TYPE Original Research
PUBLISHED 03 July 2025
DOI 10.3389/fonc.2025.1613818

Performance of large language models in the differential diagnosis of benign and malignant biliary stricture

Chenxi Kang[1], Jing Li[1], Xintian Yang[1], Gui Ren[1], Linhui Zhang[1], Wei Wang[2], Xin Liu[3], Lei Wang[4], Guochen Shang[5], Jianglong Hong[6], Bingnian Wan[7], Yu Du[8], Wei Zeng[9], Yaling Liu[1], Tongxin Li[1], Lijun Lou[1], Hui Luo[1], Shuhui Liang[1], Yong Lv[1]* and Yanglin Pan[1]*

[1]Xijing Hospital of Digestive Diseases, Air Force Medical University, Xi'an, China, [2]Department of Gastroenterology, People's Liberation Army Joint Logistics Support Force 940th Hospital, Lanzhou, Gansu, China, [3]Department of Gastroenterology, Third People's Hospital of Gansu Province, Lanzhou, Gansu, China, [4]Department of Gastroenterology, Ankang Traditional Chinese Medicine Hospital, Ankang, China, [5]Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China, [6]First Affiliated Hospital of Anhui Medical University, Hefei, Anhui, China, [7]Yantai Ludong Hospital, Shandong Provincial Hospital Group, Yantai, Shandong, China, [8]Department of Gastroenterology, Qinzhou Second People's Hospital, Qinzhou, China, [9]Xiang'an Hospital, Xiamen University, Xiamen, Fujian, China

Background: Distinguishing benign from malignant biliary strictures remains challenging. Large Language Models (LLMs) show promise in enhancing diagnostic accuracy. This study aimed to evaluate the performances of ten LLMs in the differential diagnosis of benign and malignant biliary strictures.

Methods: Consecutive patients with biliary strictures undergoing endoscopic retrograde cholangiopancreatography (ERCP) at Xijing Hospital between January and December 2024 were retrospectively analyzed. Ten LLMs were systematically prompted with standardized clinical, laboratory, and imaging data. Performance was compared against tumor markers (CA19-9, CEA), a new multivariable clinical model, and ten independent pancreaticobiliary exoerienced physicians. Subgroup analyses assessed hilar (n=29) versus non-hilar strictures. Gold-standard diagnosis relied on histopathology and ≥3-month follow-up.

Results: Among the 159 included patients (83 benign, 76 malignant), four LLMs (Kimi, Deepseek-R1, Claude-3.5S, Llama-3.1), the clinical model (AUC:0.83), and six physicians achieved >80% accuracy. Kimi demonstrated superior accuracy (87%), significantly outperforming 70% of physicians (7/10, p<0.01). Three other LLMs (Deepseek-R1:83%, Claude-3.5S:82%, Llama-3.1:81%) and the clinical model performed comparably to physicians (78-84%, p>0.05), collectively surpassing tumor markers (CA19–9 accuracy:66%, CEA:71%). Physicians demonstrated higher accuracy for hilar strictures (87% vs. 79% for non-hilar, p<0.001). LLMs showed similar performance across stricture locations (hilar:64-95%; non-hilar:62-88%, p>0.05). For hilar strictures, 7/10 physicians achieved significantly higher accuracy (87-90%) than 8/10 LLMs (64-84%, p<0.05).

EDITED BY
Xin-Rong Yang,
Fudan University, China

REVIEWED BY
Lei Xu,
The First Affiliated Hospital of Xi'an Jiaotong University, China
Balu Bhasuran,
Florida State University, United States

*CORRESPONDENCE
Yong Lv
lvyongxj@yeah.net
Yanglin Pan
yanglinpan@hotmail.com

RECEIVED 17 April 2025
ACCEPTED 18 June 2025
PUBLISHED 03 July 2025

CITATION
Kang C, Li J, Yang X, Ren G, Zhang L, Wang W, Liu X, Wang L, Shang G, Hong J, Wan B, Du Y, Zeng W, Liu Y, Li T, Lou L, Luo H, Liang S, Lv Y and Pan Y (2025) Performance of large language models in the differential diagnosis of benign and malignant biliary stricture.
Front. Oncol. 15:1613818.
doi: 10.3389/fonc.2025.1613818

Conclusions: Using clinical, lab, and imaging data, some LLMs achieved diagnostic accuracy comparable to or exceeding clinical models and experienced physicians for differentiating benign versus malignant strictures. However, for hilar strictures, LLM performance was inferior to over half of the physicians.

KEYWORDS

large language model, biliary stricture, cholangiocarcinoma, prediction model, diagnosis

Introduction

Biliary strictures, characterized by abnormal bile duct narrowing, can significantly obstruct bile flow. An estimated 54-87% of biliary strictures are malignant (1–5), arising from local or metastatic cancers. Benign biliary strictures have heterogeneous etiologies including surgical bile duct injury, chronic pancreatitis, or chronic cholangiopathies (e.g., primary sclerosing cholangitis) (6). Benign and malignant strictures differ significantly in management and prognosis. Benign strictures are typically managed by endoscopic dilation, stenting, or surgery (7–9), while malignant strictures require aggressive approaches including surgical resection, palliative drainage, and systemic therapy (10–12). Thus, accurately distinguishing between them is crucial for guiding treatment and prognostic assessment.

Characteristics of biliary strictures are typically assessed via brushing cytology, forceps biopsy, or cholangioscopic biopsy during endoscopic retrograde cholangiopancreatography (ERCP) (13). Brush cytology and forceps biopsy demonstrate diagnostic accuracies of 15–80% (14, 15), while cholangioscopic biopsy achieves 70–87% (13, 16). Endoscopic ultrasound-guided (EUS) fine needle aspiration (FNA) or fine needle biopsy (FNB) shows favorable accuracy, particularly for extrinsic mass-related strictures (1). Advanced techniques like intraductal ultrasonography (IDUS) (17), probe-based confocal laser endomicroscopy (pCLE) (18), and optical coherence tomography (OCT) (19), offer enhanced precision but are limited by invasiveness, cost, and availability.

Significant differences also exist in common clinical parameters between benign and malignant strictures, including age of onset, duration of liver function abnormalities, previous surgeries, and tumor markers. These noninvasive, accessible parameters facilitate convenient prediction. Wang et al. used CA50, CA19-9, and AFP to achieve an AUC of 0.879 (95% CI: 0.841–0.917) (20), while Zhang et al. combined MRI with inflammatory markers for an AUC of 0.802 (95% CI: 0.719–0.870) (18). Though promising, these models require further validation.

LLMs including Deepseek-R1, GPT-4T, Claude-3.5S, and Llama-3.1 show potential in improving diagnostic accuracy across medicine (21–24). Trained on vast medical data, LLMs may assist preliminary diagnosis by reducing interpretational variability. However, their utility for differentiating biliary strictures—a specialized, less common condition—remains unknown. We hypothesize that LLMs leveraging common clinical data (manifestations, blood tests, imaging) could aid this differentiation.

In this study, we aimed to evaluate the diagnostic performance of ten distinct LLMs for diagnosing benign versus malignant biliary strictures, comparing performance against tumor markers, a novel clinical model, and ten experienced pancreaticobiliary specialists.

Methods

Study design

This retrospective study evaluated the diagnostic performance of ten distinct large language models (LLMs) in differentiating between benign and malignant biliary strictures. The study protocol was approved by the Ethics Committee of Xijing Hospital. The written informed consent was obtained from all the patients or their next of kin.

Patients

Consecutive patients aged ≥18 years admitted to Xijing Hospital for biliary stricture evaluation between January and December 2024 were eligible. Inclusion criteria required a definitive etiological diagnosis confirmed by pathological examination and regular follow-up exceeding 3 months. Patients with incomplete medical records were excluded. Malignancy diagnosis relied on pathological results obtained via endoscopic retrograde cholangiopancreatography (ERCP), percutaneous transhepatic cholangiographic drainage (PTCD), endoscopic ultrasound (EUS), biopsy, or surgery. Benign strictures required confirmation by benign pathology and absence of progression over ≥3 months.

Data collection

Demographic, clinical, imaging, and pathological data were extracted from electronic medical records. An independent physician standardized data inputs for ten LLMs, ensuring unbiased differentiation. Case data was structured uniformly, excluding diagnostic conclusions, and included clinical presentation, history, imaging reports (CT, MRI, MRCP), and lab results (complete blood count, liver and renal function tests, lipids, coagulation, tumor/inflammatory markers). Any indications of benignancy or malignancy from the reports were removed. A flowchart of the overall study design is shown in Figure 1, illustrating the process from patient selection to diagnostic performance evaluation.

