Original Paper

Evaluating Multiple Input Strategies of Large Language Models for Gallbladder Polyps on Ultrasound: Comparative Study

Lin Jiang, MD; Jiaqian Yao, MD; Zebang Yang, MD; Fuqiu Tang, MM; Xin Zheng, MD; Xiaoer Zhang, MD; Xiaoyan Xie, MD, PhD; Ming Xu[*], MD, PhD; Tongyi Huang[*], MD

Department of Medical Ultrasonics, Institute of Diagnostic and Interventional Ultrasound, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Tongyi Huang, MD
Department of Medical Ultrasonics
Institute of Diagnostic and Interventional Ultrasound
The First Affiliated Hospital, Sun Yat-sen University
58 Zhongshan Er Road
Guangzhou, 510080
China
Phone: 86 020 87755766
Email: huangty26@mail.sysu.edu.cn

Abstract

Background: Gallbladder polyps have a high prevalence and are predominantly benign lesions, often detected via ultrasound. They impose diagnostic burdens on radiologists while generating substantial patient demand for report interpretation. Benign polyps include nonneoplastic polyps without malignant potential and premalignant adenomas that require cholecystectomy. Current guidelines recommending surgery for polyps ≥1.0 cm may lead to unnecessary interventions. Advanced multimodal large language models (LLMs) such as ChatGPT-4o (OpenAI) and Claude 3.5 Sonnet (Anthropic PBC) demonstrate emerging capabilities in medical image analysis. Implementing LLMs in gallbladder polyp ultrasound evaluation can potentially alleviate radiologists' workload, provide patient-accessible consultation platforms, and even reduce overtreatment.

Objective: We aimed to analyze the feasibility and conduct an early-stage evaluation of using LLMs for differentiating between adenomatous and nonneoplastic gallbladder polyps (≥1.0 cm) based on ChatGPT-4o and Claude 3.5 Sonnet, compared to assessments by radiologists and the guideline.

Methods: Ultrasound images and reports of gallbladder polyps ≥1.0 cm with pathology were retrospectively collected from a hospital between January 2011 and January 2022. LLM performance was evaluated using three input strategies: (1) direct image analysis (LLMs-image), (2) feature-based text analysis (LLMs-text), and (3) scoring model-based text analysis (LLMs-model). Both intra- and interreader agreement and diagnostic performance of LLMs were evaluated for all three strategies. The diagnostic performance metrics—including sensitivity, specificity, accuracy, area under the receiver operating characteristic curve, and unnecessary resection rate of nonneoplastic polyps of LLMs in the three strategies were compared with the guideline. Additionally, the strategy LLMs-model was specifically compared with radiologists using the same scoring system (strategy readers-model).

Results: This study included 223 patients (aged 18-72 years; 132/223, 59.2% female) as the initial cohort, with 48 adenomatous polyps and 175 nonneoplastic polyps. The external test set comprised 100 patients. The intrareader agreement coefficients for strategy LLMs-model were significantly higher than those for strategy LLMs-image and LLMs-text (all *P*<.01). The interreader agreement of the three diagnostic strategies was ranked as LLMs-model>LLMs-text>LLMs-image. The sensitivity of strategies LLMs-image and LLMs-text was significantly lower than that of the guideline (all *P*<.001). When applying a scoring model (readers/LLMs-model strategy), both radiologists and the LLMs achieved a significantly higher accuracy compared to the guideline (0.34, 0.35, and 0.34 vs 0.22, all *P*<.01), and the unnecessary resection rate of nonneoplastic polyps was significantly lower (82%, 83%, and 83% vs 100%, all *P*<.01), while the sensitivity was comparable to the guideline (0.94, 0.98, and 0.98 vs 1.00, all *P*>.05). All diagnostic performance indicators for GPT-model and Claude-model were not significantly different from those of radiologists (all *P*>.05).

Conclusions: The ability of LLMs to recognize and interpret medical images requires further improvement. The text strategy with a scoring system is currently the most appropriate diagnostic strategy for LLMs.

(JMIR Med Inform 2025;13:e71178) doi: 10.2196/71178

KEYWORDS

large language models; gallbladder polyps; ultrasound; diagnostic accuracy; medical image analysis; artificial intelligence

Introduction

Large language models (LLMs) are deep learning models trained on large amounts of text data, and their emergence has led to changes in many fields [1-3]. LLMs can understand and generate human language and have the potential to provide medical advice, making their application in the medical field of wide interest [4,5]. With the evolution of the LLMs, their capabilities range from simple summarization to complex tasks such as paper writing, medical education, and diagnosis [6-9]. Prior investigations into the diagnostic applications of LLMs have predominantly used two methodological frameworks: (1) diagnostic strategies using narrative textual inputs describing imaging findings [10], or (2) risk stratification approaches requiring LLMs to apply established scoring systems to textualized lesion characteristics [9]. Some LLMs, such as ChatGPT-4o and Claude 3.5 Sonnet, can analyze and interpret images. These two LLMs were developed by different organizations and demonstrated competitive capabilities that positioned them within the global elite of LLMs at the time of our study [11,12], showing they represent the most advanced level of general-purpose LLMs.

Gallbladder polyps are a common finding in abdominal ultrasound examination, with a reported incidence rate of 6.1%-12.1% [13,14]. They impose diagnostic burdens on radiologists and generate substantial patient demand for report interpretation. The management strategy for polypoid lesions of the gallbladder depends on their pathological type. Neoplastic polyps, including gallbladder cancer and precancerous gallbladder adenomas, require cholecystectomy [15,16]. Studies report that 28%-49.5% of gallbladder adenoma may progress to gallbladder cancer [15,17]. Nonneoplastic polyps, including cholesterol polyps, inflammatory polyps, and fibromyoadenoid polyps, rarely become malignant, and follow-up is recommended [18]. Gallbladder carcinoma can be distinguished from other polypoid lesions based on gallbladder wall continuity and contrast-enhanced patterns [19,20]. However, differentiating adenomatous polyps from nonneoplastic polyps remains challenging. Guidelines recommend cholecystectomy for polyps ≥1.0 cm in size [21]. Using these criteria, 27.1%-56% of patients undergoing cholecystectomy for gallbladder polyps are postoperatively diagnosed with nonneoplastic polyps [22,23]. Beyond financial and psychological burdens, this may result in complications that adversely affect quality of life [24,25]. Therefore, it is critical to distinguish neoplastic polyps from nonneoplastic polyps, particularly for lesions ≥1.0 cm. LLMs can potentially reduce the workload of radiologists by generating descriptions based on ultrasound images or risk stratification of gallbladder polyps, and provide medical consults for patients based on ultrasound reports. If LLMs perform better than existing guidelines or radiologists in differentiating gallbladder polyps, they might reduce unnecessary cholecystectomies. Recent studies have demonstrated that ChatGPT-4o and Claude 3.5 Sonnet exhibit superior performance compared to other LLMs in diagnostic tasks involving radiological imaging [26,27], suggesting potential for ultrasound applications. To date, no study has systematically evaluated LLMs' ability to characterize sonographic features or differentiate benign gallbladder polyps. Gallbladder polyps manifest as nonshadowing protrusions from the gallbladder wall into the anechoic lumen in ultrasound examinations. This anatomically well-defined nature with intuitive spatial localization makes them suitable for assessing LLMs' capacity in medical image interpretation. Furthermore, current literature lacks methodological comparisons of LLMs' diagnostic performance across three distinct paradigms: (1) direct image analysis, (2) text-based diagnosis, and (3) scoring-system–based risk stratification using textualized lesion characteristics.

The purpose of this study was to systematically evaluate the feasibility and conduct an early stage evaluation of LLMs across three diagnostic strategies for differential diagnosis of benign gallbladder polyps (≥1.0 cm) based on ChatGPT-4o and Claude 3.5 Sonnet, with comparison to radiologists and the joint guidelines between the European Society of Gastrointestinal and Abdominal Radiology, the European Association for Endoscopic Surgery and other Interventional Techniques, the International Society of Digestive Surgery–European Federation, and the European Society of Gastrointestinal Endoscopy.

Methods

Ethical Considerations

This study was approved by the Ethics Committee in our institute, Sun Yat-sen University (No.2016083). Due to the retrospective nature of this study, the requirement for informed consent was waived. This study adhered to the CLAIM (Checklist for Artificial Intelligence in Medical Imaging) [28] for reporting (Table S1 in Multimedia Appendix 1 [29,30]). This study provides a Reproducibility Checklist in Table S2 in Multimedia Appendix 1 to facilitate the replication of our work.

Throughout the interaction with the LLMs, all patient-identifiable information (including names, hospital ID numbers, etc) was strictly removed from the ultrasound images and reports before analysis.

Due to the retrospective design of this study, participants received no compensation.

Patient Selection

A total of 447 patients with previous imaging findings of gallbladder polyps who underwent cholecystectomy at our institution were retrospectively reviewed. The initial cohort of 312 patients (January 2011 to January 2022) was used to develop the scoring system and to evaluate the performance of the LLMs, while the subsequent cohort of 135 patients (February 2022 to March 2025) served as an external test set for the LLMs-model strategy. The inclusion criteria were as follows: (1)

transabdominal ultrasonography was performed in our hospital before cholecystectomy, and (2) there was a definite postoperative pathological diagnosis. The details of the ultrasound examination protocol are shown in Multimedia Appendix 1 (see also Multimedia Appendices 2-6). Given that the differential diagnostic focus for gallbladder polyps predominantly targets lesions ≥1.0 cm, and smaller polyps may be less clearly visualized in ultrasound images, potentially affecting the performance of LLMs, a 1.0 cm threshold was determined as one of the conditions for patient selection. Patients were excluded if they met any of the following criteria: (1) missing ultrasound images; (2) no polypoid lesions detected by the ultrasound examination in our hospital; (3) age <18 years; and (4) size of the largest polyp <1.0 cm. The patients were divided into an adenomatous polyp group and a nonneoplastic polyp group. Patients with both adenomatous polyps and nonneoplastic polyps were classified as having adenomatous polyps.

