few-shot	R2
	Category/Management Suggestion (p values)		
Zero-shot	-	0.125/0.016	0.517/0.041
Few-shot	0.125/0.016	-	0.107/0.006
R2	0.517/0.041	0.107/0.006	-

Information: p- values were derived from McNemar test,
R2: Radiologist 2

DISCUSSION

The most significant finding of our study is that the “Few-shot” prompting method substantially enhances the patient management suggestions of ChatGPT 4.5. However, the “Few-shot” prompting method slightly improved the model’s accuracy in selecting the LI-RADS categories, but this improvement was not statistically significant. This distinction underscores the potential of “Few-shot” prompting methods to enhance the management suggestions of the model rather than solely improving the diagnostic classification accuracy. Considering that one of the most important aims of LI-RADS in clinical practice is to determine management strategies, it is important to keep in mind that this improvement in the model’s management suggestions will contribute significantly to clinical decision-making and patient management.

Our results are consistent with current research demonstrating that fine-tuned or structured prompted LLMs provide greater utility in synthesizing diagnostic and management recommendations compared to purely classification-based tasks [3,21–26]. For instance, Bhayana et al. [3] evaluated ChatGPT’s performance on a radiology board-style examination and found that while the model performed well on questions assessing basic knowledge and understanding, it struggled with higher-order thinking questions involving the description of imaging findings, calculation, and classification, and concept application. Galan-Cuenca et al. [23] highlighted in their study on COVID-19 chest X-rays that integrating techniques such as data balancing, weighted loss functions, and few-shot learning into Siamese neural networks effectively mitigates data scarcity and imbalance, resulting in measurable performance improvements compared to conventional CNN models in medical imaging. Similarly, Russe et al. [21] underscored the pivotal role of prompt engineering strategies, particularly “Few-shot” and “Zero-shot” prompting, in optimizing LLM interactions in medical tasks, significantly

enhancing output precision and trustworthiness in radiology-specific applications. This suggests that although LLMs can assist in clinical decision-making, their diagnostic classification accuracy may be limited without further refinement.

Another noteworthy result is that ChatGPT 4.5 demonstrates considerable accuracy in determining LI-RADS categories in case-based questions, even in zero-shot settings. Although “Few-shot” prompting marginally improved accuracy, this improvement was not statistically significant. Prior studies have shown that LLMs successfully adhere to oncology guidelines [27–30]. Zhu et al. [27] evaluated multiple LLMs and found that ChatGPT performed the best, with most models exceeding 90% accuracy. In contrast, Coşkun et al. [28] reported suboptimal accuracy of ChatGPT in answering prostate cancer-related questions. While existing research has explored LLMs’ proficiency in various cancers, no prior study has compared multiple LLMs in LI-RADS. Our study uniquely benchmarks ChatGPT 4.5 against radiologists, offering insights into its relative efficacy in the LI-RADS guideline.

ChatGPT 4.5, the newest model of OpenAI, was released on February 27, 2025. There is no prior study evaluating the performance of this model in both radiology and medicine, making this study unique in assessing its performance in terms of LI-RADS, which is a valuable radiological guideline. ChatGPT 4.5 has comparable proficiency in LI-RADS categorization to that of a general radiologist with seven years of experience in general radiology. The model shows strong potential in LI-RADS categorization, an essential step in guiding patient management, and produces management suggestions that outperform those of the radiologist when supported by the “Few-shot” prompting method. This may be due to the fact that radiologists focus primarily on diagnosis and risk assessment of lesions rather than patient management. ChatGPT also provides appropriate management suggestions, which are of great importance in clinical practice. This result demonstrates the importance of the contribution that LLMs can provide to radiologists, not only for diagnosis or image evaluation but also for patient management suggestions, which is the most important decision and component of radiological reporting.

Limitations

This study has several limitations. First, the number of questions used in the study was limited, and all were case-based questions. Further studies are needed to investigate the text- and image-based

performance of ChatGPT on LI-RADS. Thus, the performance and potential of the model for integration into radiology practice can be better demonstrated in future studies.

Second, the radiology reports used in this study are based on fictional cases designed to standardize the input and avoid privacy concerns. Although this approach improves experimental control, it may not fully capture the complexity, ambiguity, or variability of real-world reports. Future studies should focus on using real radiology to represent the real-world capabilities of ChatGPT.

