Computational and Structural Biotechnology Journal 27 (2025) 2139–2146

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

Leveraging large language models for accurate classification of liver lesions from MRI reports

Daniel Spitzl [a,*,1], Markus Mergen [a,1], Ulrike Bauer [c], Friederike Jungmann [a,d], Keno K. Bressem [b], Felix Busch [a], Marcus R. Makowski [a], Lisa C. Adams [a], Florian T. Gassert [a]

[a] Department of Diagnostic and Interventional Radiology, TUM University Hospital, School of Medicine, Technical University of Munich, Munich, Germany
[b] Department of Cardiovascular Radiology and Nuclear Medicine, German Heart Center Munich, School of Medicine and Health, Technical University of Munich, Munich, Germany
[c] Department for Internal Medicine II, TUM University Hospital, School of Medicine, Technical University of Munich, Munich, Germany
[d] Institute for AI and Informatics, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

ARTICLE INFO

Keywords:
Large language model
Diagnostics
Liver
Clinical support system

ABSTRACT

Background & aims: The rapid advancement of large language models (LLMs) has generated interest in their potential integration in clinical workflows. However, their effectiveness in interpreting complex (imaging) reports remains underexplored and has at times yielded suboptimal results. This study aims to assess the capability of state-of-the-art LLMs to classify liver lesions based solely on textual descriptions from MRI reports, challenging the models to interpret nuanced medical language and diagnostic criteria.
Methods: We evaluated multiple LLMs, including GPT-4o, Deepseek V3, Claude 3.5 Sonnet, and Gemini 2.0 Flash, on a physician-generated fictitious dataset of 88 MRI reports designed to resemble real clinical radiology documentation. The dataset included a representative spectrum of common liver lesions, such as hepatocellular carcinoma, cholangiocarcinoma, hemangiomas, metastases, and focal nodular hyperplasia. Model performance was assessed using micro and macro F1-scores benchmarked against ground truth labels.
Results: Claude 3.5 Sonnet demonstrated the highest diagnostic accuracy among the evaluated models, achieving a micro F1-score of 0.91, outperforming other LLMs in lesion classification.
Conclusion: These findings highlight the feasibility of LLMs for text-based diagnostic support, particularly in resource-limited or high-volume clinical settings. While LLMs show promise in medical diagnostics, further validation through prospective studies is necessary to ensure reliable clinical integration. The study emphasizes the importance of rigorous benchmarking to assess model performance comprehensively.

1. Introduction

Liver lesions present a frequent diagnostic challenge in radiology, with MRI serving as the cornerstone for characterization due to its superior soft-tissue contrast [1]. Accurate differentiation between benign and malignant lesions directly informs treatment decisions, yet even experienced radiologists exhibit considerable interobserver variability in lesion classification [2]. To address these challenges, artificial intelligence (AI)-driven approaches have emerged as promising tools to enhance diagnostic consistency and efficiency [3].

Recent advances in natural language processing (NLP), particularly LLMs like GPT-4o have demonstrated promise in extracting structured insights from clinical text [4]. However, prior work has predominantly focused on Electronic Health Record (EHR) data or imaging (e.g., tumor measurements) rather than descriptive text [5–8]. MRI reports pose unique NLP challenges: they combine standardized lexicon (LI-RADS) with free-text observations, contain ambiguous modifiers (e.g., "atypical hemangioma vs. metastasis"), and often omit explicit diagnostic conclusions [9].

While convolutional neural networks have been extensively studied for lesion classification using pixel data [10], the potential of LLMs to synthesize textual evidence remains underexplored [11,12]. This gap is critical because radiologic reasoning integrates both image patterns and contextual clinical data—a process mirrored in report narratives. Early

* Correspondence to: Klinikum rechts der Isar, Technical University Munich, Ismaningerstr. 22, Munich 81675, Germany.
E-mail address: Danieljan.spitzl@mri.tum.de (D. Spitzl).
[1] These authors contributed equally to the manuscript and share first authorship

https://doi.org/10.1016/j.csbj.2025.05.019
Received 15 April 2025; Received in revised form 16 May 2025; Accepted 18 May 2025
Available online 21 May 2025

attempts at radiology report classification used rule-based systems [13] and conventional machine-learning pipelines. While these methods achieved moderate success, they often fell short in handling the nuanced and ambiguous language typical of free-text clinical narratives. In contrast, LLMs offer the potential to infer diagnostic reasoning directly from unstructured text, adapting to linguistic variability without manual feature engineering. Previous work on the extraction of TNM classification from unstructured radiology reports showed first promising results [14].

Enhanced text-based classification holds the potential to support clinical decision-making—particularly for less experienced clinicians—while improving patient management and ultimately, clinical outcomes. This work contributes to bridging the gap between computational linguistics and diagnostic radiology, responding to recent calls for AI systems that approximate human cognitive processes in medical reasoning [15].

Unlike structured EHR-based NLP approaches, our study utilizes four LLMs – ChatGPT-4o, DeepSeek V3, Claude 3.5 Sonnet and Gemini 2.0 Flash - to specifically examine free-text narratives from radiology reports, a domain characterized by variable syntax, differing conclusion structures, and clinically nuanced descriptors. By benchmarking against radiologist annotations, utilizing an advanced prompt strategy and analyzing failure modes, we provide actionable insights into deploying LLMs for radiology decision augmentation.

