

# A Foundational Multimodal Vision Language AI Assistant for Human Pathology

Ming Y. Lu<sup>1,2,3,4,6‡</sup>, Bowen Chen<sup>1,2‡</sup>, Drew F. K. Williamson<sup>1,2,3‡</sup>, Richard J. Chen<sup>1,2,3,4,5</sup>, Kenji Ikamura<sup>1,2</sup>, Georg Gerber<sup>1</sup>, Ivy Liang<sup>1,7</sup>, Long Phi Le<sup>2</sup>, Tong Ding<sup>1</sup>, Anil V Parwani<sup>8</sup>, Faisal Mahmood<sup>\*1,2,3,4,9</sup>

<sup>1</sup>*Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA*

<sup>2</sup>*Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA*

<sup>3</sup>*Cancer Program, Broad Institute of Harvard and MIT, Cambridge, MA*

<sup>4</sup>*Cancer Data Science Program, Dana-Farber Cancer Institute, Boston, MA*

<sup>5</sup>*Department of Biomedical Informatics, Harvard Medical School, Boston, MA*

<sup>6</sup>*Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA*

<sup>7</sup>*Harvard John A. Paulson School of Engineering And Applied Sciences, Harvard University, Cambridge, MA*

<sup>8</sup>*Department of Pathology, Wexner Medical Center, Ohio State University, Columbus, OH*

<sup>9</sup>*Harvard Data Science Initiative, Harvard University, Cambridge, MA*

‡ Contributed Equally

\*Corresponding author: Faisal Mahmood ([faisalmahmood@bwh.harvard.edu](mailto:faisalmahmood@bwh.harvard.edu))

## Abstract

The field of computational pathology has witnessed remarkable progress in the development of both task-specific predictive models and task-agnostic self-supervised vision encoders. However, despite the explosive growth of generative artificial intelligence (AI), there has been limited study on building general purpose, multimodal AI assistants tailored to pathology. Here we present PathChat, a vision-language generalist AI assistant for human pathology using an in-house developed foundational vision encoder pretrained on 100 million histology images from over 100,000 patient cases and 1.18 million pathology image-caption pairs. The vision encoder is then combined with a pretrained large language model and the whole system is finetuned on over 250,000 diverse disease agnostic visual language instructions. We compare PathChat against several multimodal vision language AI assistants as well as GPT4V, which powers the commercially available multimodal general purpose AI assistant ChatGPT-4. When relevant clinical context is provided with the histology image, PathChat achieved a diagnostic accuracy of 87% on multiple-choice questions based on publicly available cases of diverse tissue origins and disease models. Additionally, using open-ended questions and human expert evaluation, we found that overall PathChat produced more accurate and pathologist-preferable responses to diverse queries related to pathology. As an interactive and general vision language AI assistant that can flexibly handle both visual and natural language inputs, PathChat can potentially find impactful applications in pathology education, research, and human-in-the-loop clinical decision making.

**Video demo:** [Dropbox link](#), [Google Drive link](#)

## Introduction

The field of computational pathology has witnessed a remarkable transformation in recent years, propelled by the convergence of several key trends including increased availability and institutional adoption of digital slide scanning, rapid progress in artificial intelligence (AI) research, increasing accessibility of large datasets, and substantial high-performance computing resources<sup>1–9</sup>. With varying degrees of success, researchers have leveraged deep learning to address a diverse range of tasks, including cancer subtyping<sup>10,11</sup> and grading<sup>12,13</sup>, metastasis detection<sup>14–16</sup>, survival<sup>17–24</sup> and response-to-treatment prediction<sup>25–27</sup>, tumor site of origin prediction<sup>28,29</sup>, image search<sup>30–33</sup>, mutation prediction and biomarker screening<sup>34–37</sup>, and more<sup>38–42</sup>. At the same time, general purpose vision encoder models<sup>43–48</sup>, which are trained on vast datasets of unlabeled histopathology images and can serve as versatile task-agnostic model backbones<sup>49–56</sup>, are paving the way for further improvements across many tasks in computational pathology, both in performance and label efficiency.

However, the aforementioned developments in computational pathology do not yet reflect the important roles of natural language in pathology, as a key to unlocking rich, diverse sources of accumulated human medical knowledge, as a potential signal for model supervision, and as a unified medium for facilitating intuitive interaction between powerful AI models and end users. Notably, in general machine learning, representative works demonstrated that large-scale vision language representation learning can augment vision-only AI models with new capabilities including zero-shot image recognition<sup>57,58</sup> and text-to-image retrieval<sup>59–61</sup>. Depending on the architectural design and training data and objectives, visual language pretrained systems<sup>62–69</sup> can also often be finetuned for tailored tasks ranging from visual question answering and image captioning to object detection and semantic segmentation. In medical imaging and computational pathology, researchers have recently begun to harness diverse sources<sup>70–77</sup> of paired biomedical images and captions or reports for visual language pretraining<sup>75,78</sup>, including the development of CLIP<sup>57</sup>-like models tailored for specific domains such as pathology<sup>74,77,79,80</sup> and radiology<sup>72,81–84</sup>. In computational pathology, a few works have shown promising zero-shot performance in select diagnostic and retrieval tasks<sup>74,77,79</sup>, while other works also experimented with designing specialized models for biomedical visual question answering or captioning<sup>85–89</sup>. However, these models are not yet ready to serve as interactive assistants for pathologists, researchers using pathology image data, or pathology trainees. This is because those models such as CLIP which have been trained using pathology data are not also trained to understand and follow diverse and often complex instructions of users and have limited to no ability to generate coherent responses in natural language grounded in factual knowledge.

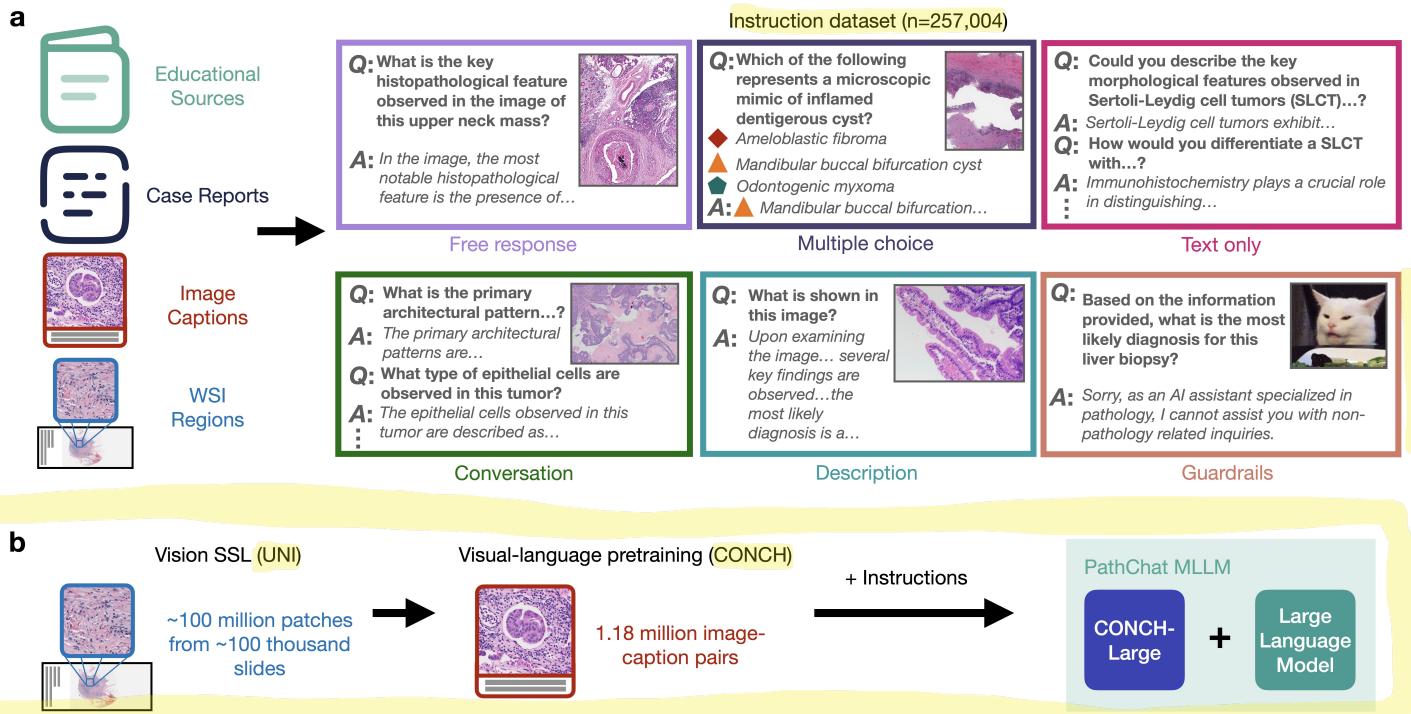
Following the rise of large language models<sup>90–103</sup> (LLMs), rapid advances in multimodal large language models<sup>104–106</sup> (MLLMs) and the broader field of generative AI<sup>107</sup> are poised to open a new frontier for computational pathology, one which emphasizes natural language and human interaction as key components of AI model design and user experience, in addition to powerful visual processing capabilities. General pur-

pose multimodal generative AI products, notably ChatGPT-4, are maturing, and have demonstrated impressive capabilities on a wide range of routine, creative, as well as professional use cases<sup>108</sup>, including coding, writing, summarization, data analysis, question answering, translation, and even image generation, all while being accessible through an intuitive and interactive user interface. While there have been attempts to investigate their performance on answering medicine-related queries<sup>109–116</sup>, their capability to assist professionals and researchers in the highly specialized but important sub-field of anatomic pathology remains relatively unexplored<sup>111, 117–120</sup>. However, the potential applications of an interactive, multimodal AI assistant for pathology are immense. The ability to understand and respond to complex queries in natural language in theory could enable such an assistant to serve as a helpful companion across various stages of human-in-the-loop clinical decision making, education, and research. For instance, in the clinic, the AI assistant might be able to ingest a histopathology image, provide an initial assessment of the morphological appearance, and identify potential features of malignancy. Subsequently, the pathologist or trainee could provide additional context about the underlying case, such as clinical parameters of the patient and the tissue site, and ask the model to suggest a differential diagnosis. If deemed reasonable, the user could then request helpful suggestions for ancillary testing and immunohistochemical stains to narrow down the differential. Finally, the results of such tests could also be provided to the model to make a final deduction to arrive at the diagnosis. In research, a multimodal AI model that can summarize the morphological features of large cohorts of histopathology images would potentially enable automated quantification and interpretation of morphological markers in large data cohorts. In medical education, an accurate, on-demand interactive AI companion could help democratize access to expert-level guidance and training in pathology, thereby narrowing the gap between regional disparities in healthcare provision.

### Vision language AI assistant for human pathology

In this article, we develop a vision language interactive AI assistant for human pathology powered by a custom, finetuned **multimodal large language model (MLLM)**. To build an MLLM-based vision language AI assistant that can reason over both visual and natural language inputs, we begin with **UNI**<sup>49</sup>, a state-of-the-art (SOTA) vision-only foundational encoder model pretrained on over 100 million histology images from over 100 thousand slides using self-supervised learning. We perform further vision language pretraining on the UNI encoder with 1.18 million pathology image caption pairs to align its image representation space with that of pathology text<sup>79</sup>. The resulting vision encoder, **CONCH-Large**, is subsequently connected to a **13 billion parameter**, pre-trained LLM<sup>98</sup> via the multimodal projector module to form the complete MLLM architecture. The MLLM is finally finetuned via a curated dataset of over 250 thousand instructions to build PathChat (**Figure 1**), a visual language AI assistant that can understand pathology images and text and respond to complex pathology-related queries. More details about data curation and model training can be found in PathChat dataset curation and PathChat model design and training section of Methods respectively.

(Llana 2)



**Figure 1: Instruction-following dataset curation and PathChat overview.** **a.** We curated the currently largest instruction finetuning dataset specialized for the domain of pathology, consisting of 257k instructions and corresponding responses covering varied formats (e.g. multi-turn conversations, multiple-choice questions, short answers; see **Extended Data Figure 1** for complete examples) from diverse sources. **b.** To build an MLLM-based vision language AI assistant that can reason over visual and natural language inputs, we begin with a SOTA vision-only self-supervised pretrained foundation encoder model, UNI, and perform further vision language pretraining analogous to CONCH. The resulting vision encoder, CONCH-Large, is subsequently connected to a 13 billion parameter, pretrained LLM via a multimodal projector module (not shown) to form the complete MLLM architecture. The MLLM is finetuned via the curated instruction-following dataset to build PathChat, a visual language AI assistant specialized for human pathology. More details about data curation and model training can be found in **PathChat dataset curation** and **PathChat model design and training** section of **Methods** respectively.

We demonstrate the capabilities of PathChat in various applications including analysis of pathology cases from diverse organ sites and practices (**Figures 2 and 3**). Additionally, we contribute, to the best of our knowledge, the first high quality open-ended pathology visual question answering benchmark curated with expert supervision (see the **PathQABench: an expert-curated pathology question-answering benchmark** section of **Methods** for more details), making it suitable for evaluating the performance of multimodal large language models in pathology and fulfilling a critical need amidst rapid advances of visual language multimodal AI research. We evaluate our custom built vision language AI assistant for pathology, named PathChat, against both LLaVA<sup>105</sup>, a state-of-the-art (SOTA) general domain open-source MLLM as well as LLaVA-Med<sup>118</sup>, which has been tailored to the biomedical domain. We also compare against a SOTA commercial solution, ChatGPT-4 (powered by GPT4V), despite our model being significantly smaller and cheaper to serve. More details about model design and training can be found in the **PathChat model design and training** section of **Methods** with hyperparameters described in **Extended Data Tables 1-3**.

# Results

## Performance on multiple-choice diagnostic questions

We began by assessing the capability of our PathChat MLLM to directly make a diagnosis from histology images. For this purpose, a board-certified pathologist manually selected salient regions of interest (ROIs) from routine H&E diagnostic whole slide images (WSIs) from both the TCGA and our in-house pathology archive (both of which are completely withheld from model pretraining or finetuning). In total, the questions cover 29 diagnoses from 9 different major pathology practices and organ sites (**Extended Data Tables 4-5**). For each organ system, the pathologist selected a set of 10 possible answers that encompasses the correct answers for all questions within that organ system as well as other relatively common diagnoses within that organ system (**Extended Data Table 6**). For each question, we consider two evaluation strategies. In the first, the image-only setting, the model is presented with only the image and the multiple choice question as input. In the second, the image with clinical context setting, designed to more closely mimic a real-world diagnostic workflow, additional relevant clinical context is provided together with the histology image, including information that may include patient age, sex, clinical history, and radiology findings as appropriate for the clinical case. In both settings, the model is assessed for its ability to accurately select the ground truth diagnosis from the set of possible options. We provide an illustrative example of the complete model input in **Figure 2a**. On all cases (denoted as “Combined” in **Figure 2b**), we compare PathChat against LLaVA 1.5, a SOTA general purpose visual language chatbot assistant, as well as LLaVA-Med, a specialized version of LLaVA finetuned for answering biomedical related queries. Additionally, on the subset of 23 cases derived from publicly available WSIs (denoted as “PathQABench-Public”), in addition to LLaVA 1.5 and LLaVA-Med, we also compare against GPT4V, which powers ChatGPT4, the current best in class, vision capability enabled commercial AI assistant developed by OpenAI. All models were evaluated “as is” without any additional task-specific finetuning, consistent with the paradigm of zero-shot transfer.