Third, this study did not assess inter-session variability in ChatGPT responses. LLMs can exhibit variability in their output, even with identical prompts across sessions. Future studies should include replicating prompting sessions to quantify and report reproducibility metrics.

Fourth, the reproducibility and temporal stability of ChatGPT responses could affect its performance. Like each LLM, ChatGPT has inherent stochasticity, whereby identical prompts can elicit subtly divergent responses across different sessions. Moreover, commercially deployed LLMs, such as ChatGPT, undergo periodic updates that may alter their underlying parameters and output behavior over time. This evolving nature introduces challenges in establishing consistent, verifiable, and clinically reliable decision support systems. To ensure clinical practice translatability and maintain rigorous validation standards, future investigations should systematically assess intra- and inter-session variability and focus on quantifying temporal drift across model versions.

Lastly, ChatGPT responses were compared with those of a single radiologist with seven years of experience, which may limit the generalizability of the observed differences in management suggestion performance. Multicenter, multi-participant future studies are needed to support our findings.

CONCLUSION

In conclusion, “Few-shot” prompting improves ChatGPT’s performance in providing clinically relevant management recommendations within LI-RADS, emphasizing the value of this prompting approach in radiology practice. While improvements in categorization accuracy remain modest, these improvements bridge the gap between diagnosis and clinical decision-making. Further studies with real clinical datasets and refined prompt

strategies may contribute to enhancing the diagnostic and decision-making capabilities such as management suggestion, of LLMs.

Acknowledgments: The content of the publication is entirely the authors’ responsibility, and the authors examined and edited it as necessary. Each author states that the submitted letter, either in full or in part, has not been previously published or is not being assessed for publication as an original article in either printed form or as digital media.

Data Availability: All data supporting the findings of this study are available within the paper and its Supplementary Materials.

Conflict of Interest: The authors declare that this letter was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest. Ethical approval is not applicable to this study.

Funding: No funding was received for this study.

Authors' Contribution: Eren Çamur: Conceptualization, Data curation, Formal, analysis, Methodology, Writing- original draft, Writing- review&editing.

Yasin Celal Güneş: Data curation, Writing- review&editing.

REFERENCES

- [1] Akinci D’Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, Cuocolo R, Cannella R, Koçak B (2024) Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol.* 30(2):80–90. <https://doi.org/10.4274/dir.2023.232417>
- [2] Kim S, Lee CK, Kim SS (2024) Large Language Models: A Guide for Radiologists. *Korean J Radiol.* 25(2):126–133. <https://doi.org/10.3348/kjr.2023.0997>
- [3] Bhayana R (2024) Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. *Radiology.* 310(1). <https://doi.org/10.1148/radiol.232756>
- [4] Zaki HA, Aoun A, Munshi S, Abdel-Megid H, Nazario-Johnson L, Ahn SH (2024) The Application of Large Language Models for Radiologic Decision Making. *J*