2. Materials and methods

Approval from an institutional review board was not required due to the use of nonidentifiable data. This study used 88 fictitious liver MRI reports generated by two radiologists and one gastroenterologist written in german. The reports were systematically created to closely resemble real-world clinical radiology reports in terms of structure, terminology, and level of detail. To ensure consistency and standardization, a predefined template was used to guide the generation of reports. The template was designed based on established radiological reporting standards and adapted to reflect common linguistic patterns observed in real clinical documentation:

1. Liver Morphology: Assessment of shape, and parenchymal architecture, with specific attention to features of cirrhosis, fibrosis, or fatty infiltration.
2. Lesion Characterization: For each hepatic lesion, the following parameters were systematically recorded:

a. Location based on liver segments,
b. Size measured in maximal axial diameter (cm),
c. Signal characteristics on T1-weighted or T2-weighted sequences,
d. Contrast enhancement behavior across dynamic phases (arterial, portal venous, and delayed), including patterns such as early arterial uptake, wash-out, and capsular enhancement where applicable

3. Lymph Node Evaluation where applicable: Documentation of the presence, size, distribution, and morphology of intra-abdominal or retroperitoneal lymph nodes, with criteria for suspicious features based on size and shape.
4. Assessment of Additional Abdominal Organs: Evaluation of the spleen, pancreas, kidneys, adrenal glands, biliary system, and bowel loops, with description of any abnormalities or incidental findings, including cystic or solid lesions, ductal dilatation, or parenchymal changes.

Some reports featured only minor pathologies—later classified as "none" —, such as cysts, to reflect the variability encountered in routine practice. The case distribution reflects a heterogeneous cohort, representative of a balanced representation of different liver lesions in MRI imaging. Each report was independently reviewed by at least one additional radiologist to ensure accuracy and adherence to the standardization criteria. Discrepancies in descriptions or classifications were resolved through consensus discussions among the three medical experts. Ambiguous phrasing was included to simulate real-world variability. The representative example reports were translated into english by a radiologist.

Radiology reports were manually entered into the respective web interfaces of each language model in German, and the corresponding outputs were retrieved directly. All models were evaluated using default settings, without parameter tuning or multiple runs to assess output variability.

A total of four large language models were tested in a zero-shot setting:

ChatGPT 4o (May 2024 version, gpt-4o-2024–05–13)
DeepSeek V3
Claude 3.5 Sonnet (claude-3–5-sonnet-20240620)
Gemini 2.0 Flash (gemini-2.0 flash experimental)
Regular prompting template:

"Given the following MRI text, identify the most likely hepatic lesion."

Advanced prompting template:

"Analyze the provided MRI report of the liver and output matching liver lesions from this list, separated by commas if multiple apply: ["Metastasis", "Hemangioma", "Hepatocellular carcinoma", "Cholangiocarcinoma", "Focal nodular hyperplasia"]. Rules: 1. Ignore all non-liver findings, incidental notes, measurements, and technical terms. 2. Match diagnoses exclusively to the options above (use synonyms if needed, e.g., "HCC" → "Hepatocellular carcinoma"). 3. If multiple terms apply (e.g., coexisting lesions), list all relevant terms alphabetically. 4. Never add explanations, probabilities, or non-listed terms. MRI report:"

The dataset size was chosen to ensure a balanced representation of different liver lesions while maintaining practical feasibility. The inclusion of 88 lesions allowed for a meaningful evaluation of classification performance. To minimize potential bias towards more frequently occurring lesions, an effort was made to achieve a representative distribution of lesions. Accuracy was defined as the proportion of cases in which the exact lesion was correctly predicted against ground truth. The Python packages NumPy (version 1.26.4), pandas (version 2.2.0), scikit-learn (version 1.4.0), statsmodels (version 0.14.1), matplotlib (version 3.8.2), and seaborn (version 0.13.2) were used for data analysis and visualization [16–20]. $P < .05$ was considered indicative of a statistically significant difference. Given data imbalances across datasets, micro and macro F1 scores were calculated.

Per-class F1 score is computed independently for each class based on its precision and recall. This highlights how well the model performs on each individual category.

Micro F1 score was calculated by aggregating the total true positives, false positives, and false negatives across all classes before computing the F1 score. This metric is weighted by class frequency and is more influenced by the model's performance on the most common classes. Macro F1 score is the arithmetic mean of all per-class F1 scores. Each class contributes equally, regardless of its frequency, making this metric sensitive to poor performance on minority classes. 95 % CIs for the F1 scores were estimated using bootstrapping with 1000 repetitions. To assess whether performance differences between models were statistically significant, we applied McNemar's test, given multiple pairwise comparisons, Benjamini-Hochberg correction was applied.

3. Results

A total of 79 liver lesions were described across 88 MRI reports. The distribution of lesion types was as follows: 36.7 % (29/79) hepatocellular carcinoma (HCC), 10.1 % (8/79) cholangiocarcinoma (CCC),

10.1 % (8/79) focal nodular hyperplasia (FNH), 16.5 % (13/79) metastases, and 26.6 % (21/79) hemangiomas. Nine reports contained no or only minor findings. This distribution reflects a diverse array of lesion types that captures the broad range of hepatic pathologies identifiable in hepatic imaging. Two representative example reports per lesion category are presented in Supplementary Table 1.

The performance of various LLMs in predicting liver lesions from MRI reports was assessed using zero-shot prompting in the first experimental setup. The experimental workflow can be seen in Fig. 1A. Claude 3.5 Sonnett demonstrated superior performance, achieving the highest F1 scores both in micro (0.91) and macro (0.78) evaluations and outperformed ChatGPT-4o ($P = 0.0029$) and Gemini 2.0 Flash ($P = 0.003$). This suggests that Claude 3.5 Sonnett was not only effective at handling the overall classification task but also maintained robust performance across individual lesion categories. DeepSeek V3 followed with respectable F1 scores (micro F1: 0.84; macro F1: 0.70), outperforming ChatGPT-4o ($P = 0.098$), which achieved micro and macro F1 scores of 0.76 and 0.63, respectively. In contrast, Gemini 2.0 Flash displayed the lowest performance metrics, with a micro F1 score of 0.69 and a macro F1 score of 0.55. These results highlight significant variability in model performance, potentially attributable to differences in architecture, training data, and language understanding capabilities. Results can be seen in Fig. 1B and Table 1. Pairwise model F1 score differences in performance in detecting different entities as well as statistical differences can be seen in Fig. 1C and 1D.