Large language models for differential diagnosis of biliary strictures

Ten LLMs were selected for evaluation for reproducibility, including 1) mainstream commercial models with proven medical reasoning capabilities in prior studies (GPT-4T, GPT-4o, Gemini-1.5 pro, Claude-3.5S), 2) models developed by Chinese companies to align with the Chinese-language clinical data (Kimi, ERNIE-4, Qwen-2, GLM-4); 3) open-source models (Deepseek-R1, Llama-3.1) (25, 26). To ensure consistency and reproducibility, a structured query approach was used. A case-specific prompt simulated consultation with an experienced pancreaticobiliary specialist: "Supposing you are an experienced physician specializing in pancreaticobiliary diseases, when encountering a case with biliary stricture, please provide a tentative judgment indicating whether the cause of the stricture is more likely to be malignant or benign." (Detailed prompt provided in Supplementary Method S1). For each LLM, a new conversation session was initiated to eliminate contextual carryover. Probabilistic LLM outputs (e.g., "likely benign," "possibly malignant") were converted into binary outcomes (benign/malignant) using predefined rules. Responses indicating malignancy (e.g., "likely malignant," "probably malignant," "suggestive of malignancy") were classified as "malignant." Responses indicating benignancy (e.g., "likely benign," "probably benign," "suggestive of a benign process") were classified as "benign." Explicit statements ("benign"/ "malignant") were directly categorized.

All queries employed identical deterministic parameters: temperature=0.0 (output consistency), max tokens=10240, and consistent system prompts. Queries were executed in January 2025 using contemporaneous model versions. The evaluated LLMs included: Deepseek-R1, GPT-4T, GPT-4o, Claude-3.5S, Gemini-1.5 Pro, Kimi, Llama-3.1 405B, ERNIE-4.0-Turbo-8K, Qwen-2-72B, and ChatGLM-4-9B (details in Supplementary Table S1). Chain-of-thought (CoT) outputs were extracted for reasoning pattern analysis. Cases were categorized by diagnostic accuracy (correct/incorrect), and reasoning traces were independently assessed by two gastroenterologists using predefined clinical logic criteria (Supplementary Table S3).

Development of a clinical prediction model

A multivariable logistic regression model incorporating key demographic, clinical, laboratory, and imaging parameters was developed. The modeling process initiated with univariable screening to identify candidate variables ($p < 0.10$), followed by multivariable regression analysis. Variables were retained multivariable models based on statistical significance ($p < 0.05$) or established clinical relevance for variables approaching significance. Collinearity was assessed using variance inflation factors (VIF) and Spearman correlation coefficients ($|\rho| > 0.6$), with clinical importance determining variable selection when collinearity was detected. Bidirectional stepwise selection (forward/backward) using Akaike (AIC) and Bayesian (BIC) information criteria optimized the model. Final model selection prioritized minimized AIC/BIC and maximized predictive performance, with internal validation implemented through bootstrapping (1,000 resamples). CA19–9 and CEA were separately assessed as independent diagnostic markers.

Experienced physicians' evaluation

Ten experienced pancreaticobiliary specialists (each with ≥10 years of clinical practice) independently evaluated comprehensive clinical summaries identical to those processed by the LLMs. All diagnostic predictions regarding benign or malignant status were made without intercommunication among physicians to ensure independent assessment.

Outcome

The primary outcome was diagnostic accuracy that was calculated as the average of sensitivity and specificity to account for class imbalance (range: 0-1, higher values superior). For LLMs, the highest accuracy from duplicate assessments was selected. Secondary performance metrics included sensitivity (proportion of true positives correctly identified), specificity (proportion of true negatives correctly identified), positive predictive value (PPV, proportion of positive predictions that were correct), negative predictive value (NPV, proportion of negative predictions that were correct), and F1-score (harmonic mean of precision and sensitivity providing balanced assessment).

Statistical analysis

Continuous variables are presented as mean ± standard deviation (normally distributed) or median [interquartile range] (non-normal distributions), while categorical variables are expressed as proportions (%) with odds ratios (OR) and 95% confidence intervals (CI) for association analyses. McNemar and DeLong tests were used to compare performance metrics between models. Confidence intervals for accuracy, sensitivity, and specificity were calculated using the Clopper-Pearson method, with proportion differences reported in pairwise comparisons. Subgroup analysis evaluated LLM diagnostic performance by stricture location (hilar vs. non-hilar). Internal concordance of LLMs was assessed through weighted Cohen's kappa (κ) with 95% CI based on duplicate assessments, interpreted as: $\kappa \le 0.20$ (slight concordance), 0.21-0.40 (fair), 0.41-0.60 (moderate), 0.61-0.80 (substantial), and 0.81-1.00 (almost perfect). Pairwise comparisons of classification accuracy employed McNemar's test with Holm-Bonferroni correction for multiple comparisons (family-wise $\alpha = 0.05$; significance threshold: adjusted $p < 0.05$). All statistical tests were two-sided. Statistical analyses were conducted using R version 4.3.1.

Results

Patient demographics and clinical characteristics

During the study period, 270 patients were diagnosed with biliary stricture, of whom 111 were excluded according to inclusion/exclusion criteria, resulting in a final cohort of 159 patients (83 benign, 76 malignant). Baseline demographic and clinical characteristics are summarized in Table 1. Compared with benign stricture patients, those with malignant lesions were older (65.13 vs. 57.19 years), exhibited higher bilirubin (149.14 vs. 78.15 μmol/L, $p < 0.001$) and CA19–9 levels (2888.33 vs. 246.32 U/mL, $p = 0.003$), a higher proportion of lymph node enlargement (47.4% vs. 19.3%, $p < 0.001$), and more frequent hilar strictures (26.3% vs. 10.8%, $p = 0.020$).

Clinical data-driven diagnostic model performance

A stepwise multivariable logistic regression was employed to develop the diagnostic model, with the final model retaining age, CA19-9, CEA, disease duration <1 month, C-reactive protein (CRP), surgical history, and lymph node enlargement (Table 2). Supplementary Figure S1 illustrates the impact of hyperparameter λ on model accuracy under L2 regularization, with the optimal λ value of 149.3 determined via 10-fold cross-validation to maximize test-set accuracy. The model achieved an AUC of 0.83 (95% CI: 0.70–0.96), accuracy of 0.83, sensitivity of 0.83, and specificity of 0.82 (Table 3, Figure 2), outperforming tumor markers alone. CA19–9 showed an AUC of 0.77 (95% CI: 0.69–0.84), accuracy of 0.66, specificity of 0.93, and sensitivity of 0.39; CEA yielded an AUC of 0.66 (95% CI: 0.58–0.75), accuracy of 0.71, specificity of 0.59, and sensitivity of 0.83. Internal validation via bootstrapping (1000 iterations) confirmed consistent AUC of 0.83 (Supplementary Figure S2).

Comparative diagnostic performance of LLMs, tumor markers, clinical model, and physicians

Four LLMs (40%), one clinical model, and six physicians (60%) achieved accuracies ≥80%. Kimi demonstrated the highest accuracy (87%), significantly outperforming 70% of physicians (7/10, $p < 0.01$) (Table 3, Figure 3, Supplementary Table S2). Top-performing LLMs including Deepseek-R1 (0.83), Claude-3.5S (0.82), Llama-3.1 (0.81), and GPT-4T (0.79) showed comparable accuracy to physicians (0.78–0.84, $p > 0.05$), while physicians outperformed lower-performing LLMs (GPT-4o, Gemini-1.5-pro). Eighty percent of LLMs exceeded CEA (0.66) and CA19-9 (0.71) in accuracy, with the clinical model (0.83) competing with top LLMs.

CA19–9 exhibited the highest sensitivity (0.93) across 23 predictive groups, significant in 18 groups (p: 0–0.044). GPT-4T, GPT-4o, and Gemini-1.5-pro showed the highest specificity (0.99), significant in 14 groups (p: 0–0.048). Kimi led in F1-score (0.87, significant in 12 groups), GPT-4T in PPV (0.98, 13 groups), and EP1 in NPV (0.85, 12 groups). Top LLMs dominated four metrics (accuracy, specificity, F1-score, PPV), while tumor markers and physicians excelled in sensitivity and NPV, respectively.