- Am Coll Radiol. 21(7):1072–8. <https://doi.org/10.1016/j.jacr.2024.01.007>
- [5] Gomez E (2025) Large Language Models with Image Processing Capabilities: An Inevitable yet Undetermined Presence in Radiology Practice and Education. *Acad Radiol.* 32(5):3103–5. <https://doi.org/10.1016/j.acra.2025.03.027>
- [6] Busch F, Hoffmann L, dos Santos DP, Makowski MR, Saba L, Prucker P, Hadamitzky M, Navab N, Kather JN, Truhn D, Cuocolo R, Adams LC, Bressen KK (2024) Large language models for structured reporting in radiology: past, present, and future. *Eur Radiol.* 35:2589–602. <https://doi.org/10.1007/s00330-024-11107-6>
- [7] Salam B, Stüwe C, Nowak S, Sprinkart AM, Theis M, Kravchenko D, Mesropyan N, Dell T, Endler C, Pieper CC, Kuetting DL, Luetkens JA, Isaak A (2025) Large language models for error detection in radiology reports: a comparative analysis between closed-source and privacy-compliant open-source models. *Eur Radiol.* 35:4549–57. <https://doi.org/10.1007/s00330-025-11438-y>
- [8] Nakaura T, Ito R, Ueda D, Nozaki T, Fushimi Y, Matsui Y, Yanagawa M, Yamada A, Tsuboyama T, Fujima N, Tatsugami F, Hirata K, Fujita S, Kamagata K, Fujioka T, Kawamura M, Naganawa S (2024) The impact of large language models on radiology: a guide for radiologists on the latest innovations in AI. *Jpn J Radiol.* 42(7):685–96. <https://doi.org/10.1007/s11604-024-01552-0>
- [9] Serapio A, Chaudhari G, Savage C, Lee YJ, Vella M, Sridhar S, Schroeder JL, Liu J, Yala A, Sohn JH (2024) An open-source fine-tuned large language model for radiological impression generation: a multi-reader performance study. *BMC Med Imaging.* 24(1):254. <https://doi.org/10.1186/s12880-024-01435-w>
- [10] Chen L, Teotia R, Verdone A, Cardall A, Tyagi L, Shen Y, Chopra S (2024) Fine-Tuning In-House Large Language Models to Infer Differential Diagnosis from Radiology Reports. <https://doi.org/10.48550/arXiv.2410.09234>
- [11] American College of Radiology. (n.d.) Liver Imaging Reporting & Data System (LI-RADS). Available from <https://www.acr.org/Clinical-Resources/Clinical-Tools-and-Reference/Reporting-and-Data-Systems/LI-RADS>
- [12] Kamaya A, Fetzer DT, Seow JH, Burrowes DP, Choi HH, Dawkins AA, Fung C, Gabriel H, Hong CW, Khurana A, McGillen KL, Morgan TA, Sirlin CB, Tse JR, Rodgers SK (2024) LI-RADS US Surveillance Version 2024 for Surveillance of Hepatocellular Carcinoma: An Update to the American College of Radiology US LI-RADS. *Radiology.* 313(3):e240169. <https://doi.org/10.1148/radiol.240169>
- [13] Choi SH, Fowler KJ, Chernyak V, Sirlin CB (2024) LI-RADS: Current Status and Future Directions. *Korean J Radiol.* 25(12):1039–46. <https://doi.org/10.3348/kjr.2024.0161>
- [14] Shi Y, Shu P, Liu Z, Wu Z, Ren H, Li Q, Liu T, Liu N, Li X (2024) MGH Radiology Llama: A Llama 3 70B Model for Radiology. <https://doi.org/10.48550/arXiv.2408.11848>
- [15] Kanemaru N, Yasaka K, Fujita N, Kanzawa J, Abe O (2024) The Fine-Tuned Large Language Model for Extracting the Progressive Bone Metastasis from Unstructured Radiology Reports. *J Imaging Inform Med.* 38(2):865–72. <https://doi.org/10.1007/s10278-024-01242-3>
- [16] Yasaka K, Kanzawa J, Kanemaru N, Koshino S, Abe O (2024) Fine-Tuned Large Language Model for Extracting Patients on Pretreatment for Lung Cancer from a Picture Archiving and Communication System Based on Radiological Reports. *J Imaging Inform Med.* 38(1):327–34. <https://doi.org/10.1007/s10278-024-01186-8>
- [17] Kanemaru N, Yasaka K, Okimoto N, Sato M, Nomura T, Morita Y, Katayama A, Kiryu S, Abe O (2025) Efficacy of Fine-Tuned Large Language Model in CT Protocol Assignment as Clinical Decision-Supporting System. *J Digit Imaging. Inform. med.* <https://doi.org/10.1007/s10278-025-01433-6>
- [18] Martín-Noguerol T, López-Úbeda P, Luna A (2024) Large language models in Radiology: The importance of fine-tuning and the fable of the luthier. *Eur J Radiol.* 178:111627. <https://doi.org/10.1016/j.ejrad.2024.111627>
- [19] SadraeiJavaeri M, Asgari E, McHardy AC, Rabiee HR (2024) SuperPos-Prompt: Enhancing Soft Prompt Tuning of Language Models with Superposition of Multi Token Embeddings. Available from <https://arxiv.org/pdf/2406.05279v1>
- [20] Bossuyt PM, Reitsma JB, Bruns DE, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HCW, Kressel HY,

- Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF (2015) STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *Radiology*. 277(3):826–32. <https://doi.org/10.1148/radiol.2015151516>
- [21] Russe MF, Reiser M, Bamberg F, Rau A (2024) Improving the use of LLMs in radiology through prompt engineering: from precision prompts to zero-shot learning. *Rofo*. 196(11). <https://doi.org/10.1055/a-2264-5631>
- [22] Nayem J, Hasan SS, Amina N, Das B, Ali MS, Ahsan MM, Raman S (2023) Few Shot Learning for Medical Imaging: A Comparative Analysis of Methodologies and Formal Mathematical Framework. In: *Data Driven Approaches Med Imaging*.
- [23] Galán-Cuenca A, Gallego AJ, Saval-Calvo M, Pertusa A (2024) Few-shot learning for COVID-19 chest X-ray classification with imbalanced data: an inter vs. intra domain study. *Pattern Anal Appl*. 27(3):1–15. <https://doi.org/10.1007/s10044-024-01285-w>
- [24] Fink A, Rau A, Kotter E, Bamberg F, Russe MF (2025) Optimized interaction with Large Language Models: A practical guide to Prompt Engineering and Retrieval-Augmented Generation. *Radiologie*. 65(4). <https://doi.org/10.1007/s00117-025-01416-2>
- [25] Pachetti E, Colantonio S (2024) A systematic review of few-shot learning in medical imaging. *Artif Intell Med*. 156:102949. <https://doi.org/10.1016/j.artmed.2024.102949>
- [26] Kaba E, (2024) Zero-, Single-, and Few-Shot Learning in Large Language Models to Identify Incidental Findings From Radiology Reports. *AJR Am J Roentgenol*. 222(3). <https://doi.org/10.2214/AJR.24.31014>
- [27] Zhu L, Mou W, Chen R (2023) Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *J Transl Med*. 21(1):1–4. <https://doi.org/10.1186/s12967-023-04123-5>
- [28] Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O (2023) Can ChatGPT, an Artificial Intelligence Language Model, Provide Accurate and High-quality Patient Information on Prostate Cancer? *Urology*. 180:35–58. <https://doi.org/10.1016/j.urology.2023.05.040>
- [29] Sorin V, Glicksberg BS, Artsi Y, Barash Y, Konen E, Nadkarni GN, Klang E (2024) Utilizing large language models in breast cancer management: systematic review. *J Cancer Res Clin Oncol*. 150(3):140. <https://doi.org/10.1007/s00432-024-05678-6>
- [30] Alasker A, Alsalamah S, Alshathri N, Almansour N, Alsalamah F, Alghafees M, AlKhamees M, Alsaikhan B (2024) Performance of large language models (LLMs) in providing prostate cancer information. *BMC Urol*. 24(1):177. <https://doi.org/10.1186/s12894-024-01570-0>

How to Cite;

Camur E, Gunes YC (2025) Evaluation of the Performance of ChatGPT 4.5 in LI-RADS Categorization and Management Suggestion: Zero-shot versus Few-shot Prompting Method. *Eur J Ther*. 31(6):403-416. <https://doi.org/10.58600/eurjther2699>

SUPPLEMENTARY MATERIAL**Case 1**

A 30-year-old male patient with known cirrhosis. There is a 10 mm diameter cystic lesion in segment 4 of the liver without contrast enhancement, smoothly circumscribed, not associated with venous structures and similar in size to the patient's examination 6 months ago.

Case 2

A 55-year-old female with hepatitis B cirrhosis. There is a 12 mm well-defined nodule in segment 7 with T1 hyperintensity and T2 hypointensity. No arterial phase hyperenhancement (APHE), no washout, no enhancing capsule. No increase in size compared to previous examination 6 months ago.

Case 3

A 62-year-old male with alcoholic cirrhosis. There is a 15 mm nodule in segment 6 with no arterial phase hyperenhancement (APHE), mild T2 hyperintensity, and restricted diffusion. No washout, no enhancing capsule. No previous imaging available for comparison.