To maximize the diagnostic accuracy of LLMs and explore potential performance improvements, we implemented advanced prompting strategies in a second experimental phase. Standard prompting may not fully leverage an LLM's reasoning capabilities, particularly when interpreting complex medical language and nuanced diagnostic criteria. By refining prompt structures—such as incorporating step-by-step reasoning, providing context-specific instructions, and using chain-of-thought techniques—we aimed to enhance model comprehension and classification accuracy. This approach was designed to reduce ambiguity, guide the models toward more precise decision-making, and mitigate inconsistencies observed in initial evaluations (Fig. 2 A).

Claude 3.5 Sonnett maintained its position as the top-performing model, achieving a Micro F1 score of 0.91 and a macro F1 score of 0.89 outperforming ChatGPT-4o ($P = 0.013$), DeepSeek V3 ($P = 0.025$) and Gemini 2.0 Flash ($P = 0.000$). This indicates that Claude 3.5 Sonnett's robust performance was not significantly influenced by the prompting strategy, suggesting inherent model strengths in processing medical imaging reports.

ChatGPT-4o, DeepSeek V3, and Gemini 2.0 Flash displayed Micro F1 scores of 0.77, 0.78, and 0.69, respectively, with corresponding macro F1 scores of 0.71, 0.72, and 0.65. While these results show slight improvements compared to the zero-shot prompting phase, the overall trend did not indicate a substantial performance enhancement attributable to advanced prompting. Results can be seen in Fig. 2B, C and D as well as Table 2.

Interestingly, performance shifts were lesion-specific. DeepSeek V3 experienced a 50 % reduction in F1 score for FNH detection, highlighting a potential sensitivity to prompt structure in this context. Conversely, Gemini 2.0 Flash showed notable improvement in detecting metastases, suggesting that specific prompt modifications can enhance performance for particular lesion types. These findings imply that while advanced prompting can influence model performance, its effects are inconsistent and not universally beneficial across all lesion categories.

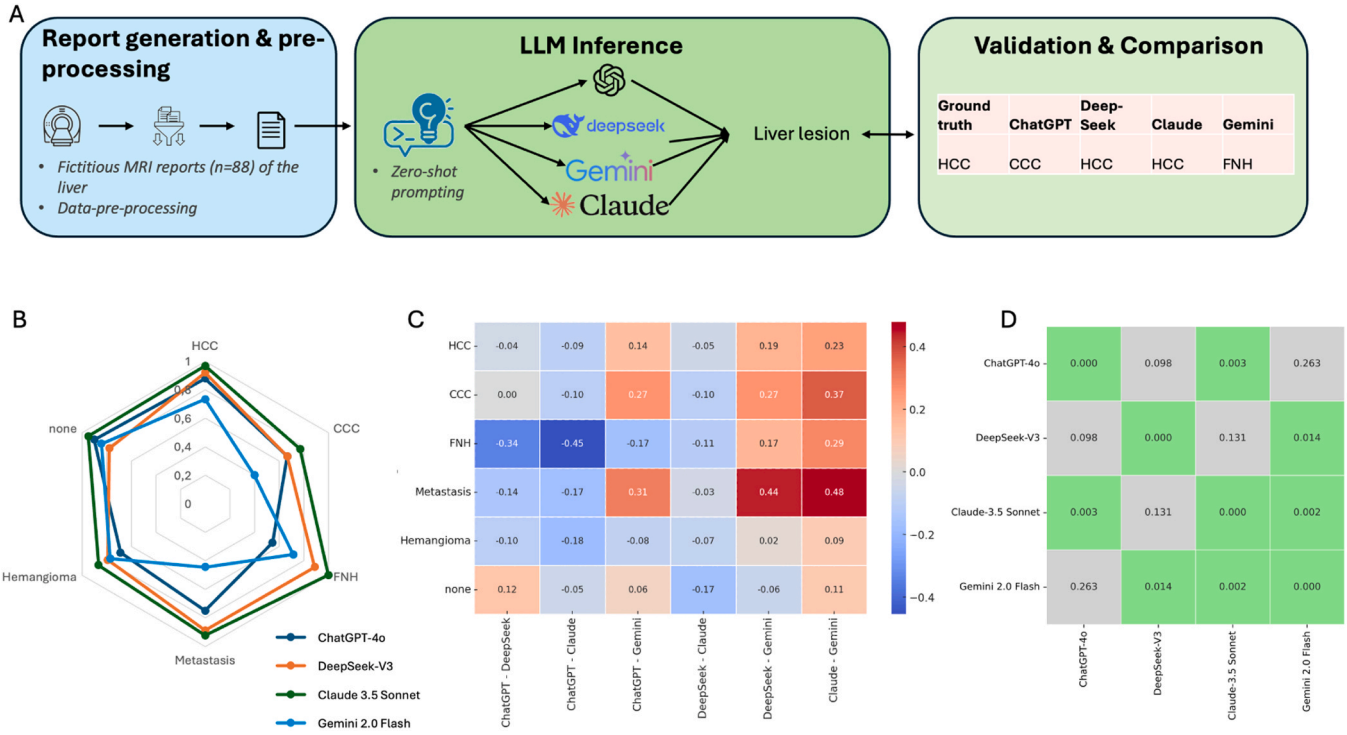

Fig. 1. (A) Schematic overview of the classification workflow for liver MRI reports. After report generation and data pre-processing, four large language models (ChatGPT-4o, DeepSeek V3, Claude 3.5 Sonnet, and Gemini 2.0 Flash) were prompted in a zero-shot manner to predict the lesion type (HCC, CCC, FNH, metastasis, hemangioma, or "none"). Validation against ground-truth labels provides a basis for performance comparison. (B) Radar chart illustrating the F1 scores for liver lesion classification by GPT-4o, DeepSeek V3, Claude 3.5 Sonnet, and Gemini 2.0 Flash across entities. Key findings include Claude 3.5 Sonnett's high F1 scores (100 %) for FNH. (C) Heatmap depicting pairwise performance differences for each lesion category among the four models. Positive (red) cells indicate a higher performance for the first named compared to the second named model, while negative (blue) cells indicate lower performance. (D) Heatmap showing pairwise statistical comparisons between model performances based on McNemar's test, with Benjamini–Hochberg correction applied for multiple testing. Each cell displays the adjusted p-value for the corresponding model pair. Green cells indicate statistically significant differences (adjusted $p < 0.05$), while gray cells indicate non-significant results.