Patient Data Collection

Preoperative clinical data of patients were collected, including demographic information, alanine aminotransferase, and aspartate aminotransferase levels. The latest preoperative ultrasound images and reports were collected. Information about polyp size and number, gallbladder wall thickness measurements, and the presence of gallstones was obtained from the reports. Other ultrasound features were evaluated independently by two radiologists (with 2 years and 10 years of experience in abdominal ultrasound, respectively) who were blinded to the patient's clinical features and pathology results. Discrepancies were discussed by the two doctors to get a consensus. When consensus could not be reached after discussion between the two radiologists, a third radiologist with seventeen years of experience in abdominal ultrasound made the final determination. If the patient had more than one gallbladder polyp, only the largest one was analyzed. In our previous study, we proposed a new index for quantifying polyp morphology, called the polyp morphology ratio (PMR). The details about the definitions of PMR and other ultrasound features are provided in Multimedia Appendix 1.

Diagnostic Strategies

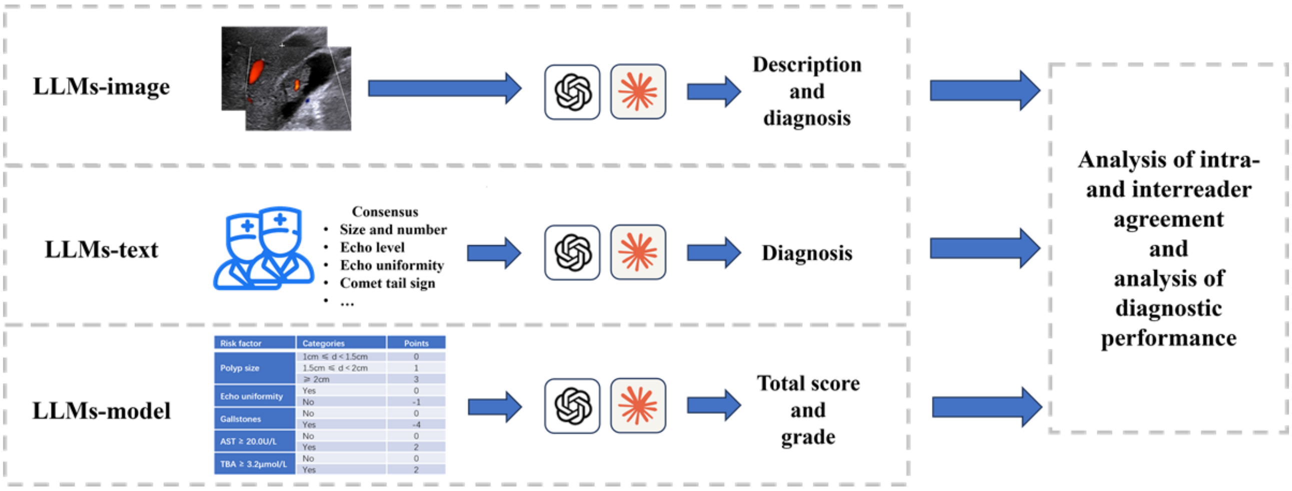
This study evaluated ChatGPT-4o and Claude 3.5 Sonnet using their standard web interfaces [31,32] between July and September 2024 according to the following strategies. All interactions were conducted without custom code, API calls, or adjustments to default inference parameters. To ensure transparency, all interactions used the platform defaults, and the complete set of prompts is detailed in Multimedia Appendix 1. The prompts input into the two LLMs were the same. Each prompt input does not include expected outcomes or class distributions, nor does it judge the correctness of the LLM's responses. For the strategy of LLMs-image, LLMs performed diagnoses based on ultrasound images. Before analysis, images were cropped to remove all patient-identifiable information. Additionally, the images were cropped to ensure the gallbladder occupied approximately 30%-40% of the frame, with lesions centered whenever possible while maintaining the structural integrity of both the gallbladder and surrounding tissues. No image enhancement or standardization was applied. The underlying vision capabilities of the LLMs are accessed directly through their vision application programming interface, rather than via a third-party wrapper. The images input into the LLMs were in JPEG format, with both horizontal and vertical resolutions of 96 dpi. The processed 2D gray-scale ultrasound image and color Doppler ultrasound image were input into LLMs, and the LLMs were required to describe the ultrasound characteristics of gallbladder polyps and provide a diagnosis.

For the strategy LLMs-text, LLMs performed diagnoses based on the text of the ultrasound description. Based on the consensus of the two radiologists, the description of the ultrasound characteristics for the gallbladder polyps was organized into a structured report. The structured report was input into the LLMs, and the LLMs were required to give the diagnosis.

For the strategy LLMs-model, LLMs performed diagnoses based on our previously developed diagnostic model for benign gallbladder polyps ≥1.0 cm [33]. Before this study, a multilevel scoring system for the differentiation between gallbladder adenomatous polyps and nonneoplastic polyps was constructed based on the same cases as those in this study (Table S3 in Multimedia Appendix 1). The relevant information was organized into structured text according to the requirements of the scoring system. Then, only the structured text was input into LLMs without the corresponding scores, and the LLMs were required to calculate the total score and indicate the corresponding grade. In addition, based on the consensus of two radiologists, the total score and the corresponding grade of each patient were calculated with the multilevel scoring system (strategy readers-model).

Figure 1 shows the flow of the diagnostic strategies for LLMs. The prompts templates in diagnostic strategies for LLMs are in Multimedia Appendix 1. The intrareader agreement, interreader agreement, and diagnostic performance of LLMs in the three diagnostic strategies were evaluated. The initial outputs from LLMs in the three strategies were used to analyze interreader agreement and diagnostic performance. From the final enrolled cases, 70 were randomly selected to assess intrareader agreement. The LLMs were asked to regenerate the output for these cases twice, that is, there were three rounds of output per strategy per LLM for each of these 70 cases, respectively.

Figure 1. Diagram of three diagnostic strategies for LLMs. LLM: large language model.



Risk factor	Categories	Points
Polyp size	1cm ⩽ d < 1.5cm	0
	1.5cm ⩽ d < 2cm	1
	⩾ 2cm	3
Echo uniformity	Yes	0
	No	-1
Gallstones	No	0
	Yes	-4
AST ≥ 20.0U/L	No	0
	Yes	2
TBA ≥ 3.2μmol/L	No	0
	Yes	2

In-Context Learning for LLMs

For the ultrasound image-based lesion characterization tasks in this study, the LLMs underwent supplemental in-context learning beyond their original architecture. LLMs were trained to recognize the degree of blood flow, blood flow pattern, and the definition of PMR and polyp base type (sessile or pedunculated) using ultrasound images of patients with gallbladder polyps in our center (excluding cases included in this study). The prompts used in in-context learning are provided in Multimedia Appendix 1. After entering the prompt word, the memory of the LLMs to was checked to ensure its correct understanding.

Statistical Analysis

The interclass correlation coefficient, Cohen κ, and weighted κ were used to evaluate the interreader agreement for continuous, unordered, and ordered categorical variables, respectively. Interclass correlation coefficient, Fleiss κ, and Kendall W coefficient were used to evaluate the intrareader agreement for continuous, unordered, and ordered categorical variables, respectively. The interreader agreement of ultrasound features was represented by a heatmap generated with the *pheatmap* package in R (R Foundation). The levels of the coefficient of the agreement analysis were defined as follows: 0.20 or less for slight agreement, 0.21-0.40 for fair agreement, 0.41-0.60 for moderate agreement, 0.61-0.80 for substantial agreement, and 0.81-1.00 for almost perfect agreement. The agreement coefficients were compared using the Z-test. The detailed sample size calculation for intrareader agreement is shown in Multimedia Appendix 1.

The diagnostic performance of LLMs was evaluated by sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, area under the receiver operating characteristic curve, and unnecessary resection rate of nonneoplastic polyps (UNRR). Sensitivity, specificity, PPV, and NPV were calculated and compared with the *stats* package and *epiR* package in R. The area under the receiver operating characteristic curves were compared using the DeLong test. UNRR was calculated as the number of nonneoplastic polyps recommended for cholecystectomy divided by the total number of nonneoplastic polyps. Chi-square test was used to compare UNRR.

For the strategy LLMs-text, univariate logistic regression was performed on ultrasound features according to the diagnosis of LLMs to analyze the basis for differential diagnosis of LLMs.

Two-tailed *P*<.05 was considered statistically significant. Statistical analysis was performed using PASS 2025 (power analysis and sample size; NCSS, LLC), SPSS 25.0 (IBM Corp), and R version 4.3.1 (R Foundation).

