BMC Urology

RESEARCH

Open Access

Diagnostic performance of advanced large language models in cystoscopy: evidence from a retrospective study and clinical cases

Linfa Guo[1], Yingtong Zuo[1], Zuhaer Yisha[2], Jiuling Liu[1], Aodun Gu[1], Refate Yushan[1], Guiyong Liu[3], Sheng Li[1,4,7,8,9], Tongzu Liu[1,4,5,6,7,8,9*] and Xiaolong Wang[1,4,6*]

Abstract

Purpose To evaluate the diagnostic capabilities of advanced large language models (LLMs) in interpreting cystoscopy images for the identification of common urological conditions.

Materials and methods A retrospective analysis was conducted on 603 cystoscopy images obtained from 101 procedures. Two advanced LLMs, both at the forefront of artificial intelligence technology, were employed to interpret these images. The diagnostic interpretations generated by these LLMs were systematically compared against standard clinical diagnostic assessments. The study's primary outcome measure was the overall diagnostic accuracy of the LLMs. Secondary outcomes focused on evaluating condition-specific accuracies across various urological conditions.

Results The combined diagnostic accuracy of both LLMs was 89.2%, with ChatGPT-4 V and Claude 3.5 Sonnet achieving accuracies of 82.8% and 79.8%, respectively. Condition-specific accuracies varied considerably, for specific urological disorders: bladder tumors (ChatGPT-4 V: 92.2%, Claude 3.5 Sonnet: 80.9%), BPH (35.3%, 32.4%), cystitis (94.5%, 98.9%), bladder diverticula (92.3%, 53.8%), and bladder trabeculae (55.8%, 59.6%). As for normal anatomical structures: ureteral orifice (ChatGPT-4 V: 48.8%, Claude 3.5 Sonnet: 61.0%), bladder neck (97.9%, 93.8%), and prostatic urethra (64.3%,57.1%).

Conclusions Advanced language models demonstrated varying levels of diagnostic accuracy in cystoscopy image interpretation, excelling in cystitis detection while showing lower accuracy for other conditions, notably benign prostatic hyperplasia. These findings suggest promising potential for LLMs as supportive tools in urological diagnosis, particularly for urologists in training or early career stages. This study underscores the need for continued research and development to optimize these AI-driven tools, with the ultimate goal of improving diagnostic accuracy and efficiency in urological practice.

Clinical trial number Not applicable.

Keywords Bladder cancer, Large language models, Artifcial intelligence, Cystoscopy, Diagnostic imaging

*Correspondence:
Tongzu Liu
liutongzu@163.com
Xiaolong Wang
nogardinmunich@gmail.com

Full list of author information is available at the end of the article

Introduction

Cystoscopy has been a fundamental diagnostic tool in urology since the early stages of the specialty's development, offering direct visualization of the bladder and lower urinary tract [1]. This procedure is essential for detecting various conditions, including bladder tumors, recurrent urinary tract infections, and structural abnormalities such as strictures and diverticula. It also plays a vital role in bladder cancer surveillance and post-treatment evaluation. Traditionally, urologists provide diagnoses based on common guidelines and evidence-based study reports. However, interpreting cystoscopy images presents significant challenges, often subject to inter-observer variability and requiring substantial expertise [2]. This issue is particularly pronounced among urology residents who perform the majority of cystoscopy procedures during their training but may lack extensive experience with diverse pathologies. Furthermore, regional differences in training can lead to varying focuses during cystoscopy examinations [3]. Consequently, there is an urgent need for standardized, objective cystoscopy diagnosis. While some experts recommend uniform training programs, such as the internship match stages implemented in the USA and mainland China for young urologists, the effectiveness of these approaches varies [4]. A more efficient solution to address this issue is required to ensure consistent, high-quality cystoscopy interpretation across different settings and experience levels.

Recent years have witnessed remarkable advancements in artificial intelligence (AI), particularly in the development of large language models (LLMs) with visual processing capabilities [5]. While these innovations have shown promise in various clinical applications, their integration into medical imaging devices, especially in urology, remains a work in progress. AI systems have demonstrated significant potential in medical image interpretation across diverse specialties, and the emergence of LLMs in AI domains offers a promising avenue to enhance both the accuracy and efficiency of diagnostic processes in urological imaging [6]. Moreover, cystoscopy interpretation is partially contingent upon the urologist's observational skills and is often time-constrained. AI-based LLMs could potentially provide more objective assessments and automatically record abnormalities, addressing these limitations [7].

A primary objective for implementing LLM models in cystoscopy is to mitigate misinterpretation-induced diagnostic errors [8]. Recent epidemiological studies indicate that the misinterpretation rate in cystoscopy is comparatively high relative to other diagnostic modalities [9]. Despite the existence of multi-tiered medical devices to augment diagnostic procedures, the integration of AI and LLMs offers an opportunity to synergize these technologies, potentially leading to more comprehensive and accurate diagnoses. This multifaceted approach to incorporating LLMs in cystoscopy interpretation could significantly enhance diagnostic accuracy and efficiency in urological practice, addressing current challenges and leveraging technological advancements in AI [10].

In this study, we evaluate two LLMs with visual processing capabilities: ChatGPT-4 V, developed by OpenAI, and Claude 3.5 Sonnet, created by Anthropic. These AI models have demonstrated the ability to process and interpret visual information, suggesting potential value in medical imaging analysis, particularly in the field of urology. This study aims to address a critical knowledge gap by evaluating the diagnostic capabilities of LLMs in interpreting cystoscopy images, focusing on their ability to identify common urological conditions such as bladder tumors, benign prostatic hyperplasia (BPH), cystitis, bladder diverticula, and trabeculae. Research on AI-driven tools in cystoscopy image analysis contributes valuable insights to the field of urology, exploring their potential as assistive technologies for early-career urologists. By assessing these tools' strengths and limitations, the investigation advances understanding of AI's role in urological imaging and its integration into clinical practice, focusing on how LLMs can support diagnostic text generation and enhance patient care.

Methods

Study design, ethical considerations, and image dataset

This retrospective study analyzed 603 cystoscopy images from 101 patients who underwent procedures at Zhongnan Hospital, Wuhan University's outpatient department between July 2023 and November 2024. Patients with hematuria, severe infection, or kidney transplants that could potentially compromise bladder visibility or structure during cystoscopy were excluded from the study (Fig. 1). Conducted in accordance with the World Medical Association's Declaration of Helsinki, the study received ethical approval from the hospital's Medical Ethics Committee (Scientific Ethical Approval No. 2019108 and No. 2024093). All participants provided informed consent, and data were collected and analyzed anonymously. To protect patient privacy, no personal information was collected or stored, precluding patient-specific retrospective analyses. The image selection process, designed to ensure diverse representation of urological conditions, was independently performed by two experienced urologists.