Table 1
Zero-shot prediction performance with regular prompt on liver MRI reports.

Result	ChatGPT-4o	DeepSeek V3	Claude 3.5 Sonnet	Gemini 2.0 Flash
Report finding				
HCC	0.88 (0.786, 0.951)	0.92 (0.842, 0.984)	0.97 (0.912, 1.000)	0.73 (0.609, 0.835)
CCC	0.67 (0.250, 0.933)	0.67 (0.222, 0.909)	0.77 (0.400, 1.000)	0.40 (0.000, 0.750)
FNH	0.55 (0.250, 0.783)	0.89 (0.667, 1.000)	1.00 (1.000, 1.000)	0.71 (0.364, 0.941)
Metastasis	0.75 (0.500, 0.923)	0.89 (0.727, 1.000)	0.92 (0.783, 1.000)	0.44 (0.111, 0.700)
Hemangioma	0.69 (0.417, 0.812)	0.79 (0.593, 0.905	0.86 (0.710, 0.950)	0.77 (0.560, 0.895)
none	0.90 (0.727, 1.000)	0.78 (0.500, 0.952)	0.95 (0.800, 1.000)	0.84 (0.571, 1.000)
MicroF1	0.76 (0.670, 0.852)	0.84 (0.761, 0.920)	0.91 (0.841, 0.966)	0.69 (0.602, 0.784)
MacroF1	0.63 (0.532, 0.809)	0.70 (0.598, 0.872)	0.78 (0.709, 0.940)	0.55 (0.440, 0.731)

Data are F1 scores with 95 %-CI in parentheses. The F1 score was calculated as the harmonic mean of precision (also known as positive predictive value) and recall (also known as sensitivity). The micro F1 score was computed by aggregating the true-positive, false-negative, and false-positive findings across all classes. The macro F1 score was computed by calculating the F1 score for each class individually and then averaging them, giving equal weight to each class regardless of its size. 95 % CIs for the F1 scores were estimated using bootstrapping with 1000 repetitions.