Results

Patient Characteristics

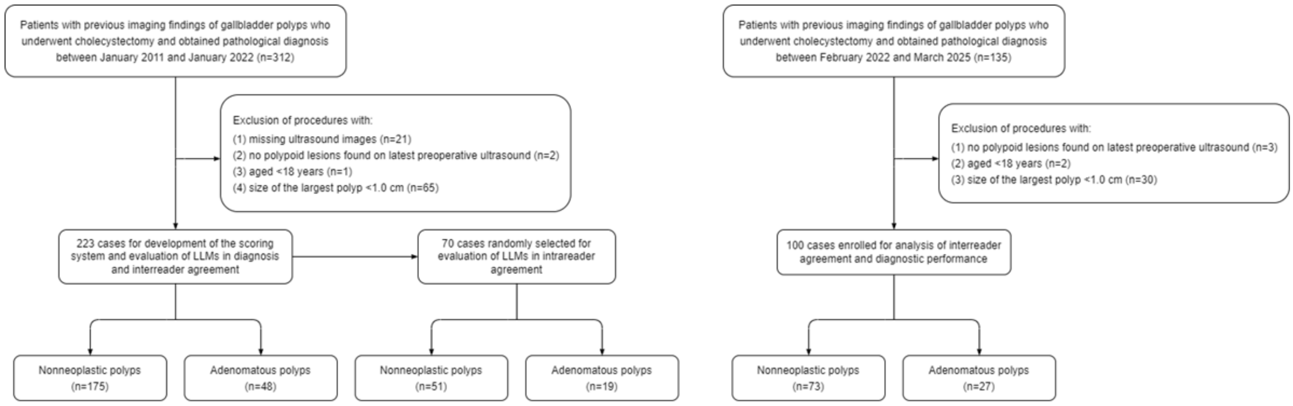
In the initial cohort, this study analyzed 223 patients aged 18-72 (median 40, IQR 34-50) years, including 132 (59.2%) females. Among these, 175 (78.5%) had nonneoplastic polyps and 48 (21.5%) had adenomatous polyps. Compared to the initial cohort, the external test set showed no significant differences in demographic characteristics or pathological type distribution (all *P* >.05; Table 1). The patient selection process is shown in Figure 2.

Table 1. Demographic characteristics and gallbladder polyp pathology of patients.

Characteristic	Initial cohort (n=223)	External test set for the LLMs[a]-model strategy (n=100)	*P* value
Sex			
Male, n (%)	91 (40.8)	43 (43)	.71
Female, n (%)	132 (59.2)	57 (57)	.71
Age (years)			.84
Median (IQR)	40 (34-50)	41 (33-52)	
Range	18-72	19-74	
Polyp pathology			
Adenomatous polyps, n (%)	48 (21.5)	27 (27)	.28
Nonneoplastic polyps, n (%)	175 (78.5)	73 (73)	.28

[a] LLM: large language model.

Figure 2. Flowchart of patient selection. LLM: large language model.


Patients with previous imaging findings of gallbladder polyps who underwent cholecystectomy and obtained pathological diagnosis between January 2011 and January 2022 (n=312)
Exclusion of procedures with:
(1) missing ultrasound images (n=21)
(2) no polypoid lesions found on latest preoperative ultrasound (n=2)
(3) aged <18 years (n=1)
(4) size of the largest polyp <1.0 cm (n=65)
223 cases for development of the scoring system and evaluation of LLMs in diagnosis and interreader agreement
70 cases randomly selected for evaluation of LLMs in intrareader agreement
Nonneoplastic polyps (n=175)
Adenomatous polyps (n=48)
Nonneoplastic polyps (n=51)
Adenomatous polyps (n=19)
Patients with previous imaging findings of gallbladder polyps who underwent cholecystectomy and obtained pathological diagnosis between February 2022 and March 2025 (n=135)
Exclusion of procedures with:
(1) no polypoid lesions found on latest preoperative ultrasound (n=3)
(2) aged <18 years (n=2)
(3) size of the largest polyp <1.0 cm (n=30)
100 cases enrolled for analysis of interreader agreement and diagnostic performance
Nonneoplastic polyps (n=73)
Adenomatous polyps (n=27)

Intrareader Agreement Analysis

The intrareader agreement coefficients of LLMs across three strategies are shown in Table 2. Compared to the intrareader agreement for grades in strategy LLMs-model, both ChatGPT-4o and Claude 3.5 Sonnet exhibited significantly lower intrareader agreement for diagnoses in strategy LLMs-text (0.58 vs 0.97, *P*<.001 for ChatGPT-4o; 0.61 vs 1.00, *P*<.001 for Claude 3.5 Sonnet) and LLMs-image (0.37 vs 0.97, *P*<.001 for ChatGPT-4o; 0.47 vs 1.00, *P*<.001 for Claude 3.5 Sonnet). However, there was no significant difference between strategy LLMs-image and LLMs-text (*P*=.15 for ChatGPT-4o and *P*=.33 for Claude 3.5 Sonnet).

For the ultrasound feature identification in strategy LLMs-image, the agreement level of blood flow degree and pattern for ChatGPT-4o was moderate to substantial (range of agreement coefficient 0.43 to 0.68, 95% CI 0.30 to 0.81). Additionally, there was only slight to fair agreement for ChatGPT-4o in other ultrasound features (range of agreement coefficient –0.12 to 0.30, 95% CI –0.25 to 0.45). However, except for the slight to fair agreement in PMR, comet tail sign and gallbladder wall thickness types (range of agreement coefficient 0.02 to 0.30, 95% CI –0.10 to 0.45), there was moderate to almost perfect agreement for Claude 3.5 Sonnet in other ultrasound features (range of agreement coefficient 0.43 to 0.97, 95% CI 0.30 to 1.00). The intrareader agreement coefficients in ultrasound features for Claude 3.5 Sonnet were all higher than those for ChatGPT-4o, except for PMR. In addition, the intraobserver agreement coefficient in diagnosis for Claude 3.5 Sonnet in strategy LLMs-image was 0.47, which was also higher than 0.7 for ChatGPT-4o.

The intraobserver agreement coefficients in diagnosis for ChatGPT-4o and Claude 3.5 Sonnet in strategy LLMs-text were 0.58, 95% CI 0.46 to 0.72, and 0.1, 95% CI 0.49 to 0.75, respectively, indicating a moderate to substantial level, which were higher than those in strategy LLMs-image.

In strategy LLMs-model, both ChatGPT-4o and Claude 3.5 Sonnet showed almost perfect intraobserver agreement for total score and grade (range of agreement coefficient 0.97 to 1.00, 95% CI 0.94 to 1.00). Almost perfect intraobserver agreement was also observed in the external test set (Table S4 in Multimedia Appendix 1).

Table 2. Intrareader agreement coefficients of LLMs[a]. Data in parentheses are 95% CIs. Except where indicated, the coefficient of agreement analysis is the Fleiss κ coefficient.

	ChatGPT-4o	Claude 3.5 Sonnet
LLMs-image		
PMR[b,c]	0.30 (0.15 to 0.45)	0.02 (–0.10 to 0.14)
Echo level[d]	0.25 (–0.03 to 0.42)	0.58 (0.45 to 0.71)
Echo uniformity	–0.12 (–0.25 to 0.01)	0.75 (0.62 to 0.88)
Comet tail sign	–0.01 (–0.14 to 0.12)	0.30 (0.19 to 0.45)
Cauliflower shape	0.00 (–0.13 to 0.12)	0.74 (0.61 to 0.87)
Pedunculated or sessile	0.18 (0.05 to 0.31)	0.43 (0.30 to 0.56)
Edge	–0.01 (–0.14 to 0.12)	1.00 (—[e])
Blood flow degree[d]	0.68 (0.55 to 0.81)	0.97 (0.94 to 1.00)
Blood flow pattern	0.43 (0.30 to 0.56)	0.85 (0.72 to 0.98)
Gallbladder wall thickness types	–0.04 (–0.17 to 0.09)	0.15 (0.02 to 0.28)
Diagnosis	0.37 (0.24 to 0.50)	0.47 (0.34 to 0.60)
LLMs-text		
Diagnosis	0.58 (0.46 to 0.72)	0.61 (0.49 to 0.75)
LLMs-model		
Total score[c]	1.00 (—)	1.00 (—)
Grade[d]	0.97 (0.94 to 1.00)	1.00 (—)

[a] LLM: large language model.

[b] PMR: polyp morphology ratio.

[c] The coefficient of agreement analysis is the interclass correlation coefficient.

[d] The coefficient of agreement analysis is the Kendall W coefficient.

[e] CIs cannot be calculated due to the perfect consistency of the three rounds' output in each case.

Interreader Agreement Analysis

Figure 3 and Table S5 in Multimedia Appendix 1 present the interreader agreement of human readers and LLMs in ultrasound features. Except for the interobserver agreement level of cauliflower shape between readers 1 and 2 being moderate (agreement coefficient=0.57), the agreement level for other features between readers 1 and 2 was substantial to almost perfect (range of agreement coefficient 0.65 to 0.99). However, the agreement level between LLMs and other observers was only slight (range of agreement coefficient –0.11 to 0.12), except for blood flow degree and pattern (range of agreement coefficient 0.34 to 0.75). As shown in Table S5 in Multimedia Appendix 1, apart from the comet tail sign, the agreement coefficient for other features between readers 1 and 2 was significantly higher than those between LLMs and other observers (*P*<.001).

The agreement coefficient for the comet tail sign could not be calculated because reader 1 considered that the comet tail sign was not present in all cases in this study.

The interobserver agreement coefficients of human readers and LLMs in diagnosis are shown in Table 3. There was slight interobserver agreement between ChatGPT-4o and Claude 3.5 Sonnet in strategy LLMs-image (Cohen κ=0.12, 95% CI –0.01 to 0.24), and fair agreement in strategy LLMs-text (Cohen κ=0.38, 95% CI 0.26 to 0.50). For the strategy readers or LLMs-model, the agreement levels were all almost perfect in readers versus ChatGPT-4o, readers versus Claude 3.5 Sonnet, ChatGPT-4o versus Claude 3.5 Sonnet (range of agreement coefficient 0.84 to 0.99, 95% CI 0.78 to 0.99). In the external test set, LLMs still showed a considerable degree of interreader agreement (Table S6 in Multimedia Appendix 1). Compared to the interreader agreement coefficient of grades between ChatGPT-4o and Claude 3.5 Sonnet in strategy LLMs-model, those for diagnosis in strategy LLMs-image (*P*<.001) and LLMs-text (*P*<.001) were significantly lower. Additionally the interreader agreement coefficient for diagnosis in strategy LLMs-text was significantly higher than that in strategy LLMs-image (*P*=.004).

