

<https://doi.org/10.1038/s41698-025-00916-7>

Evaluating the performance of large language & visual-language models in cervical cytology screening



Qi Hong^{1,6}, Shijie Liu^{2,6}, Liying Wu^{3,6}, Qiqi Lu¹, Pinglan Yang³, Dingyu Chen³, Gong Rao², Xinyi Liu¹, Hua Ye¹, Peiqi Zhuang¹, Wenxiu Yang¹, Shaoqun Zeng², Qianjin Feng^{1,4,5}, Xiuli Liu²✉, Jing Cai³✉ & Shenghua Cheng^{1,4,5}✉

Large language models (LLMs) and large visual-language models (LVLMs) have exhibited near-human levels of knowledge, image comprehension, and reasoning abilities, and their performance has undergone evaluation in some healthcare domains. However, a systematic evaluation of their capabilities in cervical cytology screening has yet to be conducted. Here, we constructed CCBench, a benchmark dataset dedicated to the evaluation of LLMs and LVLMs in cervical cytology screening, and developed a GPT-based semi-automatic evaluation pipeline to assess the performance of six LLMs (GPT-4, Bard, Claude-2.0, LLaMa-2, Qwen-Max, and ERNIE-Bot-4.0) and five LVLMs (GPT-4V, Gemini, LLaVA, Qwen-VL, and ViLT) on this dataset. CCBench comprises 773 question-answer (QA) pairs and 420 visual-question-answer (VQA) triplets, making it the first dataset in cervical cytology to include both QA and VQA data. We found that LLMs and LVLMs demonstrate promising accuracy and specialization in cervical cytology screening. GPT-4 achieved the best performance on the QA dataset, with an accuracy of 70.5% for close-ended questions and average expert evaluation score of 6.9/10 for open-ended questions. On the VQA dataset, Gemini achieved the highest accuracy for close-ended questions at 67.8%, while GPT-4V attained the highest expert evaluation score of 6.1/10 for open-ended questions. Besides, LLMs and LVLMs revealed varying abilities in answering questions across different topics and difficulty levels. However, their performance remains inferior to the expertise exhibited by cytopathology professionals, and the risk of generating misinformation could lead to potential harm. Therefore, substantial improvements are required before these models can be reliably deployed in clinical practice.

Cervical cancer is one of the most commonly diagnosed cancers and a leading cause of cancer death in women. In 2022, there were 661,021 women diagnosed with cervical cancer and 348,189 died of the disease in the world¹. Cervical cancer screening facilitates early diagnosis and timely intervention and treatment, thereby reducing the incidence and mortality of cervical cancer².

Manual screening requires doctors to identify a small number of lesion cells among tens of thousands under a microscope, which is labor-intensive

and experience-dependent³. Artificial intelligence (AI)-assisted screening systems^{3–12} can significantly improve screening efficiency and reduce doctors' workload. However, they are typically task-specific, such as lesion detection and image classification, and lack the capabilities to make cytopathological interpretation and reasoning like a cytopathologist. Recently emerging large language models (LLMs) and large visual-language models (LVLMs), such as GPT-4¹³, Bard¹⁴, Gemini¹⁵, and Claude¹⁶, have exhibited near-human or even human-level performance in medical-related tasks^{17–21}

¹School of Biomedical Engineering, Southern Medical University, Guangzhou, 510515, China. ²Britton Chance Center and MoE Key Laboratory for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics-Huazhong University of Science and Technology, Wuhan, China. ³Department of Obstetrics and Gynecology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. ⁴Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, 510515, China. ⁵Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou, 510515, China. ⁶These authors contributed equally: Qi Hong, Shijie Liu, Liying Wu. ✉e-mail: xliu@mail.hust.edu.cn; jingcai@hust.edu.cn; chengsh2023@smu.edu.cn

and are promising candidates for the application in cervical cytology screening.

The application of general-purpose LLMs and LVLMs in the medical field faces challenges related to accuracy²², bias, ethical concerns, and knowledge hallucination²³, toxicity²⁴. To ensure the reliability of LLMs and LVLMs for specific clinical tasks, it is necessary to systematically evaluate their performance using standardized datasets. Recent works on evaluating LLMs in medical related tasks indicate the supportive role of LLM in medical education²⁵, research²⁶, and clinical practice²⁷. However, in cervical cytology screening, there is a lack of a standardized benchmark dataset and systematic evaluation of the performance of different LLMs and LVLMs.

In this study, we constructed a benchmark dataset, termed CCBench, designed for LLMs and LVLMs evaluation in cervical cancer screening, and developed a GPT-based semi-automatic evaluation pipeline. CCBench consisted of 773 question-answer (QA) pairs and 420 visual-question-answer (VQA) triplets, which were constructed by utilizing the textual and image-text paired contents extracted from the widely recognized the Bethesda system (TBS) textbook. The performance of LLMs and LVLMs on CCBench was assessed using multifaceted metrics, such as accuracy, GPT-4 based G-Eval score²⁸, and expert evaluation score. We found that LLMs and LVLMs demonstrated promising accuracy and specialization in cervical cytology screening. On the QA dataset, the average accuracy across all models was 62.9% for close-ended questions, and the average expert evaluation score was 6.3/10 for open-ended questions. Notably, GPT-4 achieved the best performance, with an accuracy of 70.5% for close-ended questions and average expert evaluation score of 6.9/10 for open-ended questions. On the VQA dataset, the average accuracy across all models was 55.0% for close-ended questions, and the average expert evaluation score was 5.15/10 for open-ended questions. Gemini achieved the highest accuracy of 67.8% for close-ended questions, while GPT-4V attained the highest average expert evaluation score of 6.1/10 for open-ended questions. Besides, the abilities of LLMs and LVLMs exhibited significant variation on questions across different topics and difficulty levels. However, there remains a disparity between their responses and those of cytopathologists. The risk of generating misinformation could lead to potential harm, emphasizing the need for further refinement of these models to ensure their reliability before deployment in clinical practice. We have publicly released our dataset and codebase to promote future research and contribute to the development of the open-source community.

Results

Benchmark dataset CCBench

Due to the lack of datasets for evaluating the performance of LLM and LVLM in cervical cytology screening, we first constructed CCBench, a benchmark dataset sourcing from the TBS textbook²⁹. Through a GPT-4 based semi-automatic pipeline (Fig. 1a, Construction of QA and VQA datasets in “Methods” section, and Supplementary Note 1), we extracted 128 image-text pairs and 424 textual knowledge points from the textbook and its online atlas (<https://bethesda.soc.wisc.edu>), and created a QA sub-dataset (773 QA pairs) and a VQA sub-dataset (420 image-question-answer triplets) (Fig. 1b). The QA dataset encompasses knowledge points from the chapters of endometrial cells (7.50%), atypical squamous cells (12.16%), squamous epithelial cell abnormalities (41.01%), and glandular epithelial cell abnormalities (39.33%) (Fig. 1d). The VQA dataset covers the chapters of non-neoplastic findings (31.67%), atypical squamous cells (20.71%), squamous epithelial cell abnormalities (22.86%) and glandular epithelial cell abnormalities (15.95%), and other malignant neoplasms (8.81%). The knowledge points in each chapter are from sections of criteria, definition, explanatory notes, problematic patterns, and tables (Fig. 1d), which provide guidelines and protocols for collecting, preparing, staining, and evaluating cervical cell samples, terms for uniform terminology and classification, additional explanations about interpreting and applying the criteria, complex cases encountered in clinical practice, and standardized forms and templates for recording and reporting cervical cytopathological findings, respectively. We evaluated the performance of LLMs and LVLMs using

close-ended and open-ended questions (Fig. 1b), which were designed to simulate real-world clinical scenarios in cervical pathology screening. Figure 1c, e, f, and g showed the characteristics of the questions and answers in the QA and VQA dataset, including the first three words of questions (Fig. 1c), distribution of question length (Fig. 1e) and answer length (Fig. 1f), and the top 50 most frequent medical terms in the questions and answers (Fig. 1g).