Case 4

A 58-year-old female with HCV cirrhosis. There is an 18 mm nodule in segment 5 with nonrim arterial phase hyperenhancement (APHE), no washout, no enhancing capsule. Increased in size by 30% compared to examination 6 months ago.

Case 5

A 67-year-old male with NASH cirrhosis. There is a 12 mm nodule in segment 8 with nonrim arterial phase hyperenhancement (APHE) and nonperipheral washout. No enhancing capsule. No increase in size compared to examination 6 months ago.

Case 6

A 71-year-old male with HBV cirrhosis. There is a 16 mm nodule in segment 5 with nonrim arterial phase hyperenhancement (APHE) and enhancing capsule. No washout. Increased in size by 40% compared to examination 6 months ago.

Case 7

A 59-year-old female with HCV cirrhosis. There is a 22 mm nodule in segment 6 with nonrim arterial phase hyperenhancement (APHE). No washout, no enhancing capsule. Increased in size by 35% compared to examination 6 months ago.

Case 8

A 64-year-old male with alcoholic cirrhosis. There is a 25 mm nodule in segment 7 with nonrim arterial phase hyperenhancement (APHE) and nonperipheral washout. No enhancing capsule. Increased in size by 20% compared to examination 6 months ago.

Case 9

A 53-year-old female with HBV cirrhosis. There is a 17 mm nodule in segment 8 with nonrim arterial phase hyperenhancement (APHE), nonperipheral washout, and enhancing capsule. No increase in size compared to examination 6 months ago.

Case 10

A 60-year-old male with NASH cirrhosis. There is a 12 mm nodule in segment 5 with nonrim arterial phase hyperenhancement (APHE) and nonperipheral washout. No enhancing capsule. Increased in size by 60% compared to examination 6 months ago.

Case 11

A 72-year-old female with HCV cirrhosis. There is a 28 mm nodule in segment 4 with nonrim arterial phase hyperenhancement (APHE), nonperipheral washout, and enhancing capsule. Increased in size by 55% compared to examination 6 months ago.

Case 12

A 68-year-old male with hepatitis B and cirrhosis. There is a 35 mm mass in segment 7 with rim arterial phase hyperenhancement (APHE), peripheral washout, and delayed central enhancement. Increased in size by 25% compared to examination 6 months ago.

Case 13

A 56-year-old female with cirrhosis. There is a 30 mm infiltrative mass in segment 6 with nonrim arterial phase hyperenhancement (APHE), marked diffusion restriction, and areas of necrosis. No washout, no enhancing capsule. Increased in size by 70% compared to examination 6 months ago.

Case 14

A 77-year-old male with alcoholic cirrhosis. There is a 45 mm mass in segment 8 with targetoid appearance on diffusion-weighted imaging, rim arterial phase hyperenhancement (APHE), and peripheral washout. Increased in size by 40% compared to examination 6 months ago.

Case 15

A 63-year-old female with HCV cirrhosis. There is a 40 mm mass

in segment 5 extending into the right portal vein with enhancing soft tissue visible within the vein lumen. The mass shows nonrim arterial phase hyperenhancement (APHE) and nonperipheral washout. No increase in size of the parenchymal component compared to examination 6 months ago, but new tumor in vein.

Case 16

A 69-year-old male with HBV cirrhosis. There is a 55 mm mass in segment 7 with nonrim arterial phase hyperenhancement (APHE) and nonperipheral washout. Enhancing soft tissue is visible within the right hepatic vein. Increased in size by 30% compared to examination 6 months ago.

Case 17

A 58-year-old female with NASH cirrhosis. There is a 25 mm mass with rim arterial phase hyperenhancement (APHE) and peripheral washout. Enhancing tissue is seen extending into the main portal vein. Increased in size by 35% compared to examination 6 months ago.

Case 18

A 72-year-old male with alcoholic cirrhosis. There is a subtle 15 mm hypodense area in segment 6 that is incompletely characterized due to severe motion artifact during the arterial and portal venous phases. No previous imaging available for comparison.

Case 19

A 60-year-old female with HBV cirrhosis. There is a 20 mm lesion in segment 2 that is incompletely characterized due to omission of the arterial phase during the MRI examination. No previous imaging available for comparison.