Figure 3. Heatmap of interreader agreement in human readers and LLMs. LLM: large language model.

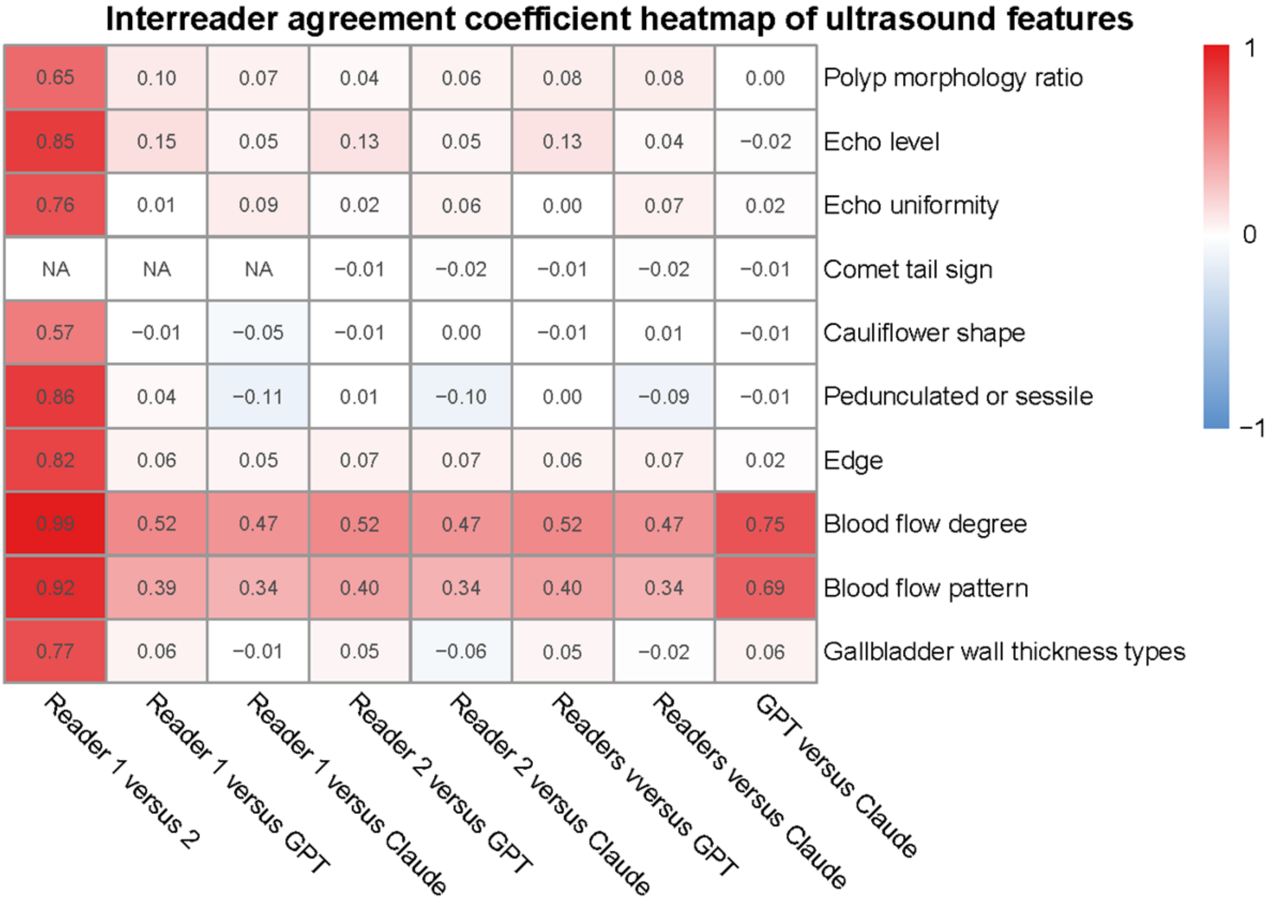


Table 3. Interreader agreement coefficients of human readers and LLMsa in diagnosis. Data in parentheses are 95% CIs. Except where indicated, the coefficient of agreement analysis is Cohen κ. The total score and grade of readers are calculated based on laboratory tests and the US features consensus of readers 1 and 2.

Diagnostic strategies	Readers versus ChatGPT-4o	Readers versus Claude 3.5 Sonnet	ChatGPT-4o versus Claude 3.5 Sonnet
LLM-image	—[b]	—	0.12 (–0.01 to 0.24)
LLM-text	—	—	0.38 (0.26 to 0.50)
Readers or LLM-model			
Total score[c]	0.97 (0.96 to 0.97)	0.98 (0.97 to 0.98)	0.99 (0.98 to 0.99)
Grade[d]	0.84 (0.78 to 0.90)	0.95 (0.92 to 0.99)	0.87 (0.82 to 0.92)

[a] LLM: large language model.

[b] Not available.

[c] The coefficient of agreement analysis is the interclass correlation coefficient.

[d] The coefficient of agreement analysis is weighted κ.

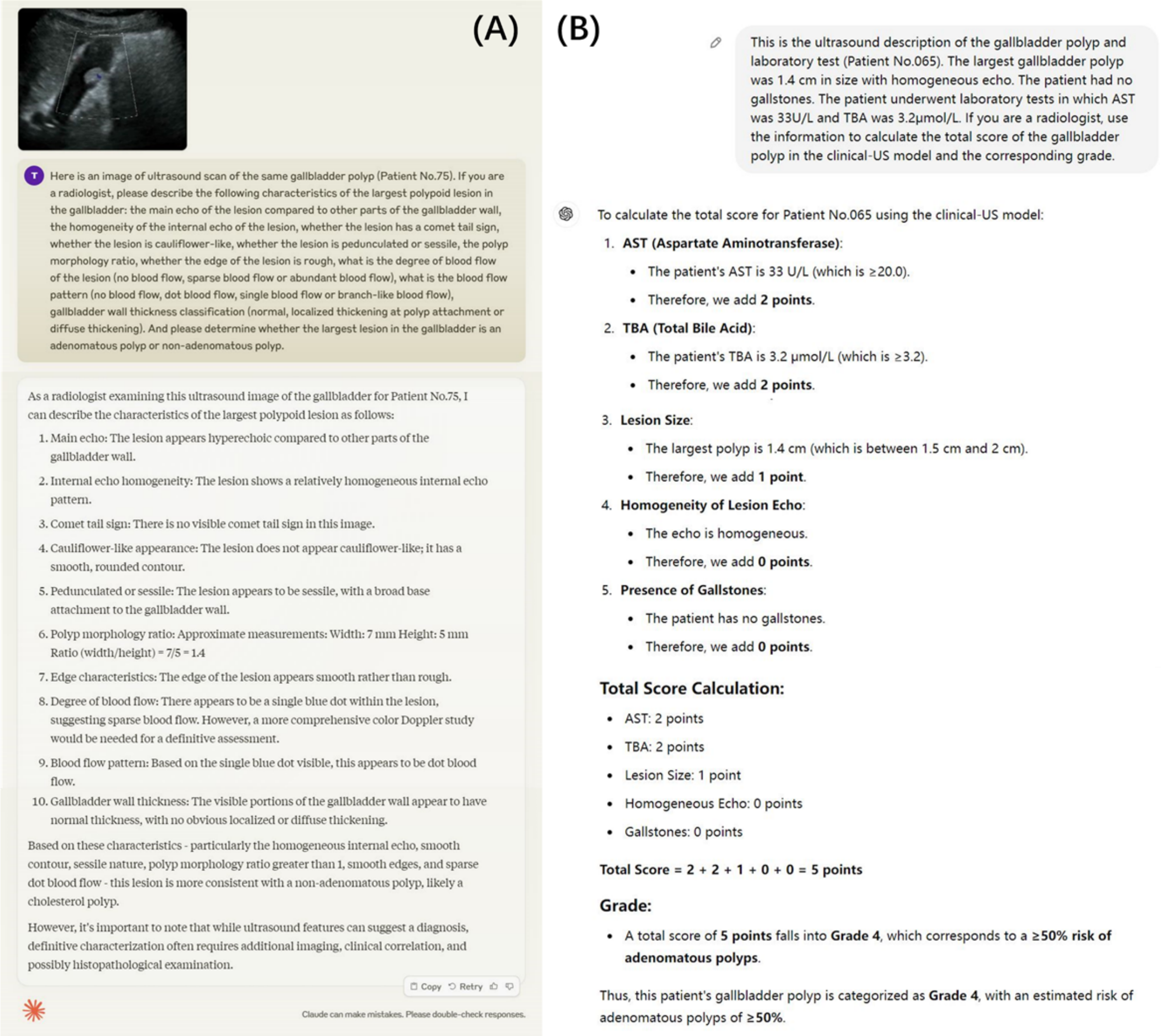
Diagnostic Performance

The diagnostic performance of all diagnostic strategies is presented in Table 4. Figure 4A and B show representative cases from strategy Claude-image and strategy GPT-model, respectively.