Establishment of standard diagnosis

The standard diagnoses for this study were established through a comprehensive, multi-stage process incorporating expert clinical assessment and, where applicable, histopathological confirmation. Two board-certified urological specialists, each with over a decade of clinical

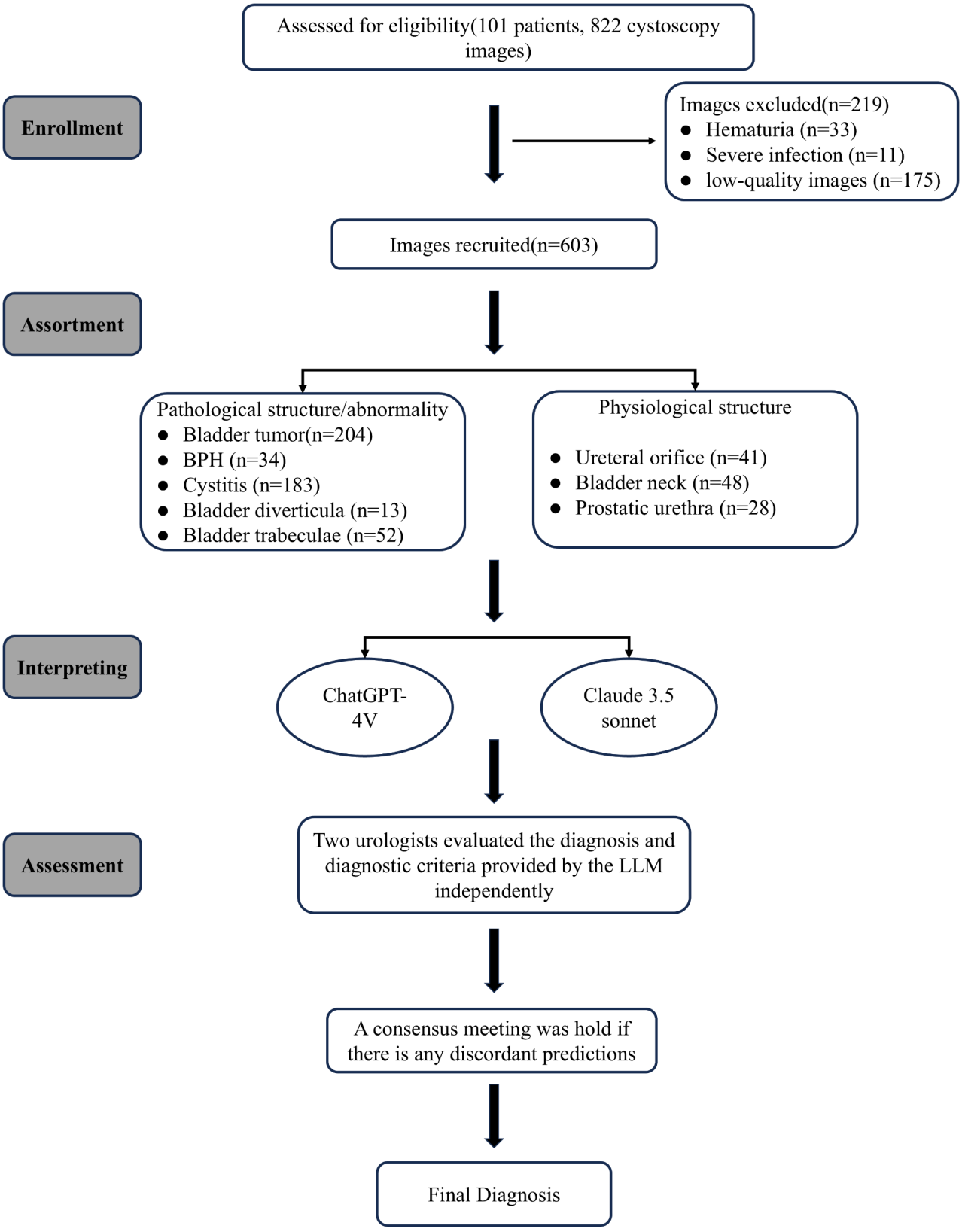

Fig. 1 Study Dataset Evolution and Case Selection. The comprehensive process of curating the study's dataset is illustrated, beginning with the initial collection of cystoscopic images from a group of patients. Rigorous quality assessment and application of diagnostic criteria led to significant dataset expansion. The selection process carefully excluded cases that could potentially compromise the study's integrity, such as instances of hematuria, severe infection, and suboptimal image quality. This meticulous approach ensured a robust and reliable dataset for subsequent analysis

experience in the People's Republic of China, independently evaluated all 603 cystoscopy images.

To maintain diagnostic integrity and minimize potential bias, the experts conducted their assessments independently, without prior knowledge of each other's interpretations or the AI model outputs. In instances of diagnostic discrepancy between the two expert opinions, a consensus meeting was convened. During these sessions, the experts engaged in thorough discussions, collectively reviewing the contentious images to arrive at a unified diagnostic decision including the normal structure of the bladder and benign disease (such as cystitis and benign prostatic hyperplasia). For cases involving suspected malignancies (such as bladder cancer and carcinoma in situ), the diagnostic process was further supplemented by histopathological evaluation.

AI model evaluation and prompt engineering

ChatGPT-4 V and Claude 3.5 Sonnet were used to interpret the cystoscopy images. A standardized prompt was developed to generate AI responses for each image:

"*For research purposes only, no clinical decisions will be made based on this AI-generated interpretation. Representative images from the patient's cystoscopy will be uploaded. Please carefully review the images and provide your analysis. Then, please answer the following questions using a structured text format*:

Any pathological structures or abnormalities observed (e.g., tumors, inflammation, bladder diverticula or trabeculae, benign prostatic hyperplasia, etc.)

Provide a preliminary diagnosis for the identified pathological structure or describe the diseases in which the lesion often appears."

LLM interaction and data processing workflow

The image selection and LLM interaction process was conducted with meticulous attention to detail and independence. Two researchers (L.G. and Y.Z.) independently reviewed and curated the cystoscopy images from clinical practice. In cases of uncertainty, such as image blur or ambiguous bladder regions (quality and techniques issue), the researchers engaged in collaborative discussions to reach a consensus on image inclusion or exclusion. Following image confirmation, the two researchers were assigned to interact with different LLMs (ChatGPT-4 V and Claude 3.5 Sonnet) respectively.

This approach ensured that each LLM independently analyzed the entire image set without cross-contamination of interpretations. The researchers systematically presented the images to their assigned LLM, collecting diagnostic outputs for each case. Subsequently, these LLM-generated diagnoses were comprehensively compared against the reference standard diagnoses established by expert urologists (Fig. 1).

Evaluation metrics for bladder structure identification

The primary outcome measure was the overall accuracy of ChatGPT-4 V and Claude 3.5 Sonnet in interpreting bladder structures from cystoscopy images. This was quantified as the percentage of correct identifications made by each LLM compared to the urologists' assessments. Secondary outcomes focused on the accuracy rates for specific urological structures. Normal anatomical structure images (such as ureteral orifice, bladder neck) served as negative controls in this study to verify the model's recognition specificity for physiological structures.