Evaluation pipeline

We developed a semi-automated evaluation pipeline (Fig. 2) to access the performance of LLMs and LVLMs and utilized prompt engineering, conversation isolation, and result post-processing to enhance the reliability of the evaluation process. Firstly, prompts and questions were simultaneously fed into models to generate answers. The prompts served to instruct models in generating specific answers and directly impacted the quality of answers. Thus, we needed to provide the models with relevant knowledge in advance to ensure that models accurately comprehended the context of the questions and responded using specialized terminology. We elaborated system prompts for each dataset, ensuring that all models receive consistent context, prior knowledge, behavioral instructions, and answer format (Supplementary Note 2). Inspired by the chain of thought framework³⁰, in these system prompts, we required the models to generate answers step by step and provide reasons for their answers, and created different response templates and question-answer examples for both close and open-ended questions (Supplementary Note 2d). Besides, Models were deployed through application programming interface (API) calls to ensure efficiency and prevent interference between dialogues and data leakage during questioning (i.e., conversation isolation). Then, we employed manual review to assess and handle unformatted answers (Supplementary Table 1 and Unformatted answers processing in “Methods” section). Finally, we used statistical metrics, LLM-based G-Eval score²⁸, and expert evaluation to assess the answers (Evaluation metrics in “Methods” section).

Performance on QA dataset

We assessed the performance of five commercial closed-source LLMs (GPT-4, Bard, Claude-2.0³¹, Qwen-Max³², and ERNIE-Bot-4.0³³) and one open-source LLM (LLaMa-2) on the QA dataset. We employed accuracy, precision, recall, F1 score, and specificity to evaluate the answers of different LLMs on the close-ended questions in the QA dataset (Fig. 3a, b). GPT-4 demonstrated the highest accuracy (0.705) (Fig. 3a), and Qwen-Max obtained the highest precision (0.764), recall (0.866), and F1 score (0.812) (Fig. 3b). While Bard and LLaMa-2 demonstrated outstanding specificity, their low recall values highlighted a significant risk of false negatives, underscoring potential limitations in their clinical reliability (Fig. 3b).

For the open-ended questions in the QA dataset, we utilized the expert evaluation scores and G-Eval scores to assess the quality of the answers of LLMs. All three experts agreed that GPT-4 achieved superior performance, earning the highest average expert evaluation score of 6.9/10 (Fig. 4a). In terms of the number of times receiving the highest ratings from experts, whether from a single expert (Fig. 4b) or multiple experts simultaneously (Fig. 4c), GPT-4 received the most recognition, followed by Claude-2.0. The consistency matrix of the three experts illustrated the degree of agreement among them in ranking the models (Fig. 4d). The G-Eval scores of different LLMs were consistent with those from expert evaluation, with GPT-4 achieving the second highest G-Eval score (Fig. 4e). The example answers from different LLMs to the same questions revealed their varying abilities to comprehend and response accurately (Fig. 4f, g). The answers from GPT-4 and Claude-2.0 were more accurate and consistent with the ground truth (Fig. 4g).

Figure 5a shows the performance of different LLMs on the QA dataset across different chapters. For the close-ended questions, LLMs exhibited varying levels of accuracy across questions on different topics (i.e., chapters). GPT-4, Claude-2.0, and Qwen-Max demonstrated high accuracy on the questions about glandular epithelial cell abnormalities (Ch.6), while all the models obtained relatively low accuracy on the questions about endometrial cell (Ch. 3). For the open-ended questions,

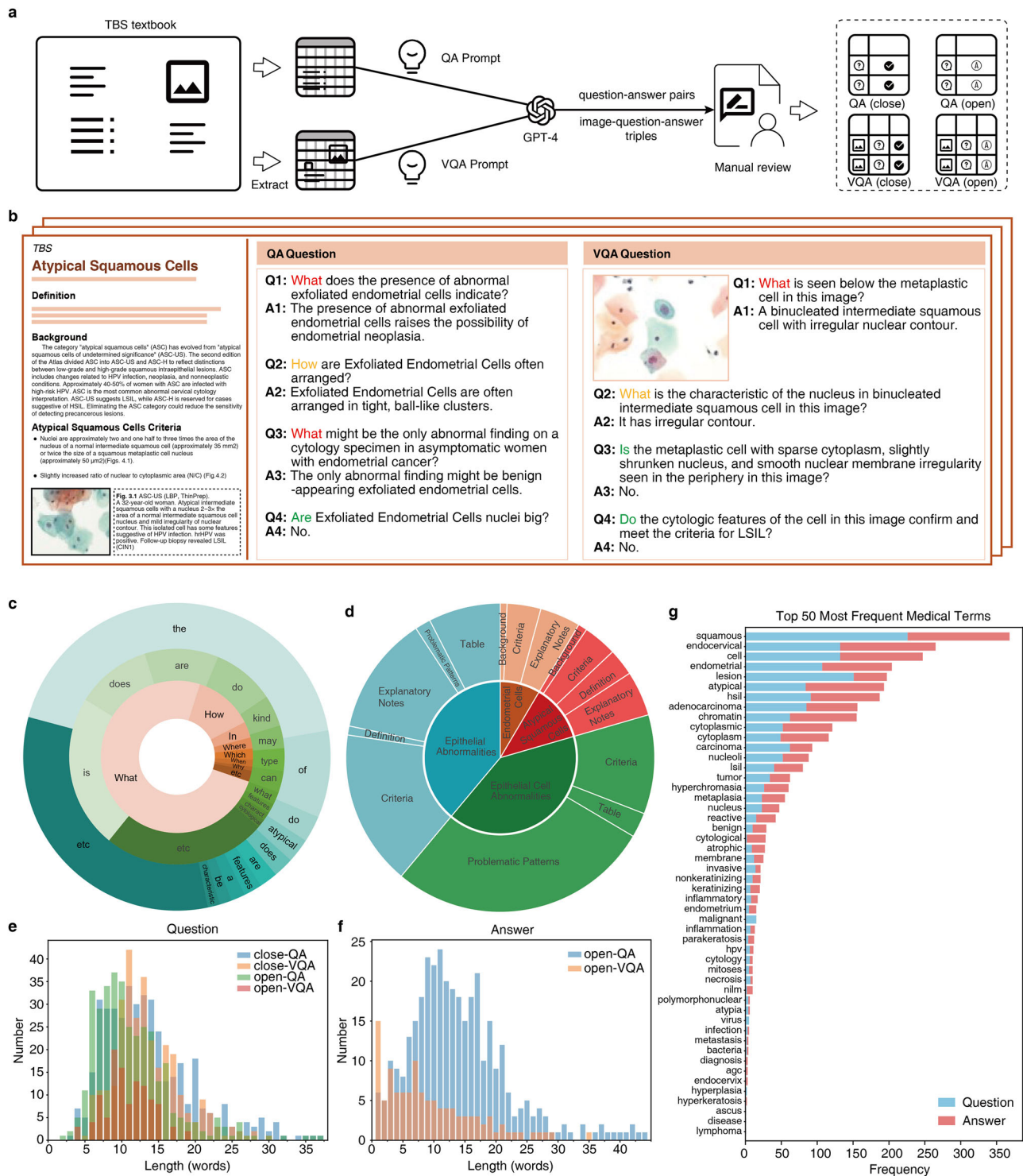


Fig. 1 | Construction pipeline of CCBench and its characteristics. **a** The GPT-4 based semi-automatic pipeline for dataset construction. Textual knowledge points and image-text pairs were extracted from the TBS textbook and its online atlas. GPT-4 was then employed to generate close/open-ended QA pairs and VQA triplets using these data, followed by a manual review to ensure their quality. **b** QA pair (middle) and VQA triplets (right) examples generated from the TBS textbook (left). **c** The

distribution of the first three words of open-ended questions in the CCBench, with the order of words radiating outward from the center. **d** The proportion of knowledge points from different chapters and sections of the textbook. **e**, **f** Distribution of question length (**e**) and answer length (**f**) in QA and VQA datasets. **g** The top 50 most frequent medical terms in the questions and answers of CCBench.