Analysis of 360 incorrect diagnoses revealed that over 80% resulted from LLMs over-relying on single data sources (Supplementary Table S6). Representative examples: Claude 3.5 misdiagnosed a benign stricture as malignant solely based on elevated CA19-9, while GPT-4T correctly integrated bilirubin trends, imaging, and histology (Supplementary Material).

Analysis of the misdiagnoses revealed >80% originated from LLMs over-relying on single data sources (Supplementary Table S3). For instance, Claude 3.5 misdiagnosed a benign stricture as malignant based solely on elevated CA19-9, whereas GPT-4T integrated bilirubin trends, imaging, and histology for correct diagnosis (Supplementary Document).

Subgroup analysis by stricture location

Performance metrics for hilar (n=29, 9 benign/20 malignant) and non-hilar subgroups are detailed in Figure 4 and Supplementary Table S4. Given the limited sample size of hilar strictures (n=29, including 9 benign and 20 malignant cases), the subgroup analysis should be interpreted with caution due to potential overfitting risks. Despite this limitation, we observed that in Physicians showed higher accuracy in hilar versus non-hilar strictures (0.87 vs. 0.79, $p < 0.001$), while LLMs had comparable accuracy in both subgroups (hilar: 0.64-0.95; non-hilar: 0.62-0.88, $p > 0.05$). In the hilar subgroup, Deepseek-R1 showed highest hilar accuracy (0.95, 95% CI:0.77-0.99), followed by Kimi (0.87) and Claude-3.5S (0.84). Notably in this exploratory analysis, 7/10 physicians achieved superior accuracy over 8/10 LLMs (87-90% vs. 64-84%, $p<0.05$). In the non-hilar subgroup, LLMs showed competitive/exceeding accuracy. Given the small sample size, these findings should be interpreted as preliminary and require validation in larger cohorts.

LLM diagnostic concordance assessment

Deepseek-R1, GPT-4T, GPT-4o, Llama-3.1, Gemini-1.5-pro, and Claude-3.5S showed near-perfect internal concordance

TABLE 1 Patient demographics, clinical characteristics, and diagnostic markers.

Variables	Overall (n=159)	Benign (n=83)	Malignant (n=76)	P value
Age (year)	60.99 (14.03)	57.19 (16.22)	65.13 (9.70)	<0.001
Male, n (%)	66 (41.5)	35 (42.2)	31 (40.8)	0.988
BMI (kg/m^2)	20.91 (3.43)	21.55 (3.20)	20.17 (3.58)	0.075
Prior surgical history *, n (%)	85 (53.5)	51 (61.4)	34 (44.7)	0.051
Disease duration < 1month †, n (%)	135 (61.4)	69 (57.0)	66 (66.7)	0.005
Traditional tumor markers				
CA125 (U/ml)	48.36 (131.48)	26.23 (51.27)	72.53 (180.03)	0.026
CA19-9 (U/ml)	1435.23 (6681.62)	246.32 (1854.95)	2888.33 (9574.67)	0.003
CEA (ng/ml)	15.00 (77.00)	3.22 (3.21)	29.40 (113.39)	0.012
AFP (ng/ml)	5.72 (17.88)	5.04 (15.03)	6.47 (20.62)	0.616
Whole blood tests				
White blood cell (cell/µL)	6570 (3050)	6410 (2420)	6740 (3620)	0.491
Hemoglobin (g/dL)	12.20 (2.05)	12.75 (1.90)	11.67 (2.06)	0.002
Platelet count (× 103/µL)	214.06 (79.79)	206.84 (76.57)	221.94 (82.95)	0.235
Liver/renal function tests				
Total bilirubin (µmol/L)	112.52 (119.36)	78.15 (105.85)	149.14 (122.78)	<0.001
Albumin (g/dL)	3.78 (0.68)	3.94 (0.66)	3.61 (0.66)	0.002
Creatinine (mg/dL)	0.85 (0.18)	0.84 (0.18)	0.85 (0.17)	0.668
Sodium (mEq/L)	141.96 (2.83)	141.92 (2.84)	142.00 (2.84)	0.858
Coagulation profiles				
International normalized ratio	1.09 (0.16)	1.10 (0.18)	1.09 (0.14)	0.605
Lipid profiles				
Total cholesterol (mg/dL)	169.50 (41.70)	171.81 (44.79)	167.18 (37.84)	0.477
Triglyceride (mg/dL)	111.50 (86.73)	110.62 (73.45)	112.39 (100.00)	0.859
Inflammatory markers				
CRP (mg/dL)	2.02 (3.18)	2.21 (3.48)	1.81 (2.83)	0.431
IL6 (pg/ml)	33.43 (363.27)	3.20 (2.72)	66.45 (525.24)	0.274
PCT (ng/ml)	1.14 (4.06)	0.60 (1.79)	1.73 (5.53)	0.08
Immune abnormalities‡, n (%)	26 (16.4)	11 (13.3)	15 (19.7)	0.174
Lymph node enlargement, n (%)	52 (32.7)	16 (19.3)	36 (47.4)	<0.001
Biliary stricture sites, n (%)				0.020
Hilar	29 (18.2)	9 (10.8)	20 (26.3)	
Non-hilar	130 (81.8)	74 (89.2)	56 (73.7)	

Data are mean (standard deviation) or numbers (percentages) unless otherwise specified. BMI, Body Mass Index; CA19-9, Carbohydrate Antigen 199; CEA, Carcinoembryonic Antigen; AFP, Alpha-Fetoprotein; CA125, Carbohydrate Antigen 125; CRP, C-Reactive Protein; IL6, Interleukin6; PCT, Procalcitonin.

*Surgical history was defined as a history of liver transplantation and biliary surgery.

†Disease duration < 1 month was defined as a period less than one month from the onset of symptoms.

‡Immune abnormalities were defined as the presence of IgG4 subclass abnormalities, autoimmune diseases, or abnormalities detected in a series of autoantibody tests.

TABLE 2 Selection of variables based on univariate and multivariate logistic regression analysis.

Variables	Univariate Logistic Regression		Multivariate Logistic Regression	
	OR (95%CI)	P	OR (95%CI)	P
Age > 55 (year)	4.53 (2.09, 9.8)	<0.001	4.68 (1.65, 14.51) §	0.005
Male	1.06 (0.56, 1.99)	0.86		
Prior Surgical history *	0.51 (0.27, 0.96)	0.04	0.74 (0.27, 1.99) §1	0.549
Disease duration < 1mouth †	2.79 (1.4, 5.57)	<0.001	3.48 (1.32, 9.79) §	0.014
Traditional tumor markers				
CA125 > 20 (U/ml)	3.79 (1.95, 7.35)	<0.001	1.74 (0.64, 4.81)	0.278
CA19-9 > 30(U/ml)	6.98 (3.33, 14.64)	<0.001	5.20 (1.81, 16.16) §	0.003
CEA > 5 (ng/ml)	8.37 (3.24, 21.63)	<0.001	4.99 (1.58, 18.14) §	0.009
AFP > 4 (ng/ml)	0.44 (0.23, 0.87)	0.02	1.32 (0.48, 3.73)	0.596
Whole blood tests				
White blood cell > 6000 (cell/μL)	0.52 (0.28, 0.99)	0.05	0.53 (0.21, 1.32)	0.179
Hemoglobin > 12.5 (g/dL)	0.50 (0.26, 0.96)	0.04	0.71 (0.30, 2.23)	0.487
Platelet count > 250 (× 103/μL)	1.51 (0.79, 2.90)	0.21		
Liver/renal function tests				
Total bilirubin > 2 (mg/dL)	5.14 (2.55, 10.39)	<0.001		
Albumin > 4 (g/dL)	0.38 (0.20, 0.75)	<0.001	0.53 (0.18, 1.49)	0.23
Creatinine > 0.75 (mg/dL)	1.67 (0.82, 3.40)	0.15		
Sodium > 140 (mEq/L)	1.51 (0.80, 2.85)	0.21		
Coagulation profiles				
International normalized ratio > 1	2.04 (0.94, 4.47)	0.07		
Inflammatory markers				
CRP > 0.6 (mg/dL)	0.42 (0.21, 0.82)	0.01	0.23 (0.08, 0.63)	0.006
IL6 > 1 (pg/ml)	2.52 (1.07, 5.92)	0.03	2.72 (0.81, 9.81)	0.112
PCT > 0.1 (ng/ml)	1.77 (0.94, 3.33)	0.08		
Immune abnormal ‡	1.61 (0.69, 3.76)	0.27		
Lymph node enlargement	3.77 (1.86, 7.64)	<0.001	2.90 (0.97, 9.16) §	0.06
Biliary stricture sites: Hilar	0.34 (0.14, 0.80)	0.01	0.50 (0.13, 1.77)	0.29

CA19-9, Carbohydrate Antigen 199; CEA, Carcinoembryonic Antigen; AFP, Alpha-Fetoprotein; CA125, Carbohydrate Antigen 125; CRP, C-Reactive Protein; IL6, Interleukin6; PCT, Procalcitonin; OR, odds ratio; CI, Confidence Interval; NA, Not Applicable.