Case 20

A 64-year-old male with HCV cirrhosis. There is an 8 mm hyperechoic nodule in segment 7 with no arterial phase hyperenhancement (APHE), no washout, no enhancing capsule. The nodule demonstrates T1 hyperintensity consistent with a siderotic nodule. No increase in size compared to examination 6 months ago.

Case 21

A 57-year-old female with alcoholic cirrhosis. There is a 14 mm nodule in segment 4 with no arterial phase hyperenhancement (APHE), no washout, no enhancing capsule, but with intralesional

fat. No increase in size compared to examination 6 months ago.

Case 22

A 66-year-old male with HBV cirrhosis. There is a 7 mm nodule in segment 8 with nonrim arterial phase hyperenhancement (APHE), no washout, no enhancing capsule. Increased in size by 40% compared to examination 6 months ago.

Case 23

A 58-year-old female with NASH cirrhosis. There is a 15 mm nodule in segment 6 with nonrim arterial phase hyperenhancement (APHE) and enhancing capsule. No washout. Increased in size by 25% compared to examination 6 months ago.

Case 24

A 70-year-old male with HCV cirrhosis. There is a 22 mm nodule in segment 7 with no arterial phase hyperenhancement (APHE) but with nonperipheral washout and enhancing capsule. Increased in size by 45% compared to examination 6 months ago.

Case 25

A 61-year-old female with HBV cirrhosis. There is an 18 mm nodule in segment 5 with nonrim arterial phase hyperenhancement (APHE) and nonperipheral washout. No enhancing capsule. Increased in size by 55% compared to examination 6 months ago.

Case 26

A 74-year-old male with alcoholic cirrhosis. There is a 30 mm nodule in segment 8 with nonrim arterial phase hyperenhancement (APHE), nonperipheral washout, and enhancing capsule. Increased in size by 20% compared to examination 6 months ago.

Case 27

A 59-year-old female with NASH cirrhosis. There is a 60 mm mass in segment 7 with peripheral rim arterial phase hyperenhancement (APHE), restricted diffusion in a targetoid pattern, and delayed central enhancement. Increased in size by 50% compared to examination 6 months ago.

Case 28

A 63-year-old male with HCV cirrhosis. There is a 40 mm infiltrative mass in segment 4 with heterogeneous arterial phase hyperenhancement (APHE) and evidence of tumor extending into the middle hepatic vein with enhancing soft tissue visible within the vein. Increased in size by 65% compared to examination 6 months ago.

Case 29

A 68-year-old female with HBV cirrhosis. There is a 16 mm nodule in segment 6 that is incompletely characterized due to severe respiratory motion artifact affecting all phases of the examination. No previous imaging available for comparison.

Case 30

A 75-year-old male with alcoholic cirrhosis. There is a 5 mm cyst in segment 3 with no arterial phase hyperenhancement (APHE), no washout, no enhancing capsule, showing typical cystic features with water density on CT and fluid signal characteristics on MRI. No increase in size compared to examination 6 months ago.

Case 31

A 51-year-old male with hepatitis C cirrhosis. There is a 6 mm hemangioma in segment 2 with peripheral nodular discontinuous enhancement, progressive centripetal fill-in, and hyperintensity on T2-weighted images. No arterial phase hyperenhancement (APHE), no washout, no enhancing capsule. No increase in size compared to examination 6 months ago.

Case 32

A 47-year-old female with alcoholic cirrhosis. There is a 12 mm geographic area of hepatic fat deposition in segment 5 without mass effect. No arterial phase hyperenhancement (APHE), no washout, no enhancing capsule. No increase in size compared to examination 6 months ago.

Case 33

A 65-year-old male with HBV cirrhosis. There is an 8 mm nodule in segment 7 with T1 hyperintensity and T2 hypointensity consistent with a siderotic nodule. No arterial phase hyperenhancement (APHE), no washout, no enhancing capsule. No increase in size compared to examination 6 months ago.

Case 34

A 59-year-old female with NASH cirrhosis. There is a 14 mm nodule in segment 4 that shows T1 hyperintensity without arterial phase hyperenhancement (APHE), no washout, no enhancing capsule. No increase in size compared to examination 6 months ago.