GPT-4, Claude-2.0, and Qwen-Max showed superior capabilities in resolving questions about atypical squamous cells (Ch. 4), achieving consistently high scores of G-Eval and expert evaluation. We retrospectively divided the questions into three difficulty levels based on the answers of LLMs (Fig. 5c). For the close-ended questions, questions with

fewer than two LLMs answering incorrectly were classified as “easy”, those with two to four LLMs answering incorrectly as “normal”, and those with more than four LLMs answering incorrectly as “hard”. For the open-ended questions, questions were sorted according to their average expert evaluation scores across all models, and the top 25% were

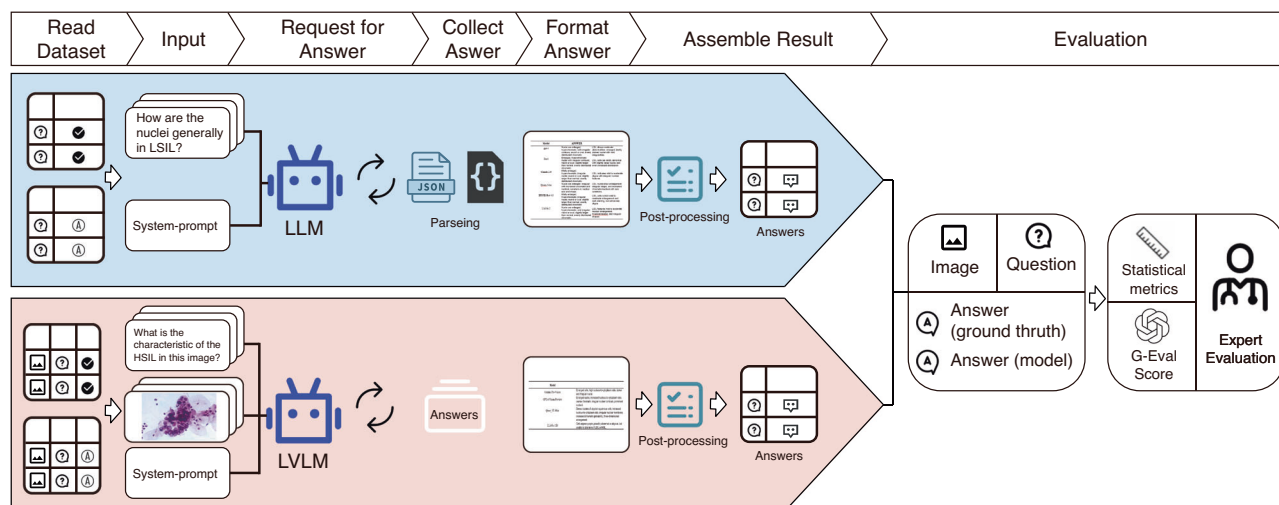


Fig. 2 | Semi-automated pipeline for LLMs and LVLMs evaluation. Fed questions and system prompts into each model, collect and format answers, and finally use statistical metrics, LLM-based metrics, and expert evaluation to assess the answers.

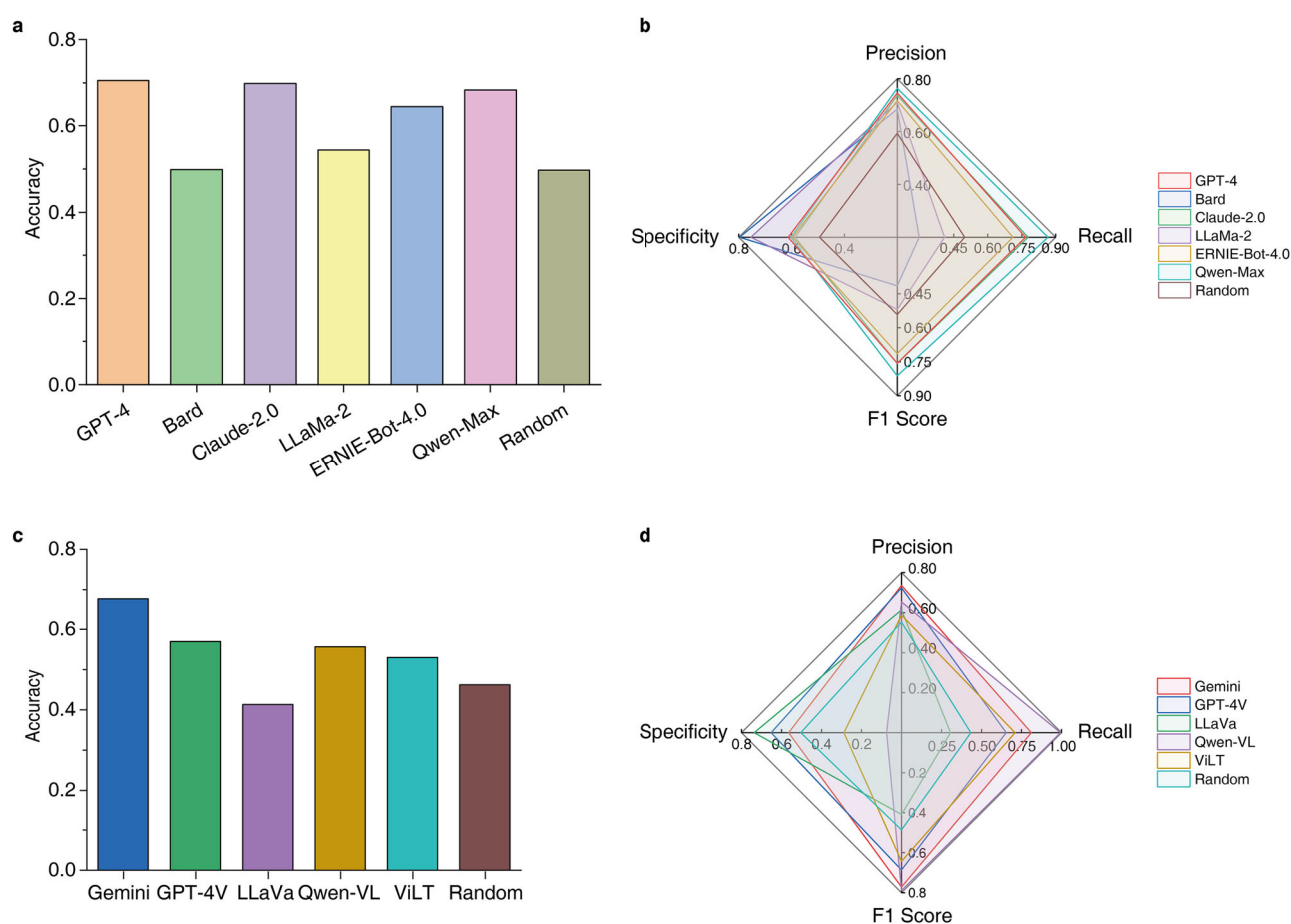


Fig. 3 | Performance of different LLMs and LVLMs on the close-ended questions in QA and VQA dataset. a, c Accuracy of different LLMs (a) and LVLMs (c). **b, d** Precision, recall, F1 score, and specificity of different LLMs (b) and LVLMs (d).

classified as “easy”, the bottom 25% as “hard”, and the remaining questions as “normal”. For the close-ended QA task, GPT-4 obtained the highest accuracy (0.682) on the normal questions, while Qwen-Max retained the highest accuracy (0.954) on the easy questions. However, all the LLMs demonstrated low accuracies when addressing the hard questions (Fig. 5c). For the open-ended QA task, GPT-4 obtained an overall higher G-Eval score and expert evaluation score across three

difficulty levels (Fig. 5c). As the difficulty level of question increased, the performance of all the LLMs decreased (Fig. 5c).

Performance on VQA dataset

We assessed the performance of three commercial closed-source LVLMs (Gemini, GPT-4V, and Qwen-VL) and two open-source LVLMs (ViLT and LLaVa) on the VQA dataset. We employed accuracy, precision, recall, F1

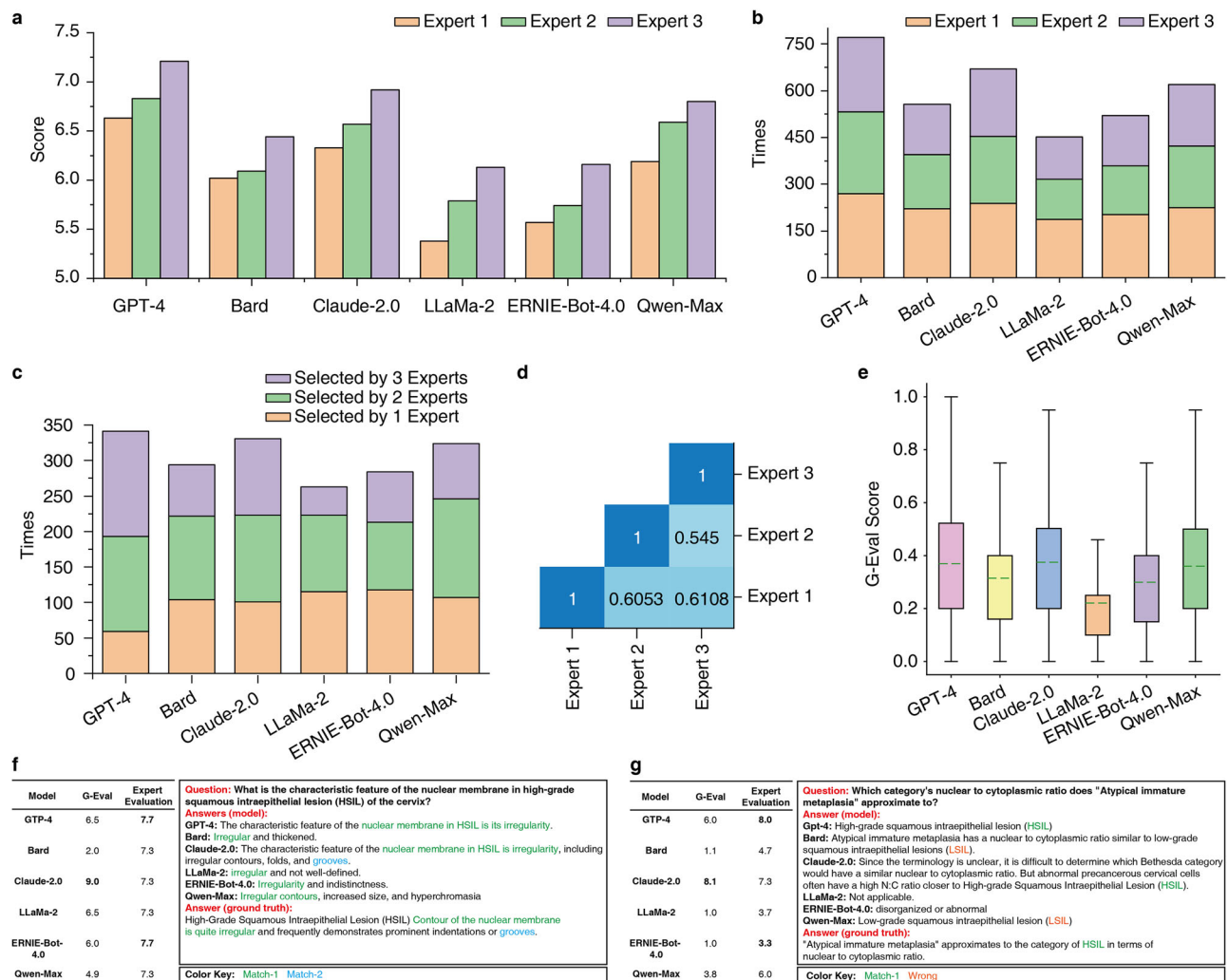


Fig. 4 | The performance of different LLMs on the open-ended questions in QA dataset. **a** The average expert evaluation scores of different LLMs from three experts. **b** The number of times that different expert recognizes each LLM performs better than all other models. **c** The number of times that each LLM is recognized as the best by 1, 2, or all 3 experts simultaneously. **d** The average Spearman's rank correlation of evaluation among experts. **e** The distribution of G-Eval scores of different LLM. The

data are shown as boxplots and whiskers (min to max) with all data points, where the upper and lower hinges represent the 25th and 75th percentiles, and the center is the average score. **f, g** Two representative questions and answers from different LLMs, as well as their corresponding G-Eval scores and expert evaluation scores. Owing to space constraints, the reasoning part of the answer was not shown.

score, and specificity to evaluate the answers of different LLMs on the close-ended questions in the VQA dataset (Fig. 3c, d). Gemini achieved the highest accuracy (0.678) (Fig. 3c) and precision (0.736), and the second highest recall (0.811) and F1 score (0.772) (Fig. 3d), demonstrating its superior overall effectiveness in accurate judgments. Although Qwen-VL stood out with the highest recall (0.994) and F1 score (0.788) (Fig. 3d), it obtained significantly lower specificity (0.074) (Fig. 3d), demonstrating its high risk of producing false positives.

For the open-ended questions in the VQA dataset, all three experts agreed that GPT-4V achieved significantly superior performance than the others (Fig. 6a). In terms of the number of times receiving the highest ratings from experts, whether from a single expert (Fig. 6b) or multiple experts simultaneously (Fig. 6c), GPT-4V received the most recognition. The consistency matrix of the three experts illustrated the degree of agreement among them in ranking the models (Fig. 6d). Additionally, GPT-4V obtained higher G-Eval scores (Fig. 6e). Although GPT-4V and Gemini occasionally produced incorrect results, they can identify certain cell subtypes in a complex smear; in contrast, LLaVa may refuse to answer, and Qwen-VL may only provide descriptions of cytological features (Fig. 6f). GPT-4V can accurately describe smear characteristics, while other LLMs

exhibited some hallucinations³⁴, introducing fabricated information that could potentially mislead clinical decision-making (Fig. 6g).

Figure 5b shows the performance of different LLMs on the VQA dataset across different chapters. For the close-ended questions, Qwen-VL obtained the highest accuracy on questions about squamous epithelial cell abnormalities (Ch. 5), glandular epithelial cell abnormalities (Ch. 6), and other malignant neoplasms (Ch. 7), but the lowest at non-neoplastic findings (Ch. 2) and atypical squamous cells (Ch. 4), which may result in the overall low accuracy (Fig. 3c). Gemini demonstrated relatively high accuracy across all topics. For the open-ended questions, GPT-4V obtained the highest expert evaluation scores on questions from all chapters, especially the questions about atypical squamous cells (Ch. 4), squamous epithelial cell abnormalities (Ch. 5), and other malignant neoplasms (Ch. 7). However, there were deviations between the results of G-Eval and expert evaluation, which may be limited to the ability of G-Eval framework. Similarly, the questions in the VQA dataset were divided into three difficulty levels based on the answers of LLMs (Fig. 5d). For the close-ended questions, questions with fewer than one LLM answering incorrectly were classified as “easy”, those with two to three LLMs answering incorrectly were deemed “normal”, and questions with more than three LLMs providing incorrect