The sensitivity of the guideline's diagnostic strategy was 1.00, indicating that no gallbladder adenomas were missed, but surgery was recommended for all nonneoplastic polyps ≥1.0 cm (UNRR=100%). In the strategies LLMs-image and LLMs-text, the sensitivity of GPT-image, Claude-image, GPT-text, and Claude-text was 0.27, 0.33, 0.56, and 0.65, respectively, which were significantly lower than the sensitivity of the guideline (all *P*<.001), indicating more gallbladder adenomas were missed.

In the strategy readers or LLMs-model, when cholecystectomy was recommended for gallbladder polyps ≥grade 2, the sensitivity of ChatGPT-4o, Claude 3.5 Sonnet, and radiologists were 0.94, 0.98, and 0.98, respectively, showing no significant difference from the sensitivity of 1.00 achieved by the guideline (all *P*<.05). However, the accuracy of ChatGPT-4o, Claude 3.5 Sonnet, and radiologists was significantly higher than that of the guideline (0.35, 0.34, and 0.34 vs 0.22, all *P*<.01), and the UNRR was significantly lower than that of the guideline (82%, 83%, and 83% vs 100%, all *P*<.01). In addition, there were no significant differences observed between GPT-model, Claude-model and readers-model in terms of sensitivity, specificity, PPV, NPV, accuracy or UNRR across all grades (all *P*>.05). Similar results were observed in the external test set (Table S7 in Multimedia Appendix 1).

Figure 4. Examples of strategy Claude-image (A) and strategy GPT-model (B). AST: aspartate aminotransferase; TBA: total bile acid.



Here is an image of ultrasound scan of the same gallbladder polyp (Patient No.75). If you are a radiologist, please describe the following characteristics of the largest polypoid lesion in the gallbladder: the main echo of the lesion compared to other parts of the gallbladder wall, the homogeneity of the internal echo of the lesion, whether the lesion has a comet tail sign, whether the lesion is cauliflower-like, whether the lesion is pedunculated or sessile, the polyp morphology ratio, whether the edge of the lesion is rough, what is the degree of blood flow of the lesion (no blood flow, sparse blood flow or abundant blood flow), what is the blood flow pattern (no blood flow, dot blood flow, single blood flow or branch-like blood flow), gallbladder wall thickness classification (normal, localized thickening at polyp attachment or diffuse thickening). And please determine whether the largest lesion in the gallbladder is an adenomatous polyp or non-adenomatous polyp.

As a radiologist examining this ultrasound image of the gallbladder for Patient No.75, I can describe the characteristics of the largest polypoid lesion as follows:

1. Main echo: The lesion appears hyperechoic compared to other parts of the gallbladder wall.
2. Internal echo homogeneity: The lesion shows a relatively homogeneous internal echo pattern.
3. Comet tail sign: There is no visible comet tail sign in this image.
4. Cauliflower-like appearance: The lesion does not appear cauliflower-like; it has a smooth, rounded contour.
5. Pedunculated or sessile: The lesion appears to be sessile, with a broad base attachment to the gallbladder wall.
6. Polyp morphology ratio: Approximate measurements: Width: 7 mm Height: 5 mm Ratio (width/height) = 7/5 = 1.4
7. Edge characteristics: The edge of the lesion appears smooth rather than rough.
8. Degree of blood flow: There appears to be a single blue dot within the lesion, suggesting sparse blood flow. However, a more comprehensive color Doppler study would be needed for a definitive assessment.
9. Blood flow pattern: Based on the single blue dot visible, this appears to be dot blood flow.
10. Gallbladder wall thickness: The visible portions of the gallbladder wall appear to have normal thickness, with no obvious localized or diffuse thickening.

Based on these characteristics - particularly the homogeneous internal echo, smooth contour, sessile nature, polyp morphology ratio greater than 1, smooth edges, and sparse dot blood flow - this lesion is more consistent with a non-adenomatous polyp, likely a cholesterol polyp.

However, it's important to note that while ultrasound features can suggest a diagnosis, definitive characterization often requires additional imaging, clinical correlation, and possibly histopathological examination.

Claude can make mistakes. Please double-check responses.

This is the ultrasound description of the gallbladder polyp and laboratory test (Patient No.065). The largest gallbladder polyp was 1.4 cm in size with homogeneous echo. The patient had no gallstones. The patient underwent laboratory tests in which AST was 33U/L and TBA was 3.2μmol/L. If you are a radiologist, use the information to calculate the total score of the gallbladder polyp in the clinical-US model and the corresponding grade.

To calculate the total score for Patient No.065 using the clinical-US model:

1. **AST (Aspartate Aminotransferase)**:
   - The patient's AST is 33 U/L (which is ≥20.0).
   - Therefore, we add **2 points**.
2. **TBA (Total Bile Acid)**:
   - The patient's TBA is 3.2 μmol/L (which is ≥3.2).
   - Therefore, we add **2 points**.
3. **Lesion Size**:
   - The largest polyp is 1.4 cm (which is between 1.5 cm and 2 cm).
   - Therefore, we add **1 point**.
4. **Homogeneity of Lesion Echo**:
   - The echo is homogeneous.
   - Therefore, we add **0 points**.
5. **Presence of Gallstones**:
   - The patient has no gallstones.
   - Therefore, we add **0 points**.

Total Score Calculation:

- AST: 2 points
- TBA: 2 points
- Lesion Size: 1 point
- Homogeneous Echo: 0 points
- Gallstones: 0 points

Total Score = 2 + 2 + 1 + 0 + 0 = 5 points

Grade:

- A total score of **5 points** falls into **Grade 4**, which corresponds to a **≥50% risk of adenomatous polyps**.

Thus, this patient's gallbladder polyp is categorized as **Grade 4**, with an estimated risk of adenomatous polyps of **≥50%**.

Table 4. Diagnostic performance of all diagnostic strategies. Data in ranges are 95% CIs. Except where indicated, *P* values are for comparison with the guideline that recommends cholecystectomy for polyps ≥1.0 cm.

Diagnostic strategies	Sensitivity	Specificity	PPV[a]	NPV[b]	Accuracy	UNRR[c,d]
Polyp size ≥1.0 cm						175/175 (100)
	1.00	0.00	0.22	—[e]	0.22	
	0.93 to 1.00	0.00 to 0.02	0.16 to 0.28	—	0.16 to 0.28	
GPT-image						54/175 (31)
	0.27	0.69	0.19	0.78	0.60	
	0.15 to 0.42	0.62 to 0.76	0.11 to 0.31	0.70 to 0.84	0.53 to 0.67	
P value	<.01	<.01	.84	—	<.01	<.001
Claude-image						80/175 (46)
	0.33	0.54	0.17	0.75	0.50	
	0.20 to 0.48	0.47 to 0.62	0.10 to 0.26	0.66 to 0.82	0.43 to 0.57	
P value	<.01	<.01	.40	—	<.01	<.001
GPT-text						108/175 (62)
	0.56	0.38	0.20	0.76	0.42	
	0.41 to 0.71	0.31 to 0.46	0.14 to 0.28	0.66 to 0.85	0.36 to 0.49	
P value	<.01	<.01	.83	—	<.01	<.001
Claude-text						127/175 (73)
	0.65	0.27	0.20	0.74	0.35	
	0.50 to 0.78	0.21 to 0.35	0.14 to 0.27	0.62 to 0.84	0.29 to 0.42	
P value	<.01	<.01	.75	—	<.01	<.001
GPT-model						
≥Grade 2						143/175 (82)
	0.94	0.18	0.24	0.91	0.35	
	0.83 to 0.99	0.13 to 0.25	0.18 to 0.31	0.77 to 0.98	0.28 to 0.41	
P value	.24	<.01	.64	—	<.01	<.001
P value[f]	.61	.78	>.99	.72	>.99	.78
≥Grade 3						84/175 (48)
	0.75	0.52	0.30	0.88	0.57	
	0.60 to 0.86	0.44 to 0.60	0.22 to 0.39	0.81 to 0.94	0.50 to 0.64	
P value	.01	<.01	.11	—	<.01	<.001
P value[f]	>.99	.45	.93	>.99	.57	.45
=Grade 4						5/175 (3)
	0.15	0.97	0.58	0.81	0.79	
	0.06 to 0.28	0.94 to 0.99	0.28 to 0.85	0.75 to 0.86	0.74 to 0.85	
P value	<.01	<.01	.01	—	<.01	<.001
P value[f]	>.99	.72	.90	>.99	.91	.72
Claude-model						
≥Grade 2						146/175 (83)
	0.98	0.17	0.24	0.97	0.34	
	0.89 to 1.00	0.11 to 0.23	0.19 to 0.31	0.83 to 1.00	0.28 to 0.41	
P value	>.99	<.01	.57	—	<.01	<.001

Diagnostic strategies	Sensitivity	Specificity	PPV[a]	NPV[b]	Accuracy	UNRR[c,d]
P value[f]	>.99	>.99	>.99	>.99	>.99	>.99
≥Grade 3						93/175 (53)
	0.73	0.47	0.27	0.86	0.53	
	0.58 to 0.85	0.39 to 0.55	0.20 to 0.36	0.78 to 0.93	0.46 to 0.59	
P value	<.01	<.01	.27	—	<.01	<.001
P value[f]	.81	>.99	.92	.85	.85	>.99
=Grade 4						5/175 (3)
	0.15	0.97	0.58	0.81	0.79	
	0.06 to 0.28	0.94 to 0.99	0.28 to 0.85	0.75 to 0.86	0.74 to 0.85	
P value	<.01	<.01	.01	—	<.01	<.001
P value[f]	>.99	.72	.90	>.99	.91	.72
Readers-model						
≥Grade 2						146/175 (83)
	0.98	0.17	0.24	0.97	0.34	
	0.89 to 1.00	0.11 to 0.23	0.19 to 0.31	0.83 to 1.00	0.28 to 0.41	
P value	.99	<.01	.57	—	<.01	<.001
≥Grade 3						92/175 (53)
	0.77	0.47	0.29	0.88	0.54	
	0.63 to 0.88	0.40 to 0.55	0.21 to 0.37	0.80 to 0.94	0.47 to 0.61	
P value	.01	<.01	.17	—	<.01	<.001
=Grade 4						3/175 (2)
	0.15	0.98	0.70	0.81	0.80	
	0.06 to 0.28	0.95 to 1.00	0.35 to 0.93	0.75 to 0.86	0.74 to 0.85	
P value	<.01	<.01	<.01	—	<.01	<.001

[a] PPV: positive predictive value.