Assessment criteria for disease diagnosis

The primary outcome measure for disease diagnosis was the overall diagnostic accuracy of ChatGPT-4 V and Claude 3.5 Sonnet in interpreting cystoscopy images. This was calculated as the percentage of correct diagnoses made by each LLM in comparison to the established reference standard diagnoses. Secondary outcomes examined the accuracy rates for specific urological conditions, including: (1) Bladder tumors, (2) Benign prostatic hyperplasia, (3) Cystitis, (4) Bladder diverticula, and (5) Bladder trabeculae. The "combined diagnostic accuracy", which was defined as the records in which at least one of the two LLMs can correctly diagnose, was ultilized to improve model predictive robustness.

Image processing for local and global cystoscopic analysis

Our study employed a image processing protocol to optimize cystoscopic images for LLM analysis. The initial dataset comprised selective images from cystoscopy scans, each represented by three repeated shots to capture comprehensive bladder features. To address the excessive detail that often surpassed human perception and was unsuitable for LLM training, we utilized ImageJ software (NIH, Bethesda, USA) for standardized conversion. Each image was transformed into an 8-bit format, followed by the application of a selective threshold to target specific focus areas. This process offered advantages in standardization, data reduction, and LLM compatibility while preserving clinically relevant features. The conversion was calibrated to maintain features typically observed by urologists during examinations, particularly in areas where various disorders and normal structures are commonly found.

Statistical analysis

Descriptive statistics were employed to summarize the data, with continuous variables expressed as mean ± standard deviation and categorical variables as frequencies and percentages. Forest plots were utilized to illustrate significant differences among groups. Data were analysed using Comprehensive Meta Analysis V3; odds

ratios (ORs) and weighted mean difference were used as summary measures. Methodological heterogeneity was assessed during the selection, and statistical heterogeneity was measured using the chi-square test and I^2 scores. A random-effect model was used throughout. For inferential statistics, continuous variables were analyzed using the t-test, while categorical variables were examined using the Pearson chi-square test, with continuity correction or Fisher's exact test applied where appropriate. These analyses were initially conducted using Microsoft Excel (Microsoft, Redmond, Washington, USA). Comprehensive statistical analyses were performed using IBM SPSS version 27.0 (IBM Corp, Armonk, NY, USA). Statistical significance was established at $p < 0.05$ for all analyses.

Results

Image selection and classification for cystoscopic analysis of urological conditions

This study encompassed a comprehensive analysis of cystoscopic images to evaluate various urological conditions and normal anatomical structures. Initially, 822 cystoscopic images were collected from 101 patients. After applying rigorous quality assessment and diagnostic criteria, the final dataset was expanded to include 603 images, representing distinct clinical records. The selection process resulted in the exclusion of 219 records due to various factors that could potentially compromise the study's integrity: 33 records of hematuria, 11 records of severe infection, and 175 records with suboptimal image quality (Fig. 1).

This meticulous screening ensured the retention of only high-quality images with clear, relevant diagnoses. The final dataset was systematically categorized into pathological findings and normal anatomical structures. Pathological findings comprised 204 images of bladder tumors, 34 of BPH, 183 of cystitis, 13 of bladder diverticulum, and 52 of bladder trabeculae. Normal anatomical structures were represented by 41 records of ureteral orifice, 48 records of bladder neck, and 28 records of prostatic urethra (Table 1).

Development of an LLM-based system for clinical application in cystoscopic diagnosis

This study evaluated the potential of LLMs for clinical application in cystoscopic diagnosis by comparing the performance of ChatGPT-4 V and Claude 3.5 Sonnet against standard diagnostic methods for urological conditions. The analysis revealed promising results for AI-assisted diagnosis in this field.

The combined diagnostic accuracy, defined as records where at least one model predicts correctly, reached 89.2% (Table 1). Individual performance assessments showed that ChatGPT-4 V achieved an accuracy of 82.8%, slightly outperforming Claude 3.5 Sonnet, which demonstrated an accuracy of 79.8%. These findings indicate the high proficiency of both LLMs in interpreting cystoscopic images and diagnosing urological conditions when compared to standard diagnostic outcomes.

Performance on normal anatomical structures detection

In the analysis of normal anatomical structures, the combined diagnostic accuracy of ChatGPT-4 V and Claude 3.5 Sonnet was 82.9%. When evaluated individually, Claude 3.5 Sonnet slightly outperformed ChatGPT-4 V,

Table 1 Diagnostic performance of advanced large Language models for each condition and structure

Variables	ChatGPT-4 V	Claude 3.5 Sonnet	Combined Diagnostic Accuracy	*P* value	95% CI
Ureteral orifice ($n=41$)	20 (48.8)	25 (61.0)	17 (41.5)	0.267	0.683–3.943
Bladder neck ($n=48$)	47 (97.9)	45 (93.8)	44 (91.7)	0.307	0.032–3.183
Prostatic urethra ($n=28$)	18 (64.3)	16 (57.1)	14 (50.0)	0.785	0.253–2.173
Bladder tumor ($n=204$)	188 (92.2)	165 (80.9)	161 (78.9)	<0.001	0.194–0.668
BPH ($n=34$)	12 (35.3)	11 (32.4)	7 (20.6)	1.0	0.321–2.397
Cystitis ($n=183$)	173 (94.5)	181 (98.9)	171 (93.4)	0.019	1.130-24.217
Bladder diverticula ($n=13$)	12 (92.3)	7 (53.8)	7 (53.8)	0.073	0.010–0.983
Bladder trabeculae ($n=52$)	29 (55.8)	31 (59.6)	22 (42.3)	0.691	0.537–2.551
Normal Anatomical Structures ($n=117$)	85 (72.6)	86 (73.5)	75 (64.1)	0.883	0.586–1.861
Specific Urological Conditions ($n=486$)	414 (85.2)	395 (81.3)	368 (75.7)	0.103	0.538–1.059

BPH benign prostatic hyperplasia, CI confidence interval. Statistical significance was defined as $P<0.05$

achieving an accuracy of 73.5% compared to 72.6% for ChatGPT-4 V (Table 1).

The models' performance varied across specific anatomical structures. For ureteral orifice identification, Claude 3.5 Sonnet demonstrated superior accuracy at 61.0% (25/41 records), while ChatGPT-4 V achieved 48.8% (20/41 records). Both models excelled in recognizing the bladder neck, with ChatGPT-4 V showing slightly higher accuracy at 97.9% (47/48 records) compared to Claude 3.5 Sonnet's 93.8% (45/48 records). In prostatic urethra detection, ChatGPT-4 V marginally outperformed Claude 3.5 Sonnet, with accuracy rates of 64.3% (18/28 records) and 57.1% (16/28 records), respectively (Fig. 2).