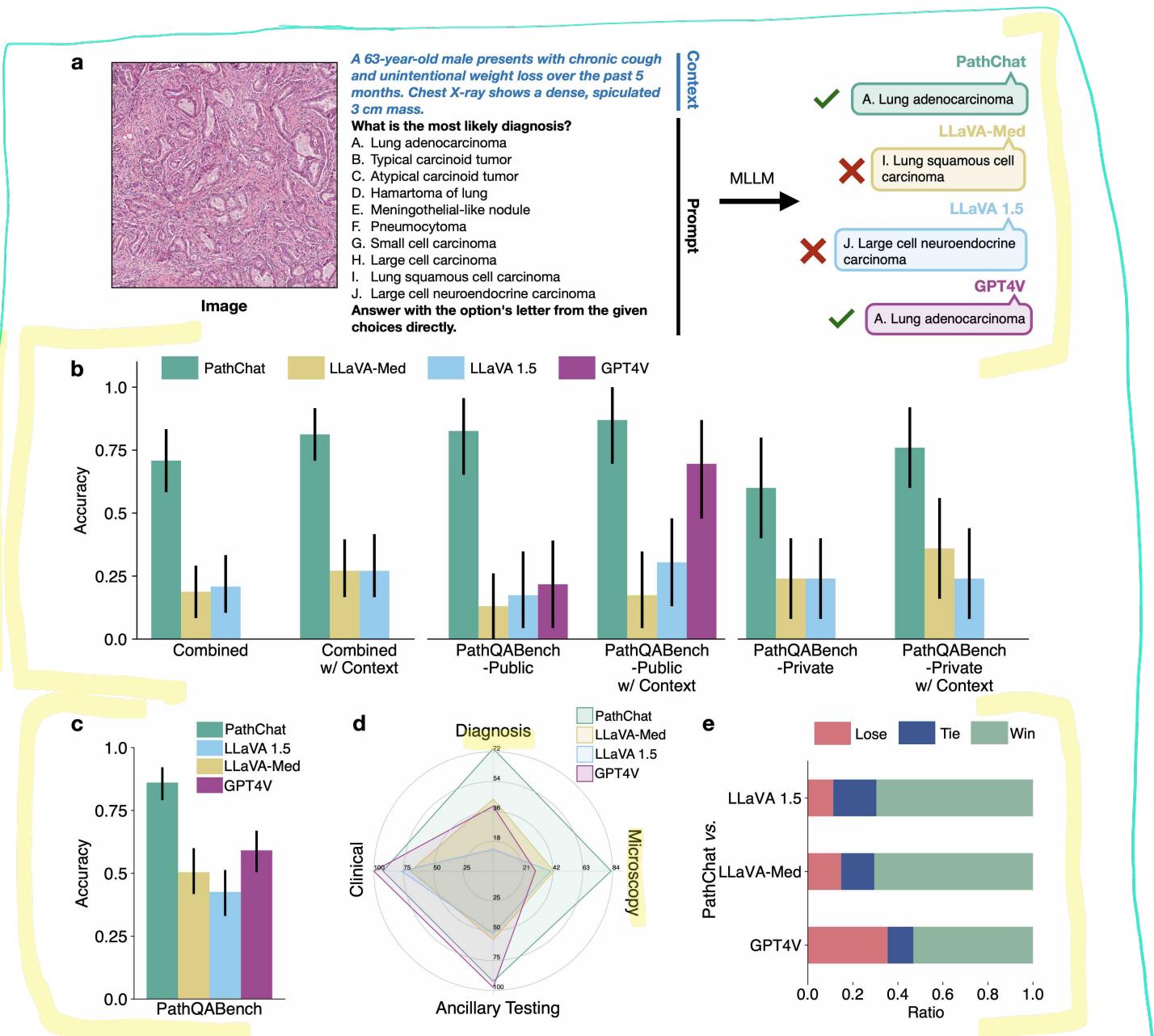
In both evaluation settings (image only and image with clinical context), PathChat convincingly outperforms the open-source baselines LLaVA 1.5 and LLaVA-Med in terms of diagnostic accuracy (**Figure 2b, Extended Data Tables 7-9**). In the image only evaluation setting, PathChat scored a 70.8% accuracy (+50% vs. LLaVA 1.5 and +52.1% vs. LLaVA-Med) on the full combined benchmark. In line with expectation, the performance of PathChat further improves to 81.2% accuracy (+54.1% vs. LLaVA 1.5 and LLaVA-Med) when additional useful clinical context is provided. Specifically, we note that the addition of clinical context consistently improves the accuracy of PathChat both on the private in-house cases (PathQABench-Private, +16%) and the public TCGA cases (PathQABench-Public, +4.4%). These findings suggest that PathChat can effectively and flexibly leverage multimodal information for more accurate diagnosis of histology images by simply providing such additional non-visual information in plain natural language without specialized data processing.

Additionally, using PathQABench-Public, which only contains cases from the publicly available TCGA WSIs, we also compare our model against the GPT-4Vision (GPT4V) model. We observe that our domain-specific PathChat MLLM achieves higher diagnostic accuracy compared to GPT4V in both evaluation settings on the cases tested, although it is much more accurate than the open-source MLLMs tested, especially when clinical context is provided. Additionally, it is worth noting that guardrails appear to have been implemented into GPT4V to sometimes prevent it from addressing queries that require examination of medical images, and it will instead inform the user that it cannot provide a pathology interpretation and instead to consult a medical professional. In such cases, we make a maximum of 2 additional submissions for the same query for a total of up to 3 attempts (see **Evaluation of GPT4V** of the **Methods** section for more details). Following this evaluation protocol, we successfully queried GPT4V for all 23 PathQABench-Public images when clinical context is included, but only 12 out of 23 questions for the image only setting. An ultimately unsuccessful query was treated as incorrect since the response did not address the question. However, we also report performance on only the subset of questions that GPT4V successfully answered (**Extended Data Table 10**). (only 23?)

### Performance on open-ended question answering

Beyond multiple-choice diagnostic questions, it is valuable to assess the ability of PathChat and other MLLMs to generate coherent, reasonable, and clinically relevant responses to open-ended pathology-related inquiries (see **PathQABench: an expert-curated pathology question-answering benchmark** section of **Methods**). Based on cases from PathQABench-Public, a board-certified anatomic pathologist carefully open-ended questions targeting a broad spectrum of topics including microscopic image description, histologic grade and differentiation status, risk factors, prognosis, treatment, diagnosis, immunohistochemical (IHC) tests, molecular alterations, and other further testing. Similar to the multiple-choice evaluation, to mimic the real-world use case of a pathology AI assistant, each question is provided to models as is, without requiring any additional model or task-specific finetuning. The responses are then evaluated by an expert pathologist both in terms of accuracy (i.e. a binary label of correct vs. incorrect), as well as ranking (from best to worst, ties allowed) based on their relevance to the question, correctness, and whether it is supplemented with correct explanation or reasoning in a succinct manner (see **MLLM evaluation** section of **Methods** for more details and **Extended Data Figures 2-4** for illustrative examples of ranked model responses). Throughout the ranking process, the pathologist, who has not had prior interaction with each model, is also blinded to which model produced which response and the responses were additionally randomly shuffled for each question to further minimize potential bias towards specific models.

Overall, we find that PathChat produces on average more accurate as well as more preferable, higher ranked responses than all other MLLMs tested. Specifically, PathChat scored an overall accuracy of 86.1% on the (**Figure 2c**, **Extended Data Table 11**), which corresponds to a 27% improvement compared to the accuracy score of 59.1% achieved by the runner up, GPT4V. Compared to the publicly available general purpose MLLM



**Figure 2: Multiple choice evaluation of PathChat.** **a.** Illustrative example of a multiple-choice style diagnostic question. The input always includes a salient histology image ROI selected by a board-certified anatomic pathologist and the instruction to select the most likely diagnosis from a set of possible choices. In the **image + clinical context evaluation setting** that is designed to more closely mimic a real-world diagnostic workflow, additional relevant clinical context (designed by the pathologist, shown in blue) is provided together with the histology image and concatenated with the original instruction. **b.** Accuracy of MLLMs on multiple choice-style diagnostic questions. Note that we only compare against GPT4V on questions based on publicly available cases (PathQABench-Public). **c.** Accuracy of MLLMs on **open-ended** questions. **b, c.** Error bars represent 95% confidence intervals. **d.** Accuracy on different categories of questions. **e.** Head-to-head records on open-ended questions for PathChat *v.s.* other MLLMs. Lose: said model is ranked higher than PathChat; Tie: PathChat is tied with the model in ranking; Win: PathChat is ranked higher than the model.

LLaVA 1.5 (accuracy of 42.6%), and the biomedicine-specialized MLLM LLaVA-Med (accuracy of 50.4%), the margin of improvement is even more substantial, at +43.5% and +35.7% respectively.

When considering head-to-head records (*e.g.* PathChat *vs.* GPT4V) for model ranking judged by a human expert, a “win” for PathChat on a question equates to PathChat’s response being ranked strictly higher than its counterpart when considering their relative ranking. Similarly, a “tie” for PathChat means the two models received the same rank while a “lose” means PathChat was ranked strictly lower. Against the runner up GPT4V, PathChat had a favorable win rate of 57.4% compared to a lose rate of just 29.6% and a tie in the remaining 13% of questions (**Figure 2e**, **Extended Data Table 12**). Once again, we observe a even larger performance gap in favor of PathChat as compared to LLaVA 1.5 (win rate of 69.6%, lose rate of 11.3%, and tie rate of 19.1%) and LLaVA-Med (win rate of 70.4%, lose rate of 14.8%, and tie rate of 14.8%).

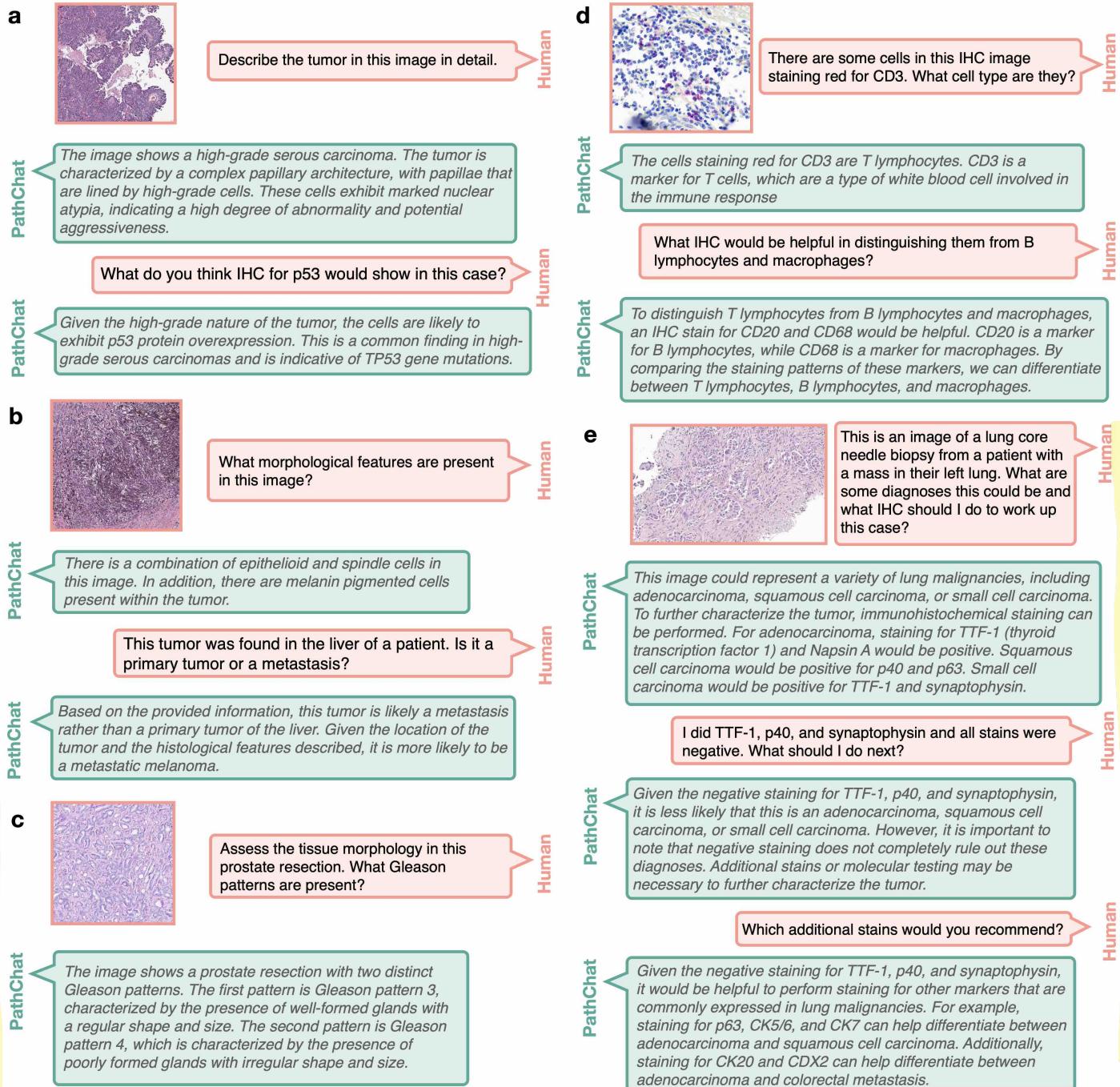
These results demonstrate that overall, PathChat generates both more accurate as well as more preferable responses to diverse pathology-related queries. Additionally, in order to better understand relative strengths and weaknesses of different models, we further analyze the performance of different models in various subgroups of questions (described in **Extended Data Tables 13-14** with examples provided in **Extended Data Figure 5**). In particular, the “**Microscopy**” category includes questions that test the ability of models to generate accurate and detailed morphological descriptions of histology microscopy images and assess clinically relevant features such as tumor differentiation and grade. Questions in the “**Diagnosis**” category test the ability of the models to directly suggest a reasonable diagnosis based on the histology image available and relevant clinical context (unlike in multiple choice questions where possible choices are provided). “**Clinical**” questions test the ability to retrieve clinically relevant background knowledge about the disease in question, including risk factors, prognosis and treatment. “**Ancillary testing**” questions test the ability of models to suggest additional testing such as IHCs and molecular in order confirm a specific diagnosis or inform prognosis and treatment. We observed that while GPT4V is the runner up to PathChat overall, PathChat’s responses are especially superior to GPT4V’s in the categories that require examination of the histology image (*i.e.* “**Microscopy**” and “**Diagnosis**”), where the accuracy is 83% and 73.9% for PathChat respectively vs. 29.8% and 39.1% for GPT4V (**Figure 2d**). Similarly, head-to-head win rate against GPT4V reaches 72.3% and 69.6% on the two categories of questions respectively, compared to the average head-to-head win-rate of 57.4%. Coupled with a lose rate against GPT4V of only 12.8% and 13% on these categories, the results imply that PathChat is better than or as good as GPT4V in nearly 85% of queries that emphasize histology image examination (**Extended Data Tables 15-16**, **Extended Data Figures 6-7**). On the flip side, we found PathChat to lag somewhat behind GPT4V on “**Clinical**” and “**Ancillary Testing**”, where PathChat achieved a respectable 92.3% and 92.5% accuracy scores compared to GPT4V’s near perfect scores of 100% and 97.5% on the two categories respectively. This is similarly reflected in the lower respective head-to-head rankings, where GPT4V’s response is preferred in approximately 65.4% and 60% of such queries. We note that we included “**Clinical**” and “**Ancillary testing**” questions in order to comprehensively assess the capabilities of AI assistant models to address pathology related queries. However, these questions frequently do not require actual examination of the histology image but instead mostly aim to test the model’s ability to recall background knowledge relevant to pathology (*e.g.* “What specific molecular

alterations are commonly found in disease X, and how might they influence the prognosis or therapeutic options?”). As a result, it is not too surprising that even general purpose multimodal AI assistants such as LLaVA 1.5 can often adequately answer questions in these categories and GPT4V especially excels here as it is presumably much larger and trained on more extensive knowledge of the internet than open-source models and PathChat. Since these queries can often readily be addressed through conventional means of querying such as internet searches or consulting a reference manual, we focus on the “Microscopy” and “Diagnosis” categories as main indicators for the utility of different models as vision language assistants for Pathology, given that the aforementioned conventional means of querying do not fill the need for AI-assisted visual question answering based on pathology images. Further breakdown of model performance by sub-categories are included in **Extended Data Tables 17-18**. We note that despite the fact that our open-ended question answering benchmark is specific to the domain of pathology, its size is largely comparable to the 140 questions used in an earlier work<sup>121</sup> for human expert evaluation of LLMs’ ability to encode general clinical knowledge.

Lastly, we note that similar to our observation in the multiple-choice evaluation, presumably due to guardrails implemented by GPT4V, only 97 of 115 questions submitted to GPT4V returned a successful answer (a maximum number of 3 attempts were made for each question, see **Evaluation of GPT4V** of the **Methods** section for more details). An unsuccessful query was treated as incorrect as the response did not address the question. However, for transparency, we record the number of unsuccessful queries for GPT4V in each question category (**Extended Data Table 19**) and also report performance on only the subset of questions that GPT4V successfully answered (**Extended Data Tables 20-25**).

### Demonstration of PathChat on diverse use cases

In addition to evaluating PathChat on multiple choice-style questions and open-ended question answering, we attempted to conceptualize and explore potential use cases for PathChat as an AI vision language assistant specialized for Pathology. We observed that PathChat can analyze and describe notable morphological details in histology images, and answer questions that require background knowledge in pathology and general biomedicine in addition to the visual input alone (**Figure 3, a-c**). The ability to analyze visual features and flexibly combine them with additional clinical context and medical knowledge (simply specified via natural language), and interpret them in the context of specific diagnostic guidelines (e.g. Gleason grading) out of the box opens PathChat to a much wider range of applications compared to task-specific visual question answering or image captioning models finetuned on relatively small datasets with a limited scope<sup>85-89</sup>. Additionally, the support for interactive, multi-turn conversation enables PathChat to potentially serve as a consultant for human-in-the-loop differential diagnosis, where an initial AI-assisted assessment can be followed up with additional clinical context, test results that are subsequently generated to narrow the differential (**Figure 3e**). This may be especially valuable in cases that involve more lengthy, complex workups such as cancers of unknown primary (CUPs) and in low-resource settings where access to experienced pathologists may be limited.



**Figure 3: Exploring additional use cases of PathChat.** Beyond evaluating PathChat on multiple choice-style questions and single turn open-ended question answering, we explore additional use cases and demonstrate examples that involve follow-up questions from users in the form of interactive, multi-turn conversations. **a** PathChat can describe tumor tissue and cell morphology, infer the diagnosis, and correctly suggest potential IHC findings grounded in relevant background knowledge about the suspected malignancy. **b**. PathChat can summarize key morphological features in the histology image and based on additional clinical context, can reasonably infer the primary origin of the tumor. **c**. PathChat understands and can attempt to follow well-known guidelines on tumor grading, in this case, the Gleason grade system for prostate adenocarcinoma. **d**. PathChat is familiar with different cell markers and can help potentially guide IHC interpretations. **e**. PathChat can potentially be consulted to perform human-in-the-loop differential diagnosis that may require multiple rounds of IHC workup.