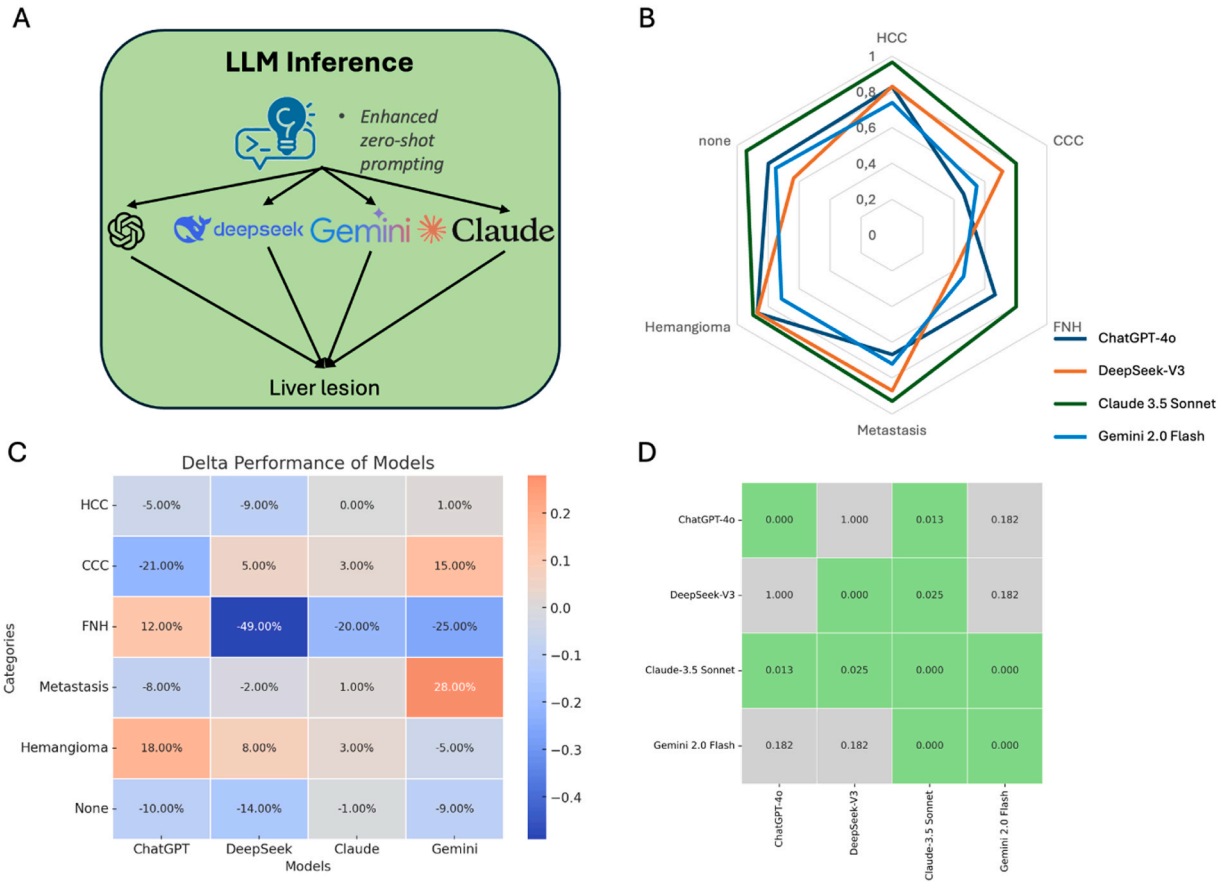

Fig. 2. (A) Schematic representation of the "enhanced zero-shot prompting" approach employed for multi-class liver lesion classification. Four large language models (ChatGPT, DeepSeek, Claude, and Gemini) receive the same input text describing MRI findings and independently predict one of six lesion categories (HCC, CCC, FNH, metastasis, hemangioma, or none). (B) Radar chart illustrating the F1 scores for liver lesion classification by GPT-4o, DeepSeek V3, Claude 3.5 Sonnett, and Gemini 2.0 Flash across entities. Key findings include Claude 3.5 Sonnett's high F1 scores (100 %) for FNH. (C) shows a heatmap depicting the differences in performance for detecting liver lesions with the regular and advanced prompt. Positive (red) cells indicate higher performance for that model-category pair with advanced prompting, while negative (blue) cells denote lower performance. (D) Heatmap showing pairwise statistical comparisons between model performances based on McNemar's test, with Benjamini–Hochberg correction applied for multiple testing. Each cell displays the adjusted p-value for the corresponding model pair. Green cells indicate statistically significant differences (adjusted $p < 0.05$), while gray cells indicate non-significant results.

3.1. Error analysis

To identify the types of mistakes made by the models, we generated confusion matrices for the advanced prompting results. ChatGPT-4o, DeepSeek V3 and Gemini 2.0 Flash frequently misclassified CCC, FNH, metastases, hemangiomas, and reports with no lesions as HCC, suggesting potential bias in the training data toward HCC detection. In contrast, Claude 3.5 Sonnet demonstrated significantly fewer errors, misclassifying only two hemangioma, FNH and CCC as well as one HCC and "none" finding. Across all models, HCC classifications were mostly accurate, though still subject to occasional labeling as metastasis or CCC. Often, DeepSeek V3 assigned a 'none' classification for hemangioma, metastasis, HCC or FNH cases (Fig. 3).

To provide insight into the model behavior beyond aggregate metrics, we present qualitative examples of model performance in Table 3. For each lesion category, one correctly and one incorrectly classified

Table 2
Zero-shot prediction performance with the advanced prompt on liver MRI reports.

Result	ChatGPT-4o	DeepSeek V3	Claude 3.5 Sonnet	Gemini 2.0 Flash
Report finding				
HCC	0.83 (0.721, 0.912)	0.83 (0.721, 0.912)	0.97 (0.909, 1.000)	0.74 (0.615, 0.838)
CCC	0.46 (0.000, 0.800)	0.71 (0.333, 0.941)	0.80 (0.499, 1.000)	0.55 (0.000, 0.857)
FNH	0.67 (0.250, 0.923)	0.40 (0.250, 0.923)	0.80 (0.500, 1.000)	0.46 (0.000, 0.778)
Metastasis	0.67 (0.364, 0.875)	0.87 (0.667, 1.000)	0.93 (0.800, 1.000)	0.72 (0.500, 0.889)
Hemangioma	0.87 (0.696, 0.952)	0.87 (0.727, 0.958)	0.90 (0.757, 0.976)	0.71 (0.474, 0.837)
none	0.80 (0.533, 0.960)	0.64 (0.308, 0.857)	0.94 (0.778, 1.000)	0.75 (0.444, 0.941)
MicroF1	0.77 (0.693, 0.852)	0.78 (0.693, 0.875)	0.91 (0.841, 0.966)	0.69 (0.602, 0.784)
MacroF1	0.71 (0.578, 0.809)	0.72 (0.590, 0.824)	0.89 (0.788, 0.958)	0.65 (0.514, 0.755)

Data are F1 scores with 95 %-CI in parentheses. The F1 score was calculated as the harmonic mean of precision (also known as positive predictive value) and recall (also known as sensitivity). The micro F1 score was computed by aggregating the true-positive, false-negative, and false-positive findings across all classes. The macro F1 score was computed by calculating the F1 score for each class individually and then averaging them, giving equal weight to each class regardless of its size. 95 % CIs for the F1 scores were estimated using bootstrapping with 1000 repetitions.