[b] NPV: negative predictive value.

[c] UNRR: unnecessary resection rate of nonneoplastic polyps.

[d] n/N (%).

[e] Not applicable.

[f] *P* values are for comparison with the same grade in the strategy readers-model.

Interpretability of the Diagnosis in Strategy LLMs-Text

The results of univariate analysis based on the diagnosis of GPT-text and Claude-text are shown in Table S8 in Multimedia Appendix 1. Both ChatGPT-4o and Claude 3.5 Sonnet considered larger size, hypoechogenicity, heterogeneous echogenicity, cauliflower shape, sessile base, and rough edge as significant factors for adenomatous polyps. Additionally, ChatGPT-4o identified higher PMR as a diagnostic indicator for adenomatous polyps. Furthermore, Claude 3.5 Sonnet also associated sparse and dot-like blood flow in color Doppler flow imaging with adenomatous polyps. Both LLMs diagnosed all lesions with abundant blood flow (6 cases) as adenomas, which was identified as a risk feature of neoplastic gallbladder polyps by a previous study, while this complete separation led to the extreme or infinite CIs.

Error Analysis in Strategy LLMs-Text

In the strategy LLMs-text, 129 cases were misdiagnosed by ChatGPT-4o, while 94 cases were correctly diagnosed. As shown in Table S9 in Multimedia Appendix 1, lesions with the following characteristics were more likely to be misdiagnosed by ChatGPT-4o: hypoechoic appearance (23% vs 11%, *P*=.047), heterogeneous echotexture (52% vs 22%, *P*<.001), cauliflower-like morphology (24% vs 13%, *P*=.04), and rough edges (56% vs 29%, *P*<.001).

For Claude 3.5 Sonnet, 144 cases were misdiagnosed by ChatGPT-4o, while 79 cases were correctly diagnosed. Table S10 in Multimedia Appendix 1 demonstrates that Claude 3.5 Sonnet was prone to misdiagnosis in lesions with these features: larger size (1.30 cm vs 1.10 cm, *P*<.001), hypoechoic appearance (21% vs 13%, *P*=.03), heterogeneous echotexture

(49% vs 23%, *P*<.001), cauliflower-like morphology (24% vs 11%, *P*=.03), and rough edges (52% vs 30%, *P*=.002).

Discussion

Principal Findings

This study evaluated the feasibility of LLMs for the differential diagnosis of gallbladder benign polyps. Our principal findings indicate that the diagnostic strategy profoundly influences LLM performance. In the strategy LLMs-image, LLMs exhibited only slight-to-moderate intrareader agreement and poor interreader consistency compared to radiologists. As the ultrasonic features associated with gallbladder polyps represent common characteristics shared by focal lesions, the finding that current LLMs have limited capabilities in recognizing these features suggests this limitation may have broad applicability. For diagnosis, the diagnostic performance of LLMs-image and the strategy LLMs-text showed significantly lower sensitivity than clinical guidelines, leading to more missed adenomas. In contrast, the strategy LLMs-model, which used text descriptions within a scoring system, demonstrated high consistency and diagnostic performance comparable to radiologists using the same model, while significantly reducing unnecessary cholecystectomies. Notably, the performance between two general-purpose LLMs, ChatGPT-4o and Claude 3.5 Sonnet, was comparable across all strategies, suggesting these findings may generalize to other general-purpose LLMs.

Comparison With Prior Work

Our results on the poor performance of general-purpose LLMs in direct image interpretation align with a growing body of evidence across various medical fields. Studies using dermoscopic images for melanoma diagnosis, orthopedic residency examination images, and musculoskeletal radiology images consistently report not only unsatisfactory diagnostic accuracy (as low as 3% to 36%) but also highlight that LLM performance is inferior to that of human specialists [34-37]. Critically, some studies report that LLMs could harm patient care by recommending unnecessary invasive procedures, such as biopsies, at a significantly higher rate than radiologists [38]. This collective evidence confirms that current general-purpose LLMs remain inadequate for reliable medical image-based diagnosis.

It is important to note that LLMs receiving specialized training in medical images have demonstrated more competent performance in specific domains, such as dermatology (SkinGPT-4) and diabetes management (DeepDR-LLM) [39,40]. Universal systems such as ChatCAD+ also show promise across multiple image types [41]. However, such specialized models are often confined to a single disease area or a limited set of tasks, and they may underperform advanced general LLMs such as GPT-4 in broader medical question-answering [42]. Other general medical LLMs capable of addressing multiple conditions, such as MedFound, do not yet support image input [43]. Given the critical needs for accessibility, generalizability, and image input capabilities in real-world clinical settings, we selected leading general-purpose LLMs (ChatGPT-4o and Claude 3.5 Sonnet) as the subjects of this study. Our findings contribute to the understanding of their current capabilities and limitations.

Regarding the diagnostic performance of text-based strategies, our findings must be viewed within the context of a highly variable literature. Previous studies report widely varying accuracy (40%-91.4%) for categorical diagnoses using text-only strategies [27,44,45]. When focusing specifically on the interpretation of radiological findings, reported diagnostic accuracy for LLMs ranges from 25% to 73% [26,46-48]. While some studies, such as one on real-world radiology reports of brain tumors, found ChatGPT-4's accuracy (73%) to be comparable with radiologists [48], the absence of standardized methods and reporting metrics in LLM studies introduces significant bias risks and hinders cross-study comparisons. In our study, the text-based strategy demonstrated superior diagnostic performance to the image-based approach, which aligns with the consensus and the reported performance levels. Nevertheless, given gallbladder adenomas' high potential for malignancy, the moderate sensitivity achieved by the text-based strategy in our study remains a critical limitation for clinical application, as higher sensitivity is paramount to avoid missed diagnoses.

The LLMs performed suboptimally in the text-based diagnostic strategy, even though the interpretability analysis of the LLMs-text approach indicated that the high-risk features of adenomas identified by the LLMs are supported by previous literature [29,49-51]. In the error analysis, we observed that LLMs showed subpar performance in diagnosing lesions with these same features, such as heterogeneous echotexture and rough edges. Essentially, this finding reflects a long-standing fundamental challenge in the imaging diagnosis of gallbladder polyps: there is significant overlap in the sonographic features between high-grade dysplastic adenomas and early-stage gallbladder cancer. Trained on broader existing literature about medical imaging, the LLMs correctly learned the strong association between these features and common malignancy risk (eg, gallbladder cancer). As indicated by the extreme odds ratio in the interpretability analysis, both LLMs assigned excessive weight to abundant blood flow. Although previous studies confirm it is indeed a feature associated with neoplastic polyps [52], the models' tendency to treat it as an absolute diagnostic rule—rather than a probabilistic indicator—reveals their limitation in performing the task of fine-grained diagnosis, a reflection of the inherent difficulty of the task itself. However, when the task was confined to the relatively idealized binary framework of "differentiating adenomatous from non-neoplastic polyps," the model's heightened sensitivity to these high-risk features created a conflict with the task objective. This may not represent a "weakness" in the model, but rather an objective reflection of complex clinical reality—the degree of dysplasia in lesions exists on a spectrum, and such ambiguity in imaging is inherent before the intervention of the pathological "gold standard." Meanwhile, the LLMs also demonstrated valuable capabilities. For instance, ChatGPT-4o correctly recognized a novel morphological index (PMR) from an unpublished study as useful for diagnosis, and Claude 3.5 Sonnet identified sparse and dot blood flow as a relevant feature, which aligns with established research [29,49]. This demonstrates their potential in tasks requiring strong information retrieval and logical reasoning.

Conversely, our study reinforces that LLMs exhibit excellent performance when working with structured text data from scoring systems. Their high accuracy and consistency in applying such systems are supported by prior research in Liver Imaging Reporting and Data System and Thyroid Imaging Reporting and Data System classification [9,53]. In addition, our findings clarify the role of general-purpose LLMs. In the LLMs-model strategy, the LLM acts as a reliable executor of clinical rules, not a simple calculator. It first understands text to find key features, then applies the scoring rule. The performance of our LLMs-model strategy, which was comparable to radiologists, suggests that leveraging LLMs to execute standardized clinical rules may be their most reliable and immediate clinical application.

Strengths and Limitations

A key strength of this study is the comprehensive evaluation of LLMs across three distinct diagnostic strategies, providing a clear understanding of their appropriate clinical roles. We also provided an in-depth analysis of their reasoning, identifying both capabilities and weaknesses.

This study has several limitations. First, the task of differentiating benign gallbladder polyps is inherently challenging, even for radiologists, which may have contributed to the LLMs’ suboptimal performance. Second, the evaluation was conducted on a single disease entity; assessing broader clinical applicability requires more diverse and complex cases. Third, we used in-context learning with only typical cases. While this approach simulates a realistic usage scenario and is highly accessible, it is inherently less stable than fine-tuning and may not capture the full spectrum of clinical presentations. Finally, the retrospective design may introduce potential selection biases.