## Discussion

The field of computational pathology has witnessed substantial progress over the years in developing increasingly accurate, task-specific predictive models based on image and/or genomics data. For histology images specifically, there has also been growing interest in building foundational task-agnostic vision encoders pre-trained with large amounts of unlabeled images, which can provide robust feature embeddings for diverse supervised and unsupervised downstream workflows. However, the explosive growth in generative AI technology and specifically MLLMs, as exemplified by the likes of ChatGPT, begins to open up a possible new frontier for both computational pathology research and real-world applications to the clinical practice of pathology. Generalist AI models equipped with natural language understanding can utilize text as a unified medium to both flexibly specify user intent (*i.e.*, in the form of a custom prompt) and to produce outputs of various levels of expressiveness (*i.e.*, from single-word to binary or multiple choice responses to coherent sentences with reasoning steps) that perform diverse tasks (*e.g.*, classification, captioning, retrieval, question answering, and more). For the field of pathology specifically, such a model can in theory have applications in a wide range of scenarios across education and research as well as human-in-the-loop clinical decision making.

In this work, we provide a proof of concept for building a general purpose visual language AI assistant tailored to human pathology. We also provide, to our best knowledge, the most extensive evaluation of such technology in the field of computational pathology by comparing our model, PathChat, against both publicly-available models developed for general machine learning and the larger domain of biomedical sciences, as well as a SOTA commercial solution, GPT4V. We created PathQABench, a high quality, expert-curated benchmark that aims to assess a diverse range of capabilities relevant to the practice of anatomic pathology, including morphological examination of histology microscopic images, making diagnoses based on both histology and clinical context, assessment of tumor grade and differentiation, suggesting further IHC/molecular testing, and understanding of risk factors, prognosis and treatment of the underlying disease. We assessed these skills through a combination of multiple-choice style diagnostic questions as well as open-ended questions and human expert evaluation. In both evaluation settings, we found PathChat to compare favorably to the current best in class commercial solution GPT4V (presumably much larger and expensive to serve than PathChat), as well as substantially outperforming the publicly available MLLMs tested in diagnostic accuracy and quality of response. Additionally, we demonstrated that the support for interactive, multi-turn conversation may enable PathChat to handle additional use cases such as complex diagnostic workups. Considering our findings, we hope PathChat can potentially find impactful applications in pathology education, research, as well as human-in-the-loop clinical decision making as the technology matures over time.

One limitation of our model is that while it has been finetuned to follow instructions, it has not undergone further alignment with human intent using techniques such as reinforcement learning from human feedback (RLHF)<sup>96</sup>. This means it may not capture certain nuances in the practice of pathology, such as occasionally not

knowing to request additional contextual information or test results when it is not possible or is difficult rule out certain morphologically similar diseases based on H&E histology alone and not understanding institutional-specific guidelines for diagnosis and treatment. It has also recently been shown that such techniques can further improve the accuracy and lower the rate of hallucination in state of the art MLLMs. The lack of human intent alignment combined with limited implementation and validation of guardrails, also means the model can output unexpected, erroneous responses when non-pathology queries or images are provided as inputs (*i.e.*, inputs from domains not encountered during training), instead of correctly identifying those queries as invalid, and refrain from answering. We hope future works can address these current limitations by using high-quality expert feedback on diverse real-world cases as well as expand the capability of PathChat-like visual language AI assistants further by supporting input and output modalities such as video and audio.

## Online Methods

### PathChat dataset curation

We curated a dataset of 257,004 instructions, which we call PathChatInstruct, for training PathChat to respond to pathology-specific queries. To ensure PathChat can generalize to a diverse range of instructions, the data encompasses several different instruction formats, including open-ended multi-turn dialogue, detailed image description, short-answer questions, multiple choice questions, and text-only questions. A diverse set of data sources were used to generate the instruction dataset, spanning image captions, educational articles, pathology case reports and regions of interests extracted from in-house WSIs. Data filtering was performed for each source individually to ensure quality and relevance for training a pathology-specific vision language assistant. Examples of frequently used heuristics for filtering include the removal of image captions that are overly short (*e.g.* < 12 words) or uninformative and overly generic (*e.g.* “An H&E image of tumor.”). We also removed captions or passages related to animal pathology (keywords include: “rat”, “pig”, *etc.*) and experimental studies (keywords include: “experimental”, “positive control”, *etc.*) using regex pattern matching. Lastly, we include basic guardrail instruction examples, where given image-specific instructions such as “Describe this histology image of a lung mass” when no image is provided, the model is expected to output the response of “Sorry, I cannot assist you since you have not uploaded any image.” Additionally, when given an image not related to pathology (sampled from MS COCO<sup>122</sup>), the model is trained to output “Sorry I can only assist you with queries related to pathology.” Similar to LLaVA<sup>105,123</sup>, for some instruction formats, we prompt general purpose LLMs<sup>98,108,123,124</sup> to structure the original source text into the desired format automatically. In those scenarios, text prompts were designed specifically for each data source and iteratively refined until a desired data quality is achieved on a randomly audited subset of the data. Lastly, we applied extensive filtering on the structured instructions to remove trivial questions that do not enhance the model’s understanding of pathology (*e.g.* “at what magnification was the image taken”) or unsuccessful queries (*e.g.* “Sorry, I cannot answer your request based on the information provided”). Overall, PathChatInstruct consists of 628,668 question answer

turns and 210,237 unique images of average dimension  $574 \times 716$  pixels. The 257,004 instructions are roughly categorized by “conversation” ( $n = 101,175$ ), “description” ( $n = 98,821$ ), “multiple choice” ( $n = 29,987$ ), “free response” ( $n = 7,981$ ), “text-only” ( $n = 3,040$ ) and “guardrail” ( $n = 16,000$ ). An illustrate example of each category is shown in **Extended Data Figure 1**.

### PathChat model design and training

Compared to text-only large language models (LLMs), a multimodal large language model (MLLM) is trained to understand and respond to user instructions in the form of natural language queries that may additionally contain inputs from other modalities such as images. The support for multi-modality is essential for the domain of histopathology since examining and interpreting visual information in high resolution microscopic images (in conjunction with other clinical information) remains the cornerstone of the discipline and extends to many aspects of disease diagnosis and management in modern medicine.

Inspired by LLaVA<sup>105,123</sup>, our MLLM, PathChat, consists of three key components: the vision encoder, the multimodal projector module and the large language model. The vision encoder is responsible for encoding the image from the original high-dimensional RGB pixel space into a low-dimensional feature representation suitable for processing by the downstream modules. The multimodal projector connects the outputs of the vision encoder to the large language model (LLM) by projecting the visual tokens to the same dimension as the LLM’s embedding space for text tokens. The LLM takes the natural language instruction as input (after it has been tokenized by a tokenizer), combines the embedded text tokens and the image token output from the multimodal projector to form the full sequence of input tokens, and predicts the desirable response via auto-regressive next word prediction. The produced response is finally decoded by the tokenizer back into natural language and presented to the end user.

For the LLM, we adopt the 13 billion parameter variant from the widely used Meta Llama 2<sup>98</sup> family of state of the art open-source LLMs, which is a decoder-only transformer-based auto-regressive language model with 40 Transformer layers, each with 40 attention heads, an embedding dimension of 5,120, a hidden dimension of 13,824 and uses rotary positional encodings, natively supporting a maximum context length of 4,096. Similar to LLaVa 1.5, we use a vision encoder based on the standard ViT-Large (ViT-L) architecture consisting of 24 Transformer multi-headed attention blocks, each with 16 attention heads, an embedding dimension of 1,024 and a feed-forward hidden dimension of 4,096. The token size is  $16 \times 16$  and we add learned absolute positional encoding to each token. The multimodal projector consists of an attention pooling layer followed by a 2-layer multi-layer perceptron (MLP). The attention pooling layer (also known as a Perceiver Resampler in some works<sup>106,125,126</sup>) uses a set of 128 learned latent queries and multiheaded cross-attention to reduce the last layer feature map of the encoder backbone into a fixed length sequence of image tokens with an initial dimension of 768 for increased training and inference efficiency, as well as to prevent the total sequence length

of tokens from potentially exceeding the context window size of the LLM. The subsequent MLP follows the design used in Llava 1.5, and consists of a single hidden layer and GeLU activation, projecting the image tokens up to the embedding dimension of the LLM (*i.e.*, 5,120 for the Llama 2 13B model). We initialize weights of the vision encoder backbone from UNI<sup>49</sup>, a SOTA vision-only SSL-pretrained general purpose encoder for H&E pathology and continue finetuning the encoder backbone together with the attention pooling module using the visual language pretraining recipe and paired image text data used to develop CONCH<sup>79</sup>, reproduced in **Extended Data Table 1**.

We follow the MLLM training recipe of LLaVa 1.5, which involves two stages of training. In the first, pretraining stage, the LLM weights are kept frozen and only the multimodal projector receives parameter updates to learn a suitable projection from the space of image tokens to the shared embedding space of the text tokens used by the LLM. For this simple purpose, the MLLM is supervised to simply predict the caption corresponding each image using roughly 100K image-caption pairs sampled from our previous dataset<sup>79</sup>, without needing to use any curated instruction data. In the second, instruction finetuning stage, both the LLM and projector are trained end-to-end to generate responses to diverse instructions that include both natural language and visual inputs, as described in the **Dataset curation** section. Specifically, given an instruction  $\mathbf{X}_{\text{instruct}}$ , the reference answer  $\mathbf{X}_{\text{ans}}$  and the image  $\mathbf{X}_{\text{img}}$ , each represented as a sequence of tokenized inputs, we maximize the likelihood of each token in  $\mathbf{X}_{\text{ans}}$ , indexed by  $1, \dots, L$ , under the chatbot (viewed as an autoregressive language model):

$$\mathcal{L}_{\text{clm}}(\theta_{\text{projector}}, \theta_{\text{llm}}) = - \sum_{i=1}^L \log p(\mathbf{X}_{\text{ans},i} | \mathbf{X}_{\text{ans},1:i-1}, \mathbf{X}_{\text{instruct}}, \mathbf{X}_{\text{img}}; \theta_{\text{projector}}, \theta_{\text{llm}}) \quad (1)$$

This instruction tuning objective easily extends to multi-turn instruction data by conditioning on all previous turns of instruction and reference answer. For instructions where there is no image present,  $\mathbf{X}_{\text{img}}$  is not defined and removed from the conditioning sequence. Similarly, if multiple images accompany a given instruction, we simply concatenate their respective image tokens, with the newline (“\n”) token inserted in-between as a separator, and treat the full sequence as  $\mathbf{X}_{\text{img}}$ . Both pretraining and finetuning were performed using an  $8 \times$  NVIDIA A100 80GB GPU node. We refer readers to **Extended Data Tables 2-3** for detailed hyperparameters used in the training process.

### PathQABench: an expert-curated pathology question-answering benchmark

The evaluation of powerful multimodal visual language AI models in histopathology is an outstanding challenge, and there is currently a lack of publicly available, high quality, expert curated, histopathology-centric QA benchmark to the best of our knowledge. One possible candidate is PathVQA, which has been used in the literature to demonstrate and evaluate the pathology image understanding capabilities of various AI models. However, our manual audit revealed numerous types of low-quality examples in the benchmark, likely due to

the lack of expert review and the automated nature of the data curation workflow used by PathVQA. We illustrate some of such examples from the evaluation (test) split of PathVQA in **Extended Data Figure 8** and use our observation as motivation to curate new a high quality QA benchmark suitable for evaluating cutting-edge multimodal large language models for pathology, described in detail below.

To evaluate PathChat, we curated a QA benchmark, named PathQABench, using high resolution, representative image ROIs hand-selected by expert pathologists from routine in total 48 H&E WSIs using the open-source QuPath digital viewer. 25 WSIs used to create PathQABench comes from in-house pathology cases at the Brigham's and Women's Hospital that have been held-out from all stages of training of PathChat, while the other 23 WSIs come from the public TCGA repository. In total, the WSIs cover 9 tissue sites and 29 diagnoses (see **Extended Data Tables 4-5**). This design choice enables us to use the subset of questions based on publicly available WSIs, referred to as PathQABench-Public, to evaluate the SOTA commercial solution GPT4V (powering ChatGPT-4 with vision capabilities) via API requests, without risk of violating institutional guidelines for handling patient data. Accordingly, the subset of questions based on private in-house WSIs, referred to as PathQABench-Private, are only used to evaluate other publicly available MLLM solutions that we can run locally inside the hospital without transmitting the data to an external server. To curate questions in PathQABench, a board-certified pathologist reviews each slide and selects a salient ROI from each WSI. Next, for each case, the pathologist invents a short clinical summary statement based on the ground truth diagnosis, which includes an appropriately devised patient age, sex and clinical symptoms and radiology findings where applicable. The summary statement is referred to as the clinical context for the corresponding case henceforth. An example clinical context is shown in **Figure 2a**. We then created both close-ended multiple choice style diagnostic questions as well as open-ended questions that aim to assess the models capabilities in assisting with diverse pathology-related queries, which cover a range of topics that include but not limited to just diagnosis (**Extended Data Table 13, Extended Data Figure 5**).

A total of 48 multiple choice questions are created using the salient ROIs (one question per ROI). In the multiple choice-style question evaluation setting, for each organ system, the pathologist selected a set of 10 possible answers that encompasses the correct answers for all questions within that organ system as well as other relatively common diagnoses within that organ system (**Extended Data Table 6**). For each multiple choice question, we consider two evaluation strategies. In the first image-only setting, the model is presented with only the image and the multiple choice question as input. In the second, image + clinical context setting that is designed to more closely mimic a real-world diagnostic workflow, the clinical context is additionally provided together with the histology image. In both settings, the model is assessed based on its ability to accurately select the ground truth diagnosis from the set of possible options.

In the open-ended question answering evaluation setting, we use the 23 cases from PathQABench-Public

(quite small datasets)

to curate 5 questions per case for a total of 115 questions. The questions can be categorized broadly under “Microscopy”, “Diagnosis”, “Clinical”, and “Ancillary testing”, as described in **Extended Data Table 13**. The “Microscopy” and “Diagnosis” questions in particular focus on targeting diagnosis and morphological examination using the histology images and additional relevant context (where applicable), which are essential skills for the practice of anatomic pathology. On the other hand, “Clinical”, and ”Ancillary testing” are often text-only questions that do not require visual examination of the image to answer, covering topics such as how to use IHCs to confirm the diagnosis and background knowledge pertaining to the underlying condition. We note that despite the fact that our open-ended question answering benchmark is specific to the domain of pathology, its size is largely comparable to the 140 questions used in an earlier work<sup>121</sup> for human expert evaluation of LLMs’ ability to encode general clinical knowledge.

## MLLM evaluation

We compare PathChat against the general purpose SOTA MLLM LLaVA 1.5<sup>123</sup> as well as the medically-focused MLLM LLaVA-Med<sup>118</sup> using the full PathQABench dataset, and only evaluate the performance of GPT-4V on cases from PathQABench-Public. The precise pretrained checkpoints for these models are specified in the **Code availability** section and the reporting summary. We use the default image processor implemented by each model and use greedy decoding during inference time when possible (not currently supported for the GPT4V API where we instead used the default arguments set by OpenAI). The evaluation of GPT4V also requires a more involved protocol due to guardrails implemented by OpenAI, which we detail in the next section (**Evaluation of GPT4V**).

For multiple choice questions, we observed that both PathChat, LLaVA 1.5 and GPT4V can output the predicted choice in a consistent and desirable format (*e.g.* “A”, “A. Lung adenocarcinoma” or “- Lung adenocarcinoma”) which can be directly used in our evaluation pipeline to compute the accuracy score. However, we found LLaVA-Med could not follow the instruction to answer in a concise and consistent format appropriate for multiple choice questions and instead would always output a full sentence. Therefore, for LLaVA-Med, we manually review each model response, extract the predicted diagnosis, assess its correctness against the ground truth and then compute the accuracy.

For the open-ended questions, we gather the prediction for each model and present them to a board-certified pathologist for human evaluation. When the model responses are presented, their order is randomly shuffled and the pathologist is blinded to which model produced which response. The responses were ranked based on, in order of importance, 1. prompt following (whether the response correctly addressed the instruction), 2. completeness of the answer, 3. succinctness and 4. use of accepted pathology terminology. Ties of two (or more) responses were allowed. Additionally, a binary correct *vs.* incorrect outcome is recorded for each response. For questions with a single best answer (*e.g.*, “What is the most likely diagnosis?”), the responses

were labeled as incorrect if the single best answer was not provided. For open ended questions (e.g., “What IHC stains would be useful in working up a glioblastoma?”), responses were labeled as incorrect if any portion of the response was hallucinated or if the response did not answer the question at all. Correct and incorrect labels were mutually exclusive and every response was labeled as correct or incorrect. The model responses on all open-ended questions can be viewed in **Supplementary Materials**.

## Evaluation of GPT4V

Evaluation of GPT4V was performed using the official API access provided by OpenAI. We observed that guardrails appear to have been implemented into GPT4V to often prevent it from addressing queries that require examination of histopathology images. In such instances, it will inform the user that it cannot provide an interpretation of the pathology image and he or she should instead consult a trained medical professional. Queries for which the response returned by GPT4V obviously refused to address the given instructions were deemed “unsuccessful”. In such instances, we make a maximum of 2 additional re-submissions for the same query up to a total of 3 attempts. Following this evaluation protocol, we recorded 12 out of 23 successful queries in the multiple choice diagnostic assessment evaluation of PathQABench-Public cases when no additional clinical context is provided as part of each question while all 23 out of 23 queries were eventually successful when the clinical context is included. Using the same protocol, in the open-ended QA evaluation section of PathQABench-Public, we obtained 97 out of 115 successful queries. A breakdown of successful queries by category is provided in **Extended Data Table 19**.