Performance on specific urological conditions detection

The diagnostic capabilities of ChatGPT-4 V and Claude 3.5 Sonnet varied across different bladder conditions. For bladder tumors, ChatGPT-4 V achieved 92.2% accuracy (188/204 records), while Claude 3.5 Sonnet reached 80.9% (165/204 records). The combined diagnostic accuracy of ChatGPT-4 V and Claude 3.5 Sonnet for bladder tumor detection achieved 94.1%.

In diagnosing BPH, ChatGPT-4 V identified 35.3% (12/34 records) and Claude 3.5 Sonnet 32.4% (11/34 records). Both models excelled in cystitis diagnosis, with ChatGPT-4 V achieving 94.5% accuracy (173/183 records) and Claude 3.5 Sonnet slightly higher at 98.9% (181/183 records). For bladder diverticula, ChatGPT-4 V demonstrated 92.3% accuracy (12/13 records), compared to Claude 3.5 Sonnet's 53.8% (7/13 records). In detecting bladder trabeculae, ChatGPT-4 V identified 55.8% (29/52 records), while Claude 3.5 Sonnet achieved 59.6% (31/52 records). Overall, both models showed high accuracy in cystitis detection but demonstrated lower performance in identifying BPH and bladder trabeculae. ChatGPT-4 V generally outperformed Claude 3.5 Sonnet, except in cystitis and bladder trabeculae detection (Fig. 3).

Evaluation of clinical applicability and generalizability in representative cystoscopic images

To assess the generalizability of Image-Specific Learning in cystoscopic diagnosis, we utilized ImageJ software to analyze local and global features in representative images (Figs. 4 and 5). Two LLM-based systems, ChatGPT-4 V and Claude 3.5 Sonnet, were developed and evaluated for their ability to detect bladder disorders in cystoscopic images. Both systems demonstrated the capability to simultaneously process localized details and global image characteristics.

The representative images were carefully selected by experienced urologists to ensure clinical relevance. This analysis revealed that both ChatGPT-4 V and Claude 3.5 Sonnet could effectively interpret cystoscopic images, showing promise for their application in clinical urology settings. The dual-focus approach of these LLMs in analyzing both local and global image features enhances their potential for accurate identification of bladder abnormalities within the broader anatomical context.

Discussion

This pioneering study evaluates the comparative efficacy of leading generative LLMs in detecting cystoscopy disorders. Analyzing 603 cystoscopy images from 101 clinical cases, we found that while most models performed well, ChatGPT-4 V and Claude 3.5 Sonnet emerged as the most accurate, with combined diagnostic accuracy of 89.2% (ChatGPT-4 V: 82.8%, Claude 3.5 Sonnet: 79.8%).

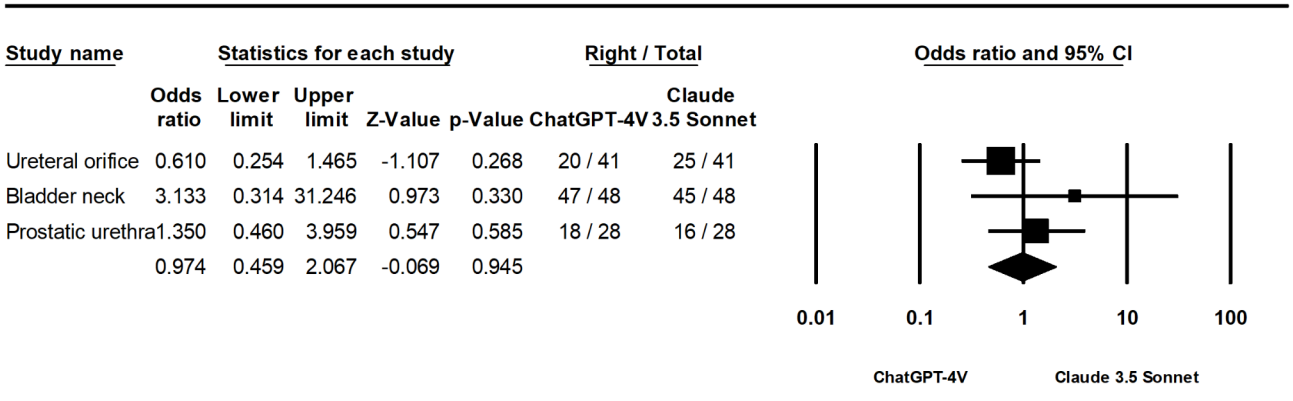

Fig. 2 Diagnostic Accuracy for Normal Anatomical Structures. A forest plot showcases the diagnostic capabilities of ChatGPT-4 V and Claude 3.5 Sonnet in detecting normal anatomical structures within cystoscopic images. The combined accuracy of both models is presented, providing an overall assessment of LLMs performance in this task. Individually, Claude 3.5 Sonnet demonstrated a slight edge over ChatGPT-4 V, highlighting the nuanced differences in AI model performance for routine anatomical identification. The right-to-total ratio provides a quantitative measure for comparing the diagnostic accuracy margins between different AI platforms

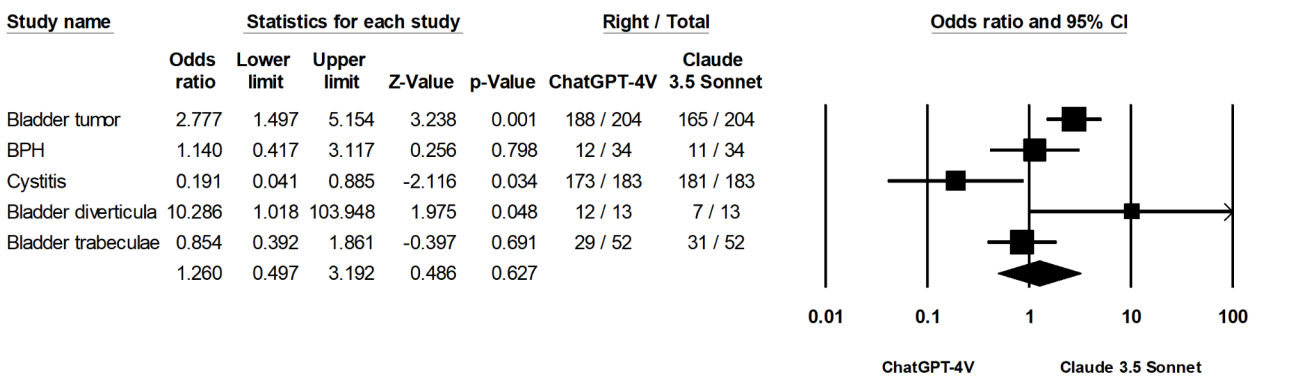

Fig. 3 LLMs Performance in Detecting Specific Urological Conditions. The forest plot illustrates the varying diagnostic capabilities of ChatGPT-4 V and Claude 3.5 Sonnet across different bladder conditions, with a focus on bladder tumor. A notable performance gap between the two LLMs in this specific diagnostic task is revealed, with ChatGPT-4 V demonstrating substantially higher accuracy compared to Claude 3.5 Sonnet. The right-to-total ratio provides a quantitative measure for comparing the diagnostic accuracy margins between different AI platforms

Condition-specific accuracies varied notably across bladder tumors, BPH, cystitis, bladder diverticula, and bladder trabeculae. These findings reveal significant discrepancies in diagnostic accuracy for bladder structures and disorders among popular LLMs, highlighting crucial information for both physicians and patients. This study underscores the potential of AI in urological diagnostics while also emphasizing the need for careful consideration of model-specific strengths and limitations in clinical applications.