*Surgical history was defined as a history of liver transplantation and biliary tract surgery.

†Disease duration was defined as a period less than one month from the onset of symptoms.

‡Immune abnormal was defined as the presence of immunoglobulin subclass 4 abnormalities, or having an autoimmune disease, or abnormalities in a series of autoantibody tests.

§Variables included in the clinical model.

§1Variables included based on clinical relevance despite borderline univariate p-values (p<0.05).

(κ=0.81-0.97). Kimi, ERNIE-4, and GLM-4 demonstrated substantial concordance, while Qwen-2 showed fair concordance. When compared to true values, no model reached almost perfect concordance: Deepseek-R1, Llama-3.1, Claude-3.5S, and Kimi showed substantial concordance; ERNIE-4, GLM-4, Qwen-2 moderate; Gemini-1.5 pro and GPT-4o fair (Supplementary Table S6, Supplementary Figure S3).

Discussion

The differentiation between benign and malignant biliary strictures remains clinically challenging. Current endoscopic techniques show limited sensitivity: ERCP-based brushing/biopsy (0.45-0.67) (27–29) and cholangioscopy biopsy (0.43-0.74) (2, 30) often prove inadequate. Our study demonstrates that select large

TABLE 3 Performance metrics of LLMs, EPs, clinical model and tumor markers.

Predictors	Acc, (95%CI)	Sens	Spec	F1	PPV	NPV	AUC, (95%CI)
LLMs							
GPT-4T	0.79 (0.71, 0.84)	0.59 (0.51, 0.64)	0.99 (0.92, 1.00)	0.74	0.98	0.69	–
GPT-4o	0.66 (0.57, 0.72)	0.34 (0.30, 0.39)	0.99 (0.92, 1.00)	0.50	0.97	0.58	–
Claude-3.5S	0.82 (0.74, 0.87)	0.71 (0.65, 0.76)	0.92 (0.88, 0.97)	0.80	0.91	0.74	–
Gemini-1.5 pro	0.62 (0.52, 0.68)	0.25 (0.19, 0.30)	0.99 (0.92, 1.00)	0.40	0.95	0.55	–
Kimi	0.87 (0.81, 0.92)	0.83 (0.77, 0.89)	0.91 (0.85, 0.96)	0.87	0.91	0.83	–
ERNIE-4	0.79 (0.71, 0.85)	0.66 (0.61, 0.72)	0.92 (0.88, 0.95)	0.76	0.90	0.71	–
Llama-3.1	0.81 (0.73, 0.86)	0.65 (0.60, 0.72)	0.97 (0.91, 0.99)	0.78	0.96	0.72	–
Qwen-2	0.76 (0.68, 0.82)	0.65 (0.60, 0.71)	0.87 (0.82, 0.92)	0.73	0.84	0.69	–
GLM-4	0.77 (0.69, 0.83)	0.69 (0.62, 0.75)	0.86 (0.82, 0.90)	0.75	0.84	0.71	–
Deepseek-R1	0.83 (0.76, 0.89)	0.81 (0.77, 0.86)	0.86 (0.81, 0.91)	0.83	0.86	0.80	–
Experienced Physicians							
EP1	0.84 (0.78, 0.90)	0.87 (0.82, 0.91)	0.82 (0.77, 0.88)	0.85	0.84	0.85	–
EP2	0.80 (0.73, 0.86)	0.80 (0.73, 0.85)	0.80 (0.75, 0.86)	0.80	0.81	0.78	–
EP3	0.79 (0.71, 0.84)	0.64 (0.59, 0.71)	0.93 (0.85, 0.96)	0.75	0.91	0.70	–
EP4	0.81 (0.74, 0.87)	0.84 (0.79, 0.88)	0.78 (0.72, 0.85)	0.82	0.80	0.82	–
EP5	0.81 (0.74, 0.87)	0.81 (0.77, 0.87)	0.82 (0.76, 0.88)	0.82	0.83	0.79	–
EP6	0.80 (0.73, 0.86)	0.76 (0.70, 0.82)	0.84 (0.79, 0.88)	0.80	0.84	0.76	–
EP7	0.79 (0.71, 0.85)	0.70 (0.65, 0.76)	0.88 (0.82, 0.93)	0.77	0.87	0.73	–
EP8	0.80 (0.72, 0.85)	0.66 (0.61, 0.72)	0.93 (0.88, 0.96)	0.77	0.92	0.72	–
EP9	0.78 (0.70, 0.84)	0.71 (0.66, 0.78)	0.84 (0.79, 0.89)	0.77	0.83	0.73	–
EP10	0.79 (0.72, 0.85)	0.77 (0.71, 0.83)	0.82 (0.78, 0.89)	0.80	0.82	0.77	–
Clinical Model							
Clinical Model	0.83 (0.69, 0.92)	0.83 (0.77, 0.88)	0.82 (0.76, 0.89)	0.83	0.83	0.82	0.83, (0.70, 0.96)
Tumor markers							
CA19-9	0.66 (0.59, 0.75)	0.93 (0.88, 0.97)	0.39 (0.35, 0.44)	0.75	0.63	0.83	0.77, (0.69, 0.84)
CEA	0.71 (0.63, 0.77)	0.59 (0.55, 0.64)	0.83 (0.78, 0.59)	0.68	0.79	0.65	0.66, (0.58, 0.75)
P *	0.016	0.201	0.769	0.678	1.000	0.769	0.387
P †	<0.001	0.013	0.646	0.029	0.009	0.646	0.031

CA19-9, Carbohydrate Antigen 199; CEA, Carcinoembryonic Antigen; AUC, Area Under the Curve; Acc, Accuracy; Sens, Sensitivity; Spec, Specificity; F1, F1 score; PPV, Positive Predictive Value; NPV, Negative Predictive Value; CI, Confidence Interval; EP, Experienced Physician.

*p value of Logistic Prediction Model versus CA19-9.

†p value of Logistic Prediction Model versus CEA.

language models (LLMs) achieve diagnostic accuracy rivaling or exceeding human expertise. Kimi attained the highest accuracy (87%), significantly outperforming 70% of experienced physicians (7/10, p<0.01). Three additional LLMs (Deepseek-R1:83%, Claude-3.5S:82%, Llama-3.1:81%) and our clinical prediction model (83%) performed comparably to physicians (78-84%, p>0.05). Collectively, 80% of LLMs surpassed conventional tumor markers (CA19–9 accuracy:66%; CEA:71%).

Notably, this is the first study to demonstrate that select large language models (LLMs) match or exceed the accuracy of a clinical model and experienced physicians. These findings suggest LLMs could serve as accessible, real-time diagnostic aids, particularly in resource-constrained settings where specialist expertise is limited. The variation among LLM performance reveals clinically meaningful insights. While GPT-4o and Gemini-1.5-Pro lead general benchmarking tasks (31), these models underperformed in our specific diagnostic application (accuracies 0.66 and 0.62 respectively). These underperforming models exhibited extreme specificity (0.99) and PPV (0.95-0.97) but critically low sensitivity (0.25-0.34), likely reflecting excessive safety prioritization during