## Computing hardware and software

We used Python (version 3.10.13) for all experiments and analyses in the study. For all model training, we used 8×80GB NVIDIA A100 GPUs configured for multi-GPU training using the popular open-source deep learning framework PyTorch (version 2.0.1, CUDA 11.8). All inference jobs were performed using 24GB NVIDIA 3090 GPUs. We use the implementation of MLLM training and inference provided by LLaVA (version 1.1.3) and incorporate our own custom vision encoder and multimodal projector implemented in Timm (version 0.9.2) and Pytorch. Flash Attention (version 2.3.3) and DeepSpeed (version 0.9.5) were used to enable accelerated training of PathChat MLLM. Gradio (version 3.35.2) is used to build the interactive demo. The demo video recording is played at 1.5× speed but text prompts, input images, and model outputs are presented as is without editing. We used images from PathQABench and other real world cases not used for model training. Pandas (version 2.1.3) was used for random sampling in computing 95% confidence intervals for reported metrics by nonparametric bootstrapping (n = 1,000 iterations). Matplotlib (version 3.7.1) and Seaborn (version 0.12.2) were used to create plots and figures. Other miscellaneous libraries used are listed in the **Reporting Summary**.

## Data availability.

PathQABench data was curated using a combination of WSIs from the TCGA and from in-house pathology

database at the Brigham and Women's Hospital. The PathQABench-Public subset of the QA benchmark will be released upon publication to serve as a potential resource for the purpose of researching and evaluating the capabilities of multimodal generative AI models for the field of pathology. The original TCGA WSIs and associated clinical metadata are available from the NIH genomic data commons (portal.gdc.cancer.gov). The PathQABench-Private subset was curated with institutional permission through IRB approval for the current study and thus cannot be made publicly available. All requests for data collected or curated in-house will be evaluated based on institutional and departmental policies to determine whether the data requested is subject to intellectual property or patient privacy obligations. Instruction data was curated from image-caption pairs in educational resources, in-house patient data and PubMed. Educational resources are subject to copyright terms of publishers and will not be made available. The unprocessed PubMed Central Open Access dataset is available from the NIH PubMed Central website (pmc/tools/openftlist/).

## Code availability

MLLMs used for comparisons can be accessed via their respective official repository: LLaVA 1.5 (LLaVA) and LLaVA-Med (LLaVA-Med). Similarly, the model weights can be downloaded at: Llava 1.5 (llava-v1.5-13b) and Llava-Med (llava\_med\_in\_text\_60k). Code to train PathChat will be made available for non-commercial academic use once the article is published. We have documented all technical deep learning methods and software libraries used in the study while ensuring the paper is accessible to the broader scientific audience.

## Author contributions

M.Y.L., B.C., D.F.K.W. and F.M. conceived the study and designed the experiments. M.Y.L., B.C., D.F.K.W., R.J.C., K.I., I.L., and T.D. performed data collection and processing. M.Y.L and B.C. performed model development. M.Y.L., B.C., D.F.K.W. and R.J.C. performed experimental analysis and interpreted the results. R.J.C., K.I., G.G., I.L., T.D., L.P.L and A.V.P. provided feedback on the analysis. M.Y.L., B.C., D.F.K.W. and F.M. prepared the manuscript with input from all co-authors. F.M. supervised the research.

## Acknowledgements

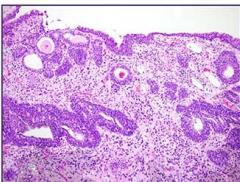
This work was supported in part by the BWH president's fund, BWH & MGH Pathology. R.J.C. was also supported by the NSF Graduate Fellowship. We thank Timothy Janicki, Richard Kenny, Abe Ahmed and the system administration staff at the MGB Enterprise Research Infrastructure & Services (ERIS) Research Computing Core for their dedicated support in providing and maintaining access to NVIDIA A100 computing resources.

## Competing interests

A provisional patent corresponding this work has been filed by Mass General Brigham.



### Multiple choice

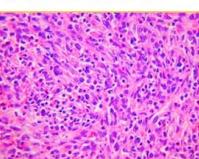


A 75 year old man presents with a polypoid bladder lesion on cystoscopy. Which feature could be present in nested variant of invasive urothelial carcinoma but should not be found in this entity?

- A. Background inflammation;
- B. Expression of CK7 and CK20;
- C. Invasion into muscularis propria;
- D. Irregular nests variably distributed within lamina propria;
- E. Mild cytologic atypia

Invasion into muscularis propria

### Conversation



What is the predominant cell type in this image?

The predominant cell type in this image appears to be spindled cells.

What other types of cells are present in the background?

In the background of this image, there are lymphocytes, plasma cells, and scattered eosinophils. These cells are likely part of the inflammatory infiltrate.

Can you describe the arrangement or distribution of the spindled cells?

The spindled cells appear to be scattered throughout the image. It does not seem to be a clustered or cohesive arrangement.

What can be inferred from the presence of lymphocytes, plasma cells, and eosinophils in the background?

The presence of lymphocytes, plasma cells, and eosinophils suggests that there is an inflammatory response in the area. This could indicate an immune reaction to a foreign substance or an ongoing infection. It is important to further evaluate the clinical history and other diagnostic tests to determine the underlying cause of this inflammatory infiltrate.

### Text only

**What is urothelial carcinoma in situ, and what are the typical clinical presentations of this condition?**

Urothelial carcinoma in situ or high-grade intraurothelial neoplasia (HG IUN) is a flat, non-invasive lesion composed of cells in the mid to upper epithelium exhibiting high cytologic grade. As this carcinoma is confined to the urothelial layer and doesn't invade the lamina propria, no mass will be present. Clinically, patients with urothelial carcinoma in situ often present symptoms akin to cystitis, with hematuria being a common feature. The lesion may appear erythematous, granular or cobblestone-like, involving substantial areas of the bladder mucosa, ureters, and urethra.

**From a histological perspective, how would you describe the features of urothelial carcinoma in situ cells?**

Histologically, the cells in urothelial carcinoma in situ, barring those of the small cell pattern, are typically large with irregular, hyperchromatic nuclei. They exhibit prominent nuclear pleomorphism and a high nucleus-to-cytoplasm ratio. Mitotic figures often appear in the mid to upper epithelium. Notably, the nuclear size is usually around five times that of lymphocytes, contrasting with the nuclear size of normal urothelial cells which is approximately twice that of lymphocytes.

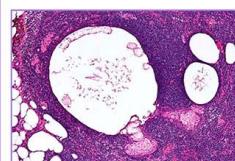
**What is meant by the involvement of von Brunn nests in carcinoma in situ, and how can this be confusing for pathologists?**

Von Brunn nests represent a benign, reactive change in the bladder wall where the urothelium invaginates into the lamina propria. In urothelial carcinoma in situ, these nests can be involved, leading to the presence of nests of neoplastic cells within the lamina propria. This can create an illusion of invasion, complicating the histologic assessment. However, unlike true invasive carcinoma, von Brunn nests typically have a rounded contour and lack the stromal changes associated with invasion, aiding differentiation.

**What are the treatment options available for patients with urothelial carcinoma in situ, bearing in mind its potential to become invasive?**

Patients with urothelial carcinoma in situ are often treated with nonsurgical methods, including a procedure known as transurethral resection (TUR) of the bladder tumor, followed by an immediate instillation of intravesical chemotherapy - either bacillus Calmette-Guerin (BCG) or mitomycin-C. If left untreated, there is a fair risk, about 20%, of bladder CIS cases advancing to invasive carcinoma. Therefore, immediate intervention is crucial.

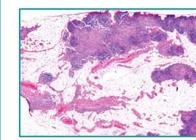
### Free response



**Describe the histologic features observed in this parotid gland mass. What are the two main structural components of the lesion?**

The histologic image of this parotid gland mass shows a distinct large and cystic sebaceous gland, which is similar in appearance to those found in the skin. This gland is surrounded by a lymphoid stroma, a second main structural component of this lesion. Additionally, there is adipose tissue present at the left side of the image.

### Description



**Describe this image.**

The image displays a mounted section of a phantom hyperechoic nodule that was resected from the thyroid gland. Based on histological examination using hematoxylin-eosin staining, the tissue has been confirmed to be ectopic thymic tissue. The ectopic thymic tissue is composed of lymphoid cells, epithelial nests, and adipose tissue.

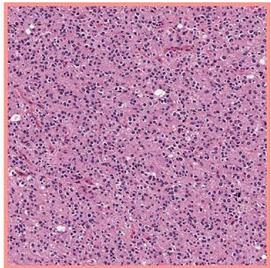
### Guardrails



**Based on the information provided, what is the most likely diagnosis for this liver biopsy?**

Sorry, as an AI assistant specialized in pathology, I cannot assist you with non-pathology related inquiries.

**Extended Figure 1: Examples of instructions for finetuning MLLM.** An example of each of six different types of instructions to develop PathChat via instruction finetuning is illustrated. **Bolded** texts represent instructions provided to the model while **italicized** texts represent the reference outputs the model is expected to output during training. More details on dataset curation are provided in the **PathChat dataset curation** section of **Methods**.



Ground truth diagnosis (not provided to the model): **Oligodendroglioma**

A 53-year-old male presents with rapid cognitive decline, seizures, and a poorly circumscribed lesion on MRI. Based on the observed histopathological findings in the image, what is the most probable diagnosis?

Human

PathChat

The image shows a diffuse infiltration of the brain parenchyma by a monomorphic population of oligodendrocytes. These cells have round nuclei and clear cytoplasm, which is characteristic of oligodendroglioma. The presence of these features, along with the patient's clinical presentation, suggests a diagnosis of oligodendroglioma.

LLaVA-Med

Based on the histopathological findings in the image, the most probable diagnosis for the 53-year-old male with rapid cognitive decline, seizures, and a poorly circumscribed lesion on MRI is a glioblastoma multiforme (GBM). GBM is an aggressive type of brain tumor that arises from glial cells in the brain. It is important to note that a definitive diagnosis should be made by a healthcare professional after considering the patient's clinical history, symptoms, and additional diagnostic tests.

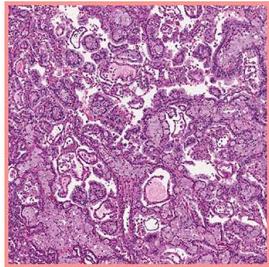
LLaVA 1.5

The most probable diagnosis for the patient with the observed symptoms and the lesion on MRI is a glioblastoma multiforme (GBM). The image shows a close-up of a purple and red brain tissue with a large number of small, round, and irregularly shaped cells, which is consistent with the histopathological features of GBM. GBM is a highly aggressive and fast-growing brain tumor that typically presents with symptoms such as rapid cognitive decline, seizures, and a poorly circumscribed lesion on MRI.

GPT4V

The image appears to be a histological slide of brain tissue, which is presumably stained to show cellular details used for pathological diagnosis... The scenario you describe -- rapid cognitive decline, seizures, and an MRI-detected brain lesion combined with the histopathology from a biopsy or resection -- suggests a possible diagnosis of a malignant brain tumor, such as glioblastoma multiforme (GBM). GBM is an aggressive type of primary brain tumor that typically appears in adults and is characterized by poorly circumscribed lesions on imaging and a variety of histopathological features including necrosis, vascular proliferation, and cellular atypia...

**Extended Figure 2: Comparing model outputs on open-ended question answering, example 1.** An example question in PathQABench-Public, for which the response by PathChat is ranked higher (considered more preferable by the expert pathologist) than other models as it clearly and correctly addresses the query that asks for the most probable diagnosis as well as provides a reasonable description of the image. The other models give the same incorrect diagnosis of glioblastoma multiforme with outdated terminology. For this example, an expert pathologist ranked the PathChat output first followed by the other three models ranked equally.



Ground truth diagnosis (not provided to the model): **Papillary renal cell carcinoma**

The image is from a 60-year-old man with flank pain and a left renal mass. In light of the clinical history and histological features, what would be your diagnosis? What elements potentially supported your conclusion?

Human

PathChat

*Given the patient's clinical history of flank pain and a left renal mass, along with the histological features of the image, the most likely diagnosis is papillary renal cell carcinoma. The presence of papillary structures, which are characteristic of this type of renal cell carcinoma, supports this conclusion.*

LLaVA-Med

*Based on the histological features and the clinical history, my diagnosis would be papillary renal cell carcinoma (RCC). The image shows a solid mass with papillary architecture, which is a characteristic feature of this type of cancer. The presence of a left renal mass and flank pain in the context of the patient's history supports this diagnosis. However, it is important to remember that a definitive diagnosis should be made by a healthcare professional after considering all relevant clinical information and test results.*

GPT4V

*Based on the image, the diagnosis is likely to be a renal cell carcinoma (RCC) due to the presence of a renal mass and the surrounding tissue. The histological features, such as the irregular shape, size, and color of the mass, along with the surrounding tissue, suggest that it is a malignant tumor. The age of the patient, a 60-year-old man, is also a risk factor for developing RCC. The flank pain experienced by the patient further supports the diagnosis, as it is a common symptom associated with renal tumors.*

*The image appears to be a histological slide of renal tissue, likely stained with hematoxylin and eosin (H&E). Given the clinical history of a 60-year-old man with flank pain and a left renal mass, and based on the visual features of the histological slide, a possible diagnosis could be renal cell carcinoma (RCC).*

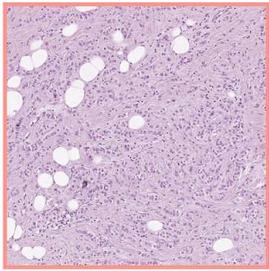
*Histological features that might support this conclusion include:*

1. The presence of clear or eosinophilic cells within the renal parenchyma, which can be characteristic of certain subtypes of RCC, such as clear cell RCC or papillary RCC.
2. Disorganization and variation in cell size and shape (pleomorphism), which indicates malignant cellular changes.
3. Abnormal architecture, such as loss of normal tubular structures or acinar pattern which are seen in healthy renal tissue.
4. Evidence of necrosis or hemorrhage, which can be associated with more aggressive disease.

*However, an accurate diagnosis would depend on a range of diagnostic modalities, including a thorough histological examination by a pathologist, and may also involve immunohistochemistry, molecular tests, and clinical correlation with imaging studies. It's also essential to note that there are various subtypes of RCC, each with its unique histological features and clinical presentations.*

*Given the suspected diagnosis of RCC, important considerations for prognosis and treatment would involve the stage of the cancer, which is determined by the tumor size, presence of metastasis, and involvement of nearby lymph nodes...*

**Extended Figure 3: Comparing model outputs on open-ended question answering, example 2.** An example question in PathQABench-Public for which the responses by PathChat and LLaVA-Med were considered to be of comparable quality by an expert pathologist for producing the correct diagnosis (papillary renal cell carcinoma) along with reasonable supporting evidence. The responses by LLaVA 1.5 and GPT4V were ranked lower as “renal cell carcinoma” is not a sufficiently specific and valid diagnosis. For this example, an expert pathologist ranked PathChat and LLaVA-Med tied in first place and LLaVA 1.5 and GPT4V tied for last.



Ground truth diagnosis (not provided to the model): **Invasive lobular carcinoma**

Describe the key histopathological features evident in the given breast tissue's image. How do the morphologic features of the neoplastic cells affect their arrangement within the tissue?

Human

PathChat

The image demonstrates neoplastic cells with a low nuclear grade, which means that the nuclei are relatively uniform in size and shape, and the chromatin is evenly distributed. These cells are arranged in a cribriform pattern, which is characterized by the formation of small, round, gland-like spaces within the tissue.

LLaVA-Med

In the histopathological image of the breast tissue, the key morphologic features of the neoplastic cells include hyperchromatic nuclei, scant cytoplasm, and rounded or oval shapes. These features can be observed in the low-power view of the image. The arrangement of these neoplastic cells varies depending on the type of lesion. In the case of lobular carcinoma, the neoplastic cells are arranged in cords or fascicles, while in the case of ductal carcinoma, they are arranged in a more trabecular or sheet-like pattern. This variation in arrangement is an important aspect of the histopathological examination, as it helps to differentiate between different types of breast lesions and guides further diagnostic and treatment decisions.

LLaVA 1.5

The image shows a close-up view of a breast tissue sample, which appears to be a biopsy of a tumor. The tissue is composed of a dense network of cells, including neoplastic cells, which are arranged in a disorganized manner. The neoplastic cells are scattered throughout the tissue, with some cells being larger and more prominent than others. The presence of these neoplastic cells indicates that the tissue is affected by cancerous growth, which can lead to the formation of a tumor. The disorganized arrangement of the cells within the tissue suggests that the cancerous cells are not following the normal organization and structure of the breast tissue, which can be a characteristic feature of cancerous growths.