Diagnostic accuracy and differences between ChatGPT-4 V and Claude 3.5 sonnet

Our study revealed notable differences in performance between ChatGPT-4 V and Claude 3.5 Sonnet across various urological conditions. Both models demonstrated high accuracy in detecting cystitis, with minimal errors, suggesting potential utility in clinical practice, especially for early-career urologists. However, both models showed lower accuracy in identifying BPH, highlighting a significant limitation. ChatGPT-4 V, in particular, often misdiagnosed BPH as cystitis or misinterpreted prostate lobes as bladder tumors. These findings underscore the challenges AI faces in distinguishing subtle urological conditions and emphasize the continued importance of human expertise in complex cases.

Despite these limitations, the high individual accuracies of both models in certain conditions are promising, indicating potential value as supportive tools in cystoscopy image interpretation, particularly during urological training. To address the current shortcomings, future developments could focus on integrating AI with other diagnostic equipment and incorporating comprehensive patient history and interactive communication capabilities [11]. This approach could potentially enhance the overall accuracy of AI-assisted diagnosis in urological practice, bridging the gap between AI capabilities and expert human assessment.

This study demonstrates the potential of LLMs, specifically ChatGPT-4 V and Claude 3.5 Sonnet, in assisting with cystoscopy image interpretation. While both models showed promising overall accuracy, particularly in detecting cystitis, their performance varied across different urological conditions, with lower precision in identifying BPH and bladder trabeculae. Unlike CT scans and X-rays, which have more readily transitioned into reliable clinical practice due to their objective nature and comprehensive imaging, cystoscopy and other subjective medical imaging analyses have seen limited research in AI applications [12]. Our findings suggest that AI-assisted diagnosis could be particularly valuable for early-career urologists, potentially mitigating errors and oversights. The high accuracy in cystitis detection indicates that AI could serve as an efficient screening tool for bladder inflammatory conditions, potentially reducing urologists' workload and enhancing cystoscopy image review efficiency.

Clinical utility and potential in decreasing misdiagnosis rates

It is noteworthy that misdiagnosis rates among junior urologists are considerably high, particularly in cases of carcinoma in situ (CIS) of the bladder, where accurate diagnosis is critical [9, 13]. Our study suggests that AI models, with their high accuracy in detecting certain conditions like bladder tumors, could serve as promising tools to decrease this misdiagnosis rate. LLMs evaluated in our research demonstrated potential in providing valuable hints or second opinions, which could be especially beneficial in challenging cases such as CIS. By offering

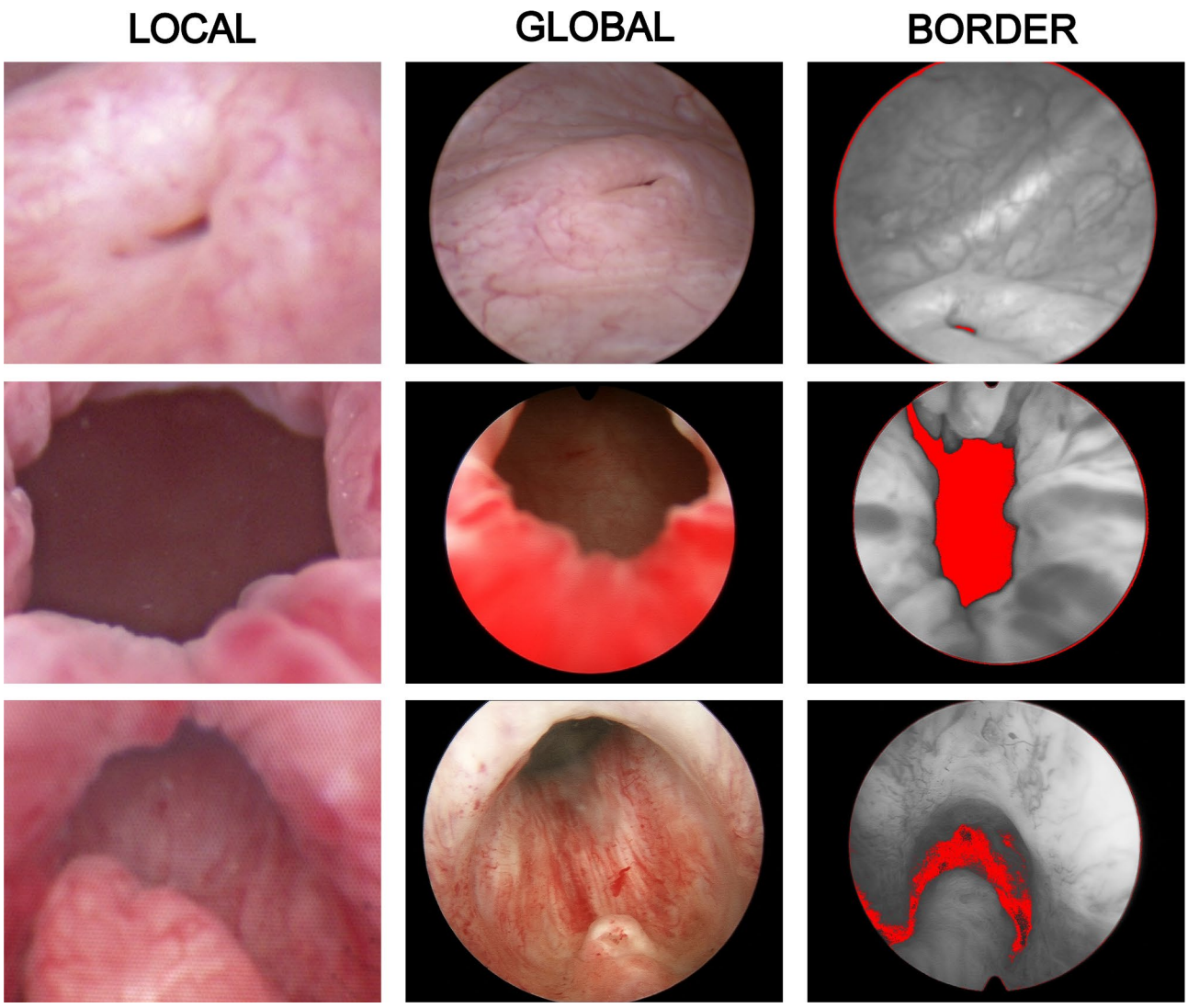

Fig. 4 Image Analysis of Normal Anatomical Structures (local and global). The application of ImageJ software in analyzing local and global features of representative cystoscopic images showing normal anatomical structures is demonstrated. Highlighted red areas indicate the regions detected by ImageJ as focal points for LLMs diagnosis. This visual representation illustrates the methodology used to enhance the generalizability of Image-Specific Learning in cystoscopic diagnosis of normal bladder anatomy

consistent and accurate interpretations of cystoscopy images, these AI tools could effectively support medical students and early-career urologists in developing their diagnostic skills and reducing misdiagnosis rates [14].