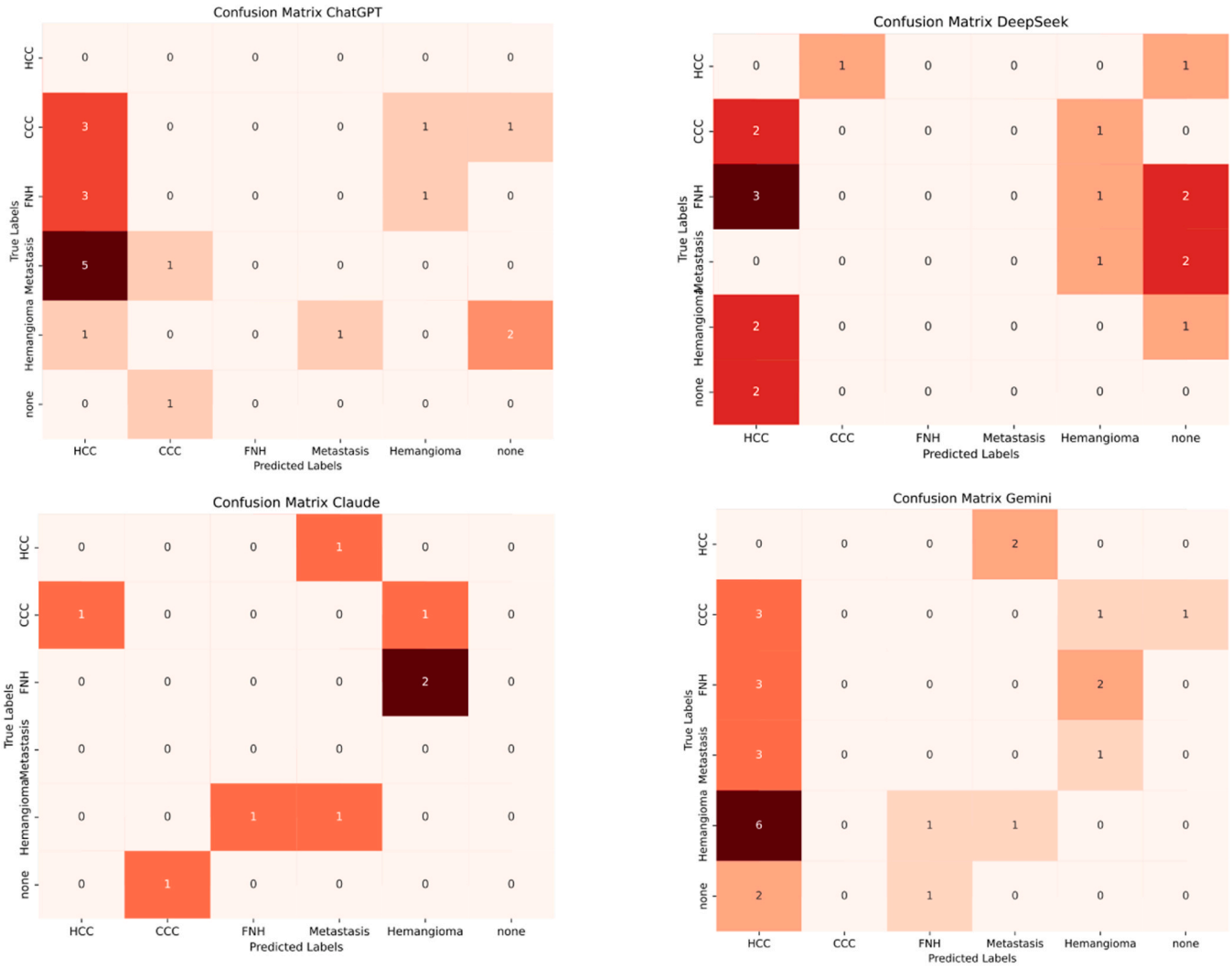

Fig. 3. Confusion matrices comparing the multi-class classification performance of four LLM-based approaches—ChatGPT-4o, DeepSeek V3, Claude 3.5 Sonnet, and Gemini 2.0 Flash—for six hepatic lesion categories: HCC, CCC, FNH, metastasis, hemangioma, and "none". Each matrix plots the true labels on the y-axis against the predicted labels on the x-axis, with color intensity corresponding to the number of instances in each cell (darker shades indicate higher counts). Correct classifications appear on the main diagonal, while off-diagonal cells indicate misclassifications.

case are shown, alongside the corresponding report excerpt, model output, and a brief explanation. Correct predictions typically relied on classical radiological descriptors—for example, arterial hyperenhancement with washout and a capsule for HCC, or a central scar with hepatobiliary contrast retention for FNH. In contrast, misclassifications were often associated with ambiguous phrasing, overlapping features across lesion types (e.g., hypervascularity in both HCC and metastases), or atypical presentations such as diffusion restriction in hemangiomas. Some models also failed to integrate contextual cues, including known primary malignancies or lack of correlates on other sequences. These

Table 3
Representative examples of correct and incorrect model classifications across lesion categories.

Lesion	Report excerpt	Model	Output	Explanation
HCC	A **centrally located arterially hypervascular** lesion, demonstrating **washout in the portal venous phase** and a surrounding **pseudocapsule**	all	HCC	Classic description of an HCC; matches LI-RADS major features (APHE, washout, capsule)
HCC	**diffusion restriction**, **arterial hyperperfusion**, and **venous washout** are noted, centrally **fluid-isointense** components	Gemini 2.0 Flash	Met.	Model was likely misled by "fluid-isointense components" and missed the classical HCC vascular pattern
CCC	**centrally located**, pronounced **peripheral contrast enhancement** with **central non-enhancing areas** across all contrast phases, **diffusion restriction**	all	CCC	Classic description of CCC; key phrases like "peripheral enhancement" and "central necrosis" correctly triggered the classification
CCC	subcapsular **cystic lesion** and **near the hepatic bifurcation** a soft tissue component extending dorsally up to 12 mm, accompanied by a **circumferential contrast-enhancing wall** thickening	ChatGPT−4o	none	Model ignored the solid lesion and focused on the cystic appearance, likely due to ambiguous phrasing and lesion proximity
FNH	**T2-hyperintense, with peripheral contrast uptake** following contrast administration, marked **arterial-phase contrast enhancement with central contrast sparing** corresponding to the fibrotic/scarred areas	all	FNH	Classic FNH description with all hallmark signs: central scar, strong arterial enhancement, and hepatocyte contrast uptake
FNH	**faintly T2-hyperintense lesion**, demonstrate **early arterial enhancement**, in hepatobiliary phase **progressive contrast retention with central scar-like** components is observed	DeepSeek V3	Hemangi-oma	Model likely confused this with hemangioma due to progressive fill-in and hyperintensity, missing the central scar significance
Metastasis	**T2-weighted hyperintense** lesions, showing **diffusion restriction** and **peripheral contrast enhancement**	all	Metastasis	Classical metastatic pattern—multiple DWI-positive lesions with rim enhancement—correctly identified by all models
Metastasis	pancreatic head **neuroendocrine tumor**, Multiple **hypervascular intrahepatic lesions**, with a distinct correlate on **diffusion-weighted imaging**	ChatGPT−4o	HCC	Model overemphasized the hypervascularity and failed to integrate clinical context of known extrahepatic primary
Hemangi-oma	a sharply demarcated **T2-hyperintense** lesion demonstrating **peripheral nodular contrast enhancement** in the arterial phase and **washout** of hepatocyte-specific contrast in the delayed phase	all	Hemangi-oma	Matches classic hemangioma pattern except for atypical delayed washout; still classified correctly likely due to dominant T2 and nodular pattern
Hemangi-oma	**T2-hyperintense, diffusion-restricted** hepatic lesion, peripheral **nodular enhancement** in the portal venous phase with **progressive contrast uptake in the delayed phase**	Claude 3.5 Sonnet	Metastasis	Model was likely misled by diffusion restriction (uncommon in hemangioma) and peripheral pattern suggesting metastasis
none	A **cystic lesion** is noted, **without evidence of contrast enhancement**	all	none	Correctly classified; "cystic," "no enhancement" clearly excludes pathology
none	**faint arterial hyperperfusion** centrally in segment VII **without diffusion restriction or correlating abnormality** on other sequences	Claude 3.5 Sonnet	CCC	Model overemphasized arterial hyperperfusion while ignoring absence of correlates on other sequences, leading to misclassification