GPT4V

The image appears to be a histological section of breast tissue stained with hematoxylin and eosin (H&E), a common staining method that enables the identification of cells and extracellular matrix structures. As I am an AI and do not have the ability to diagnose, the following description is based on the visual aspects generally associated with neoplastic transformations observed in pathology:

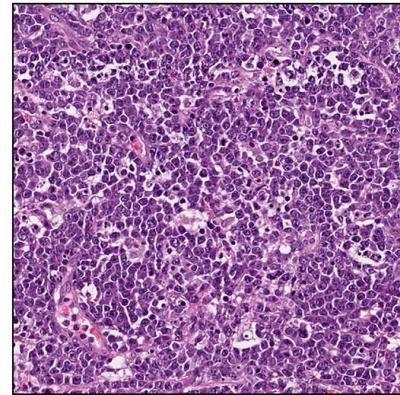
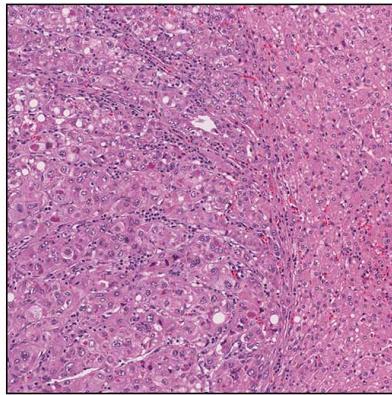
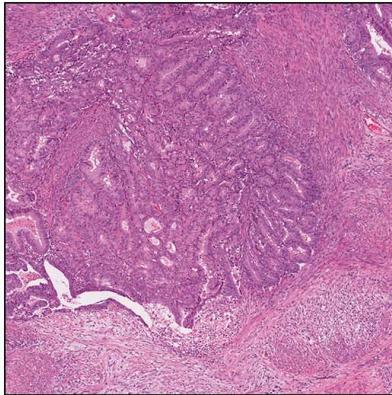
Key histopathological features that might be visible in a breast tissue sample with neoplastic changes often include:

1. Cellular atypia: This refers to the presence of cells that have abnormal appearances, such as variations in size (anisocytosis), shape (pleomorphism), and staining properties (hyperchromasia of nuclei). Atypical cells may have increased nuclear-to-cytoplasmic ratios and prominent nucleoli.
2. Increased mitotic activity: Neoplastic tissue often shows an elevated number of mitotic figures, indicating increased cellular proliferation.
3. Architectural disturbances: In breast cancer, for example, the normal glandular structure of the breast can be replaced by invasive nests or sheets of malignant cells. The arrangement and structure of ducts may be disrupted, and there may be evidence of infiltration into the surrounding stroma.
4. Loss of polarity: Neoplastic cells often lose the uniform orientation seen in normal tissue, which results in a disorganized appearance.
5. Desmoplastic reaction: This refers to the fibrous tissue response to an invasive tumor, which can appear as increased connective tissue in the stroma.
6. Necrosis and tumor heterogeneity: Advanced neoplasms may present with areas of necrosis and varying cellular morphology throughout the lesion.

The morphologic features of the neoplastic cells can affect their arrangement within the tissue by disrupting normal tissue architecture. This can lead to irregular growth patterns, loss of the proper organization of cell layers, and invasion of surrounding tissues. The relationship between cells and the extracellular matrix is also altered, potentially facilitating the invasion and metastasis of malignant cells...

**Extended Figure 4: Comparing model outputs on open-ended question answering, example 3.** An example question in PathQABench-Public for which the responses by all models were considered to be low quality by an expert pathologist. PathChat describes a cribriform pattern of cells which is not present in the image. LLaVA-Med does not give a specific description of the arrangement of neoplastic cells in the image and incorrectly claims that the image is low-power. LLaVA 1.5 describes a “dense network of cells” that is not found in the image and gives a very general description that mostly does not pertain to the image. GPT4V gives no correct specific description about the image at all. For this example, all models are ranked equally by an expert pathologist.

## Visual Question Answering



**Case ID:** TCGA\_15  
**Broad category:** Microscopy  
**Sub-category:**  
Microscopic description

**Question:** Describe the morphological features seen in this uterine sample. What significant features might hint at the aggressive nature of the tumor?

**Case ID:** TCGA\_18  
**Broad category:** Microscopy  
**Sub-category:**  
Differentiation / Grading

**Question:** Considering the observed histopathological features, what do the large pleiomorphic nuclei and eosinophilic cytoplasm in the tumor cells tell us about the grade and differentiation of the tumor?

**Case ID:** TCGA\_8  
**Broad category:** Diagnosis  
**Sub-category:**  
Diagnosis

**Question:** Given the clinical information of a 34-year-old male patient with a painless testicular mass, along with the cell morphology observed in this image, what would be the most probable diagnosis?

## Text-only Question Answering

**Case ID:** TCGA\_8  
**Broad category:** Clinical  
**Sub-category:**  
Risks / Treatment / Prognosis

**Question:** Considering the diagnosis of hepatocellular carcinoma, what are its common risk factors and what significance does this have for the patient's prognosis and treatment?

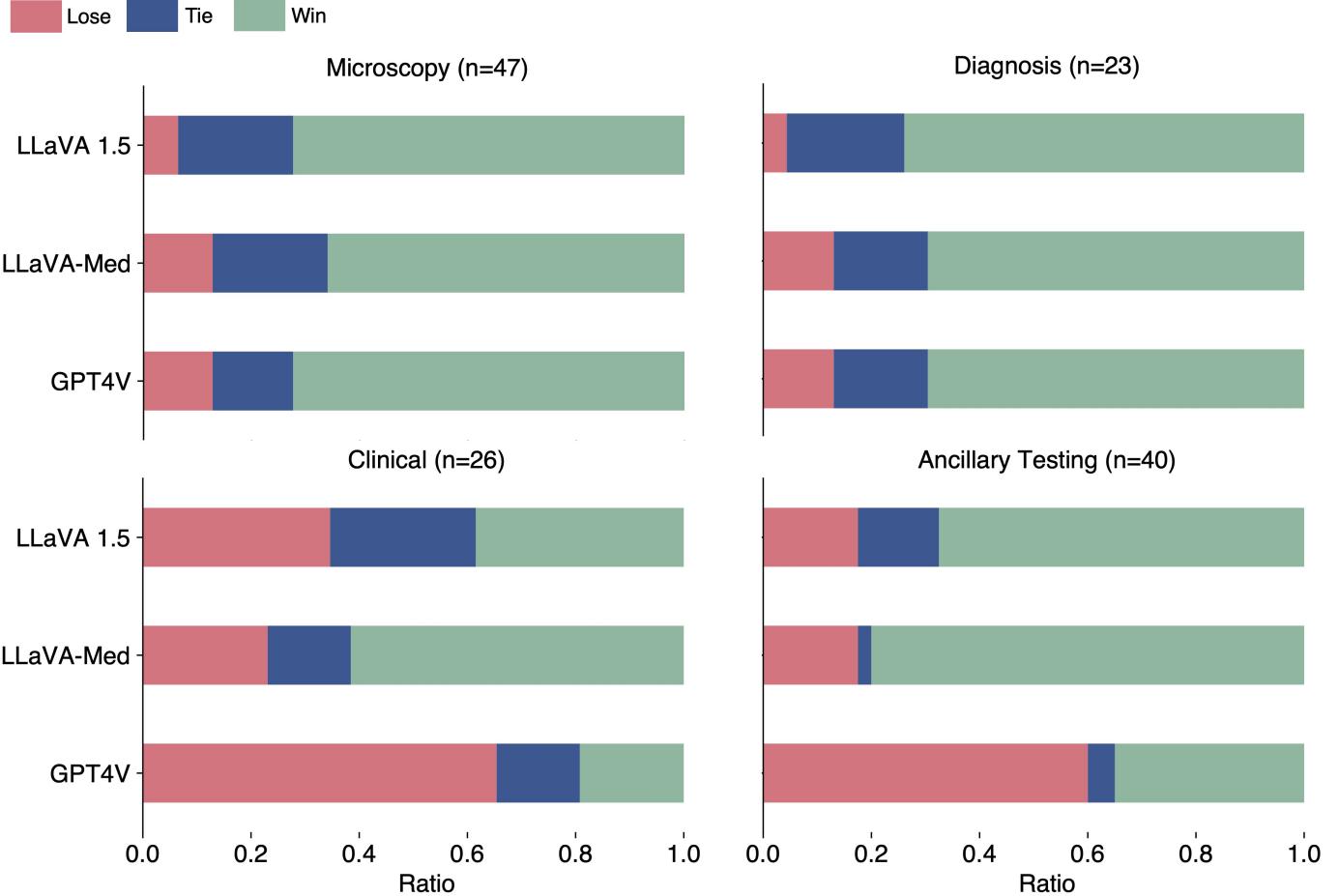
**Case ID:** TCGA\_15  
**Broad category:** Ancillary Testing  
**Sub-category:**  
IHC / Molecular

**Question:** What molecular and immunohistochemical markers are typically associated with uterine endometrioid carcinoma and how might they aid in diagnosis?

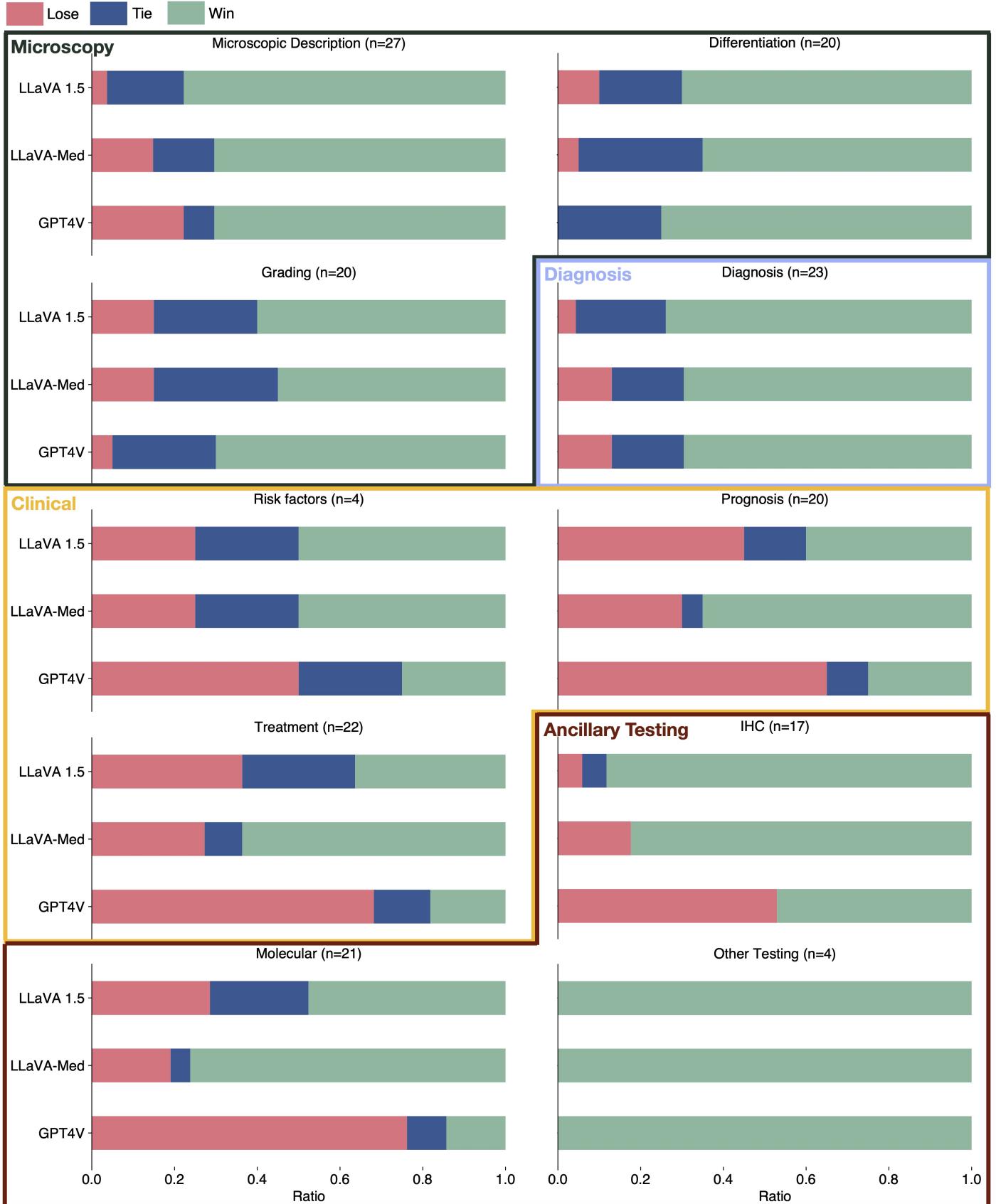
**Case ID:** TCGA\_19  
**Broad category:** Ancillary Testing  
**Sub-category:**  
Further Testing

**Question:** What further diagnostic tests can confirm a diagnosis of oropharyngeal squamous cell carcinoma?

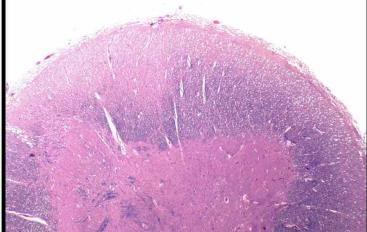
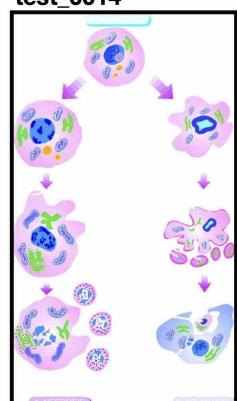
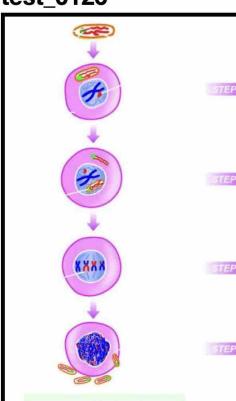
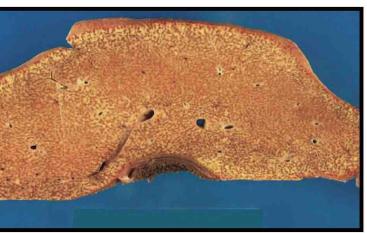
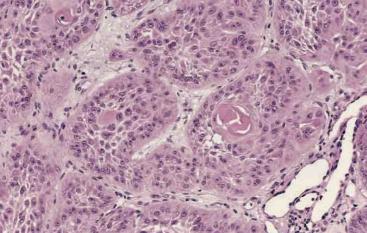
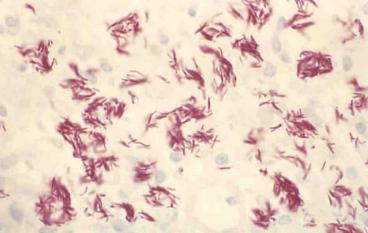
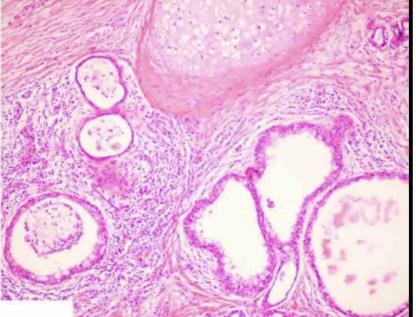
**Extended Figure 5: Example questions from PathQABench-Public.** PathQABench contains 115 high quality, expert reviewed, open-ended questions created using cases from PathQABench-Public, aimed at assessing a wide range of skills relevant to the practice of pathology. Each question is assigned one or more broad and sub-category based on the topics and skills that it aims to assess. The broad categories are “Microscopy”, “Diagnosis”, “Clinical” and “Ancillary testing”. A detailed description of each category is included in **Extended Data Table 13**. Examples from each broad category are illustrated here.



**Extended Figure 6: Performance on PathQABench open-ended questions stratified by broad categories.** We analyze the head-to-head performance of PathChat against other MLLMs in each broad category of questions. For each competing model (LLaVA 1.5, LLaVA-Med, GPT4V), we compute the win/tie/lose rate of PathChat against said model. Win (green): PathChat is ranked higher than the model; Tie (blue): PathChat is tied with the model in ranking; Lose (blue): PathChat is ranked lower than the model.



**Extended Figure 7: Performance on PathQABench open-ended questions stratified by sub-categories.** We further analyze the head-to-head performance of PathChat against other MLLMs in each sub-category of questions. For each competing model (LLaVA 1.5, LLaVA-Med, GPT4V), we compute the win/tie/lose rate of PathChat against said model. Win (green): PathChat is ranked higher than the model; Tie (blue): PathChat is tied with the model in ranking; Lose (blue): PathChat is ranked lower than the model.

 <p><b>test_0404</b></p> <p><b>Q:</b> What is present? <b>A:</b> nervous</p> <p>Vague question Poor ground truth answer</p>	 <p><b>test_0014</b></p> <p><b>Q:</b> What is cell necrosis identified by? <b>A:</b> homogeneous</p>	 <p><b>test_0126</b></p> <p><b>Q:</b> What are released into the cytosol? <b>A:</b> viral rna genome</p> <p>Blurry diagram, not specific to pathology</p>	
 <p><b>test_0172</b></p> <p><b>Q:</b> Is endoscopy small (700 g), bile-stained, soft, and congested? <b>A: no</b></p> <p><b>Q:</b> Is endoscopy small (700 g), bile-stained, soft, and congested? <b>A: yes</b> Contradictory answer</p>	<p><b>test_0235</b></p>  <p><b>Q:</b> How are the tumor cells? Vague question <b>A:</b> strikingly similar to normal squamous epithelial cells</p>	<p><b>test_0165</b></p>  <p><b>Q:</b> What shows an irreversibly sickled cell in the center? <b>A:</b> higher magnification Poor ground truth answer</p>	<p><b>test_0080</b></p>  <p><b>Q:</b> What is AIDS? <b>A:</b> acquired immunodeficiency syndrome Not specific to pathology</p>
<p><b>test_0076</b></p>   <p><b>Q:</b> What shows characteristic lining of the cyst wall by epidermis and its appendages ? <b>A:</b> microscopy Poor ground truth answer</p> <p><b>Q:</b> What is also seen? Vague question <b>A:</b> islands of mature cartilage</p>			

**Extended Figure 8: Examples of low-quality question answer pairs from PathVQA.** While PathVQA has been widely used as a benchmark to assess the pathology image understanding capabilities of AI models, we found that it consists of many low-quality examples and therefore cannot reliably serve its purpose. Here we show just a mere handful of low-quality examples from the PathVQA test set to illustrate numerous types of issues with the benchmark. **Q:** question; **A:** ground truth answer provided in PathVQA. As shown, low-quality questions are poorly constructed, vague, non-specific to pathology or paired with non-pathology images, while low-quality ground truths can similarly be vague, non-specific to pathology, and sometimes even nonsensical or contradictory. Overall, due to the lack of expert review and the automated nature of the data curation workflow of PathVQA, questions are frequently phrased in an unnatural tone, while answers are brief, mostly without explanation, do not explore alternative possibilities to open-ended questions, and frequently erroneous. These observations motivated us to curate PathQABench as a high-quality evaluation benchmark for pathology VQA.

Hyperparameter	Value
Automatic mixed precision	FP16
Batch size	192
Gradient accumulation steps	4
Learning rate scheduler	Cosine
Warmup steps	250
Peak learning rate	1e-4
AdamW $\beta$	(0.9, 0.999)
AdamW $\epsilon$	1e-8
Weight decay	0.2
Softmax temperature	Learned
Epochs	20

**Extended Data Table 1: Hyperparameters used in visual-language pretraining.** *Batch size* refers to the total batch size across GPUs. Effective batch size used for optimization is  $batch\ size \times gradient\ accumulation\ steps$ . Learning rate is increased from zero linearly to the *peak learning rate* over the course of *warmup steps* and decays back to zero following the *learning rate scheduler*. The maximum sequence length for captions is set to 128. Non-squared images are first padded to square and then resized to  $448 \times 448$ . The same image preprocessor is used in all subsequent stages of model training and evaluation.

Hyperparameter	Value
Automatic mixed precision	BF16
DeepSpeed ZeRO	Stage 3
LLM max. context length	4,096
Batch size	128
Learning rate scheduler	Cosine
Warmup ratio	0.03
Peak learning rate	1e-3
AdamW $\beta$	(0.9, 0.999)
AdamW $\epsilon$	1e-8
Weight decay	0.
Gradient clipping max. norm	1.0
Training epochs	1
Gradient checkpointing	Yes
TF32	Yes

**Extended Data Table 2: Hyperparameters used in PathChat model pretraining.** 8 × 80GB NVIDIA A100 GPUs were used for training. *Batch size* refers to the effective batch size (*i.e.*, the total batch size across GPUs). The learning rate is increased from zero linearly to the *peak learning rate* over the course of *total number batches* × *warmup ratio* steps and decays back to zero following the *learning rate scheduler*.

Hyperparameter	Value
Automatic mixed precision	BF16
DeepSpeed ZeRO	Stage 3
LLM max. context length	4,096
Batch size	64
Gradient accumulation steps	2
Learning rate scheduler	Cosine
Warmup ratio	0.03
Peak learning rate	2e-5
AdamW $\beta$	(0.9, 0.999)
AdamW $\epsilon$	1e-8
Weight decay	0.
Gradient clipping max. norm	1.0
Training epochs	1
Gradient checkpointing	Yes
TF32	Yes

**Extended Data Table 3: Hyperparameters used in PathChat model finetuning.** 8 × 80GB NVIDIA A100 GPUs were used for training. *Batch size* refers to the total batch size across GPUs. The effective batch size used for optimization is *batch size* × *gradient accumulation steps*. The learning rate is increased from zero linearly to the *peak learning rate* over the course of *total number batches* × *warmup ratio* steps and decays back to zero following the *learning rate scheduler*.

Organ	PathQABench-Public	PathQABench-Private	Total
GU	7	8	15
PHB	3	1	4
Brain	2	3	5
Lung	2	3	5
GI	2	2	4
H&N	2	3	5
GYN	2	2	4
Breast	2	2	4
Skin	1	1	2
Total	23	25	48

**Extended Data Table 4: Organ distribution of cases in PathQABench.** GI: Gastrointestinal, GU: Genitourinary, GYN: Gynecology, H&N: Head and Neck, PHB: Pancreaticohepatobiliary.

Organ	Included diagnoses
Brain	Ependymoma, Glioblastoma, Oligodendrogloma, Pilocytic astrocytoma
Breast	Invasive ductal carcinoma, Invasive lobular carcinoma
GI	Colon adenocarcinoma, Stomach adenocarcinoma
GU	Chromophobe renal cell carcinoma, Clear cell renal cell carcinoma, Mixed germ cell tumor, Papillary renal cell carcinoma, Prostatic adenocarcinoma, Renal oncocyroma, Seminoma, Testicular lymphoma, Urothelial carcinoma
GYN	High-grade serous ovarian carcinoma, Uterine endometrioid carcinoma
H&N	Adenoid cystic carcinoma, Oropharyngeal squamous cell carcinoma, Papillary thyroid carcinoma
Lung	Lung adenocarcinoma, Lung squamous cell carcinoma, Typical carcinoid tumor
PHB	Cholangiocarcinoma, Hepatocellular carcinoma, Pancreatic adenocarcinoma
Skin	Cutaneous melanoma

**Extended Data Table 5: Included diagnoses for PathQABench by organ.** GI: Gastrointestinal, GU: Genitourinary, GYN: Gynecology, H&N: Head and Neck, PHB: Pancreaticohepatobiliary.

Organ	Multiple choice options
Brain	Pilocytic astrocytoma, Ependymoma, Glioblastoma, Pleomorphic xanthoastrocytoma, Subependymal giant cell astrocytoma, Oligodendrolioma, Subependymoma, Ganglioglioma, Medulloblastoma, Neurrocytoma
GI	Colon adenocarcinoma, Stomach adenocarcinoma, Esophageal squamous cell carcinoma, Peutz-jeghers polyp, Sessile serrated adenoma, Hyerplastic polyp, Malt lymphoma, Ulcerative colitis, Neuroendocrine tumor, Gastrointestinal stromal tumor
GU	Chromophobe renal cell carcinoma, Clear cell renal cell carcinoma, Papillary renal cell carcinoma, Renal oncocytoma, Prostatic stromal sarcoma, Urothelial carcinoma, Prostatic adenocarcinoma, Testicular lymphoma, Seminoma, Mixed germ cell tumor
GYN and Breast	High-grade serous ovarian carcinoma, Low-grade serous ovarian carcinoma, Mucinous cystadenoma, Uterine endometrioid carcinoma, Invasive ductal carcinoma, Invasive lobular carcinoma, Phyllodes tumor, Cervical squamous cell carcinoma, Dcis, Paget disease
Head & Neck	Oropharyngeal squamous cell carcinoma, Acinic cell carcinoma, Adenoid cystic carcinoma, Mucoepidermoid carcinoma, Papillary thyroid carcinoma, Medullary thyroid carcinoma, Pleomorphic adenoma, Mucosal melanoma, Squamous papilloma, Nut carcinoma
Lung	Lung adenocarcinoma, Lung squamous cell carcinoma, Typical carcinoid tumor, Atypical carcinoid tumor, Hamartoma of lung, Meningothelial-like nodule, Pneumocytoma, Small cell carcinoma, Large cell carcinoma, Large cell neuroendocrine carcinoma
PHB	Hepatocellular carcinoma, Pancreatic adenocarcinoma, Cholangiocarcinoma, Pancreatic neuroendocrine tumor, Gallbladder adenocarcinoma, Cirrhosis, Hepatoblastoma, Intraductal papillary mucinous neoplasm, Mucinous cystic neoplasm, Fibrolamellar carcinoma
Skin	Cutaneous melanoma, Basal cell carcinoma, Squamous cell carcinoma, Merkel cell carcinoma, Mycosis fungoides, Verruca vulgaris, Dermatofibroma, Blue nevus, Dysplastic nevus, Hidradenoma

**Extended Data Table 6: Options for multiple choice diagnostic questions in PathQABench by organ.** For each organ system, the pathologist selected a set of 10 possible answers that encompasses the correct answers for all questions within that organ system as well as other relatively common diagnoses within that organ system. Finally, when each multiple choice question is constructed using the above options, the order in which the options appear is randomized to ensure the correct answer is equally likely to be slotted into any location among the possible choices. GI: Gastrointestinal, GU: Genitourinary, GYN: Gynecology, H&N: Head and Neck, PHB: Pancreaticohepatobiliary.

Model	Combined	Combined w/ Context
PathChat	<b>0.708 (0.583, 0.833)</b>	<b>0.812 (0.708, 0.917)</b>
LLaVA-Med	0.188 (0.083, 0.292)	0.271 (0.167, 0.396)
LLaVA 1.5	0.208 (0.104, 0.333)	0.271 (0.166, 0.417)

**Extended Data Table 7: Performance on PathQABench multiple-choice diagnostic questions.** Accuracy is reported on the full set of PathQABench multiple choice questions ( $n = 48$ ) in both the image only evaluation setting and the image + clinical context evaluation setting (denoted as “w/ Context”). 95% confidence intervals from bootstrapping are included in parentheses. For more details see **MLLM evaluation** in **Methods**.

Model	PathQABench-Public	PathQABench-Public w/ Context
PathChat	<b>0.826 (0.652, 0.957)</b>	<b>0.870 (0.696, 1.000)</b>
GPT4V	* 0.217 (0.043, 0.391)	0.696 (0.478, 0.870)
LLaVA-Med	0.130 (0.000, 0.261)	0.174 (0.043, 0.348)
LLaVA 1.5	0.174 (0.043, 0.348)	0.304 (0.130, 0.479)

**Extended Data Table 8: Performance on PathQABench-Public multiple-choice diagnostic questions.** Accuracy is reported on the PathQABench-Public multiple choice questions ( $n = 23$ ) in both the image only evaluation setting and the image + clinical context evaluation setting (denoted as “w/ Context”). 95% confidence intervals from bootstrapping are included in parentheses. For more details see **MLLM evaluation** in **Methods**. \*Note that due to guardrails implemented by GPT4V, only 12 / 23 questions submitted to the API yielded successful answers for PathQABench-Public in the image only setting (a maximum number of 3 attempts were made for each question). An unsuccessful query was treated as incorrect since the response did not address the question. We also report performance just on the subset of questions that yielded successful queries for GPT4V in **Extended Data Table 10**. In the image + clinical context setting, all questions were successfully answered by GPT4V. For more details see **Evaluation of GPT4V** in **Methods**.

Model	PathQABench-Private	PathQABench-Private w/ Context
PathChat	<b>0.600 (0.400, 0.800)</b>	<b>0.760 (0.600, 0.920)</b>
LLaVA-Med	0.240 (0.080, 0.400)	0.360 (0.160, 0.560)
LLaVA 1.5	0.240 (0.080, 0.400)	0.240 (0.080, 0.440)

**Extended Data Table 9: Performance on PathQABench-Private multiple-choice diagnostic questions.** Accuracy is reported on the PathQABench-Private multiple choice questions ( $n = 25$ ) in both the image only evaluation setting and the image + clinical context evaluation setting (denoted as “w/ Context”). 95% confidence intervals from bootstrapping are included in parentheses. For more details see **MLLM evaluation** in **Methods**.

Model	PathQABench-Public	PathQABench-Public w/ Context
PathChat	<b>0.667 (0.417, 0.917)</b>	<b>0.870 (0.696, 1.000)</b>
GPT4V	0.417 (0.167, 0.750)	0.696 (0.478, 0.870)
LLaVA-Med	0.083 (0.000, 0.250)	0.174 (0.043, 0.348)
LLaVA 1.5	0.167 (0.000, 0.417)	0.304 (0.130, 0.479)

**Extended Data Table 10: Performance on PathQABench-Public multiple-choice diagnostic questions, restricted to successful GPT4V queries.** Accuracy is reported on the subset of PathQABench-Public multiple choice questions for which GPTV successfully answered the question (*i.e.*, did not refuse to give an answer within the maximum number of 3 attempted API calls), for both the image only evaluation setting ( $n = 12$ ) and the image + clinical context evaluation setting (denoted as “w/ Context”,  $n = 23$ ). 95% confidence intervals from bootstrapping are included in parentheses. For more details see **MLLM evaluation** in **Methods**.

Model	Accuracy
PathChat	<b>0.861 (0.791, 0.922)</b>
GPT4V*	0.591 (0.504, 0.670)
LLaVA 1.5	0.426 (0.330, 0.513)
LLaVA-Med	0.504 (0.417, 0.600)

**Extended Data Table 11: Proportion of open-ended questions in PathQABench correctly answered by each model.** 95% confidence intervals from bootstrapping are included in parentheses. See **Extended Data Table 15 and 17** for accuracy stratified by category. \*Note that due to guardrails implemented by GPT4V, only 97 / 115 questions submitted to the API yielded successful answers for PathQABench-Public (a maximum number of 3 attempts were made for each question). An unsuccessful query was treated as incorrect since the response did not address the question. We also report performance just on the subset of questions that yielded successful queries for GPT4V in **Extended Data Table 20**. For more details see **Evaluation of GPT4V in Methods**.

PathChat vs. model	Lose	Tie	Win
GPT4V*	0.296 (0.209, 0.374)	0.130 (0.078, 0.191)	0.574 (0.478, 0.661)
LLaVA-Med	0.148 (0.087, 0.217)	0.148 (0.087, 0.217)	0.704 (0.617, 0.791)
LLaVA 1.5	0.113 (0.061, 0.174)	0.191 (0.122, 0.270)	0.696 (0.609, 0.783)

**Extended Data Table 12: Head-to-head performance of PathChat against other MLLMs on PathQABench open-ended questions.** For each competing model (GPT4V, LLaVA-Med, LLaVA 1.5), we compute the lose/tie/win rate ( $n = 115$ ) of PathChat against said model. Lose: said model is ranked higher than PathChat; Tie: PathChat is tied with the model in ranking; Win: PathChat is ranked higher than the model. 95% confidence intervals from bootstrapping are included in parentheses. For more details see **PathChat model evaluation** in **Methods**. \*Note that due to guardrails implemented by GPT4V, only 97 / 115 queries to the API yielded successful answers for PathQABench-Public (a maximum number of 3 attempts were made for each question). An unsuccessful query was treated as incorrect (and therefore ranked as last or tied for last) since the response did not address the question. We also report performance just on the subset of questions that yielded successful queries for GPT4V in **Extended Data Table 23**. For more details see **Evaluation of GPT4V in Methods**.

Broad Category	Description
Microscopy	Questions test the ability of models to generate accurate and detailed morphological descriptions of histology microscopy images and assess clinically relevant features such as tumor differentiation and grade Sub-categories: Microscopic description, Differentiation, Grading
Diagnosis	Questions test the ability of models to directly suggest a reasonable diagnosis based on the histology image available and relevant clinical context Sub-categories: Diagnosis
Clinical	Questions test the ability of models to retrieve clinically relevant background knowledge about the disease in question, including risk factors, prognosis and treatment. Sub-categories: Risks, Prognosis, Treatment
Ancillary testing	Questions test the ability of models to suggest additional testing such as IHCs and molecular to confirm a specific diagnosis Sub-categories: IHC, Molecular, Further testing

**Extended Data Table 13: Categorization of open-ended questions in PathQABench.** Number of questions in each category is summarized in **Extended Data Table 14**. Some questions may fit the description of more than one category or sub-category. Examples of each category can be found in **Extended Data Figure 5**.

Broad category	Count	Narrow category	Count
Microscopy	47	Microscopic Description	27
		Differentiation	20
		Grading	20
Diagnosis	23	Diagnosis	23
Clinical	26	Risk Factors	4
		Prognosis	20
		Treatment	22
Ancillary Testing	40	IHC	17
		Molecular	21
		Other Testing	4

**Extended Data Table 14: Broad and sub-categories for PathQABench open-ended questions.** In total of 115 questions were curated and reviewed by a board-certified anatomic pathologist from 23 cases in PathQABench-Public. Each question may fall under more than one category.

Category	PathChat	GPT4V*	LLaVA-Med	LLaVA 1.5
Microscopy	<b>0.830 (0.702, 0.936)</b>	0.298 (0.170, 0.426)	0.426 (0.298, 0.574)	0.404 (0.277, 0.532)
Diagnosis	<b>0.739 (0.565, 0.913)</b>	0.391 (0.217, 0.609)	0.435 (0.260, 0.652)	0.130 (0.000, 0.304)
Clinical	0.923 (0.808, 1.000)	<b>1.000 (1.000, 1.000)</b>	0.692 (0.500, 0.846)	0.769 (0.614, 0.923)
Ancillary Testing	0.925 (0.850, 1.000)	<b>0.975 (0.925, 1.000)</b>	0.575 (0.425, 0.725)	0.525 (0.350, 0.675)

**Extended Data Table 15: Proportion of open-ended questions in PathQABench correctly answered by each model, stratified by broad categories.** See **Extended Data Table 17** for accuracy stratified by category.

\*Note that due to guardrails implemented by GPT4V, only 97 / 115 questions submitted to the API yielded successful answers for PathQABench-Public (a maximum number of 3 attempts were made for each question). An unsuccessful query was treated as incorrect since the response did not address the question. We also report performance just on the subset of questions that yielded successful queries for GPT4V in **Extended Data Table 21**. For more details see **Evaluation of GPT4V in Methods**.

Category	PathChat vs. model	Lose	Tie	Win
Microscopy	GPT4V*	0.128 (0.043, 0.234)	0.149 (0.064, 0.255)	0.723 (0.596, 0.851)
	LLaVA-Med	0.128 (0.043, 0.234)	0.213 (0.106, 0.319)	0.660 (0.511, 0.787)
	LLaVA 1.5	0.064 (0.021, 0.149)	0.213 (0.106, 0.340)	0.723 (0.596, 0.851)
Diagnosis	GPT4V*	0.130 (0.043, 0.304)	0.174 (0.043, 0.348)	0.696 (0.478, 0.870)
	LLaVA-Med	0.130 (0.043, 0.261)	0.174 (0.043, 0.316)	0.696 (0.522, 0.870)
	LLaVA 1.5	0.043 (0.043, 0.174)	0.217 (0.086, 0.391)	0.739 (0.565, 0.913)
Clinical	GPT4V*	0.654 (0.500, 0.846)	0.154 (0.038, 0.308)	0.192 (0.077, 0.346)
	LLaVA-Med	0.231 (0.077, 0.423)	0.154 (0.038, 0.308)	0.615 (0.423, 0.808)
	LLaVA 1.5	0.346 (0.154, 0.538)	0.269 (0.115, 0.423)	0.385 (0.192, 0.577)
Ancillary Testing	GPT4V*	0.600 (0.425, 0.750)	0.050 (0.025, 0.125)	0.350 (0.200, 0.525)
	LLaVA-Med	0.175 (0.075, 0.300)	0.025 (0.025, 0.098)	0.800 (0.675, 0.900)
	LLaVA 1.5	0.175 (0.075, 0.300)	0.150 (0.050, 0.275)	0.675 (0.525, 0.825)

**Extended Data Table 16: Head-to-head performance of PathChat against other MLLMs on PathQABench open-ended questions, stratified by broad categories.** For each competing model (LLaVA 1.5, LLaVA-Med, GPT4V), we compute the lose/tie/win rate of PathChat against said model. Within each category, Lose: said model is ranked higher than PathChat; Tie: PathChat is tied with the model in ranking; Win: PathChat is ranked higher than the model. 95% confidence intervals from bootstrapping are included in parentheses. For more details see **PathChat model evaluation in Methods**. \*Note that due to guardrails implemented by GPT4V, only 97 / 115 questions submitted to the API yielded successful answers for PathQABench-Public (a maximum number of 3 attempts were made for each question). An unsuccessful query was treated as incorrect (and therefore ranked as last or tied for last) since the response did not address the question. We also report performance just on the subset of questions that yielded successful queries for GPT4V in **Extended Data Table 24**. For more details see **Evaluation of GPT4V in Methods**.

Category	PathChat	GPT4V*	LLaVA-Med	LLaVA 1.5
Microscopic	<b>0.852 (0.704, 0.963)</b>	0.333 (0.148, 0.519)	0.370 (0.185, 0.556)	0.333 (0.148, 0.519)
Differentiation	<b>0.800 (0.600, 0.950)</b>	0.250 (0.100, 0.450)	0.450 (0.200, 0.651)	0.450 (0.250, 0.650)
Grading	<b>0.750 (0.550, 0.950)</b>	0.150 (0.000, 0.300)	0.450 (0.250, 0.650)	0.450 (0.250, 0.650)
Diagnosis	<b>0.739 (0.565, 0.913)</b>	0.391 (0.217, 0.609)	0.435 (0.260, 0.652)	0.130 (0.000, 0.304)
Risk factors	<b>1.000 (1.000, 1.000)</b>	<b>1.000 (1.000, 1.000)</b>	0.750 (0.250, 1.000)	0.750 (0.250, 1.000)
Prognosis	0.950 (0.850, 1.000)	<b>1.000 (1.000, 1.000)</b>	0.750 (0.550, 0.950)	0.850 (0.700, 1.000)
Treatment	0.955 (0.864, 1.000)	<b>1.000 (1.000, 1.000)</b>	0.727 (0.545, 0.909)	0.773 (0.591, 0.955)
IHC	0.882 (0.706, 1.000)	<b>0.941 (0.824, 1.000)</b>	0.529 (0.294, 0.765)	0.235 (0.059, 0.412)
Molecular	0.952 (0.857, 1.000)	<b>1.000 (1.000, 1.000)</b>	0.524 (0.332, 0.714)	0.619 (0.429, 0.810)
Other Testing	<b>1.000 (1.000, 1.000)</b>	<b>1.000 (1.000, 1.000)</b>	0.750 (0.250, 1.000)	<b>1.000 (1.000, 1.000)</b>

**Extended Data Table 17: Proportion of open-ended questions in PathQABench correctly answered by each model, stratified by sub-categories.** 95% confidence intervals from bootstrapping are included in parentheses. \*Note that due to guardrails implemented by GPT4V, only 97 / 115 questions submitted to the API yielded successful answers for PathQABench-Public (a maximum number of 3 attempts were made for each question). An unsuccessful query was treated as incorrect since the response did not address the question. We also report performance just on the subset of questions that yielded successful queries for GPT4V in **Extended Data Table 22**. For more details see **Evaluation of GPT4V** in **Methods**.

Category	PathChat vs. model	Lose	Tie	Win
Microscopic	GPT4V*	0.222 (0.074, 0.407)	0.074 (0.037, 0.185)	0.704 (0.519, 0.852)
	LLaVA-Med	0.148 (0.037, 0.296)	0.148 (0.037, 0.296)	0.704 (0.519, 0.852)
	LLaVA 1.5	0.037 (0.037, 0.111)	0.185 (0.037, 0.333)	0.778 (0.593, 0.926)
Differentiation	GPT4V*	0.000 (0.000, 0.000)	0.250 (0.100, 0.450)	0.750 (0.550, 0.900)
	LLaVA-Med	0.050 (0.000, 0.200)	0.300 (0.100, 0.500)	0.650 (0.450, 0.850)
	LLaVA 1.5	0.100 (0.000, 0.250)	0.200 (0.050, 0.400)	0.700 (0.500, 0.900)
Grading	GPT4V*	0.050 (0.050, 0.150)	0.250 (0.100, 0.450)	0.700 (0.500, 0.900)
	LLaVA-Med	0.150 (0.050, 0.300)	0.300 (0.100, 0.500)	0.550 (0.350, 0.800)
	LLaVA 1.5	0.150 (0.050, 0.350)	0.250 (0.100, 0.450)	0.600 (0.400, 0.800)
Diagnosis	GPT4V*	0.130 (0.043, 0.304)	0.174 (0.043, 0.348)	0.696 (0.478, 0.870)
	LLaVA-Med	0.130 (0.043, 0.261)	0.174 (0.043, 0.316)	0.696 (0.522, 0.870)
	LLaVA 1.5	0.043 (0.043, 0.174)	0.217 (0.086, 0.391)	0.739 (0.565, 0.913)
Risk Factors	GPT4V*	0.500 (0.000, 1.000)	0.250 (0.000, 0.750)	0.250 (0.000, 0.750)
	LLaVA-Med	0.250 (0.000, 0.750)	0.250 (0.000, 0.750)	0.500 (0.000, 1.000)
	LLaVA 1.5	0.250 (0.000, 0.750)	0.250 (0.000, 0.750)	0.500 (0.000, 1.000)
Prognosis	GPT4V*	0.650 (0.450, 0.850)	0.100 (0.050, 0.250)	0.250 (0.100, 0.450)
	LLaVA-Med	0.300 (0.100, 0.500)	0.050 (0.000, 0.160)	0.650 (0.449, 0.850)
	LLaVA 1.5	0.450 (0.250, 0.650)	0.150 (0.050, 0.300)	0.400 (0.200, 0.600)
Treatment	GPT4V*	0.682 (0.500, 0.864)	0.136 (0.045, 0.273)	0.182 (0.045, 0.364)
	LLaVA-Med	0.273 (0.091, 0.455)	0.091 (0.045, 0.227)	0.636 (0.455, 0.818)
	LLaVA 1.5	0.364 (0.182, 0.591)	0.273 (0.091, 0.455)	0.364 (0.182, 0.545)
IHC	GPT4V*	0.529 (0.235, 0.765)	0.000 (0.000, 0.000)	0.471 (0.235, 0.765)
	LLaVA-Med	0.176 (0.059, 0.353)	0.000 (0.000, 0.000)	0.824 (0.647, 1.000)
	LLaVA 1.5	0.059 (0.059, 0.176)	0.059 (0.000, 0.176)	0.882 (0.706, 1.000)
Molecular	GPT4V*	0.762 (0.571, 0.905)	0.095 (0.048, 0.238)	0.143 (0.048, 0.286)
	LLaVA-Med	0.190 (0.048, 0.352)	0.048 (0.048, 0.143)	0.762 (0.571, 0.952)
	LLaVA 1.5	0.286 (0.095, 0.476)	0.238 (0.048, 0.429)	0.476 (0.286, 0.714)
Other Testing	GPT4V*	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	1.000 (1.000, 1.000)
	LLaVA-Med	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	1.000 (1.000, 1.000)
	LLaVA 1.5	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	1.000 (1.000, 1.000)

**Extended Data Table 18: Head-to-head performance of PathChat against other MLLMs on PathQABench open-ended questions, stratified by sub-categories.** For each competing model (LLaVA 1.5, LLaVA-Med, GPT4V), we compute the lose/tie/win rate of PathChat against said model. Within each category, Lose: said model is ranked higher than PathChat; Tie: PathChat is tied with the model in ranking; Win: PathChat is ranked higher than the model. 95% confidence intervals from bootstrapping are included in parentheses. For more details see **PathChat model evaluation** in **Methods**. \*Note that due to guardrails implemented by GPT4V, only 97 / 115 questions submitted to the API yielded successful answers for PathQABench-Public (a maximum number of 3 attempts were made for each question). An unsuccessful query was treated as incorrect (and therefore ranked as last or tied for last) since the response did not address the question. We also report performance just on the subset of questions that yielded successful queries for GPT4V in **Extended Data Table 25**. For more details see **Evaluation of GPT4V in Methods**.

Broad category	Count	Narrow category	Count
Microscopy	33 / 47	Microscopic Description	21 / 27
		Differentiation	12 / 20
		Grading	12 / 20
Diagnosis	19 / 23	Diagnosis	19 / 23
Clinical	26 / 26	Risk Factors	4 / 4
		Prognosis	20 / 20
		Treatment	22 / 22
Ancillary Testing	39 / 40	IHC	16 / 17
		Molecular	21 / 21
		Other Testing	4 / 4

**Extended Data Table 19: Broad and sub-categories for PathQABench open-ended questions successfully answered by GPT4V** In total of 115 questions were curated and reviewed by a board-certified anatomic pathologist from 23 cases in PathQABench-Public. Each question may fall under more than one category. For each category, we indicate how many of the total questions were successfully answered by GPT4V within a maximum of 3 attempts (see **Evaluation of GPT4V** of Methods).

Model	Accuracy
PathChat	<b>0.875 (0.802, 0.938)</b>
GPT4V	0.708 (0.615, 0.792)
LLaVA-Med	0.542 (0.448, 0.646)
LLaVA 1.5	0.458 (0.354, 0.562)

**Extended Data Table 20: Proportion of open-ended questions in PathQABench correctly answered by each model, restricted to successful GPT4V queries.** 95% confidence intervals from bootstrapping are included in parentheses. See **Extended Data Table 21** for accuracy stratified by category. Accuracy is reported on the subset of PathQABench-Public open-ended questions ( $n = 97$ ) for which GPT4V successfully answered the question (*i.e.*, did not refuse to give an answer within the maximum number of 3 attempted API calls). For more details see **Evaluation of GPT4V** in Methods.

Category	PathChat	GPT4V	LLaVA-Med	LLaVA 1.5
Microscopy	<b>0.812 (0.656, 0.938)</b>	0.438 (0.281, 0.595)	0.500 (0.312, 0.688)	0.438 (0.281, 0.625)
Diagnosis	<b>0.789 (0.632, 0.947)</b>	0.474 (0.263, 0.684)	0.368 (0.158, 0.580)	0.158 (0.000, 0.316)
Clinical	0.923 (0.808, 1.000)	<b>1.000 (1.000, 1.000)</b>	0.692 (0.500, 0.846)	0.769 (0.614, 0.923)
Ancillary Testing	0.923 (0.821, 1.000)	<b>1.000 (1.000, 1.000)</b>	0.590 (0.436, 0.744)	0.538 (0.385, 0.692)

**Extended Data Table 21: Proportion of open-ended questions in PathQABench correctly answered by each model, stratified by broad categories and restricted to successful GPT4V queries.** 95% confidence intervals from bootstrapping are included in parentheses. See **Extended Data Table 22** for accuracy stratified by sub-category. Accuracy is reported on the subset of PathQABench-Public open-ended questions ( $n = 97$ ) for which GPT4V successfully answered the question (*i.e.*, did not refuse to give an answer within the maximum number of 3 attempted API calls). For more details see **Evaluation of GPT4V** in Methods.

Category	PathChat	GPT4V	LLaVA-Med	LLaVA 1.5
Microscopic	<b>0.800 (0.600, 0.950)</b>	0.450 (0.250, 0.650)	0.350 (0.150, 0.550)	0.400 (0.200, 0.650)
Differentiation	<b>0.833 (0.583, 1.000)</b>	0.417 (0.167, 0.669)	0.667 (0.417, 0.917)	0.417 (0.167, 0.667)
Grading	<b>0.750 (0.500, 1.000)</b>	0.250 (0.000, 0.500)	0.667 (0.417, 0.917)	0.500 (0.250, 0.750)
Diagnosis	<b>0.789 (0.632, 0.947)</b>	0.474 (0.263, 0.684)	0.368 (0.158, 0.580)	0.158 (0.000, 0.316)
Risk factors	<b>1.000 (1.000, 1.000)</b>	<b>1.000 (1.000, 1.000)</b>	0.750 (0.250, 1.000)	0.750 (0.250, 1.000)
Prognosis	0.950 (0.850, 1.000)	<b>1.000 (1.000, 1.000)</b>	0.750 (0.550, 0.950)	0.850 (0.700, 1.000)
Treatment	0.955 (0.864, 1.000)	<b>1.000 (1.000, 1.000)</b>	0.727 (0.545, 0.909)	0.773 (0.591, 0.955)
IHC	0.875 (0.688, 1.000)	<b>1.000 (1.000, 1.000)</b>	0.562 (0.312, 0.812)	0.250 (0.062, 0.500)
Molecular	0.952 (0.857, 1.000)	<b>1.000 (1.000, 1.000)</b>	0.524 (0.332, 0.714)	0.619 (0.429, 0.810)
Other Testing	<b>1.000 (1.000, 1.000)</b>	<b>1.000 (1.000, 1.000)</b>	0.750 (0.250, 1.000)	<b>1.000 (1.000, 1.000)</b>

**Extended Data Table 22:** Proportion of open-ended questions in PathQABench correctly answered by each model, stratified by sub-categories and restricted to successful GPT4V queries. 95% confidence intervals from bootstrapping are included in parentheses. Accuracy is reported on the subset of PathQABench-Public open-ended questions ( $n = 97$ ) for which GPT4V successfully answered the question (*i.e.*, did not refuse to give an answer within the maximum number of 3 attempted API calls). For more details see **Evaluation of GPT4V in Methods**.

PathChat vs. model	Lose	Tie	Win
GPT4V	0.351 (0.268, 0.454)	0.113 (0.052, 0.175)	0.536 (0.433, 0.629)
LLaVA-Med	0.144 (0.082, 0.216)	0.134 (0.072, 0.206)	0.722 (0.629, 0.804)
LLaVA 1.5	0.113 (0.052, 0.175)	0.206 (0.124, 0.299)	0.680 (0.588, 0.773)

**Extended Data Table 23:** Head-to-head performance of PathChat against other MLLMs on PathQABench open-ended questions, restricted to successful GPT4V queries For each competing model (GPT4V, LLaVA-Med, LLaVA 1.5), we compute the lose/tie/win rate of PathChat against said model on the subset of questions ( $n = 97$ ) for which GPT4V successfully answered the question (*i.e.*, did not refuse to give an answer within the maximum number of 3 attempted API calls). Lose: said model is ranked higher than PathChat; Tie: PathChat is tied with the model in ranking; Win: PathChat is ranked higher than the model. 95% confidence intervals from bootstrapping are included in parentheses. For more details see **MLLM model evaluation in Methods**. For more details see **Evaluation of GPT4V in Methods**.

Category	PathChat vs. model	Lose	Tie	Win
Microscopy	GPT4V*	0.182 (0.061, 0.333)	0.152 (0.030, 0.273)	0.667 (0.515, 0.818)
	LLaVA-Med	0.121 (0.030, 0.242)	0.212 (0.091, 0.364)	0.667 (0.515, 0.818)
	LLaVA 1.5	0.030 (0.030, 0.091)	0.303 (0.182, 0.455)	0.667 (0.515, 0.818)
Diagnosis	GPT4V*	0.158 (0.000, 0.316)	0.105 (0.053, 0.263)	0.737 (0.526, 0.895)
	LLaVA-Med	0.105 (0.000, 0.263)	0.158 (0.053, 0.316)	0.737 (0.578, 0.895)
	LLaVA 1.5	0.053 (0.000, 0.158)	0.158 (0.053, 0.316)	0.789 (0.632, 0.947)
Clinical	GPT4V*	0.654 (0.500, 0.846)	0.154 (0.038, 0.308)	0.192 (0.077, 0.346)
	LLaVA-Med	0.231 (0.077, 0.423)	0.154 (0.038, 0.308)	0.615 (0.423, 0.808)
	LLaVA 1.5	0.346 (0.154, 0.538)	0.269 (0.115, 0.423)	0.385 (0.192, 0.577)
Ancillary Testing	GPT4V*	0.615 (0.462, 0.744)	0.051 (0.026, 0.128)	0.333 (0.205, 0.487)
	LLaVA-Med	0.179 (0.077, 0.308)	0.026 (0.026, 0.103)	0.795 (0.667, 0.923)
	LLaVA 1.5	0.179 (0.077, 0.308)	0.154 (0.051, 0.282)	0.667 (0.513, 0.821)

**Extended Data Table 24: Head-to-head performance of PathChat against other MLLMs on PathQABench open-ended questions, stratified by broad categories, restricted to successful GPT4V queries.** For each competing model (LLaVA 1.5, LLaVA-Med, GPT4V), we compute the lose/tie/win rate of PathChat against said model on the subset of questions ( $n = 97$ ) for which GPT4V successfully answered the question (*i.e.*, did not refuse to give an answer within the maximum number of 3 attempted API calls). Within each category, Lose: said model is ranked higher than PathChat; Tie: PathChat is tied with the model in ranking; Win: PathChat is ranked higher than the model. 95% confidence intervals from bootstrapping are included in parentheses. For more details see **MLLM model evaluation** in **Methods**. For more details see **Evaluation of GPT4V** in **Methods**.

Category	PathChat vs. model	Lose	Tie	Win
Microscopic	GPT4V*	0.286 (0.095, 0.476)	0.095 (0.048, 0.238)	0.619 (0.429, 0.810)
	LLaVA-Med	0.095 (0.048, 0.238)	0.190 (0.048, 0.381)	0.714 (0.524, 0.905)
	LLaVA 1.5	0.048 (0.048, 0.143)	0.238 (0.095, 0.429)	0.714 (0.524, 0.905)
Differentiation	GPT4V*	0.000 (0.000, 0.000)	0.250 (0.083, 0.500)	0.750 (0.500, 1.000)
	LLaVA-Med	0.083 (0.000, 0.250)	0.250 (0.083, 0.500)	0.667 (0.333, 0.917)
	LLaVA 1.5	0.000 (0.000, 0.000)	0.333 (0.083, 0.667)	0.667 (0.333, 0.917)
Grading	GPT4V*	0.083 (0.000, 0.250)	0.250 (0.083, 0.500)	0.667 (0.417, 0.917)
	LLaVA-Med	0.167 (0.000, 0.417)	0.333 (0.083, 0.583)	0.500 (0.250, 0.750)
	LLaVA 1.5	0.083 (0.000, 0.250)	0.417 (0.167, 0.667)	0.500 (0.250, 0.750)
Diagnosis	GPT4V*	0.158 (0.000, 0.316)	0.105 (0.053, 0.263)	0.737 (0.526, 0.895)
	LLaVA-Med	0.105 (0.000, 0.263)	0.158 (0.053, 0.316)	0.737 (0.578, 0.895)
	LLaVA 1.5	0.053 (0.000, 0.158)	0.158 (0.053, 0.316)	0.789 (0.632, 0.947)
Risk Factors	GPT4V*	0.500 (0.000, 1.000)	0.250 (0.000, 0.750)	0.250 (0.000, 0.750)
	LLaVA-Med	0.250 (0.000, 0.750)	0.250 (0.000, 0.750)	0.500 (0.000, 1.000)
	LLaVA 1.5	0.250 (0.000, 0.750)	0.250 (0.000, 0.750)	0.500 (0.000, 1.000)
Prognosis	GPT4V*	0.650 (0.450, 0.850)	0.100 (0.050, 0.250)	0.250 (0.100, 0.450)
	LLaVA-Med	0.300 (0.100, 0.500)	0.050 (0.000, 0.160)	0.650 (0.449, 0.850)
	LLaVA 1.5	0.450 (0.250, 0.650)	0.150 (0.050, 0.300)	0.400 (0.200, 0.600)
Treatment	GPT4V*	0.682 (0.500, 0.864)	0.136 (0.045, 0.273)	0.182 (0.045, 0.364)
	LLaVA-Med	0.273 (0.091, 0.455)	0.091 (0.045, 0.227)	0.636 (0.455, 0.818)
	LLaVA 1.5	0.364 (0.182, 0.591)	0.273 (0.091, 0.455)	0.364 (0.182, 0.545)
IHC	GPT4V*	0.562 (0.312, 0.812)	0.000 (0.000, 0.000)	0.438 (0.188, 0.688)
	LLaVA-Med	0.188 (0.062, 0.375)	0.000 (0.000, 0.000)	0.812 (0.625, 1.000)
	LLaVA 1.5	0.062 (0.062, 0.250)	0.062 (0.000, 0.188)	0.875 (0.688, 1.000)
Molecular	GPT4V*	0.762 (0.571, 0.905)	0.095 (0.048, 0.238)	0.143 (0.048, 0.286)
	LLaVA-Med	0.190 (0.048, 0.352)	0.048 (0.048, 0.143)	0.762 (0.571, 0.952)
	LLaVA 1.5	0.286 (0.095, 0.476)	0.238 (0.048, 0.429)	0.476 (0.286, 0.714)
Other Testing	GPT4V*	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	1.000 (1.000, 1.000)
	LLaVA-Med	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	1.000 (1.000, 1.000)
	LLaVA 1.5	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	1.000 (1.000, 1.000)

**Extended Data Table 25: Head-to-head performance of PathChat against other MLLMs on PathQABench open-ended questions, stratified by sub-categories, restricted to successful GPT4V queries.** For each competing model (LLaVA 1.5, LLaVA-Med, GPT4V), we compute the lose/tie/win rate of PathChat against said model on the subset of questions ( $n = 97$ ) for which GPT4V successfully answered the question (*i.e.*, did not refuse to give an answer within the maximum number of 3 attempted API calls). Within each category, Lose: said model is ranked higher than PathChat; Tie: PathChat is tied with the model in ranking; Win: PathChat is ranked higher than the model. 95% confidence intervals from bootstrapping are included in parentheses. For more details see **MLLM evaluation** in **Methods**.

## References

1. Madabhushi, A. & Lee, G. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical image analysis* **33**, 170–175 (2016).
2. Shmatko, A., Ghaffari Laleh, N., Gerstung, M. & Kather, J. N. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nature Cancer* **3**, 1026–1038 (2022).
3. Song, A. H. *et al.* Artificial intelligence for digital and computational pathology. *Nature Reviews Biotechnology* **1**–20 (2023).
4. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology* **16**, 703–715 (2019).
5. Heinz, C. N., Echle, A., Foersch, S., Bychkov, A. & Kather, J. N. The future of artificial intelligence in digital pathology—results of a survey across stakeholder groups. *Histopathology* **80**, 1121–1127 (2022).
6. Cui, M. & Zhang, D. Y. Artificial intelligence and computational pathology. *Laboratory Investigation* **101**, 412–422 (2021).
7. Abels, E. *et al.* Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *The Journal of pathology* **249**, 286–294 (2019).
8. Waqas, A. *et al.* Revolutionizing digital pathology with the power of generative artificial intelligence and foundation models. *Laboratory Investigation* **100** 255 (2023).
9. Lipkova, J. *et al.* Artificial intelligence for multimodal data integration in oncology. *Cancer cell* **40**, 1095–1110 (2022).
10. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine* **24**, 1559–1567 (2018).
11. Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* **5**, 555–570 (2021).
12. Bulten, W. *et al.* Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology* **21**, 233–241 (2020).
13. Nagpal, K. *et al.* Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *NPJ digital medicine* **2**, 48 (2019).
14. Huang, S.-C. *et al.* Deep neural network trained on gigapixel images improves lymph node metastasis detection in clinical settings. *Nature communications* **13**, 3347 (2022).
15. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine* **25**, 1301–1309 (2019).
16. Ehteshami Bejnordi, B. *et al.* Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* **318**, 2199–2210 (2017).
17. Beck, A. H. *et al.* Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science translational medicine* **3** (2011).
18. Mobadersany, P. *et al.* Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences* **115**, E2970–E2979 (2018).

19. Chen, R. J. *et al.* Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* **40**, 865–878 (2022).
20. Lee, Y., Park, J., Oh, S. *et al.* Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. *Nat. Biomed. Eng* (2022).
21. Courtiol, P. *et al.* Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine* **25**, 1519–1525 (2019).
22. Lu, C. *et al.* A prognostic model for overall survival of patients with early-stage non-small cell lung cancer: a multicentre, retrospective study. *The Lancet Digital Health* **2**, e594–e606 (2020).
23. Amgad, M. *et al.* A population-level digital histologic biomarker for enhanced prognosis of invasive breast cancer. *Nature Medicine* 1–13 (2023).
24. Boehm, K. M. *et al.* Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nature cancer* **3**, 723–733 (2022).
25. Sammut, S.-J. *et al.* Multi-omic machine learning predictor of breast cancer therapy response. *Nature* **601**, 623–629 (2022).
26. Vanguri, R. S. *et al.* Multimodal integration of radiology, pathology and genomics for prediction of response to pd-(l) 1 blockade in patients with non-small cell lung cancer. *Nature cancer* **3**, 1151–1164 (2022).
27. Huang, Z. *et al.* Artificial intelligence reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images. *NPJ Precision Oncology* **7**, 14 (2023).
28. Lu, M. Y. *et al.* Ai-based pathology predicts origins for cancers of unknown primary. *Nature* **594**, 106–110 (2021).
29. Zhu, L. *et al.* An accurate prediction of the origin for bone metastatic cancer using deep learning on digital pathological images. *EBioMedicine* **87** (2023).
30. Chen, C. *et al.* Fast and scalable search of whole-slide images via self-supervised deep learning. *Nature Biomedical Engineering* **6**, 1420–1434 (2022).
31. Kalra, S. *et al.* Yottixel—an image search engine for large archives of histopathology whole slide images. *Medical Image Analysis* **65**, 101757 (2020).
32. Hegde, N. *et al.* Similar image search for histopathology: Smily. *NPJ digital medicine* **2**, 56 (2019).
33. Wang, X. *et al.* Retcl: clustering-guided contrastive learning for whole-slide image retrieval. *Medical image analysis* **83**, 102645 (2023).
34. Kather, J. N. *et al.* Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature cancer* **1**, 789–799 (2020).
35. Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature cancer* **1**, 800–810 (2020).
36. Saldanha, O. L. *et al.* Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. *NPJ Precision Oncology* **7**, 35 (2023).
37. Wagner, S. J. *et al.* Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell* **41**, 1650–1661 (2023).
38. Yala, A. *et al.* Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nature medicine* **28**, 136–143 (2022).

39. Zhou, Y. *et al.* Multi-site cross-organ calibrated deep learning (musclD): Automated diagnosis of non-melanoma skin cancer. *Medical image analysis* **84**, 102702 (2023).
40. Laleh, N. G. *et al.* Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Medical image analysis* **79** (2022).
41. Graham, S. *et al.* Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis* **58**, 101563 (2019).
42. Graham, S. *et al.* One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. *Medical Image Analysis* **83**, 102685 (2023).
43. Caron, M. *et al.* Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660 (2021).
44. Oquab, M. *et al.* Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).
45. He, K. *et al.* Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009 (2022).
46. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607 (PMLR, 2020).
47. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738 (2020).
48. Zhou, J. *et al.* Image bert pre-training with online tokenizer. In *International Conference on Learning Representations* (2021).
49. Chen, R. J. *et al.* A general-purpose self-supervised model for computational pathology. *arXiv preprint arXiv:2308.15474* (2023).
50. Wang, X. *et al.* Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis* **81**, 102559 (2022).
51. Lai, J. *et al.* Domain-specific optimization and diverse evaluation of self-supervised models for histopathology. *arXiv preprint arXiv:2310.13259* (2023).
52. Vorontsov, E. *et al.* Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778* (2023).
53. Azizi, S. *et al.* Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering* 1–24 (2023).
54. Campanella, G. *et al.* Computational pathology at health system scale—self-supervised foundation models from three billion images. *arXiv preprint arXiv:2310.07033* (2023).
55. Kang, M., Song, H., Park, S., Yoo, D. & Pereira, S. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3344–3354 (2023).
56. Ciga, O., Xu, T. & Martel, A. L. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications* **7**, 100198 (2022).
57. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PMLR, 2021).

58. Jia, C. *et al.* Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916 (PMLR, 2021).
59. Li, G., Duan, N., Fang, Y., Gong, M. & Jiang, D. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 11336–11344 (2020).
60. Lu, J., Batra, D., Parikh, D. & Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* **32** (2019).
61. Chen, Y.-C. *et al.* Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120 (Springer, 2020).
62. Li, X. *et al.* Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, 121–137 (Springer, 2020).
63. Wang, J. *et al.* Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100* (2022).
64. Zhang, P. *et al.* Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5579–5588 (2021).
65. Hu, X. *et al.* Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17980–17989 (2022).
66. Wang, W. *et al.* Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442* (2022).
67. Yu, J. *et al.* Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).
68. Li, J., Li, D., Xiong, C. & Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900 (PMLR, 2022).
69. Li, J. *et al.* Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* **34**, 9694–9705 (2021).
70. Schaumberg, A. J. *et al.* Interpretable multimodal deep learning for real-time pan-tissue pan-disease pathology search on social media. *Modern pathology* **33**, 2169–2185 (2020).
71. Schuhmann, C. *et al.* Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* **35**, 25278–25294 (2022).
72. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, 2–25 (PMLR, 2022).
73. Pelka, O., Koitka, S., Rückert, J., Nensa, F. & Friedrich, C. M. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings* 3, 180–189 (Springer, 2018).
74. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. & Zou, J. Leveraging medical twitter to build a visual–language foundation model for pathology ai. *bioRxiv* 2023–03 (2023).

75. Zhang, S. *et al.* Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915* (2023).
76. Gamper, J. & Rajpoot, N. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16549–16559 (2021).
77. Ikezogwo, W. O. *et al.* Quilt-1m: One million image-text pairs for histopathology. *arXiv preprint arXiv:2306.11207* (2023).
78. Lin, W. *et al.* PMC-CLIP: Contrastive language-image pre-training using biomedical documents. In *Lecture Notes in Computer Science*, 525–536 (Springer Nature Switzerland, 2023). URL [https://doi.org/10.1007/978-3-031-43993-3\\_51](https://doi.org/10.1007/978-3-031-43993-3_51).
79. Lu, M. Y. *et al.* Towards a visual-language foundation model for computational pathology. *arXiv preprint arXiv:2307.12914* (2023).
80. Lu, M. Y. *et al.* Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19764–19775 (2023).
81. Tiu, E. *et al.* Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering* **6**, 1399–1406 (2022).
82. Huang, S.-C., Shen, L., Lungren, M. P. & Yeung, S. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3942–3951 (2021).
83. Boecking, B. *et al.* Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, 1–21 (Springer, 2022).
84. Eslami, S., de Melo, G. & Meinel, C. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906* (2021).
85. Zhang, H. *et al.* Pathnarratives: Data annotation for pathological human-ai collaborative diagnosis. *Frontiers in Medicine* **9**, 1070072 (2023).
86. Tsuneki, M. & Kanavati, F. Inference of captions from histopathological patches. In *International Conference on Medical Imaging with Deep Learning*, 1235–1250 (PMLR, 2022).
87. Zhang, R., Weber, C., Grossman, R. & Khan, A. A. Evaluating and interpreting caption prediction for histopathology images. In *Machine Learning for Healthcare Conference*, 418–435 (PMLR, 2020).
88. Naseem, U., Khushi, M. & Kim, J. Vision-language transformer for interpretable pathology visual question answering. *IEEE Journal of Biomedical and Health Informatics* **27**, 1681–1690 (2022).
89. He, X. Towards visual question answering on pathology images. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing*, vol. 2 (2021).
90. Vaswani, A. *et al.* Attention Is All You Need. In *Neural Information Processing Systems (NeurIPS)* (2017).
91. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. *et al.* Improving language understanding by generative pre-training (2018).
92. Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).

93. Brown, T. *et al.* Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020).
94. Hoffmann, J. *et al.* An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems* **35**, 30016–30030 (2022).
95. Wei, J. *et al.* Emergent abilities of large language models. *Transactions on Machine Learning Research* (2022).
96. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **35**, 27730–27744 (2022).
97. Touvron, H. *et al.* Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
98. Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
99. Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* **21**, 5485–5551 (2020).
100. Zhao, W. X. *et al.* A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
101. Chowdhery, A. *et al.* Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).
102. Anil, R. *et al.* Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).
103. Moor, M. *et al.* Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189* (2023).
104. Li, C. *et al.* Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020* (2023).
105. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. In *NeurIPS* (2023).
106. Alayrac, J.-B. *et al.* Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems* **35**, 23716–23736 (2022).
107. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
108. Bubeck, S. *et al.* Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
109. Alsentzer, E. *et al.* Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78 (2019).
110. Singhal, K. *et al.* Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617* (2023).
111. Tu, T. *et al.* Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334* (2023).
112. Jiang, L. Y. *et al.* Health system-scale language models are all-purpose prediction engines. *Nature* 1–6 (2023).
113. Ayers, J. W. *et al.* Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* (2023).

114. Nori, H. *et al.* Can generalist foundation models outcompete special-purpose tuning? case study in medicine (2023). *arXiv*:2311.16452.
115. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).
116. Buckley, T., Diao, J. A., Rodman, A. & Manrai, A. K. Accuracy of a vision-language model on challenging medical cases. *arXiv preprint arXiv:2311.05591* (2023).
117. Sun, Y. *et al.* Pathasst: Redefining pathology through generative foundation ai assistant for pathology. *arXiv preprint arXiv:2305.15072* (2023).
118. Li, C. *et al.* Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890* (2023).
119. Wu, C. *et al.* Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909* (2023).
120. Oon, M. L., Syn, N. L., Tan, C. L., Tan, K.-B. & Ng, S.-B. Bridging bytes and biopsies: A comparative analysis of chatgpt and histopathologists in pathology diagnosis and collaborative potential. *Histopathology* (2023).
121. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
122. Lin, T.-Y. *et al.* Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, 740–755 (Springer, 2014).
123. Liu, H., Li, C., Li, Y. & Lee, Y. J. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744* (2023).
124. Chiang, W.-L. *et al.* Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality (2023). URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
125. Zeng, Y. *et al.* What matters in training a gpt4-style language model with multimodal inputs? *arXiv preprint arXiv:2307.02469* (2023).
126. Jaegle, A. *et al.* Perceiver: General perception with iterative attention. In *International conference on machine learning*, 4651–4664 (PMLR, 2021).