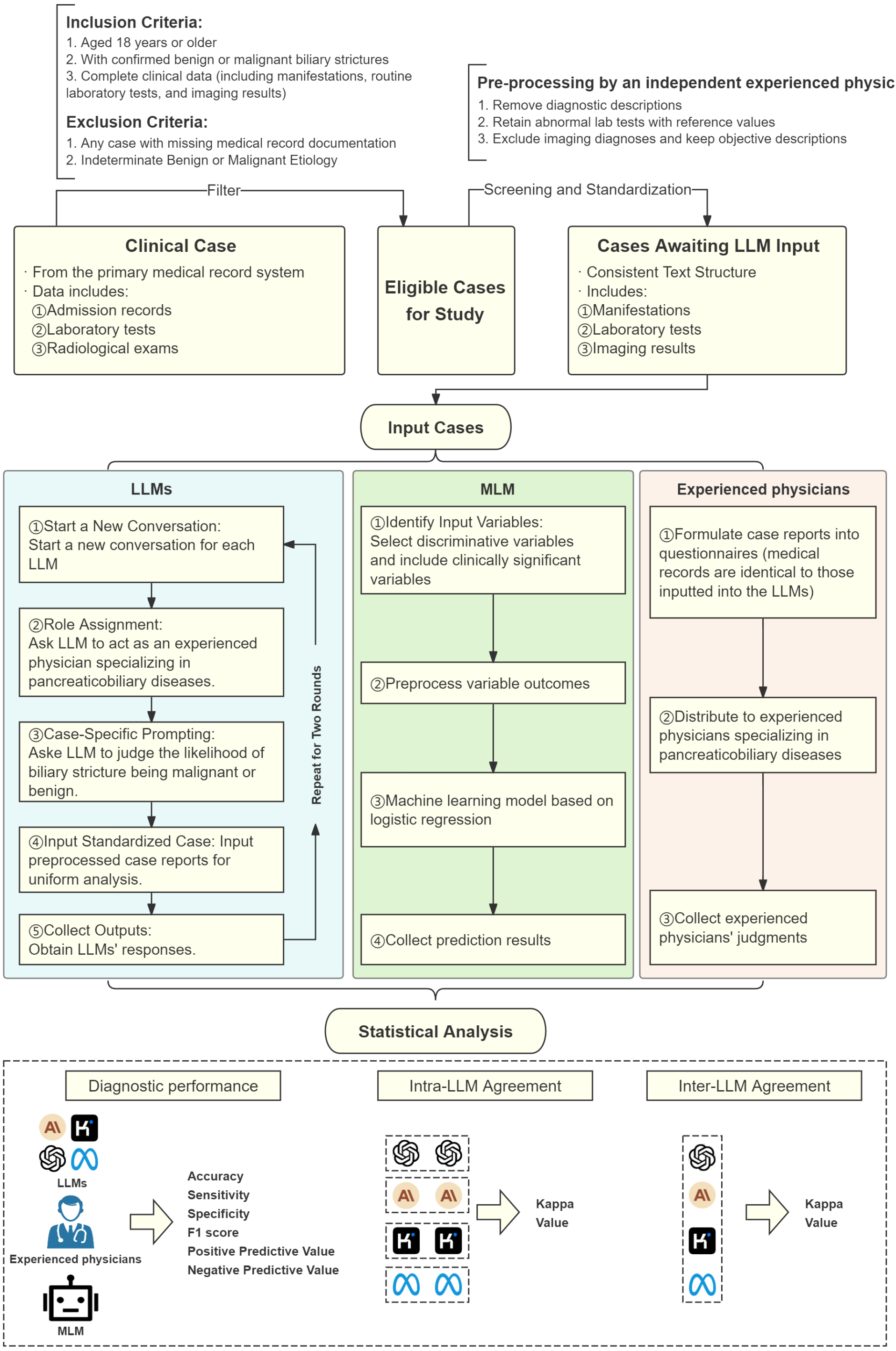
Inclusion Criteria:
1. Aged 18 years or older
2. With confirmed benign or malignant biliary strictures
3. Complete clinical data (including manifestations, routine laboratory tests, and imaging results)
Exclusion Criteria:
1. Any case with missing medical record documentation
2. Indeterminate Benign or Malignant Etiology
Pre-processing by an independent experienced physician
1. Remove diagnostic descriptions
2. Retain abnormal lab tests with reference values
3. Exclude imaging diagnoses and keep objective descriptions
Filter
Screening and Standardization
Clinical Case
· From the primary medical record system
· Data includes:
①Admission records
②Laboratory tests
③Radiological exams
Eligible Cases for Study
Cases Awaiting LLM Input
· Consistent Text Structure
· Includes:
①Manifestations
②Laboratory tests
③Imaging results
Input Cases
LLMs
①Start a New Conversation: Start a new conversation for each LLM
②Role Assignment: Ask LLM to act as an experienced physician specializing in pancreaticobiliary diseases.
③Case-Specific Prompting: Aske LLM to judge the likelihood of biliary stricture being malignant or benign.
④Input Standardized Case: Input preprocessed case reports for uniform analysis.
⑤Collect Outputs: Obtain LLMs' responses.
Repeat for Two Rounds
MLM
①Identify Input Variables: Select discriminative variables and include clinically significant variables
②Preprocess variable outcomes
③Machine learning model based on logistic regression
④Collect prediction results
Experienced physicians
①Formulate case reports into questionnaires (medical records are identical to those inputted into the LLMs)
②Distribute to experienced physicians specializing in pancreaticobiliary diseases
③Collect experienced physicians' judgments
Statistical Analysis
Diagnostic performance
LLMs
Experienced physicians
MLM
Accuracy
Sensitivity
Specificity
F1 score
Positive Predictive Value
Negative Predictive Value
Intra-LLM Agreement
Kappa Value
Inter-LLM Agreement
Kappa Value

FIGURE 1
Flowchart of overall study design. LLM, Large Language Model; MLM, Machine Learning Model.

training protocols. This pattern necessitates caution when using such models for biliary stricture assessment. Conversely, Kimi's superior performance (87%) highlights how task-specific optimization can yield exceptional diagnostic capability irrespective of general benchmarking performance.

Our subgroup analysis revealed physicians outperformed LLMs in the evaluation of hilar strictures (n=29), with 7/10 physicians achieving significantly higher accuracy than 8/10 LLMs (87-90% vs. 64-84%, p<0.05). This finding aligns with established clinical knowledge that >90% of hilar strictures are malignant (1), suggesting experienced clinicians better integrate this epidemiological context. The performance gap may indicate incomplete learning of clinical nuances by current LLMs. However, targeted fine-tuning with medical knowledge or Retrieval-Augmented Generation (RAG) (32–34) could

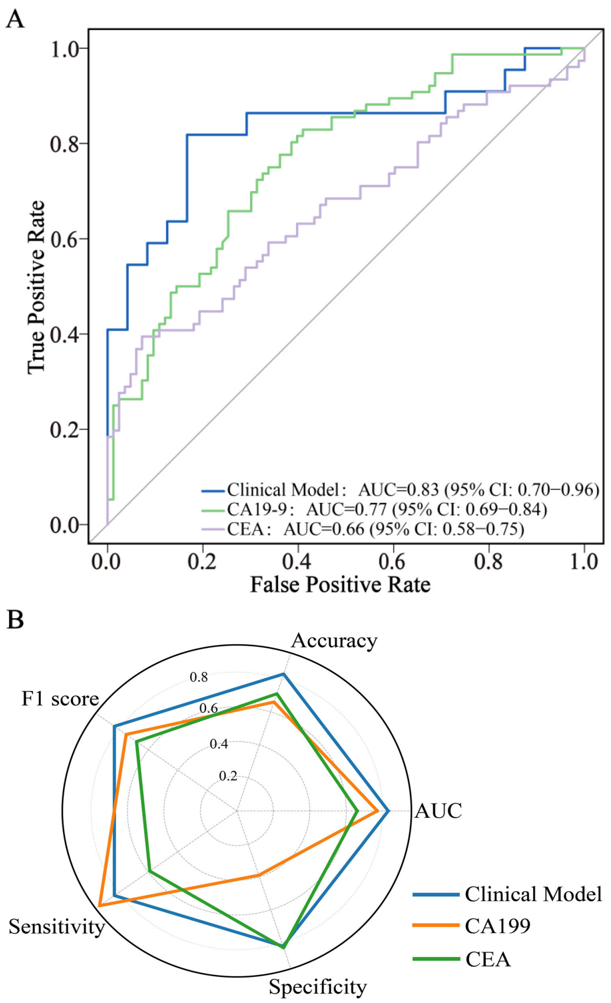

FIGURE 2
ROC curve analysis **(A)** and radar chart **(B)** of diagnostic model and tumor markers. CA19-9, Carbohydrate Antigen 199; CEA, Carcinoembryonic Antigen; AUC, Area Under the Curve; CI, Confidence Interval; ROC, Receiver Operating Characteristic.

potentially bridge this gap in future iterations. Clinically, this underscores the continued value of expert judgment in anatomically complex presentations, while suggesting LLMs may currently serve best as diagnostic aids for non-hilar strictures where they demonstrated parity with physicians. However, due to due to small sample size in the subgroup of hilar strictures, this analysis is exploratory and requires validation.