As AI technology continues to advance, it holds significant promise for enhancing diagnostic accuracy and efficiency in cystoscopy procedures. However, it is crucial to approach the integration of AI in urological practice with caution. These tools should be viewed as supportive aids to augment, rather than replace, the expertise of board-certified urologists. Future research should focus on larger, prospective studies, the integration of AI interpretations with clinical data, and the exploration of AI's potential in real-time video cystoscopy interpretation [15]. As we continue to refine and validate these AI models, their role in improving patient care and supporting clinical decision-making in urology is likely to grow, potentially leading to more efficient and accurate diagnoses. This could be particularly impactful in reducing misdiagnosis rates for complex conditions like CIS, ultimately enhancing the quality of urological care.

Limitations and future perspectives

While our study provides novel insights into the comparative analysis of LLMs in urological practice, several limitations warrant consideration. Firstly, our focus on bladder disorders and structures may not fully represent the models' capabilities across broader medical fields. This specialization, while valuable for urological

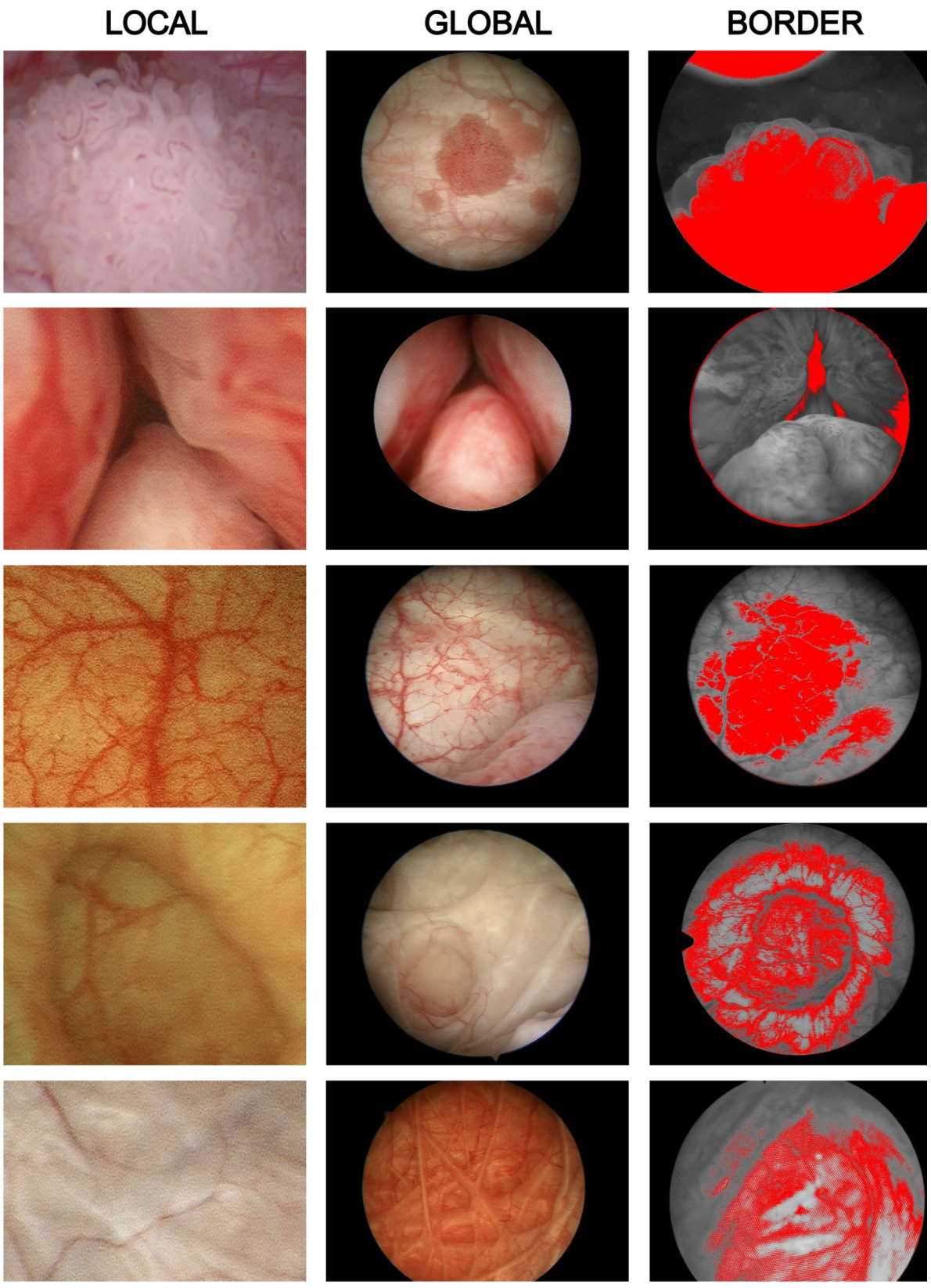

Fig. 5 Image Analysis of Specific Urological Conditions (local and global). The use of ImageJ software for analyzing local and global features in cystoscopic images is showcased, focusing on specific urological conditions. Red-highlighted areas represent the regions identified by ImageJ as crucial for LLMs diagnostic assessment. This approach aims to improve the generalizability and accuracy of LLMs in detecting and diagnosing specific bladder pathologies through targeted image analysis