This table provides illustrative radiology report excerpts from the synthetic dataset, showcasing one correctly and one incorrectly classified example for each lesion type. Each row includes the ground truth label, the large language model used, its output, and a brief explanation of the model's reasoning or failure mode. Excerpts are limited to the relevant lesion description within each report. These examples highlight both classical diagnostic patterns (e.g., "arterial hyperenhancement with washout" for HCC) and common sources of misclassification, such as ambiguous language, atypical imaging features, or failure to consider clinical context.

findings highlight both the promise and the current limitations of zero-shot large language models in handling complex, nuanced radiological narratives.

4. Discussion

Our findings clearly demonstrate the substantial potential of LLMs to accurately classify liver lesions based solely on MRI report text. Among the models evaluated, Claude 3.5 Sonnett achieved the highest overall performance. This suggests that LLMs can serve as highly valuable tools in augmenting radiologists' workflows. The ability of LLMs to assist in such scenarios may significantly enhance diagnostic efficiency and support clinical decision-making. However, it is important to note that the performance variation observed across different models highlights both the immense promise and the current limitations of LLMs in the nuanced task of radiological text interpretation.

Previous studies focusing on AI-driven medical text classification have primarily concentrated on EHR data or relied on rule-based NLP approaches tailored for radiology reports [5,13]. Our study extends this existing body of work by systematically evaluating state-of-the-art LLMs in a particularly challenging diagnostic task. Unlike structured datasets, we leveraged free-text MRI reports, which often contain complex, ambiguous language, uncertain medical terminology, and diagnoses that are implied rather than explicitly stated [21]. Previous research based on free-text has illustrated the potential of LLMs in a range of radiological classification applications. For example, one study [11] demonstrated that the LI-RADS score can be automatically derived from radiology reports, enhancing consistency in liver lesion evaluations. In another study [14], LLMs were employed to conduct TNM classification for NSCLC by extracting key tumor characteristics from free-text CT reports and translating them into standardized staging information. Similarly, in the context of brain tumor diagnosis [22], LLMs have demonstrated the ability to synthesize complex diagnostic content from radiology reports, effectively supporting clinical decision-making.

In contrast, there are currently no studies that specifically address the classification of liver lesions using LLMs based solely on MRI reports.

The superior performance of Claude 3.5 Sonnett, when compared to models such as DeepSeek V3, ChatGPT-4o, and Gemini 2.0 Flash, suggests that features like larger context windows and refined language comprehension capabilities may play a crucial role in improving classification accuracy within clinical NLP tasks. This finding highlights the evolving sophistication of LLMs and their potential to outperform traditional models in complex, real-world medical applications.

The noticeable performance disparity observed among the evaluated models underscores several key challenges inherent in clinical NLP and radiology report interpretation. Claude 3.5 Sonnett's higher Micro F1 score (0.91) and macro F1 score (0.78) indicate that it generalizes effectively across a diverse range of liver lesion types, showcasing robust adaptability. In contrast, Gemini 2.0 Flash exhibited the lowest performance metrics, particularly struggling to accurately distinguish between HCC and FNH. This discrepancy is likely attributable to differences in the models' training data, specifically regarding the extent of exposure to medical corpora and the fine-tuning strategies employed. Such variability highlights the importance of tailored model training to optimize performance in specialized medical contexts. By utilizing an advanced prompt, we sought to enhance diagnostic accuracy, however the strategy's value was mostly formal rather than interpretive and further improvements may require more sophisticated prompting approaches,

such as multi-step reasoning or expert-augmented feedback.

While our findings highlight the promising potential of large language models (LLMs) in radiological applications, several important limitations must be acknowledged. A key challenge lies in the inherent linguistic variability of MRI reports, which often blend structured classification systems such as LI-RADS with unstructured, free-text observations [11]. This variability is further compounded by differences in individual radiologists' reporting styles, which can introduce inconsistencies and impact model performance [23]. Another significant limitation stems from the fundamental nature of text-based diagnoses. Unlike imaging-based AI systems that directly analyze pixel-level features, LLMs must infer diagnostic meaning from descriptive language—language that frequently lacks the specificity and granularity needed for confident classification [21]. While zero-shot prompting enabled efficient large-scale analysis without the need for extensive task-specific fine-tuning, this approach inherently limits the model's alignment with clinical nuance. Incorporating few-shot learning or domain-adapted prompting in future work may improve both accuracy and contextual understanding.

Moreover, our study did not include a blinded human reader comparison, preventing a direct benchmark of LLM performance against expert radiologists. Including such a comparison would provide valuable context for interpreting the clinical relevance of model outputs and should be prioritized in future research.

Furthermore, generalizability remains an ongoing concern. Although our study utilized a physician-generated fictitious dataset of 88 MRI reports, providing a robust foundation for model evaluation, prospective validation within real-time clinical decision-making environments is essential to ensure model robustness across diverse clinical settings. While these simulated reports were designed to closely emulate real-world structures, terminologies, and complexities, some degree of authenticity may be lost compared to live clinical settings. However, this approach allowed us to (1) create a controlled yet realistic environment for testing initial feasibility and (2) ensure no patient privacy concerns.

The demonstrated ability of LLMs to assist in the classification of liver lesions from MRI reports suggests a wide range of promising applications within clinical decision support systems. In resource-limited settings, these models could serve as invaluable aids to non-specialist clinicians, flagging potentially malignant lesions for further expert review [24], facilitating the triage of complex cases to subspecialists, and supporting retrospective quality assurance processes. Additionally, LLMs may serve as educational tools for junior radiologists by offering structured guidance, highlighting inconsistencies in reporting, and reinforcing the use of standardized interpretation frameworks.