Future Directions

Based on our findings, future research should focus on several key areas. First, the development and evaluation of more versatile and medically tuned multimodal LLMs are crucial. Second, the performance of LLMs is heavily dependent on the quality of the underlying scoring system; therefore, integrating them with more robust and validated clinical models could maximize their diagnostic utility. Finally, standardizing evaluation methods and reporting metrics across LLM studies is urgently needed to enable meaningful comparisons. Given their strong performance in information retrieval and logical reasoning, future work should also explore the role of LLMs in medical education [54,55] and self-management counseling for patients with chronic conditions [56].

Conclusions

In conclusion, current general-purpose LLMs have poor reproducibility and diagnostic performance in image-based diagnosis of gallbladder polyps, limiting their direct clinical application. However, they demonstrate significant potential when used in a text-based strategy that uses a clinical scoring system, achieving performances comparable to radiologists. This model-based approach currently represents the most appropriate diagnostic strategy for LLMs in this domain.

Acknowledgments

Generative artificial intelligence (GenAI) was not used to generate any part of this paper. In this study, GenAI is only a part of the research design. All prompt templates sent to the GenAI are detailed in Multimedia Appendix 1.

Funding

This work was supported by the National Natural Science Foundation of China (No. 82572248 & No. 82102057).

Data Availability

The original ultrasound images and clinical records analyzed during this study are not publicly available due to patient confidentiality and privacy protection regulations. However, deidentified data underlying the reported results can be made available from the corresponding author upon reasonable request. Requests should include a detailed research proposal outlining the intended use of the data and must be approved by the institutional ethics committee that oversaw the original study.

Authors' Contributions

Conceptualization: MX, TH, LJ
Funding acquisition: TH, MX
Investigation: LJ, TH, JY, ZY, FT, X Zheng, X Zhang, XX
Methodology: LJ, TH, JY, ZY
Resources: MX, TH, XX
Supervision: MX, TH
Writing - original draft: LJ
Writing - review & editing: TH, MX

Conflicts of Interest

None declared.

Multimedia Appendix 1

Study protocols, prompt templates, statistical details, and extended results.
[DOCX File , 339 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Example No.1. An example of dot and sparse blood flow.
[PNG File , 244 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Example No.2. An example of single and sparse blood flow.
[PNG File , 254 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Example No.3. An example of branch-like and abundant blood flow.
[PNG File , 166 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Example No.4. An example of a pedunculated polyp.
[PNG File , 377 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Example No.5. An example of a sessile polyp.
[PNG File , 277 KB-Multimedia Appendix 6]

References

1. Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, et al. Large language models: a survey. arXiv. Preprint posted online on March 23, 2025. [doi: 10.48550/arXiv.2402.06196]
2. Voultsiou E, Vrochidou E, Moussiades L, Papakostas GA. The potential of large language models for social robots in special education. Prog Artif Intell. 2025;14(2):165-189. [doi: 10.1007/s13748-025-00363-2]
3. Hang CN, Wei Tan C, Yu P. MCQGen: a large language model-driven MCQ generator for personalized learning. IEEE Access. 2024;12:102261-102273. [doi: 10.1109/access.2024.3420709]
4. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
5. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res. 2023;25:e48568. [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
6. Chintagunta B, Katariya N, Amatriain X, Kannan A. Medically aware GPT-3 as a data generator for medical dialogue summarization. Association for Computational Linguistics; 2021. Presented at: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations; November 30, 2025:66-76; Online. [doi: 10.18653/v1/2021.nlpmc-1.9]
7. Margetts TJ, Karnik SJ, Wang HS, Plotkin LI, Oblak AL, Fehrenbacher JC, et al. Use of AI language engine ChatGPT 4.0 to write a scientific review article examining the intersection of Alzheimer's disease and bone. Curr Osteoporos Rep. 2024;22(1):177-181. [FREE Full text] [doi: 10.1007/s11914-023-00853-z] [Medline: 38225472]
8. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. JMIR Med Educ. 2023;9:e50945. [FREE Full text] [doi: 10.2196/50945] [Medline: 37578830]
9. Wu S, Tong W, Li M, Hu H, Lu X, Huang Z, et al. Collaborative enhancement of consistency and accuracy in US diagnosis of thyroid nodules using large language models. Radiology. 2024;310(3):e232255. [doi: 10.1148/radiol.232255] [Medline: 38470237]
10. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. Radiology. 2023;308(1):e231040. [doi: 10.1148/radiol.231040] [Medline: 37462501]
11. Chiang W, Zheng L, Sheng Y, Angelopoulos A, Li T, Li D. Chatbot Arena: an open platform for evaluating LLMs by human preference. arXiv. Preprint posted online on March 7, 2024. [doi: 10.48550/arXiv.2403.04132]
12. Li T, Chiang W, Song Y, Jain N, Dunlap L, Li D. Chatbot Arena Categories. 2024. URL: https://blog.lmarena.ai/blog/2024/arena-category/ [accessed 2025-12-04]

13. Lin W, Lin D, Tai D, Hsieh S, Lin C, Sheen I, et al. Prevalence of and risk factors for gallbladder polyps detected by ultrasonography among healthy Chinese: analysis of 34 669 cases. J Gastroenterol Hepatol. 2008;23(6):965-969. [doi: 10.1111/j.1440-1746.2007.05071.x] [Medline: 17725602]
14. Heitz L, Kratzer W, Gräter T, Schmidberger J, EMIL study group. Gallbladder polyps - a follow-up study after 11 years. BMC Gastroenterol. 2019;19(1):42. [FREE Full text] [doi: 10.1186/s12876-019-0959-3] [Medline: 30885181]
15. Roa I, de Aretxabala X, Araya JC, Roa J. Preneoplastic lesions in gallbladder cancer. J Surg Oncol. 2006;93(8):615-623. [doi: 10.1002/jso.20527] [Medline: 16724345]
16. Branch of Biliary Surgery, Chinese Society of Surgery, Chinese Medical Association, Chinese Medical Doctor Association in Chinese Committee of Biliary Surgeons. [Consensus on the surgical management of benign gallbladder diseases(2021 edition)]. Zhonghua Wai Ke Za Zhi. 2022;60(1):4-9. [doi: 10.3760/cma.j.cn112139-20210811-00373] [Medline: 34839607]
17. Albores-Saavedra J, Chablé-Montero F, González-Romo MA, Ramírez Jaramillo M, Henson DE. Adenomas of the gallbladder. Morphologic features, expression of gastric and intestinal mucins, and incidence of high-grade dysplasia/carcinoma in situ and invasive carcinoma. Hum Pathol. 2012;43(9):1506-1513. [doi: 10.1016/j.humpath.2011.11.011] [Medline: 22386521]
18. Taskin OC, Bellolio E, Dursun N, Seven IE, Roa JC, Araya JC, et al. Non-neoplastic polyps of the gallbladder: a clinicopathologic analysis of 447 cases. Am J Surg Pathol. 2020;44(4):467-476. [FREE Full text] [doi: 10.1097/PAS.0000000000001405] [Medline: 31725469]
19. Yuan H, Cao J, Kong W, Xia H, Wang X, Wang W. Contrast-enhanced ultrasound in diagnosis of gallbladder adenoma. Hepatobiliary Pancreat Dis Int. 2015;14(2):201-207. [doi: 10.1016/s1499-3872(15)60351-4] [Medline: 25865694]
20. Zhang H, Bai M, Gu J, He Y, Qiao X, Du L. Value of contrast-enhanced ultrasound in the differential diagnosis of gallbladder lesion. World J Gastroenterol. 2018;24(6):744-751. [FREE Full text] [doi: 10.3748/wjg.v24.i6.744] [Medline: 29456413]
21. Foley KG, Lahaye MJ, Thoeni RF, Soltes M, Dewhurst C, Barbu ST, et al. Management and follow-up of gallbladder polyps: updated joint guidelines between the ESGAR, EAES, EFISDS and ESGE. Eur Radiol. 2022;32(5):3358-3368. [FREE Full text] [doi: 10.1007/s00330-021-08384-w] [Medline: 34918177]
22. Wennmacker SZ, van Dijk AH, Raessens JHJ, van Laarhoven CJHM, Drenth JPH, de Reuver PR, et al. Polyp size of 1 cm is insufficient to discriminate neoplastic and non-neoplastic gallbladder polyps. Surg Endosc. 2019;33(5):1564-1571. [FREE Full text] [doi: 10.1007/s00464-018-6444-1] [Medline: 30203209]
23. Pickering O, Pucher PH, Toale C, Hand F, Anand E, Cassidy S, et al. Prevalence and sonographic detection of gallbladder polyps in a western European population. J Surg Res. 2020;250:226-231. [doi: 10.1016/j.jss.2020.01.003] [Medline: 32106001]
24. Vetrhus M, Berhane T, Søreide O, Søndenaa K. Pain persists in many patients five years after removal of the gallbladder: observations from two randomized controlled trials of symptomatic, noncomplicated gallstone disease and acute cholecystitis. J Gastrointest Surg. 2005;9(6):826-831. [doi: 10.1016/j.gassur.2005.01.291] [Medline: 15985239]
25. Farrugia A, Attard JA, Khan S, Williams N, Arasaradnam R. Postcholecystectomy diarrhoea rate and predictive factors: a systematic review of the literature. BMJ Open. 2022;12(2):e046172. [FREE Full text] [doi: 10.1136/bmjopen-2020-046172] [Medline: 35177439]
26. Kurokawa R, Ohizumi Y, Kanzawa J, Kurokawa M, Sonoda Y, Nakamura Y, et al. Diagnostic performances of Claude 3 Opus and Claude 3.5 sonnet from patient history and key images in radiology's "diagnosis please" cases. Jpn J Radiol. 2024;42(12):1399-1402. [doi: 10.1007/s11604-024-01634-z] [Medline: 39096483]
27. Sonoda Y, Kurokawa R, Nakamura Y, Kanzawa J, Kurokawa M, Ohizumi Y, et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "diagnosis please" cases. Jpn J Radiol. 2024;42(11):1231-1235. [doi: 10.1007/s11604-024-01619-y] [Medline: 38954192]
28. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. CLAIM 2024 Update Panel. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. Radiol Artif Intell. 2024;6(4):e240300. [FREE Full text] [doi: 10.1148/ryai.240300] [Medline: 38809149]
29. Fei X, Li N, Zhu L, Han P, Jiang B, Tang W, et al. Value of high frame rate contrast-enhanced ultrasound in distinguishing gallbladder adenoma from cholesterol polyp lesion. Eur Radiol. 2021;31(9):6717-6725. [doi: 10.1007/s00330-021-07730-2] [Medline: 33569621]
30. Gupta P, Dutta U, Rana P, Singhal M, Gulati A, Kalra N, et al. Gallbladder reporting and data system (GB-RADS) for risk stratification of gallbladder wall thickening on ultrasonography: an international expert consensus. Abdom Radiol (NY). Feb 2022;47(2):554-565. [doi: 10.1007/s00261-021-03360-w] [Medline: 34851429]
31. ChatGPT. OpenAI. URL: https://chatgpt.com/ [accessed 2025-12-04]
32. Claude. ANTHROPIC PBC. URL: https://claude.ai/new [accessed 2025-12-05]
33. Jiang L, Xu M, Yao J, Yang Z, Zhang X, Wu W, et al. Multilevel scoring systems based on ultrasound for differentiating between gallbladder adenomatous polyps and non-neoplastic polyps. Clin Radiol. 2025;89:107024. [doi: 10.1016/j.crad.2025.107024] [Medline: 40795450]
34. Shifai N, van Doorn R, Malvehy J, Sangers TE. Can ChatGPT vision diagnose melanoma? An exploratory diagnostic accuracy study. J Am Acad Dermatol. 2024;90(5):1057-1059. [FREE Full text] [doi: 10.1016/j.jaad.2023.12.062] [Medline: 38244612]

35. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. J Am Acad Orthop Surg. 2023;31(23):1173-1179. [FREE Full text] [doi: 10.5435/JAAOS-D-23-00396] [Medline: 37671415]
36. Horiuchi D, Tatekawa H, Oura T, Shimono T, Walston SL, Takita H, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. Eur Radiol. 2025;35(1):506-516. [doi: 10.1007/s00330-024-10902-5] [Medline: 38995378]
37. Huppertz MS, Siepmann R, Topp D, Nikoubashman O, Yüksel C, Kuhl CK, et al. Revolution or risk?-Assessing the potential and challenges of GPT-4V in radiologic image interpretation. Eur Radiol. 2025;35(3):1111-1121. [doi: 10.1007/s00330-024-11115-6] [Medline: 39422726]
38. Chen Z, Chambara N, Wu C, Lo X, Liu SYW, Gunda ST, et al. Assessing the feasibility of ChatGPT-4o and Claude 3-Opus in thyroid nodule classification based on ultrasound images. Endocrine. 2025;87(3):1041-1049. [doi: 10.1007/s12020-024-04066-x] [Medline: 39394537]
39. Zhou J, He X, Sun L, Xu J, Chen X, Chu Y, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. Nat Commun. 2024;15(1):5649. [FREE Full text] [doi: 10.1038/s41467-024-50043-3] [Medline: 38969632]
40. Li J, Guan Z, Wang J, Cheung CY, Zheng Y, Lim L, et al. Integrated image-based deep learning and language models for primary diabetes care. Nat Med. 2024;30(10):2886-2896. [doi: 10.1038/s41591-024-03139-8] [Medline: 39030266]
41. Zhao Z, Wang S, Gu J, Zhu Y, Mei L, Zhuang Z, et al. ChatCAD+: toward a universal and reliable interactive CAD using LLMs. IEEE Trans Med Imaging. 2024;43(11):3755-3766. [doi: 10.1109/TMI.2024.3398350] [Medline: 38717880]
42. Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, et al. MEDITRON-70B: scaling medical pretraining for large language models. arXiv. Preprint posted online on November 27, 2023. [doi: 10.48550/arXiv.2311.16079]
43. Liu X, Liu H, Yang G, Jiang Z, Cui S, Zhang Z, et al. A generalist medical language model for disease diagnosis assistance. Nat Med. 2025;31(3):932-942. [doi: 10.1038/s41591-024-03416-6] [Medline: 39779927]
44. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res. 2023;25:e48659. [FREE Full text] [doi: 10.2196/48659] [Medline: 37606976]
45. Giuffrè M, Kresevic S, You K, Dupont J, Huebner J, Grimshaw AA, et al. Systematic review: the use of large language models as medical chatbots in digestive diseases. Aliment Pharmacol Ther. 2024;60(2):144-166. [doi: 10.1111/apt.18058] [Medline: 38798194]
46. Fervers P, Hahnfeldt R, Kottlors J, Wagner A, Maintz D, Pinto Dos Santos D, et al. ChatGPT yields low accuracy in determining LI-RADS scores based on free-text and structured radiology reports in German language. Front Radiol. 2024;4:1390774. [FREE Full text] [doi: 10.3389/fradi.2024.1390774] [Medline: 39036542]
47. Cesur T, Güneş YC. Optimizing diagnostic performance of ChatGPT: the impact of prompt engineering on thoracic radiology cases. Cureus. 2024;16(5):e60009. [FREE Full text] [doi: 10.7759/cureus.60009] [Medline: 38854352]
48. Mitsuyama Y, Tatekawa H, Takita H, Sasaki F, Tashiro A, Oue S, et al. Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors. Eur Radiol. 2025;35(4):1938-1947. [doi: 10.1007/s00330-024-11032-8] [Medline: 39198333]
49. Sadamoto Y, Oda S, Tanaka M, Harada N, Kubo H, Eguchi T, et al. A useful approach to the differential diagnosis of small polypoid lesions of the gallbladder, utilizing an endoscopic ultrasound scoring system. Endoscopy. 2002;34(12):959-965. [doi: 10.1055/s-2002-35859] [Medline: 12471539]
50. Wang Y, Peng J, Liu K, Sun P, Ma Y, Zeng J, et al. Preoperative prediction model for non-neoplastic and benign neoplastic polyps of the gallbladder. Eur J Surg Oncol. 2024;50(2):107930. [FREE Full text] [doi: 10.1016/j.ejso.2023.107930] [Medline: 38159390]
51. Liu J, Qian Y, Yang F, Huang S, Chen G, Yu J, et al. Value of prediction model in distinguishing gallbladder adenoma from cholesterol polyp. J Gastroenterol Hepatol. 2022;37(10):1893-1900. [doi: 10.1111/jgh.15928] [Medline: 35750491]
52. Kim SY, Cho JH, Kim EJ, Chung DH, Kim KK, Park YH, et al. The efficacy of real-time colour Doppler flow imaging on endoscopic ultrasonography for differential diagnosis between neoplastic and non-neoplastic gallbladder polyps. Eur Radiol. 2018;28(5):1994-2002. [doi: 10.1007/s00330-017-5175-3] [Medline: 29218621]
53. Gu K, Lee JH, Shin J, Hwang JA, Min JH, Jeong WK, et al. Using GPT-4 for LI-RADS feature extraction and categorization with multilingual free-text reports. Liver Int. 2024;44(7):1578-1587. [doi: 10.1111/liv.15891] [Medline: 38651924]
54. Hui Z, Zewu Z, Jiao H, Yu C. Application of ChatGPT-assisted problem-based learning teaching method in clinical medical education. BMC Med Educ. 2025;25(1):50. [FREE Full text] [doi: 10.1186/s12909-024-06321-1] [Medline: 39799356]
55. Ch'en PY, Day W, Pekson RC, Barrientos J, Burton WB, Ludwig AB, et al. GPT-4 generated answer rationales to multiple choice assessment questions in undergraduate medical education. BMC Med Educ. 2025;25(1):333. [FREE Full text] [doi: 10.1186/s12909-025-06862-z] [Medline: 40038669]
56. Bazzari AH, Bazzari FH. Assessing the ability of GPT-4o to visually recognize medications and provide patient education. Sci Rep. 2024;14(1):26749. [FREE Full text] [doi: 10.1038/s41598-024-78577-y] [Medline: 39501020]

Abbreviations

CLAIM: Checklist for Artificial Intelligence in Medical Imaging
LLM: large language model
NPV: negative predictive value
PASS: power analysis and sample size
PMR: polyp morphology ratio
PPV: positive predictive value
UNRR: unnecessary resection rate of nonneoplastic polyps

Edited by A Benis; submitted 11.Jan.2025; peer-reviewed by B Jiang, CN Hang, X Liu, I Taskent, S Sivarajkumar; comments to author 13.Mar.2025; revised version received 19.Nov.2025; accepted 30.Nov.2025; published 23.Dec.2025

Please cite as:
Jiang L, Yao J, Yang Z, Tang F, Zheng X, Zhang X, Xie X, Xu M, Huang T
Evaluating Multiple Input Strategies of Large Language Models for Gallbladder Polyps on Ultrasound: Comparative Study
JMIR Med Inform 2025;13:e71178
URL: https://medinform.jmir.org/2025/1/e71178
doi: 10.2196/71178
PMID: