Supplementary Information

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

Qi Hong[1,#], Shijie Liu[2,#], Liying Wu[3,#], Qiqi Lu[1], Pinglan Yang[3], Dingyu Chen[3], Gong Rao[2], Xinyi Liu[1], Hua Ye[1], Peiqi Zhuang[1], Wenxiu Yang[1], Shaoqun Zeng[2], Qianjin Feng[1], Xiuli Liu[2,*], Jing Cai[3,*], Shenghua Cheng[1,*]

[1]School of Biomedical Engineering and Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, China

[2]Britton Chance Center and MoE Key Laboratory for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics-Huazhong University of Science and Technology, Wuhan, China

[3]Department of Obstetrics and Gynecology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

*Correspondence: chengsh2023@smu.edu.cn (S. Cheng), jingcai@hust.edu.cn (J. Cai), xlliu@mail.hust.edu.cn (X. Liu)

Supplementary Notes

Supplementary Note 1 Prompts used for dataset construction

a. Prompts used for QA dataset construction

As a cytopathology researcher, you need to convert the following input text (mostly declarative sentences) into multiple questions and answers.

First, the question transformation phase:

Objective:

Identify sentences describing cytology in the input text to create target statements, typically describing smears, slides, single cells/cell groups, or patients. From these descriptions, generate questions while retaining sentences that imply cytological descriptions. For instance, a sentence like 'Admixture of superficial and intermediate squamous cells' is primarily about the whole smear and thus should be retained. Conversely, sentences serving educational purposes (describing certain cells or metaplastic phenomena) should be omitted.

(Context-Aware Prompting Criteria)

Instructions:

- Omit descriptions related to histopathology: If text exclusively revolves around histopathology, leave the output template entirely blank.
- Retain relevant input sentences: Do not use sentences assessing cell types or smear types as question sources but retain them in the input information.
- Address sentences lacking subjects: Deduce the subject from context and integrate it into the sentence.
- Resolve ambiguous pronouns: For sentences using pronouns like "these" without clear reference, deduce and replace them with appropriate subjects.
- Extract main statements: For sentences beginning with "note" or imperative clauses, focus on extracting the primary detail (e.g., about cell morphology).
- Remove irrelevant follow-up sentences: Omit sentences containing the word "follow-up."

Second, converting target statements into questions:

Initially, convert sentences into general inquiries, e.g., "The inset reveals a characteristic superficial cell at high magnification." becomes "Does the inset reveal a characteristic superficial cell at high magnification?" Subsequently, convert these into specific types of questions (What/When/How much/many/Whose/Which) based on the components mentioned in the sentences.

For example:

"What does the inset reveal at high magnification?"

(Specific Prompting)

When converting text into questions:

- What-type questions: Extract any data or time entity at the start of the sentence and use it as the response's timing information.

- How much/many-type questions: Extract any numerical markers and use them as the quantity information.
- Whose-type questions: Extract possessive pronouns and the subjects they refer to.
- Which-type questions: Focus on phrases starting with prepositions indicating location.
- What-type questions: Extract phrases or objects and use them as the subjects of the response.

(Few-shot Examples)

Type	Original Sentence	Question
What	The end of the long bone is expanded in the region of epiphysis.	What is expanded in the region of epiphysis?
Where	The left ventricle is on the lower right in this apical view.	Where is the left ventricle in this apical view?
When	After 1 year of abstinence, most scars are gone.	When are most scars gone?
How many	Two multi-faceted gallstones are present in the lumen.	How many gallstones are present in the lumen?
Whose	The tumor cells and their nuclei are fairly uniform.	Whose nuclei are fairly uniform?
How	The trabecular bone shows traceable osteoblastic activity.	How does the trabecular bone show traceability?

Finally, generate answer and reason.

(Output Format)

Generated Sentence Format:

Simplified Sentences (English):

1.

2.

...

(Replace with a newline)

Questions (English, preferably no more than three pairs of questions/answers per input text, the generated question types should be yes/no general questions or What-type questions if possible):

1.

...

(Replace with a newline)

Answers (English):

1.

...

(Replace with a newline)

Issues to Report (Chinese, if any, answer in points):

1.

...

(Replace with a newline)

Reason for Generating Sentences (Explain each question-answer pair, in points, Chinese):
1.
...
(Replace with a newline)

Notes:
1. If the information is insufficient, do not say anything else. Leave simplified sentences, questions, answers, doubts, and reasons empty. Leave newlines as required; otherwise, the program will read incorrectly.
2. There should be a new line between each paragraph and a new line between each number in the paragraphs.
3. If there are no issues to report, leave it empty, do not write "None."
4. Since each split/simplified sentence will be used as an independent question for posing, if there are pronouns or transition words like "they," "these," "however," and "while" without context or clear antecedent, abandon that question. If this input text can only pose that question, abandon the sentence.
5. Delete the parenthetical instructions in the output format.

(Reference Materials)
knowledge point: {user input} <EOS>

b. Prompts used for VQA dataset

(Role-play Prompting)
As a cell pathology research specialist, you are tasked with transforming the text provided (mostly declarative sentences) into several questions and answers. These will be used to test the capabilities of a multimodal model in the field of cell pathology through VQA (Visual Question Answering) assessments. Please adhere strictly to the instructions provided in Section 4 without adding additional elements.

Step One: the question transformation phase
Objective
Identify sentences describing cytology in the input to create target statements, typically describing smears, slides, single cells/cell types, or patients. From these descriptions, generate questions while retaining sentences that imply cytological descriptions. For instance, a sentence like "A mixture of superficial and intermediate squamous epithelial cells is primarily about the whole smear and thus should be retained." Conversely, sentences serving educational purposes (describing certain cells or metaphasic phenomena) should not be included.

Instructions
(Context-Aware Prompting Criteria)
1. Omit descriptions unrelated to histopathology; those are not relevant to cytopathology. If all text revolves around this, leave the output template entirely blank.

2. Do not use sentences assigning cell types or smear types as question sources, but retain them as input references.
3. For sentences lacking a subject, deduce the subject from the context and insert it into the sentence.
4. For sentences using pronouns like “these” without a clear reference, apply the same rule as above. Deduce and replace the pronoun based on the previous context.
5. For sentences beginning with “note” or imperative sentences, extract the main statement. If it only elaborates on cell morphology without descriptive details, default to omitting the statement.
6. Remove sentences containing the phrase “this explains.”

Second: converting target statements into questions

- Convert the sentence to a general interrogative sentence.
  For example: The inset reveals a characteristic superficial cell at high magnification. Does the inset reveal a characteristic superficial cell at high magnification?
- Then convert based on the components into specific interrogative sentences like Who/What/How many/Whose/Which, for example:
  What does the inset reveal at high magnification?

(Specific Rules)

- When-Type Questions: Extract date and time details from the sentence and use them as the answer’s time information.
- Where-Type Questions: Extract time periods from phrases like “in/around the point of” and use them as the answer’s time information.
- How Many/How-Type Questions: Extract numeric markers and use them as the answer’s quantity information.
- Whose-Type Questions: Extract the subject referred to by possessive pronouns and use them as the answer’s owner information.
- Which-Type Questions: Extract location information from phrases like “in the inset/where” and use them as the answer’s location information.
- What-Type Questions: Extract main objects and use them as the answer’s object or content information.

(Few-shot Examples)

Type	Original Sentence	Question
What	The inset reveals a characteristic superficial cell.	What does the inset reveal?
Where	Abnormal cells are present in the apical view.	Where are the abnormal cells located?
When	After staining, abnormal keratinocytes are seen.	When are abnormal keratinocytes observed?
How many	Three non-keratinized spindle cells were identified.	How many spindle cells were identified?
Whose	The spindle cell morphology was noted for its uniformity.	Whose morphology was noted for uniformity?
Which	In the image, the epithelial layers are disrupted.	Which layers are disrupted in the image?

(Output Format)
Questions: Write complete questions in English using various types, ideally starting with "what," "how," etc.
Answers: Provide specific responses in English or "Yes/No" where applicable.
Points of Inquiry: Include any applicable inquiry details from the input.
Rationale for Generated Sentences: Explain step-by-step how each transformation from text to question occurred.

Notes:
1. If required, write "Not for the citations, pieces of inquiry/questions, if there is a need for a simpler template here."
2. If the input text allows free-formulation of a question about "then," "however," "when," and other connective discourse sequences, questions to templates must prioritize this. Skip template-valid datasets entirely.
3. Since I cannot send pictures, the absence of an accompanying image is not a criterion for abandoning the question transformation.
4. Remove any bracketed text as per the output rules.

(Reference Materials)
Fig description: {user input} <EOS>

Supplementary Note 2 The system prompts used at the "request for answer" stage of LLMs and LVLMs evaluation

a. The system prompt used at the "request for answer" stage of LLMs evaluation

<BOS> You are a cytopathologist currently working with the cytopathologic notion of cervical cancer and need to develop standardized answers to the questions below and explain the reason.

- **Rule**: {rule}
- **Example**: {example}
- **Response Template**: {response-template}
- **Knowledge**: {knowledge}
- **Question**: {question}
- **Response**: {response}<EOS>

b. The system prompt used at the "request for answer" stage of LVLMs evaluation

<BOS> You are a cytopathologist currently working with images of cervical cancer cytopathology and need to develop a standard answer to these images and the questions about them.

- **Rule**: {rule}
- **Example**: {example}
- **Response Template**: {response-template}
- **Knowledge**: {knowledge}

- **Image**: {image-token}
- **Question**: {question}
- **Response**: {response}<EOS>

c. Rules

1) QA task

Please use your common sense to make a judgment. Answer the question with a conclusion that is supported by the knowledge you have, rather than giving a result based on the tendency of the question, e.g., is it impossible for a human to turn into another animal? This question may sound too absolute, but it is correct from the current point of view. Cell location, if not specified, is usually in the cervical portion of the uterus.

Whether the user input is a question or a statement, it needs to be answered as a question. Although the question is generally in the form of "Does cell A have characteristic B?" it is possible that the error is not only that the cells in the diagram do not have this feature, but also that there are no cells of type A in the diagram at all, so you need to be careful what you look for.

2) VQA task

When analyzing images, treat each image (large or small) independently as per the instruction's emphasis. Ignore letters like A, B, and arrows that are meant for sequencing images and not relevant to cytopathological analysis. Unless otherwise specified, cells are generally located in the cervical area. Whether the user's input is a question or a statement, it should be treated as a question and answered accordingly.
The image may contain arrows or other annotation symbols. Check if there are any instructions regarding these symbols in the question; if not, ignore these symbols and do not mistake them for cellular structures. For questions involving area, length, and other physical dimensions, it is impossible to provide magnification parameters. Therefore, the estimate is based on the normal size fluctuations of the cells or organelles indicated in the question.

Each question is related to the input image; the answer is based on the image and does not rely on general knowledge about the characteristics of this type of cell. Although the question format is generally "Does cell A have feature B?" the error might not only be that the cell in the image lacks this feature. It is also possible that cell A is not present in the image at all, so careful observation is needed. Terms like "in this image," "on this (conventional/liquid-based) smear," "on this cell," "on this slide," "in this sample," "on this background," or "in this example" are all describing the input image.

d. Response templates and question-answer examples

	QA	VQA
Close	**Response template:** {reason: “”, answer: “”} **Example:** reason: The background is white, not black answer: Yes	**Response template:** {reason: “”, answer: “”} **Example:** reason: The background is white, not black answer: Yes
Open	**Response template:** {reason: “”, answer: “”} **Example:** reason: Explain your thought process and reason for your judgment. answer: Freely answer the question based on your knowledge of cervical cancer cytopathology images but remember you can explain the details of the reason and keep the answer as simple and short as you can.	**Response template:** {reason: “”, answer: “”} **Example:** reason: Based on the image, the cytoplasm is close to the nucleus answer: Approximately 1

e. knowledge

Common Guidance

If the question involves the subclass or classification of cells, please refer to The Bethesda System for Reporting Cervical Cytology: Definitions, Third Edition, for classification. The class types are Pos and Neg, and the following table is a reference for the common types of subclasses.

Classification	Description
Normal (NILM)	Negative (Neg). Normal cervical cell morphology, no abnormal cells.
Atypical Squamous Cells (ASC)	Positive (Pos). Mild abnormalities in cervical cell morphology, but not definitively diagnosed as LSIL or HSIL.
Atypical Squamous Cells of Uncertain Significance (ASC-US)	Positive (Pos). Mild abnormalities in cervical cell morphology, but LSIL or HSIL cannot be excluded.
Atypical Squamous Cells not excluding high-grade squamous intraepithelial lesions (ASC-H)	Positive (Pos). Moderate abnormalities in cervical cell morphology; possibly HSIL but not definitively diagnostic.
Low-grade squamous intraepithelial lesion (LSIL)	Positive (Pos). Moderate abnormalities in cervical cell morphology, but no involvement of the basement membrane.
High-grade squamous intraepithelial lesion (HSIL)	Positive (Pos). Severe abnormalities in cervical cell morphology and involvement of the basement membrane.
Squamous cell carcinoma (SCC)	Positive (Pos). Malignant cervical cell morphology, invasive.
Adenocarcinoma (AGC)	Positive (Pos). Cervical glandular epithelial cells are cancerous and invasive.