Our clinical prediction model, incorporating established risk factors (age, CA19-9, CEA, disease duration <1 month, CRP, surgical history, lymphadenopathy), achieved an AUC of 0.83 (95% CI:0.70-0.96), aligning with prior reports (AUC 0.75-0.83) (35–38) while maintaining practical clinical utility. CA19–9 demonstrated an expected AUC (0.77) matching literature reports (0.759-0.783) (35, 38), but presented a diagnostic paradox with high sensitivity (0.93) yet poor specificity (0.39). This suggests potential utility as a rule-out screening tool requiring subsequent confirmation, while CEA demonstrated weaker discriminative capacity (AUC 0.66) than some prior studies (39–42), emphasizing context-dependent variability.

Despite promising diagnostic capabilities, clinical implementation of LLMs faces significant barriers requiring strategic resolution. Error analysis demonstrated that >80% of misdiagnoses originated from LLMs over-relying on isolated data elements rather than multimodal integration. Representative examples included Claude 3.5 misclassifying a benign stricture as malignant based solely on elevated CA19-9, contrasting with GPT-4T’s accurate diagnosis achieved through synthesizing bilirubin trends, imaging findings, and histology. This significant limitation persists despite recent demonstrations of LLMs outperforming physicians in controlled diagnostic settings (43), underscoring a critical challenge in translating artificial intelligence capabilities to clinical practice where multimodal reasoning is essential. Text-based implementation currently constrains LLMs, but emerging multimodal capabilities in radiology (44–46) and dermatology (47) suggest promising diagnostic extensions. Future integration of CT/MRCP imaging could substantially enhance biliary stricture evaluation, though diagnosing rare conditions requires specialized training approaches. To address these constraints, we propose a

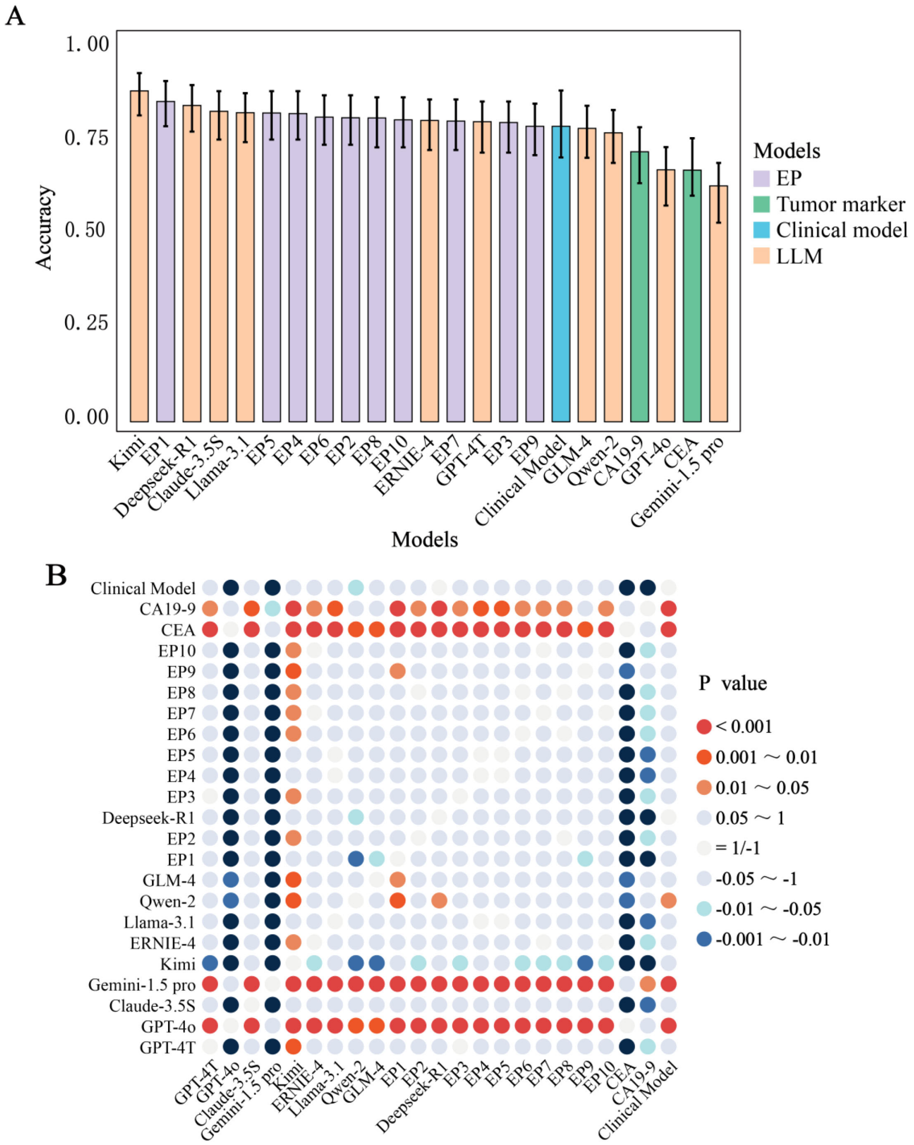

FIGURE 3
Comparative performance evaluation. **(A)** Accuracy among Different LLMs, Experienced Physicians, Clinical Model and Tumor Markers. **(B)** A Comparative Analysis of Diagnostic Accuracy and Significance Testing among models. The p-values (Holm-adjusted) were from the comparison of accuracy between the predictive groups listed along the horizontal axis and those on the vertical axis. A positive p-value indicates that the accuracy of the group on the horizontal axis is statistically greater than that of the group on the vertical axis, whereas a negative p-value signifies the opposite. CA19-9, Carbohydrate Antigen 199; CEA, Carcinoembryonic Antigen; EP, Experienced physician.

structured workflow comprising: 1) electronic Medical Record-integrated real-time malignancy probability scoring, and 2) automatic referral to targeted multidisciplinary review for cases with LLM confidence scores below 80%. This structure preserves physician oversight while optimizing diagnostic efficiency, particularly valuable in resource-limited settings. However, clinical deployment of LLMs demands addressing several critical ethical considerations: 1) Accountability through legal frameworks addressing liability for diagnostic errors; 2) Hallucination mitigation requiring detection protocols (evidenced in 12% of erroneous outputs); 3) Patient acceptance considerations, with survey data showing 67% rejection of AI-exclusive diagnoses for cancer-related decisions; 4) Equity concerns including documented performance disparities in elderly populations; 5) Transparency requirements for interpretable decision pathways; and 6) Privacy mandates demanding robust data anonymization. Essential mitigation strategies include human-AI collaborative diagnostic models, algorithmic bias correction techniques, and targeted patient education initiatives clarifying LLMs' assistive role.

Our study exhibits several limitations that warrant consideration. First, the retrospective single-center design introduces potential selection bias. Second, modest sample size limiting statistical power to address heterogeneity in biliary stricture presentations, potentially restricting generalizability across diverse healthcare settings; Third, the small hilar stricture subgroup (n = 29) limits statistical power for physician-LLM