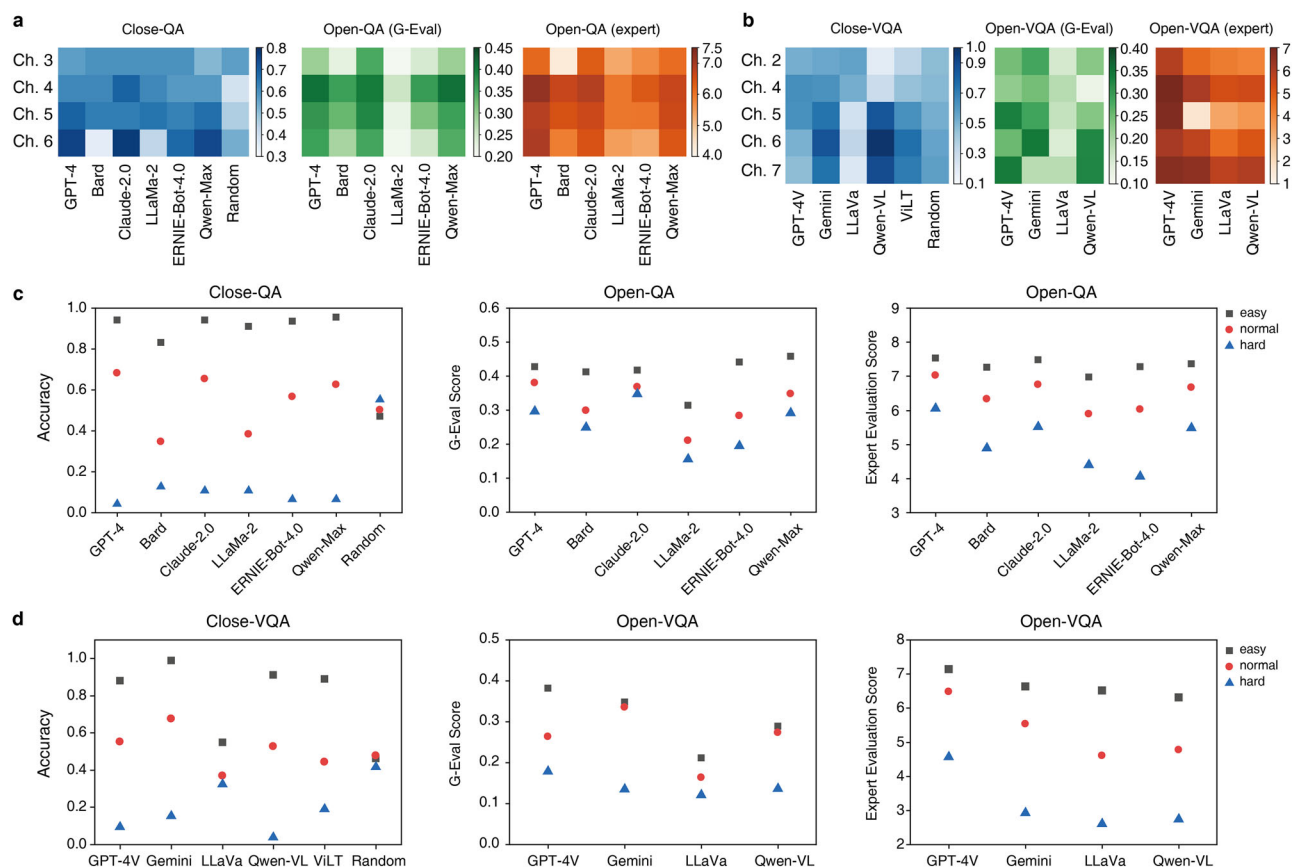


Fig. 5 | Performance of different LLMs and LVMs across different chapters and difficulty levels. a, b The performance of different models on QA (a) and VQA (b) datasets across different chapters. **c, d** The performance of different models on QA (c) and VQA (d) datasets across three difficulty levels: “easy”, “normal”, and “hard”.

The titles of the chapters are as follows: non-neoplastic findings (Ch. 2), endometrial cell (Ch. 3), atypical squamous cells (Ch. 4), squamous epithelial cell abnormalities (Ch. 5), glandular epithelial cell abnormalities (Ch. 6), and other malignant neoplasms (Ch. 7).

answers were categorized as “hard”. The open-ended questions were designated as “easy” or “hard” if they fell within the top 30% or bottom 30% of the average expert evaluation scores across all models, and the remaining questions as “normal”. As the difficulty level of question increased, the performance of all the LVMs decreased (Fig. 5d). For the close-ended questions, Gemini obtained the highest accuracy on easy and normal questions (Fig. 5d). However, all the LVMs obtained low accuracies (below 0.5) on hard questions (Fig. 5d). For the open-ended questions, GPT-4V consistently obtained higher expert evaluation scores across all three difficulty levels (Fig. 5d).

Error pattern analysis in LVMs

To better understand the reasoning processes of different models in cervical cytology interpretation, we analyzed error patterns in their responses across VQA datasets. Our analysis revealed several common error patterns shared across different models, providing insights into their limitations when handling specialized medical tasks (Fig. 7). We identified two main categories of errors: knowledge-based errors and hallucination-based errors. Knowledge-based errors included misclassification/feature identification errors and answer matching failures, which manifested when models generated responses that were grammatically correct and potentially factually accurate in isolation but failed to directly address the question posed. For example, both GPT-4V and Gemini produced well-formatted responses that appeared professional but did not identify the key cytomorphological features described in the reference answer (Fig. 7a). They missed critical diagnostic elements such as glandular architecture, irregular chromatin distribution, and prominent macronucleoli, which are essential features for identifying endocervical adenocarcinoma. Misclassification and feature identification errors also occurred when

models failed to correctly identify or classify cytomorphological features critical for diagnosis.

Among the hallucination-based errors, logical reasoning failures were particularly concerning. Although both GPT-4V and Gemini correctly identified some cellular features like nuclear enlargement and hyperchromasia, they made unsupported diagnostic leaps to wrong subclass such as high-grade squamous intraepithelial lesion (HSIL) or squamous cell carcinoma (SCC) (Fig. 7b). This tendency to reach premature conclusions without sufficient evidence represents a critical limitation in medical diagnosis contexts. In addition, instruction drifting was observed when models failed to adhere to the specific requirements of the prompt; for instance, when asked to identify the components of a three-dimensional cluster, multiple models instead attempted to provide a diagnosis, completely diverging from the instructional intent (Fig. 7b). Fluent nonsense generation was notably observed in LLaVa’s responses across both examples. LLaVa made illogical inferences about cell abnormality based solely on color attributes—stating that cells were “predominantly blue and purple in color, indicating that they are likely abnormal or atypical”—a conclusion with no cytopathological validity. Similarly (Fig. 7a), LLaVa produced text that, while grammatically coherent, offered no meaningful description of the cellular morphology shown in the image (Fig. 7b). The prevalence of these error patterns across different models suggests common limitations possibly stemming from similar training approaches or knowledge gaps, and the parallel reasoning failures observed indicate fundamental challenges in medical reasoning that require targeted improvements beyond general model capabilities. These findings underscore the need for careful assessment and potentially domain-specific refinement before deploying such models in clinical cytopathology settings.