Case 35

A 62-year-old male with HCV cirrhosis. There is a 16 mm nodule in segment 6 with no arterial phase hyperenhancement (APHE),

mild T2 hyperintensity, and restricted diffusion. No washout, no enhancing capsule. Increased in size by 15% compared to examination 6 months ago.

Case 36

A 54-year-old female with alcoholic cirrhosis. There is a 19 mm nodule in segment 8 with no arterial phase hyperenhancement (APHE) but with hepatobiliary phase hypointensity. No washout, no enhancing capsule. No increase in size compared to examination 6 months ago.

Case 37

A 70-year-old male with HBV cirrhosis. There is a 9 mm nodule in segment 5 with nonrim arterial phase hyperenhancement (APHE) and nonperipheral washout. No enhancing capsule. Increased in size by 20% compared to examination 6 months ago.

Case 38

A 56-year-old female with NASH cirrhosis. There is a 17 mm nodule in segment 7 with nonrim arterial phase hyperenhancement (APHE) but no washout. Enhancing capsule present. Increased in size by 30% compared to examination 6 months ago.

Case 39

A 68-year-old male with HCV cirrhosis. There is a 24 mm nodule in segment 2 with no arterial phase hyperenhancement (APHE) but with nonperipheral washout and enhancing capsule. Increased in size by 35% compared to examination 6 months ago.

Case 40

A 53-year-old female with alcoholic cirrhosis. There is a 10 mm nodule in segment 6 with nonrim arterial phase hyperenhancement (APHE) and nonperipheral washout. No enhancing capsule. Increased in size by 70% compared to examination 6 months ago.

Case 41

A 66-year-old male with HBV cirrhosis. There is a 16 mm nodule in segment 8 with nonrim arterial phase hyperenhancement (APHE), nonperipheral washout, and enhancing capsule. No increase in size compared to examination 6 months ago.

Case 42

A 59-year-old female with HCV cirrhosis. There is a 21 mm nodule in segment 5 with nonrim arterial phase hyperenhancement (APHE) and nonperipheral washout. No enhancing capsule. Increased in size by 50% compared to examination 6 months ago.

Case 43

A 73-year-old male with NASH cirrhosis. There is a 32 mm mass in segment 4 with targetoid appearance on diffusion-weighted imaging, rim arterial phase hyperenhancement (APHE), and peripheral washout. No increase in size compared to examination 6 months ago.

Case 44

A 61-year-old female with HBV cirrhosis. There is a 28 mm infiltrative mass in segment 7 with heterogeneous arterial phase hyperenhancement (APHE), marked diffusion restriction, and areas of necrosis. No washout, no enhancing capsule. Increased in size by 60% compared to examination 6 months ago.

Case 45

A 57-year-old male with alcoholic cirrhosis. There is a 36mm mass in segment 6 with rim arterial phase hyperenhancement (APHE), delayed central enhancement, and targetoid appearance in the transitional phase. Increased in size by 45% compared to examination 6 months ago.

Case 46

A 69-year-old female with HCV cirrhosis. There is a 45 mm mass in segment 7 with nonrim arterial phase hyperenhancement (APHE) and nonperipheral washout. Enhancing soft tissue is visible within the right portal vein. Increased in size by 25% compared to examination 6 months ago.

Case 47

A 64-year-old male with HBV cirrhosis. There is a 30 mm infiltrative mass in segment 5 with heterogeneous arterial phase hyperenhancement (APHE). Enhancing soft tissue is visible within the middle hepatic vein. Increased in size by 40% compared to examination 6 months ago.

Case 48

A 71-year-old male with NASH cirrhosis. There is a 25 mm mass in segment 8 with rim arterial phase hyperenhancement (APHE) and enhancing tissue extending into the inferior vena cava. Increased in size by 55% compared to examination 6 months ago.

Case 49

A 56-year-old female with HCV cirrhosis. There is a 17 mm nodule in segment 4 that is incompletely characterized due to significant susceptibility artifacts from surgical clips in proximity to the observation. No previous imaging available for comparison.

Case 50

A 63-year-old male with alcoholic cirrhosis. There is a 22 mm lesion in segment 6 that is incompletely characterized due to premature termination of the MRI exam due to patient claustrophobia, resulting in omission of post-contrast sequences. No previous imaging available for comparison.