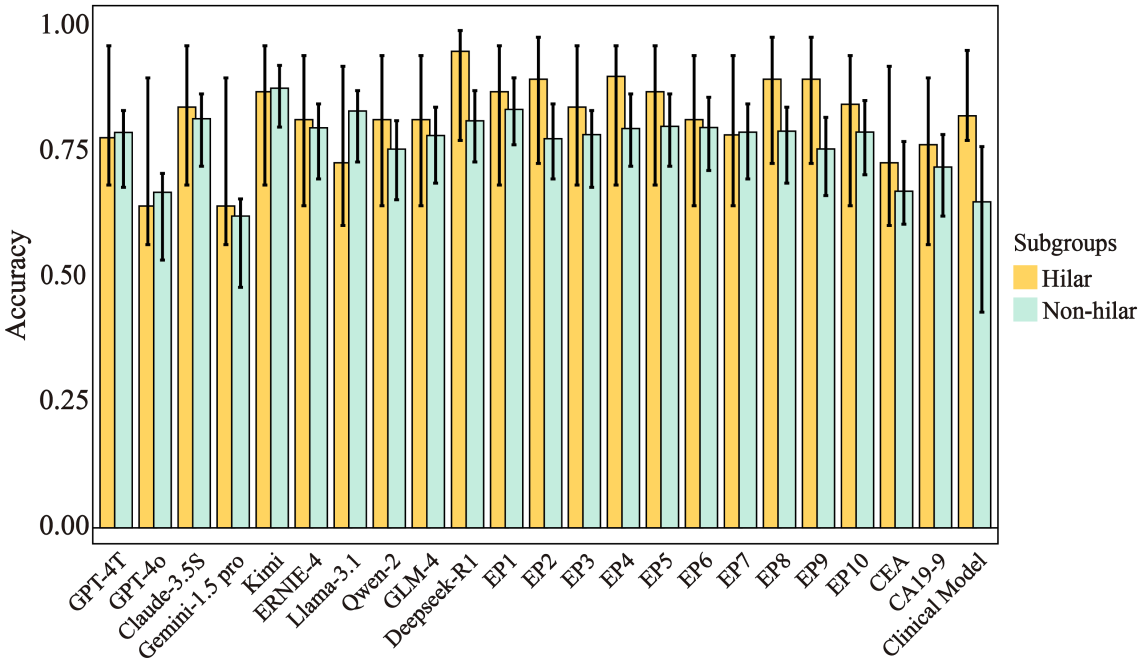

FIGURE 4
Accuracy and 95% CIs across subgroups for different prediction models. Error bars represent 95% confidence intervals. Hilar strictures (n=29: 9 benign, 20 malignant), non-hilar strictures (n=130: 74 benign, 56 malignant). CA19-9, Carbohydrate Antigen 199; CEA, Carcinoembryonic Antigen; EP, Experienced physician.

comparisons in this anatomically complex subset. Fourth, version-specific LLM evaluation restricts generalizability to updated iterations. Fifth, variability in physician experience levels may impact human performance benchmarks. Sixth, absence of external validation constrains generalizability.

Conclusion

In conclusion, this study demonstrates select LLMs (Kimi, Deepseek-R1, Claude-3.5S, Llama-3.1) achieve diagnostic accuracy comparable to or exceeding clinical models and physicians for biliary strictures, though hilar cases remain challenging. Their optimal implementation involves augmenting clinical judgment rather than replacing it, especially valuable for non-hilar strictures where performance matched physicians.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving humans were approved by the Ethics Committee of Xijing Hospital. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

CK: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. JL: Data curation, Formal analysis, Visualization, Writing – review & editing, Resources. XY: Data curation, Formal analysis, Visualization, Writing – review & editing, Software. GR: Investigation, Validation, Writing – review & editing. LZ: Investigation, Validation, Writing – review & editing. WW: Investigation, Writing – review & editing. XL: Investigation, Writing – review & editing. LW: Investigation, Writing – review & editing. GS: Investigation, Writing – review & editing. JH: Investigation, Writing – review & editing. BW: Investigation, Writing – review & editing. YD: Investigation, Writing – review & editing. WZ: Investigation, Writing – review & editing. YLL: Methodology, Writing – review & editing. TL: Validation, Writing – review & editing. LL: Validation, Writing – review & editing. HL: Investigation, Writing – review & editing. SL: Investigation, Writing

– review & editing. YL: Conceptualization, Methodology, Supervision, Writing – review & editing. YP: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This study was supported by National Key R&D Program of China (2022YFC2505100) and grants from the National Natural Science Foundation of China (82373117&82370619).

Conflict of interest

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

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2025.1613818/full#supplementary-material

References

1. Elmunzer BJ, Maranki JL, Gómez V, Tavakkoli A, Sauer BG, Limketkai BN, et al. ACG clinical guideline: diagnosis and management of biliary strictures. *Am J Gastroenterol.* (2023) 118:405–26. doi: 10.14309/ajg.0000000000002190

2. Wen L-J, Chen J-H, Xu H-J, Yu Q, Liu K. Efficacy and safety of digital single-operator cholangioscopy in the diagnosis of indeterminate biliary strictures by targeted biopsies: A systematic review and meta-analysis. *Diagn Basel Switz.* (2020) 10:666. doi: 10.3390/diagnostics10090666

3. de Moura DTH, Ryou M, de Moura EGH, Ribeiro IB, Bernardo WM, Thompson CC. Endoscopic ultrasound-guided fine needle aspiration and endoscopic retrograde cholangiopancreatography-based tissue sampling in suspected Malignant biliary strictures: A meta-analysis of same-session procedures. *Clin Endosc.* (2020) 53:417–28. doi: 10.5946/ce.2019.053

4. De Moura DTH, Moura EGHD, Bernardo WM, De Moura ETH, Baraca FI, Kondo A, et al. Endoscopic retrograde cholangiopancreatography versus endoscopic ultrasound for tissue diagnosis of Malignant biliary stricture: Systematic review and meta-analysis. *Endosc Ultrasound.* (2018) 7:10–9. doi: 10.4103/2303-9027.193597

5. Tummala P, Munigala S, Eloubeidi MA, Agarwal B. Patients with obstructive jaundice and biliary stricture ± mass lesion on imaging: prevalence of Malignancy and potential role of EUS-FNA. *J Clin Gastroenterol.* (2013) 47:532–7. doi: 10.1097/MCG.0b013e3182745d9f

6. Hu B, Sun B, Cai Q, Wong Lau JY, Ma S, Itoi T, et al. Asia-Pacific consensus guidelines for endoscopic management of benign biliary strictures. *Gastrointest Endosc.* (2017) 86:44–58. doi: 10.1016/j.gie.2017.02.031

7. Ramchandani M, Lakhtakia S, Costamagna G, Tringali A, Püspöek A, Tribl B, et al. Fully covered self-expanding metal stent vs multiple plastic stents to treat benign biliary strictures secondary to chronic pancreatitis: A multicenter randomized trial. *Gastroenterology.* (2021) 161:185–95. doi: 10.1053/j.gastro.2021.03.015

8. Sato T, Kogure H, Nakai Y, Ishigaki K, Hakuta R, Saito K, et al. A prospective study of fully covered metal stents for different types of refractory benign biliary strictures. *Endoscopy.* (2020) 52:368–76. doi: 10.1055/a-1111-8666

9. Ponsioen CY, Arnelo U, Bergquist A, Rauws EA, Paulsen V, Cantú P, et al. No superiority of stents vs balloon dilatation for dominant strictures in patients with primary sclerosing cholangitis. *Gastroenterology.* (2018) 155:752–759.e5. doi: 10.1053/j.gastro.2018.05.034

10. Franssen S, Van Driel LMJW, Moelker A, Groot Koerkamp B. Primary percutaneous stenting above the ampulla for palliative biliary drainage of Malignant hilar biliary obstruction. *J Clin Oncol.* (2023) 41:527–7. doi: 10.1200/JCO.2023.41.4_suppl.527

11. Jang S, Stevens T, Parsi MA, Bhatt A, Kichler A, Vargo JJ. Superiority of self-expandable metallic stents over plastic stents in treatment of Malignant distal biliary strictures. *Clin Gastroenterol Hepatol.* (2022) 20:e182–95. doi: 10.1016/j.cgh.2020.12.020

12. Wang AY, Yachimski PS. Endoscopic management of pancreatobiliary neoplasms. *Gastroenterology.* (2018) 154:1947–63. doi: 10.1053/j.gastro.2017.11.295

13. Wallace MB, Wang KK, Adler DG, Rastogi A. Recent advances in endoscopy. *Gastroenterology.* (2017) 153:364–81. doi: 10.1053/j.gastro.2017.06.014

14. Khamaysi I, Firman R, Martin P, Vasilyev G, Boyko E, Zussman E. Mechanical perspective on increasing brush cytology yield. *ACS Biomater Sci Eng.* (2024) 10:1743–52. doi: 10.1021/acsbiomaterials.3c00935

15. Wang J, Xia M, Jin Y, Zheng H, Shen Z, Dai W, et al. More endoscopy-based brushing passes improve the detection of Malignant biliary strictures: A multicenter randomized controlled trial. *Am J Gastroenterol.* (2022) 117:733–9. doi: 10.14309/ajg.0000000000001666

16. Bang JY, Navaneethan U, Hasan M, Sutton B, Hawes R, Varadarajulu S. Optimizing outcomes of single-operator cholangioscopy-guided biopsies based on a randomized trial. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc.* (2020) 18:441–448.e1. doi: 10.1016/j.cgh.2019.07.035