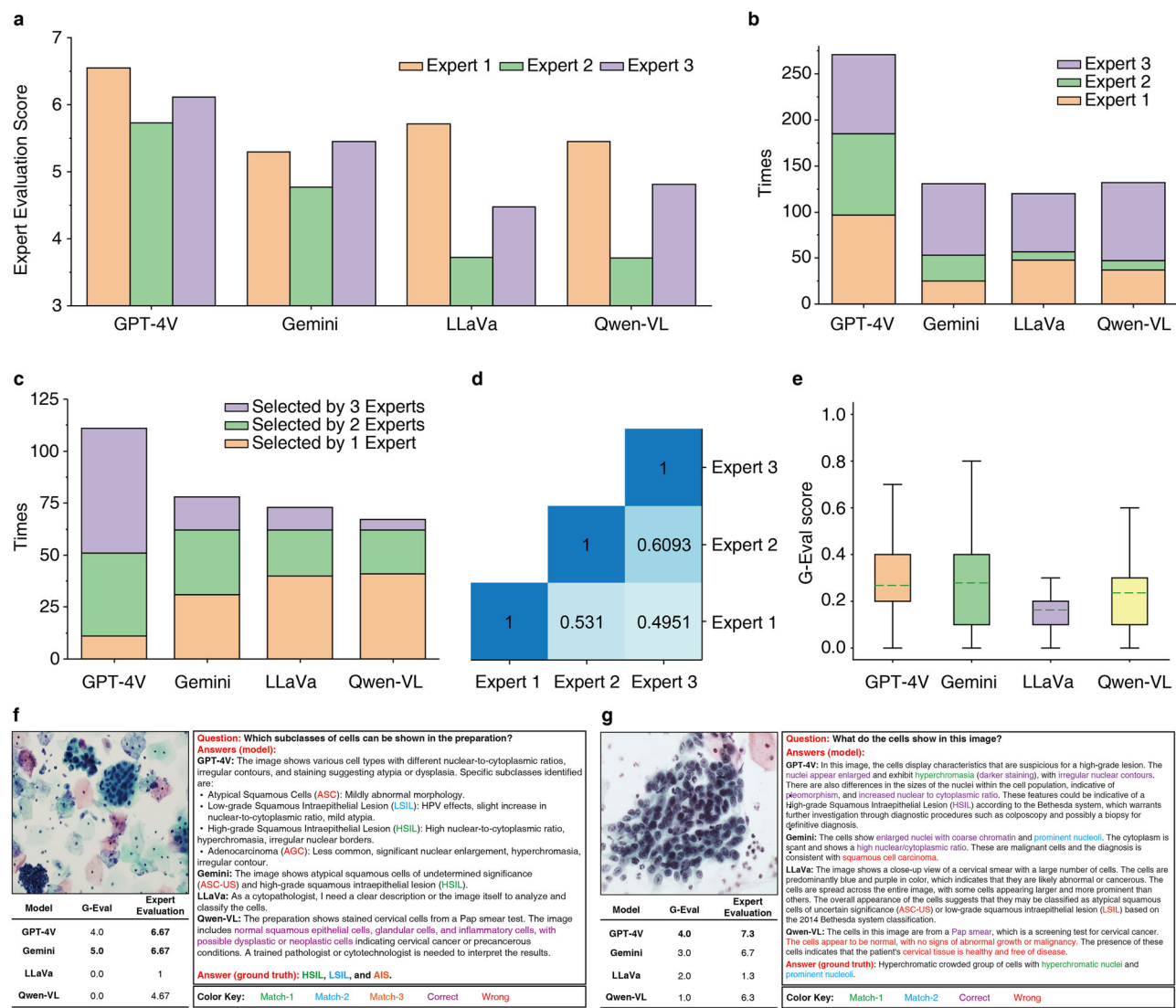


Fig. 6 | The performance of different LVLMs on the open-ended questions in VQA dataset. **a** The average expert evaluation scores of different LVLMs from three experts. **b** The number of times that different expert recognizes each LVM performs better than all other models. **c** The number of times that each LVM is recognized as the best by 1, 2, or all 3 experts simultaneously. **d** The average Spearman's rank correlation of evaluation among experts. **e** The distribution of G-Eval scores of different LVM. The data are presented in the form of boxplots and whiskers, encompassing the entire range from minimum to maximum values, with the upper

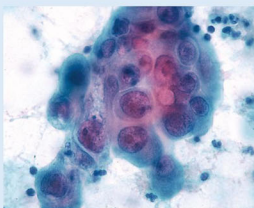




and lower hinges corresponding to the 25th and 75th percentiles, respectively. The dashed line denotes the average score. **f, g** Two representative questions and answers from different LVLMs, as well as their corresponding G-Eval scores and expert evaluation scores. The reasoning part of the answer was not shown due to space constraints. Due to the context length limitation, ViLT could not process questions with a length acceptable by other models during open-ended question evaluations, and therefore, it was not assessed.

Discussion

It is of great concern to evaluate the performance of LLMs and LVLMs on specific clinical tasks before their deployment in clinical practice. This study constructed a dedicated benchmark dataset CCBench (Fig. 1) and developed a GPT-based semi-automated evaluation pipeline (Fig. 2) to assess the performance of popular LLMs and LVLMs in cervical cytology screening. Across the QA tasks, GPT-4 demonstrated better overall performance than the other LLMs, with higher accuracy for close-ended questions (Fig. 3a) and more recognition from the experts for open-ended questions (Fig. 4). In the more challenging VQA tasks, Gemini exhibited better overall performance for close-ended questions (Fig. 3c, d), while GPT-4V obtained significantly more recognition from the experts for the open-ended questions (Fig. 6). Our study highlighted the advanced text comprehension capability of GPT-4 and the superior cytopathological image understanding and reasoning abilities of both Gemini and GPT-4V, suggesting they may serve as promising alternatives for the cervical cytology screening application.

Moreover, the benchmark dataset CCBench and the well-designed evaluation pipeline can serve as valuable tools for assessing LLMs and LVLMs before their deployment in cervical cytology screening application.

Although the evaluated LLMs and LVLMs (such as GPT-4, GPT-4V, and Gemini) have shown promising results in text and image understanding and reasoning tasks for cervical cancer screening, there are still defects hindering their application in clinical practice. They may generate correct answers mixed with erroneous information (Figs. 4g, 6f, g), which could negatively impact medical diagnoses. Besides, they are obviously incompetent in solving relatively difficult questions, and their performance on close-ended tasks is even lower than the random baseline (Fig. 5c, d). Despite prior knowledge about cervical cancer screening being provided in system prompts (Supplementary Note 2e), it is still limited and insufficient compared to the knowledge reserve of cytopathologists. Thus, utilizing more specialized data about cervical cancer screening to fine-tune the models may further enhance their performance and reliability.

<div>a</div> <div></div>	<div>Class: Pos</div> <div>SubClass: MMMT</div> <div>Explanatory Notes: Pos. Malignant Müllerian mixed tumor (MMMT). Three-dimensional cluster of large epithelioid cells with round but pleomorphic nuclei, coarse granular chromatin, macronucleoli, and a moderate amount of cytoplasm.</div> <div>Cytomorphologic Criteria: None</div>
<div>Question: What does the three-dimensional cluster consist of in this image?</div>	
<div>Ground Truth: Large epithelioid cells with round but pleomorphic nuclei, coarse granular chromatin, macronucleoli, and a moderate amount of cytoplasm.</div>	
<div><div> ChatGPT</div><div>The 3D cell cluster shows cells with enlarged hyperchromatic nuclei and high nuclear-to-cytoplasmic ratio. The irregular chromatin pattern and nuclear <u>variation</u> suggest possible dysplasia or neoplasia. These features resemble <u>HSIL</u> or <u>SCC</u>, but definitive diagnosis requires <u>clinical correlation and additional testing</u>.</div></div>	<div>Logic Leap, Instruction Drift</div>
<div><div> Gemini</div><div>The three-dimensional cluster in this image consists of malignant cells with enlarged, hyperchromatic, and <u>pleomorphic</u> nuclei, <u>consistent with a diagnosis</u> of high-grade squamous intraepithelial lesion (HSIL).</div></div>	<div>Logic Leap, Instruction Drift</div>
<div><div> LLaVA</div><div>The three-dimensional cluster in the image consists of various types of cervical cells, including <u>normal, atypical, and abnormal cells</u>, ...s to be a mixture of different cell types, indicating that the image may be capturing a transitional phase or a snapshot of the cervical cells' development.</div></div>	<div>Fluent Nonsense Generation</div>
<div><div> Qwen</div><div>The three-dimensional cluster in the image consists of <u>cancerous cells</u>.</div></div>	<div>Logic Error</div>

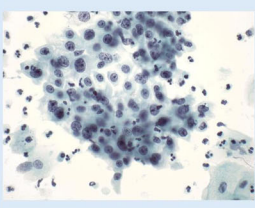




<div>b</div> <div></div>	<div>Class: Pos</div> <div>SubClass: Endocervical adenocarcinoma</div> <div>Explanatory Notes: Adenocarcinoma, endocervical. Cell group demonstrates glandular architecture and large nuclei, irregular chromatin distribution, and prominent macronucleoli. This group shows well-defined cytoplasmic boundaries mimicking reparative change.</div> <div>Cytomorphologic Criteria: Glandular cells with large nuclei, irregular chromatin distribution and prominent nucleoli.</div>
<div>Question: What are the characteristics of the smear in the image?</div>	
<div>Ground Truth: Cell group demonstrates glandular architecture and large nuclei, irregular chromatin distribution, and prominent macronucleoli. This group shows well-defined cytoplasmic boundaries mimicking reparative change.</div>	
<div><div> ChatGPT</div><div>The Pap smear reveals cervical cells with increased cell density, variable morphology, enlarged hyperchromatic nuclei with irregular contours, and increased nuclear-to-cytoplasmic ratios against a clean background, suggesting potential cellular abnormalities...</div></div>	<div>Unmatch Error</div>
<div><div> Gemini</div><div>The smear in the image is composed of a cluster of cells with enlarged, hyperchromatic nuclei and increased nuclear to cytoplasmic ratio.</div></div>	<div>Unmatch Error</div>
<div><div> LLaVA</div><div>The image shows a smear of cervical cells, which appears to be a liquid-based smear. <u>The cells are predominantly blue and purple in color, indicating that they are likely abnormal or atypical.</u> The cells are scattered throughout the image, with some located closer to the center and others towards the edges. The overall appearance of the smear suggests that it may contain abnormal or precancerous cells, which would require further investigation and potentially a biopsy for a definitive diagnosis.</div></div>	<div>Logic Confusion Error</div>
<div><div> Qwen</div><div>The smear in the image is a Pap smear, which is a type of medical test used to detect abnormal cells in the cervix. The cells in the smear appear to be blue and round, with some appearing to be clustered together. This is a characteristic feature of a Pap smear, where the cells are stained and visualized under a microscope to identify any abnormalities.</div></div>	<div>Low quality analyze Error</div>

Fig. 7 | Error pattern analysis across different LVLs. **a** Examples of error patterns in identifying endocervical adenocarcinoma characteristics. **b** Error patterns in identifying three-dimensional cell cluster features.