Fig description: {user input} <EOS> (Role-play Prompting)

Supplementary Note 3 Performance on private clinical dataset

To further assess the capability of LVLMs in real-world clinical scenarios and mitigate potential data contamination concerns from the TBS textbook, we constructed a private

VQA dataset using clinical pathological smear images collected from hospitals, along with their cytomorphological descriptions and diagnostic interpretations. This private dataset comprised 198 cervical cytology images with an equal distribution across three diagnostic categories: ASCUS, HSIL, and LSIL. We created 99 close-ended and 99 open-ended question-answer pairs using the semi-automatic pipeline described earlier (Fig. 1b). For close-ended tasks, we balanced the distribution of "yes" and "no" answers, while for open-ended tasks, we generated cytomorphological feature descriptions and TBS subclass classification questions. All questions and answers were manually reviewed by experienced cytopathologists to ensure clinical accuracy.

For model testing on this private dataset, we made minimal adjustments to accommodate the deprecation of some previously tested models. We used GPT-4o-2024-05-13 (replacing GPT-4V) and Gemini 1.5 Pro (replacing Gemini pro vision), selecting versions closest to the original models in release timing and performance. ViLT, LLaVA, and Qwen-VL-Max maintained identical configurations to those used in the CCBench evaluation. We disabled the network access of all internet-enabled model APIs during the evaluation and kept the hyperparameters at their default settings to reflect the actual capabilities of each model.

We evaluated the performance of four commercial LVLMs and one open-source LVLM on this private dataset (Supplementary Fig. 1). For the close-ended questions, Gemini achieved the highest accuracy (0.616), followed closely by GPT-4o (0.606) and Qwen-VL (0.596), while LLaVA (0.485) performed below the random baseline (0.515). For the open-ended questions, we employed the G-Eval methodology described earlier to assess answer quality. Qwen-VL obtained the highest mean G-Eval score (0.579), followed by GPT-4o (0.531) and Gemini (0.443), while LLaVA obtained the lowest score (0.230). These results on the private clinical dataset largely aligned with our findings on the CCBench dataset (Fig. 3c and Fig. 6e), confirming the robustness of our evaluation methodology and the relative capabilities of the tested models in real-world clinical applications.

Supplementary Tables

Supplementary Table 1 The validity of the responses from different models.

	Model	Formatted Answer	Unformatted Answer		
			Blank	Unexpected return	Refuse to answer
LLM	**Bard**	415		3	
	Claude-2.0	284		134	
	GPT-4	418			
	LLaMa-2	403	4	11	
	Qwen-Max	418			
	ERNIE-Bot-4.0	268		141	9
LVLM	**Gemini**	292	2	13	
	GPT-4V	268		33	6
	LLaVA-1.5	260		47	
	ViLT	284	20	3	
	Qwen-VL-max	259		47	1

Supplementary Table 2 Overview of LLMs and LVLMs.

	Model	Max Input Token	Max Output Token	Version
LLM	GPT-4	128k	4096	gpt-4-0125-preview
	Bard	8196	1024	chat-bison@002[2]
	Claude-2.0	100K	100K	claude-2.0
	Qwen-Max	8K	8K	qwen-max-0428
	ERNIE-Bot-4.0	5K	2K	ERNIE-4.0-Turbo-8K
	LLaMa-2	32K	32K	llama-2-13b-chat
LVLM	Gemini	16384	2048	Gemini 1.0 Pro Vision
	GPT-4V	128K	8K	gpt-4-1106-vision-preview
	Qwen-VL	8K	8K	Qwen-VL- Max
	ViLT	40	40	vilt-b32-finetuned-vqa
	LLaVA	4096	4096	llava-v1.5-vicuna-13b