17. Chen L, Lu Y, Kulkarni P. Tu1008 a meta-analysis of the value of Intraductal Ultrasound (IDUS) in differentiating Malignant from benign biliary strictures. *Gastroenterology.* (2020) 158:S–1003-S-1004. doi: 10.1016/S0016-5085(20)33183-8

18. Han S, Kahaleh M, Sharaiha RZ, Tarnasky PR, Kedia P, Slivka A, et al. Probe-based confocal laser endomicroscopy in the evaluation of dominant strictures in patients with primary sclerosing cholangitis: results of a U.S. multicenter prospective trial. *Gastrointest Endosc.* (2021) 94:569–576.e1. doi: 10.1016/j.gie.2021.03.027

19. Joshi V, Patel SN, Vanderveldt H, Oliva I, Raijman I, Molina C, et al. Mo1963 A pilot study of safety and efficacy of directed cannulation with a low profile catheter (LP) and imaging characteristics of bile duct wall using optical coherance tomography (OCT) for indeterminate biliary strictures initial report on *in-vivo* evaluation during ERCP. *Gastrointest Endosc.* (2017) 85:AB496–7. doi: 10.1016/j.gie.2017.03.1150

20. Wang Y, Wang W, Zhang S, Cai W, Song R, Mei T, et al. Diagnostic value of carbohydrate antigen 50 in biliary tract cancer: a large-scale multicenter study. *Cancer Med.* (2024) 13:e7388. doi: 10.1002/cam4.7388

21. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. *Lancet Digit Health.* (2024) 6:e555–61. doi: 10.1016/S2589-7500(24)00097-9

22. Peng L, Cai S, Wu Z, Shang H, Zhu X, Li X. MMGPL: multimodal medical data analysis with graph prompt learning. *Med Image Anal.* (2024) 97:103225. doi: 10.1016/j.media.2024.103225

23. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med.* (2023) 29:1930–40. doi: 10.1038/s41591-023-02448-8

24. Santoki A, Jones R, Peter M, Mathew A. Implementing large language model-based artificial intelligence (AI) technology in proposing effective treatment plans in

patients with cancer. *J Clin Oncol*. (2024) 42:e13660–0. doi: 10.1200/JCO.2024.42.16_suppl.e13660

25. Qiu P, Wu C, Zhang X, Lin W, Wang H, Zhang Y, et al. Towards building multilingual language model for medicine. *Nat Commun*. (2024) 15:8384. doi: 10.1038/s41467-024-52417-z

26. Zhou S, Luo X, Chen C, Jiang H, Yang C, Ran G, et al. The performance of large language model powered chatbots compared to oncology physicians on colorectal cancer queries. *Int J Surg*. (2024) 110(10):6509–17. doi: 10.1097/JS9.0000000000001850

27. Yoon SB, Moon S-H, Ko SW, Lim H, Kang HS, Kim JH. Brush cytology, forceps biopsy, or endoscopic ultrasound-guided sampling for diagnosis of bile duct cancer: A meta-analysis. *Dig Dis Sci*. (2022) 67:3284–97. doi: 10.1007/s10620-021-07138-4

28. Jeon TY, Choi MH, Yoon SB, Soh JS, Moon S-H. Systematic review and meta-analysis of percutaneous transluminal forceps biopsy for diagnosing Malignant biliary strictures. *Eur Radiol*. (2022) 32:1747–56. doi: 10.1007/s00330-021-08301-1

29. Navaneethan U, Njei B, Lourdusamy V, Konjeti R, Vargo JJ, Parsi MA. Comparative effectiveness of biliary brush cytology and intraductal biopsy for detection of Malignant biliary strictures: a systematic review and meta-analysis. *Gastrointest Endosc*. (2015) 81:168–76. doi: 10.1016/j.gie.2014.09.017

30. Sun X, Zhou Z, Tian J, Wang Z, Huang Q, Fan K, et al. Is single-operator peroral cholangioscopy a useful tool for the diagnosis of indeterminate biliary lesion? A systematic review and meta-analysis. *Gastrointest Endosc*. (2015) 82:79–87. doi: 10.1016/j.gie.2014.12.021

31. Chiang W-L, Zheng L, Sheng Y, Angelopoulos AN, Li T, Li D, et al. *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference* (2024). Available online at: http://arxiv.org/abs/2403.04132 (Accessed August 26, 2024).

32. Tran H, Yang Z, Yao Z, Yu H. BioInstruct: instruction tuning of large language models for biomedical natural language processing. *J Am Med Inform Assoc*. (2024) 31:1821–32. doi: 10.1093/jamia/ocae122

33. Ge J, Sun S, Owens J, Galvez V, Gologorskaya O, Lai JC, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. *Hepatol Baltim Md*. (2024) 80(5):1158–68. doi: 10.1097/HEP.0000000000000834

34. Du D, Zhong F, Liu L. Enhancing recognition and interpretation of functional phenotypic sequences through fine-tuning pre-trained genomic models. *J Transl Med*. (2024) 22:756. doi: 10.1186/s12967-024-05567-z

35. Wen N, Peng D, Xiong X, Liu G, Nie G, Wang Y, et al. Cholangiocarcinoma combined with biliary obstruction: an exosomal circRNA signature for diagnosis and early recurrence monitoring. *Signal Transduct Target Ther*. (2024) 9:107. doi: 10.1038/s41392-024-01814-3

36. Liang B, Zhong L, He Q, Wang S, Pan Z, Wang T, et al. Diagnostic accuracy of serum CA19–9 in patients with cholangiocarcinoma: A systematic review and meta-analysis. *Med Sci Monit*. (2015) 21:3555–63. doi: 10.12659/MSM.895040

37. Liu DSK, Prado MM, Giovannetti E, Jiao LR, Krell J, Frampton AE. MOY 2 microRNAs as bile based biomarkers for pancreaticobiliary cancers (MIRABILE). *Br J Surg*. (2023) 110:znad241.005. doi: 10.1093/bjs/znad241.005

38. Gao L, Lin Y, Yue P, Li S, Zhang Y, Mi N, et al. Identification of a novel bile marker clusterin and a public online prediction platform based on deep learning for cholangiocarcinoma. *BMC Med*. (2023) 21:294. doi: 10.1186/s12916-023-02990-9

39. Loosen SH, Roderburg C, Kauertz KL, Koch A, Vucur M, Schneider AT, et al. CEA but not CA19–9 is an independent prognostic factor in patients undergoing resection of cholangiocarcinoma. *Sci Rep*. (2017) 7:16975. doi: 10.1038/s41598-017-17175-7

40. Li Y, Li D-J, Chen J, Liu W, Li J-W, Jiang P, et al. Application of joint detection of AFP, CA19-9, CA125 and CEA in identification and diagnosis of cholangiocarcinoma. *Asian Pac J Cancer Prev*. (2015) 16:3451–5. doi: 10.7314/APJCP.2015.16.8.3451

41. Tuzun Ince A, Yildiz K, Baysal B, Danalioglu A, Kocaman O, Tozlu M, et al. Roles of serum and biliary CEA, CA19-9, VEGFR3, and TAC in differentiating between Malignant and benign biliary obstructions. *Turk J Gastroenterol*. (2014) 25:162–9. doi: 10.5152/tjg.2014.6056

42. Macias RIR, Kornek M, Rodrigues PM, Paiva NA, Castro RE, Urban S, et al. Diagnostic and prognostic biomarkers in cholangiocarcinoma. *Liver Int*. (2019) 39:108–22. doi: 10.1111/liv.14090

43. Brodeur PG, Buckley TA, Kanjee Z, Goh E, Ling EB, Jain P, et al. Superhuman performance of a large language model on the reasoning tasks of a physician. *arXiv* (2024). doi: 10.48550/ARXIV.2412.10849

44. Brin D, Sorin V, Barash Y, Konen E, Glicksberg BS, Nadkarni GN, et al. Assessing GPT-4 multimodal performance in radiological image analysis. *Eur Radiol*. (2025) 35(4):1959–65. doi: 10.1007/s00330-024-11035-5

45. Kaczmarczyk R, Wilhelm TI, Martin R, Roos J. Evaluating multimodal AI in medical diagnostics. *NPJ Digit Med*. (2024) 7:205. doi: 10.1038/s41746-024-01208-3

46. Beşler MS. The accuracy of the multimodal large language model GPT-4 on sample questions from the interventional radiology board examination. *Acad Radiol*. (2024) 31:3476. doi: 10.1016/j.acra.2024.03.023

47. 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) 15:5649. doi: 10.1038/s41467-024-50043-3