We acknowledge the limitations of this study. The CCbench was constructed from the representative data in the TBS textbook, which ensured its reliability and standardization. However, in clinical practice, the variations in staining and imaging characteristics³⁵ pose additional challenges for LVLMS. To further assess the capability of LVLMS in real-world clinical scenarios and mitigate potential data contamination concerns, we constructed a private VQA dataset using real-world clinical pathological smear images. The evaluation results on our private clinical dataset (Supplementary Note 3, Supplementary Fig. 1) aligned with our findings on the CCbench dataset, confirming the reliability of our evaluation method and the capabilities of the evaluated models. Besides, incorporating a larger volume of high-quality, heterogeneous data from real-world clinical scenarios, such as data extracted from clinical cases, could further enhance the reliability of the evaluation results and help identify potential issues and limitations of LLMs and LVLMS.

Methods

Construction of QA and VQA datasets

Knowledge points extraction. The TBS standard²⁹ is a widely recognized gold standard for cytopathological interpretation in cervical cytology, widely accepted by cytopathologists worldwide. The TBS textbook establishes a grading system for cervical cytopathology, clearly defines the interpretation criteria for each grade, provides numerous complex case examples and expert interpretations of smear images, and standardizes the terminology used in cytological evaluation. Moreover, there is a large number of high-quality image-text paired data in the TBS textbook and its online atlas (<https://bethesda.soc.wisc.edu>). The online atlas offers the higher-resolution versions of the images presented in the textbook. Thus, we used the TBS textbook and its online atlas as the data source and extracted key knowledge points and images (the high-resolution versions in the online atlas were used), which can serve as a benchmark for assessing whether LLMs or LVLMs possess the capability of cytopathological interpretation. We manually extracted sentences from the background, criteria, definition, explanatory notes, and problematic patterns sections, as well as from tables and figure captions. To avoid loss of contextual information, we replaced the pronouns in the

sentences using content to which they refer. Each image was saved in ‘png’ format, and its corresponding caption and associated text were saved as ‘txt’ file. In each sentence, we replaced pronouns, conjunctions, and elements that imply connections to surrounding sentences with the elements they actually referred to, and removed footnotes and citation information. Sentences were manually revised to enhance their fluency and coherence. For the image-text pairs, we removed pairs unrelated to pathology-related topics, corrected any mismatches, split grouped images into individual images, and removed marks (e.g., figure numbers and arrows) from the images.

QA pair and VQA triplet generation. We developed a GPT-4-based semi-automated pipeline (Fig. 1a) to convert the extracted texts and images into QA pairs and VQA triplets. For the QA dataset construction, the extracted sentences were commonly long and complex. Thus, utilizing prompt engineering (Supplementary Note 1), we decomposed and simplified the sentences to make each sentence focus on a single knowledge point. We retained sentences describing cell morphology and grading, supplemented incomplete sentences, and excluded those containing the term “follow-up”. Then, we defined rules for GPT-4 to generate questions and answers based on these sentences. To ensure the variety of questions, we required the generated questions to include the terms “what”, “where”, “when”, “how much”, “how many” or “is”. The questions and answers were mandated to be coherent and prohibited from using pronouns like “these”. Besides, the system prompts also included few-shot examples, specific prompting rules, and context-aware prompting standards to help GPT-4 generate compliant QA pairs more effectively. Finally, the pipeline used regular expressions to match and collect the correctly formatted QA pairs, their reasoning, and any doubts regarding the knowledge points in the input sentences. The final QA pairs were determined through manual review and verification, ensuring their conformity to the rules specified in the prompts (Supplementary Note 1).

For the VQA dataset construction, since the sentences in captions were generally simpler than those for QA dataset construction, we did not perform sentence simplification. We only split a single caption into multiple sentences using GPT-4. We formulated additional rules (Supplementary

Note 1b) to exclude descriptions unrelated to cytopathology and sentences that evaluate cell types or smear types, and extracted the subject from the context for the sentence lacking a subject. Following the same pipeline used in QA dataset construction, firstly, we generated QA pairs, their underlying reasoning, and any doubts regarding the knowledge points in the input sentences. Then, the QA pairs and their corresponding images constituted the VQA dataset.

For the close-ended questions in both datasets, we used GPT-4 to insert “not” into some questions to convert their answers from “yes” to “no,” aiming for a balanced distribution of “yes” and “no” responses. All the QA pairs and VQA triplets in QA and VQA datasets were manually reviewed to identify and correct logic, grammar, context, question phrasing, and type categorization errors.

Competing models

We selected the state-of-the-art commercial closed-source models Claude-2.0, GPT-4, Bard, Qwen-Max, and ERNIE-Bot-4.0 for text-based language tasks, and GPT-4V, Qwen-VL, and Gemini for multimodal tasks. We chose three popular open-source models: ViLT, LLaMa-2, and LLaVA (the multimodal extension of LLaMa-2). We used the latest and most commonly used version of each model at the time of evaluation (Supplementary Table 2). For the Gemini model, we selected the version accessible via API. We disabled the network access of all internet-enabled model APIs during the evaluation. In addition, in order to reflect the actual capabilities of each model and ensure that there were no hyperparameters that affected the smooth progress of the experiment, the hyperparameters of each model were kept at their default settings without any modifications.

Unformatted answers processing

We instructed the models to answer in a specific format for each task (Supplementary Note 2d) and subsequently performed a validity check on the format of answers (Supplementary Table 1). For close-ended and open-ended tasks in QA datasets, as well as close-ended tasks in the VQA dataset, we required the answers to be formatted in JSON, consisting of two key-value pairs: “answer” and “reason”. Due to the context length limitation of ViLT and the capability constraint of Qwen-VL, we were unable to collect answers from them in JSON format and only requested unformatted answers. Subsequently, we converted these unformatted answers into the correct format using GPT-4. For open-ended tasks in the VQA dataset, imposing strict formatting requirements could negatively affect the quality of the answers due to the complexity of the task. Therefore, all models provide answers in an unformatted form. There were four types of unformatted answers: formatting issues, blank responses, unexpected returns, and refuse to answer (Supplementary Table 1), which corresponded to cases where the model did not adhere to the prescribed JSON format, produced blank responses for unknown reasons, exhibited incorrect answer format due to a misunderstanding of the question (e.g., answering open-ended questions in a close-ended format or generating irrelevant contents), and avoided directly answering the question, respectively. We only retested answers with formatting issues and blank responses. Since answers with formatting issues were automatically retested by an error correction program within the pipeline, we did not record their occurrence. The remaining two types of answers reflected the comprehension capability of the model, and thus, no additional processing was performed.

Evaluation metrics

We used accuracy, specificity, precision, recall, and F1 score to evaluate the performance of different LLMs and LVLMs on close-ended questions. For the open-ended questions, we conducted extensive expert evaluation to assess the quality of answers. Three experienced experts manually scored the answers according to custom criteria, which comprehensively evaluated the accuracy, completeness, logic, precision, risk awareness, and conciseness of the answers (Supplementary Tables 3 and 4). The maximum expert evaluation score is set to 10 points. Besides, we employed G-Eval score²⁸ to assess the quality of answers generated by models automatically. G-Eval is a

framework that uses LLM (we used GPT-4 in this study) to evaluate LLM outputs based on custom criteria. The custom evaluation criteria of G-Eval were consistent with those used in expert evaluation (Supplementary Tables 3 and 4). The maximum value of the G-Eval score was set to 1 point. To avoid the rater bias between different experts, we employed the mean Spearman’s rank correlation coefficient³⁶ across all the questions to assess the consistency of experts’ rankings derived from their evaluation scores.