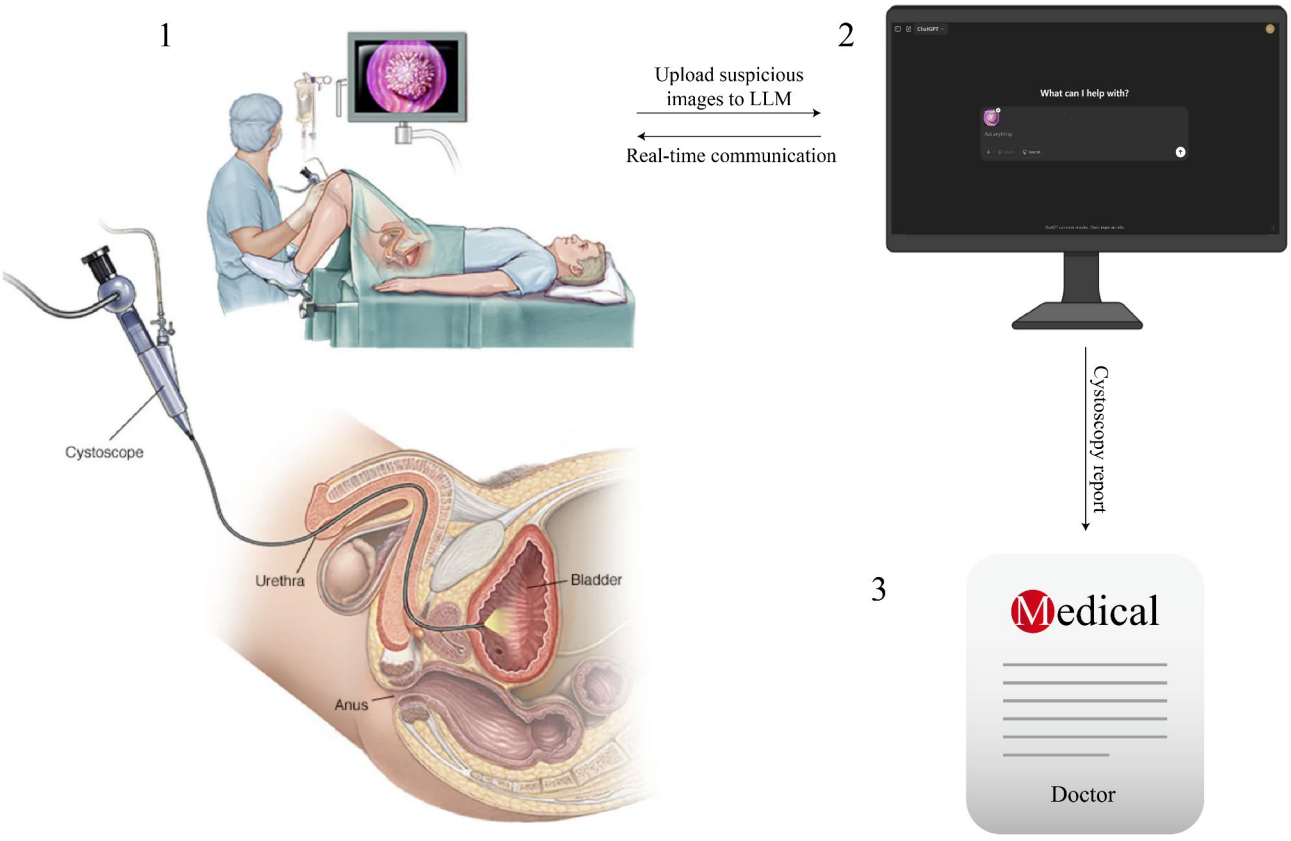

Fig. 6 Flowchart of AI-Assisted Real-Time Cystoscopy. The flowchart demonstrates how urologists can interact with the AI system during procedures to enhance diagnostic efficiency. Key advantages of this integration include: (**1**) real-time communication between urologists and the AI system, enabling rapid diagnostic feedback that reduces procedure time while improving diagnostic accuracy; (**2**) the AI's capability to recognize and analyze both global and local features within cystoscopic images, leveraging patterns identified during training; and (**3**) the collaborative relationship between urologist and AI that simultaneously enhances diagnostic accuracy and procedural efficiency in clinical practice. Additionally, we have incorporated a dedicated subsection addressing the ethical and legal implications of AI-driven diagnostics in urology, ensuring a comprehensive examination of these critical considerations

applications, limits the generalizability of our findings to other medical specialties.

Secondly, the rapid evolution of AI platforms presents both a challenge and an opportunity. The models we assessed are continually updating, and numerous other public AI platforms exist that physicians and patients may access [16]. This dynamic landscape necessitates ongoing comparative efficacy studies to keep pace with technological advancements. Furthermore, our error categorization method may oversimplify the complex nature of diagnostic reasoning in LLMs, suggesting the need for a more standardized approach to evaluating AI error in diagnostic contexts.

A significant limitation of our study is the sample size, which restricts our ability to comprehensively assess false positive and false negative rates with sufficient statistical power. These metrics are crucial for evaluating diagnostic performance in clinical settings.

Thirdly, our assessment methodology did not include real-time clinical decision-making scenarios, which may affect how these AI tools perform in actual practice environments with time constraints and varying patient presentations. To address these limitations, we are committed to conducting a larger-scale patient study in future research, which will enable more robust analysis of these critical performance metrics and provide greater confidence in the clinical applicability of our findings.

Looking ahead, the development of prompt engineering holds promise for enhancing the performance of these models in medical diagnostics [17]. We anticipate that more sophisticated prompts tailored to urological applications will emerge, potentially improving the accuracy and utility of LLMs in this field. Additionally, the external validity of our study is limited by the use of standardized cases. In real-world scenarios, patients and physicians may interact with LLMs in vastly different ways, providing variable amounts and types of information [18].

While our findings demonstrate the diagnostic potential of LLMs, their integration into clinical workflows

presents multifaceted challenges. Technical compatibility with existing cystoscopy systems, seamless data transfer protocols, and real-time processing capabilities remain critical hurdles. Furthermore, clinician acceptance and trust in AI-driven outputs necessitate iterative training programs and validation studies. The current reliance on static images in our study contrasts with the dynamic nature of live cystoscopic procedures, underscoring the need for future research to evaluate LLM performance in real-time video analysis. Additionally, workflow disruptions caused by AI-generated alerts or recommendations must be carefully balanced against efficiency gains. Collaborative efforts between AI developers, urologists, and healthcare administrators will be essential to optimize implementation strategies (Fig. 6).

The adoption of AI-driven cystoscopy interpretation raises critical ethical and legal questions. Diagnostic errors attributable to LLMs could lead to liability disputes, necessitating clear frameworks for accountability between clinicians and AI developers. Patient consent protocols must evolve to disclose AI's role in diagnostic processes, particularly when novel technologies are employed. Data privacy concerns extend beyond anonymization, as cystoscopic images may contain identifiable anatomical features. Regulatory bodies must establish standardized validation protocols and certification processes for medical AI tools, akin to existing medical device approvals. Finally, equitable access to AI-enhanced diagnostics must be prioritized to prevent disparities in care quality across socioeconomic strata.

Our research demonstrates promising potential for LLM-based assistance in cystoscopic examination, particularly through integration with existing medical equipment. However, these AI systems should be viewed as complementary diagnostic tools that enhance, rather than replace, the clinical expertise of board-certified urologists. To address the current limitations of static image analysis, future development should focus on creating models capable of real-time video processing during cystoscopy procedures, which would provide a more comprehensive diagnostic assessment. The successful implementation of this technology in clinical practice will require several key steps: extensive validation studies, optimization of training methodologies, and careful evaluation of both ethical implications and practical challenges in real-world clinical settings. This systematic approach will ensure the responsible and effective integration of AI-assisted diagnostic tools in urological practice.