The tool developed in this study enables diagnostic training by generating lesion classifications based solely on textual descriptions, thereby encouraging pattern recognition and reasoning skills in a report-driven context. However, before such systems can be reliably integrated into routine clinical workflows, further model fine-tuning and validation on larger, multi-institutional datasets are essential. Additionally, hybrid AI approaches that combine LLMs' strength in text interpretation with vision models capable of analyzing image data may offer even greater diagnostic reliability. These integrated systems hold the potential to bridge the gap between textual and visual information, supporting more accurate and comprehensive clinical decision-making.

5. Conclusion

This study highlights the emerging and transformative role of LLMs in the field of diagnostics by demonstrating their capability to classify liver lesions using only MRI report text. While Claude 3.5 Sonnett outperformed other evaluated models, the observed performance variability and the inherent challenges associated with textual interpretation underscore the need for continued refinement. Accurate classification is vital, as any misdiagnosis, whether of HCC, cholangiocarcinoma, or benign lesions, can directly affect treatment plans and patient outcomes. Future research should focus on model fine-tuning with diverse, medically relevant datasets, integration with imaging-based AI models, and rigorous real-world clinical validation. These steps are critical to ensure the safe, effective, and robust deployment of LLMs in radiology practice, ultimately enhancing diagnostic workflows and patient outcomes across various healthcare settings.

CRediT authorship contribution statement

Florian T. Gassert: Writing – review & editing, Writing – original draft, Supervision, Methodology, Data curation, Conceptualization. **Lisa C. Adams:** Writing – review & editing, Conceptualization. **Keno K. Bressem:** Writing – review & editing, Conceptualization. **Friederike Jungmann:** Writing – review & editing, Validation, Methodology. **Marcus R. Makowski:** Writing – review & editing. **Felix Busch:** Writing – review & editing, Conceptualization. **Markus Mergen:** Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Daniel Spitzl:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ulrike Bauer:** Writing – review & editing, Formal analysis.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used ChatGPT in order to to improve readability and language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication

Funding statement

None

Declaration of Competing Interest

None

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.05.019.

References

[1] Gatti M, et al. Benign focal liver lesions: the role of magnetic resonance imaging. World J Hepatol 2022;14(5):923–43.
[2] Cerny M, et al. LI-RADS for MR imaging diagnosis of hepatocellular carcinoma: performance of major and ancillary features. Radiology 2018;288(1):118–28.
[3] Ying H, et al. A multicenter clinical AI system study for detection and diagnosis of focal liver lesions. Nat Commun 2024;15(1):1131.
[4] Menezes MCS, et al. The potential of Generative Pre-trained Transformer 4 (GPT-4) to analyse medical notes in three different languages: a retrospective model-evaluation study. Lancet Digit Health 2025;7(1):e35–43.
[5] Bhattarai K, et al. Leveraging GPT-4 for identifying cancer phenotypes in electronic health records: a performance comparison between GPT-4, GPT-3.5-turbo, Flan-T5, Llama-3-8B, and spaCy's rule-based and machine learning-based methods. JAMIA Open 2024;7(3):ooae060.
[6] Ford E, et al. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inf Assoc 2016;23(5):1007–15.
[7] Wu J, et al. Radiological tumor classification across imaging modality and histology. Nat Mach Intell 2021;3:787–98.
[8] Oh Y, et al. LLM-driven multimodal target volume contouring in radiation oncology. Nat Commun 2024;15(1):9186.
[9] Moura Cunha G, et al. Up-to-date role of CT/MRI LI-RADS in hepatocellular carcinoma. J Hepatocell Carcinoma 2021;8:513–27.
[10] Yasaka K, et al. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. Radiology 2018;286(3):887–96.

[11] Fervers P, et al. ChatGPT yields low accuracy in determining LI-RADS scores based on free-text and structured radiology reports in German language. Front Radio 2024;4:1390774.
[12] Matute-Gonzalez M, et al. Utilizing a domain-specific large language model for LI-RADS v2018 categorization of free-text MRI reports: a feasibility study. Insights Imaging 2024;15(1):280.
[13] Wagholikar A, et al. Automated classification of limb fractures from free-text radiology reports using a clinician-informed gazetteer methodology. Austral Med J 2013;6(5):301–7.
[14] Lee JE, et al. Lung cancer staging using chest CT and FDG PET/CT free-text reports: comparison among three ChatGPT large language models and six human readers of varying experience. AJR Am J Roentgenol 2024;223(6):e2431696.
[15] Rajpurkar P, et al. AI in health and medicine. Nat Med 2022;28(1):31–8.
[16] Harris CR, MK, van der Walt SJ, et al. Array programming with NumPy. Nature 2020;585(7825):357–62.
[17] McKinney W. Data structures for statistical computing in Python. In: van der Walt S, MJ, editor. Proceedings of the 9th Python in Science Conference; 2010. p. 56–61.
[18] Pedregosa F VG, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12(85):2825–30.
[19] Seabold S., P.J., statsmodels: Econometric and statistical modeling with Python. Proceedings of the 9th Python in Science Conference, 2010.
[20] Waskom M. Seaborn: statistical data visualization. J Open Source Softw 2021;6.
[21] Larson DB, et al. Improving consistency in radiology reporting through the use of department-wide standardized structured reporting. Radiology 2013;267(1):240–50.
[22] Kanzawa J, et al. Automated classification of brain MRI reports using fine-tuned large language models. Neuroradiology 2024;66(12):2177–83.
[23] Shinagare AB, et al. Radiologist preferences, agreement, and variability in phrases used to convey diagnostic certainty in radiology reports. J Am Coll Radio 2019;16(4 Pt A):458–64.
[24] Driver CN, et al. Artificial intelligence in radiology: a call for thoughtful application. Clin Transl Sci 2020;13(2):216–8.