Institutional review board statement

The collection of data for the private VQA dataset (Supplementary Note 3) was approved by the Medical Ethics Committee of Medical Ethics Committee of Tongji Medical College at Huazhong University of Science and Technology. The collection and analysis were conducted in accordance with the Declaration of Helsinki.

Data availability

We release CCBech dataset at Google Drive (https://drive.google.com/drive/folders/1TBIUX74JKIdaU5C3YrTGMO7QJ_t0pVdK) The source data underlying Figs. 1c–g, 3, 4a–e, 5, and 6a–e, are provided as a Source Data File.

Code availability

We release all code used to produce the main experiments. The code can be found at the GitHub repository: <https://github.com/systemoutprintnlhello/world/CCBench>.

Received: 17 January 2025; Accepted: 19 April 2025;

Published online: 23 May 2025

References

- Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **74**, 229–263 (2024).
- Peto, J., Gilham, C., Fletcher, O. & Matthews, F. E. The cervical cancer epidemic that screening has prevented in the UK. *Lancet Lond. Engl.* **364**, 249–256 (2004).
- Zhu, X., et al. Hybrid AI-assistive diagnostic model permits rapid TBS classification of cervical liquid-based thin-layer cell smears. *Nat. Commun.* **12**, 3541 (2021).
- Cheng, S., et al. Robust whole slide image analysis for cervical cancer screening using deep learning. *Nat. Commun.* **12**, 5639 (2021).
- Bai, X. et al. Assessment of Efficacy and Accuracy of Cervical Cytology Screening With Artificial Intelligence Assistive System. *Mod. Pathol.* **37**, 100486 (2024).
- Wang, J., et al. Artificial intelligence enables precision diagnosis of cervical cytology grades and cervical cancer. *Nat. Commun.* **15**, 4369 (2024).
- Jiang, P. et al. A systematic review of deep learning-based cervical cytology screening: from cell identification to whole slide image analysis. *Artif. Intell. Rev.* **56**, 2687–2758 (2023).
- Xue, P., et al. Improving the Accuracy and Efficiency of Abnormal Cervical Squamous Cell Detection With Cytologist-in-the-Loop Artificial Intelligence. *Mod. Pathol. J. U.S. Can. Acad. Pathol. Inc.* **36**, 100186 (2023).
- Cao, M. et al. Patch-to-Sample Reasoning for Cervical Cancer Screening of Whole Slide Image. *IEEE Trans. Artif. Intell.* **5**, 2779–2789 (2024).
- Jiang, H. et al. Deep learning for computational cytology: A survey. *Med. Image Anal.* **84**, 102691 (2023).
- Song, Y. et al. A deep learning based framework for accurate segmentation of cervical cytoplasm and nuclei. in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 2903–2906 <https://doi.org/10.1109/EMBC.2014.6944230> (2014).
- Lin, H. et al. Dual-path network with synergistic grouping loss and evidence driven risk stratification for whole slide cervical image analysis. *Med. Image Anal.* **69**, 101955 (2021).

13. OpenAI et al. GPT-4 Technical Report. Preprint at <https://doi.org/10.48550/arXiv.2303.08774> (2024).
14. An important next step on our AI journey. Google <https://blog.google/technology/ai/bard-google-ai-search-updates/> (2023).
15. Gemini Team et al. Gemini: A Family of Highly Capable Multimodal Models. Preprint at <https://doi.org/10.48550/arXiv.2312.11805> (2024).
16. Meet Claude. <https://www.anthropic.com/claude>.
17. Liu, F. et al. Application of large language models in medicine. *Nat. Rev. Bioeng.* **3**, 103–115 (2025).
18. Safavi-Naini, S. A. A. et al. Vision-Language and Large Language Model Performance in Gastroenterology: GPT, Claude, Llama, Phi, Mistral, Gemma, and Quantized Models. Preprint at <https://doi.org/10.48550/arXiv.2409.00084> (2024).
19. Pal, A. & Sankarasubbu, M. Gemini Goes to Med School: Exploring the Capabilities of Multimodal Large Language Models on Medical Challenge Problems & Hallucinations. *Proceedings of the 6th Clinical Natural Language Processing Workshop* <https://doi.org/10.18653/v1/2024.clinicalnlp-1.3> (2024).
20. Cozzi, A. et al. BI-RADS Category Assignments by GPT-3.5, GPT-4, and Google Bard: A Multilanguage Study. *Radiology* **311**, e232133 (2024).
21. Luo, X. et al. Large language models surpass human experts in predicting neuroscience results. *Nat. Hum. Behav.* 1–11 <https://doi.org/10.1038/s41562-024-02046-9> (2024).
22. Xu, T. et al. Current Status of ChatGPT Use in Medical Education: Potentials, Challenges, and Strategies. *J. Med. Internet Res.* **26**, e57896 (2024).
23. Du, X. et al. Generative Large Language Models in Electronic Health Records for Patient Care Since 2023: A Systematic Review. <https://doi.org/10.1101/2024.08.11.24311828> (2024).
24. Bedi, S. et al. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA*. <https://doi.org/10.1001/jama.2024.21700> (2024).
25. Kung, T. H. et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit. Health* **2**, 1–12 (2023).
26. Hou, W. & Ji, Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat. Methods* **21**, 1462–1465 (2024).
27. Williams, C. Y. K., Miao, B. Y., Kornblith, A. E. & Butte, A. J. Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. *Nat. Commun.* **15**, 8236 (2024).
28. Liu, Y. et al. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. Preprint at <http://arxiv.org/abs/2303.16634> (2023).
29. *The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes*. (Springer International Publishing, Cham, 2015). <https://doi.org/10.1007/978-3-319-11074-5>.
30. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. https://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.
31. Claude 2. <https://www.anthropic.com/news/claude-2>.
32. Bai, J. et al. Qwen Technical Report. Preprint at <https://doi.org/10.48550/arXiv.2309.16609> (2023).
33. ERNIE-4.0-Turbo-8K - ModelBuilder. <https://cloud.baidu.com/doc/WENXINWORKSHOP/s/7lxwwtafi>.
34. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (ACM, Virtual Event Canada, 2021). <https://doi.org/10.1145/3442188.3445922>.
35. Bethesda Interobserver Reproducibility Study-2 (BIRST-2) Bethesda System 2014. *J. Am. Soc. Cytopathol.* **6**, 131–144 (2017).
36. Rotello, C. M., Myers, J. L., Well, A. D. & Jr, R. F. L. in *Research Design and Statistical Analysis* (Taylor & Francis, 2025).

Acknowledgements

This work is supported by the National Natural Science Foundation of China (grant 62471212, 62375100, and 62201221), China Postdoctoral Science Foundation (grant 2021M701320, 2022T150237), and Science and Technology Projects in Guangzhou (grant 2024A04J4960).

Author contributions

S.C., J.C., and X.L. conceived the project. Q.H., S.L., X.Y.L., H.Y., P.Z., W.Y., S. C., X.L., and J.C. constructed the CCBench dataset. Q.H., S.L., and S. C. developed the semi-automatic evaluation pipeline. J.C., L.W., P.Y., D.C., Q.H., and G.R. assessed the performance of the LLM and LVLM models including expert evaluation score and algorithm evaluation metric. Q.H., Q.L., S.L., S.C., X.L., and J.C. analyzed the evaluation results with conceptual advice from S.Z. and Q.F. Q.H., Q.L., S.C., and S.L. wrote the manuscript with input from all authors. S.C. and J.C. supervised the project. All authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41698-025-00916-7>.

Correspondence and requests for materials should be addressed to Xiuli Liu, Jing Cai or Shenghua Cheng.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025