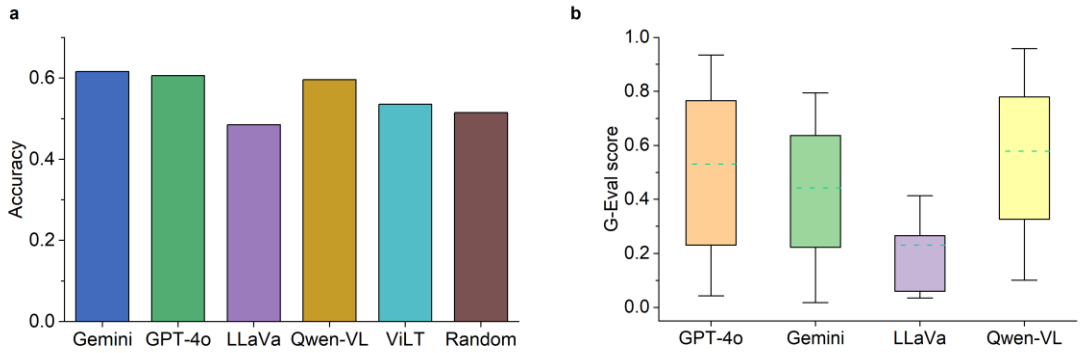
Supplementary Table 3 Open-QA evaluation standard.

Scoring Item	Description	Scoring Details
Accuracy and Information Accuracy	Does the model's response match the standard answer or medical facts, ensuring accuracy?	• Compared with the answer, the total score is 70 points. • If the correct answer can be divided into points, start from a base score of 10 points. All points equally share the remaining 60 points, rounding up. • If the answer is short, try to subdivide it into multiple points for scoring.
Completeness, Detail, and Explanation	Does the model provide all relevant information and explain the reasons for the answer?	• Judged based on personal knowledge, the total score is 6 points. • Incorrect explanations do not count. • Logical explanations count, merely explaining terms do not count.
Logic Reasoning	Is the reasoning process logical and coherent?	• Read each model's answer to judge, the total score is 6 points. • Each logical discrepancy deducts 2 points.
Precision and Use of Medical Terminology	Does the model's response use precise medical terminology, avoiding ambiguity?	• Judged based on personal knowledge, the total score is 6 points.
Risk Awareness	Awareness of any potential risks or uncertainties in the provided information.	• Read each model's answer to judge, and a total score of 6 points.
Conciseness and Efficiency	The brevity of the model's reasons, avoiding unnecessary details while maintaining comprehensiveness.	• Compare each model's reasons to judge, with a total score of 6 points. • Compare the amount of irrelevant content in each model's answer. • Compare the length of each sentence; the shorter, the better.

Supplementary Table 4 Open-VQA evaluation standard.

Scoring Item	Description	Scoring Details
Accuracy and Information Accuracy	Does the model's response match the standard answer or medical facts, ensuring accuracy?	• Compared with the answer, the total score is 70 points. • If the correct answer can be divided into points, start from a base score of 10 points. All points equally share the remaining 60 points, rounding up. • If the answer is short, try to subdivide it into multiple points for scoring.
Completeness, Detail, and Explanation	Does the model provide all relevant information and explain the reasons for the answer? Does the explanation effectively incorporate the information provided in the image?	• Judged based on personal knowledge, the total score is 6 points. • Incorrect explanations do not count. • Logical explanations count, merely explaining terms do not count.
Logic Reasoning	Is the reasoning process logical and coherent?	• Read each model's answer to judge, the total score is 6 points. • Each logical discrepancy deducts 2 points.
Precision and Use of Medical Terminology	Does the model's response use precise medical terminology, avoiding ambiguity?	• Judged based on personal knowledge, the total score is 6 points.
Risk Awareness	Awareness of any potential risks or uncertainties in the provided information.	• Read each model's answer to judge, and a total score of 6 points.
Conciseness and Efficiency	The brevity of the model's reasons, avoiding unnecessary details while maintaining comprehensiveness.	• Compare each model's reasons to judge, with a total score of 6 points. • Compare the amount of irrelevant content in each model's answer. • Compare the length of each sentence; the shorter, the better.

Supplementary Fig. 1



Supplementary Fig. 1 Performance of different LVLMs on the private clinical dataset. a Accuracy of different LVLMs on close-ended questions. **b** Distribution of G-Eval scores for different LVLMs on open-ended questions. The data are presented as boxplots and whiskers (min to max), with the upper and lower hinges representing the 25th and 75th percentiles, respectively. The dashed line denotes the average score.