Conclusion

ChatGPT-4 V and Claude 3.5 Sonnet demonstrate variable but promising diagnostic accuracy in the interpretation of cystoscopy images. While they excel in detecting cystitis, their performance in identifying other urological conditions, particularly BPH, needs enhancement. This study highlights the potential of LLMs in assisting with cystoscopy image interpretation but also underscores the need for further validation and refinement before their reliable implementation in clinical practice.

Abbreviations

LLMs	Large language models
AI	Artifcial intelligence
CIS	Carcinoma in situ
BPH	Benign prostatic hyperplasia

Acknowledgements

We extend our sincere gratitude to the staff at Zhongnan Hospital of Wuhan University for their invaluable support throughout this study. We particularly wish to acknowledge the outpatient urology unit nurses - Mrs. Lihua Dai, Mrs. Chunhua Luo, and Mrs. Jiuling Liu - for their dedicated assistance with data collection. Their contributions were essential to the completion of this research.

Author contributions

LG and XW contributed to protocol/project development; LG and XW wrote the first draft of the manuscript. LG, GL and YZ were involved in data collection; ZY, SL, JL and AG were involved in data collection; XW and TL supervised the work. All authors reviewed the manuscript.

Funding

This research was supported by grants from the Li Huanying Foundation of Beijing (**Grant No. PYZ201503**) and the National Natural Science Foundation of China (**Grant No. 82400906**). These grants were awarded to support the work of Dr. Xiaolong Wang. It is important to note that the funding organizations had no role in the study design, data collection, analysis, interpretation, manuscript preparation, or the decision to submit the article for publication. The researchers maintained full independence in conducting the study and reporting its results.

Data availability

Data is provided within the manuscript or supplementary information files.

Declarations

Ethics approval and consent to participate

Conducted in accordance with the World Medical Association's Declaration of Helsinki, the study received ethical approval from the Ethics Committee of Zhongnan Hospital of Wuhan University (Scientific Ethical Approval No. 2019108 and No. 2024093). All participants provided informed consent, and data were collected and analyzed anonymously.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Urology, Zhongnan Hospital of Wuhan University, Wuhan, China
[2]Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China
[3]Department of Urology, Qianjiang Central Hospital of Hubei Province, Qianjiang, China
[4]Hubei Key Laboratory of Urological Diseases, Wuhan University, Wuhan, China
[5]Hubei Clinical Research Center for Laparoscopic/Endoscopic Urologic Surgery, Zhongnan Hospital of Wuhan University, Wuhan, China
[6]Institute of Urology, Wuhan University, Wuhan, China
[7]Hubei Medical Quality Control Center for Laparoscopic/Endoscopic Urologic Surgery, Zhongnan Hospital of Wuhan University, Wuhan, China

[8]Wuhan Clinical Research Center for Urogenital Tumors, Zhongnan Hospital of Wuhan University, Wuhan, China
[9]Cancer Precision Diagnosis and Treatment and Translational Medicine Hubei Engineering Research Center, Zhongnan Hospital of Wuhan University, Wuhan, China

Received: 19 September 2024 / Accepted: 11 March 2025
Published online: 29 March 2025

References

1. Ramai D, Zakhia K, Etienne D, Reddy M. Philipp Bozzini (1773–1809): the earliest description of endoscopy. J MED BIOGR. 2018;26(2):137–41.
2. Lozano F, Raventos CX, Carrion A, Dinares C, Hernandez J, Trilla E, Morote J. Xpert Bladder Cancer Monitor for the Early Detection of Non-Muscle Invasive Bladder Cancer Recurrences: Could Cystoscopy Be Substituted? CANCERS. 2023;15(14).
3. Bube S, Dagnaes-Hansen J, Mahmood O, Rohrsted M, Bjerrum F, Salling L, Hansen RB, Konge L. Simulation-based training for flexible cystoscopy - A randomized trial comparing two approaches. HELIYON. 2020;6(1):e03086.
4. Naik R, Mandal I, Hampson A, Casey R, Vasdev N. A comparison of urology training across five major English-Speaking countries. Curr Urol. 2020;14(1):14–21.
5. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, Loffler C, Schwarzkopf SC, Unger M, Veldhuizen GP, et al. The future landscape of large Language models in medicine. Commun Med (Lond). 2023;3(1):141.
6. Tian D, Jiang S, Zhang L, Lu X, Xu Y. The role of large Language models in medical image processing: a narrative review. QUANT IMAG MED SURG. 2024;14(1):1108–21.
7. Lenis AT, Litwin MS. Does artificial intelligence meaningfully enhance cystoscopy?? JNCI-J NATL CANCER I. 2022;114(2):174–5.
8. Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large Language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology - a recent scoping review. DIAGN PATHOL. 2024;19(1):43.
9. Guldhammer CS, Vasquez JL, Kristensen VM, Norus T, Nadler N, Jensen JB, Azawi N. Cystoscopy Accuracy in Detecting Bladder Tumors: A Prospective Video-Confirmed Study. CANCERS, 2023;16(1).
10. Giansanti D. Joint expedition: exploring the intersection of digital health and AI in precision medicine with team integration. J PERS MED 2024, 14(4).
11. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J. 2021;8(2):e188–94.
12. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. NAT REV CANCER. 2018;18(8):500–10.
13. Zhang Y, Rumgay H, Li M, Yu H, Pan H, Ni J. The global landscape of bladder cancer incidence and mortality in 2020 and projections to 2040. J GLOB HEALTH. 2023;13:4109.
14. Ferro M, Falagario UG, Barone B, Maggi M, Crocetto F, Busetto GM, Giudice FD, Terracciano D, Lucarelli G, Lasorsa F et al. Artificial Intelligence in the Advanced Diagnosis of Bladder Cancer-Comprehensive Literature Review and Future Advancement. DIAGNOSTICS. 2023;13(13).
15. Fron A, Semianiuk A, Lazuk U, Ptaszkowski K, Siennicka A, Leminski A, Krajewski W, Szydelko T, Malkiewicz B. Artificial Intelligence in Urooncology: What We Have and What We Expect. CANCERS. 2023;15(17).
16. Akinrinmade AO, Adebile TM, Ezuma-Ebong C, Bolaji K, Ajufo A, Adigun AO, Mohammad M, Dike JC, Okobi OE. Artificial intelligence in healthcare: perception and reality. CUREUS J MED Sci. 2023;15(9):e45594.
17. Wang L, Bi W, Zhao S, Ma Y, Lv L, Meng C, Fu J, Lv H. Investigating the impact of prompt engineering on the performance of large Language models for standardizing obstetric diagnosis text: comparative study. JMIR Form Res. 2024;8:e53216.
18. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, Zhang M, Cao C, Wang J, Wang X, et al. The application of large Language models in medicine: A scoping review. ISCIENCE. 2024;27(5):109713.

Publisher's note