

SlideChat is a multimodal generative artificial intelligence assistant for whole-slide computational pathology across cancer types

Received: 27 July 2025

Accepted: 13 July 2026

Published online: 01 September 2026

 Check for updates

Ying Chen^{1,2,3,12}, Chenglong Ma^{1,2,12}, Qiongqiong Li⁴, Fang Yan¹, Yirong Chen¹, Tianbin Li¹, Jin Ye¹, Ming Hu¹, Yuxiang Lin⁵, Yanjun Li¹, Guoan Wang¹, Huihui Xu¹, Hui Dong⁶, Xiang Wang⁷, Xiaoxiao Xu⁴, Yanyan Zhou⁴, Xia Zhu⁸, Sen Yang⁹, Xiyue Wang⁹, Lu Zhang¹⁰, Yu Qiao^{1,2,11}, Rongshan Yu^{3,5}, Junjun He^{1,2}✉ & Yuanfeng Ji⁹✉

Multimodal artificial intelligence, albeit showing great potential in computational pathology, remains limited to isolated patch-level interpretation and often fails to analyze gigapixel-scale whole-slide images (WSIs) essential for clinical utility. Here we present SlideChat, a multimodal generative artificial intelligence assistant for whole-slide computational pathology across cancer types. SlideChat integrates patch-level and slide-level pathology encoders with a pretrained large language model. Using 274,233 multimodal instruction samples, SlideChat is trained to learn the associations between WSIs and diagnostic reports and interpret complex queries in clinical practice. Evaluated on 8,836 closed-ended questions, 129 open-ended questions and 3,149 WSI reports from five cohorts spanning 31 cancer types, SlideChat outperformed leading baselines by 19.1% in closed-ended accuracy and by 7.7% in report-generation Metric for Evaluation of Translation with Explicit Ordering score and received the highest expert ratings across five dimensions in open-ended question answering, showing the potential to enhance diagnostic workflows, medical education and clinical decision-making.

Computational pathology is pivotal in modern diagnostics by enabling precise, quantitative analysis of digitized histopathological slides, which are fundamental for cancer diagnosis, prognosis and treatment planning^{1–4}. Recently, this field has gained rapid momentum with advances in visual foundation models^{5–7} and vision–language models^{8–11}. These models learn generalizable representations from large-scale data and perform effectively across diverse pathological tasks^{12,13}. Incorporation of large language models (LLMs) into visual systems has further driven the emergence of multimodal large language models (MLLMs) for pathology^{14–16}, enabling human–artificial

intelligence (AI) interactions across clinical decision-making, medical education and biomedical research.

Despite these advances, existing pathology MLLMs remain constrained by patch- or region-level analysis^{14–16}, focusing on isolated areas of the slide while lacking awareness of broader tissue context. Such approaches are inadequate for real-world clinical pathology, which demands slide-level and even patient-level analysis and integration of information across multiple regions, scales and contexts in the whole-slide image (WSI), to support complete diagnostic workflows, including subtyping, differential diagnosis and pathology report

generation. However, models capable of holistic, WSI-based interpretation are still largely unexplored, with several unresolved challenges. First, despite the availability of large-scale WSI datasets with diagnostic labels or raw reports, data suitable for training language-driven, whole-slide MLLMs are scarce. Effective training data should include natural language dialogues grounded in slide content and clinical workflow decision points. Second, the gigapixel scale and complexity of WSIs present computational and informational challenges^{17,18}. Models must efficiently handle whole-slide inputs while selectively integrating features relevant to a given clinical query, which remains beyond the reach of existing patch-level MLLMs.

In this Article, we present SlideChat, a vision–language assistant designed for end-to-end WSI analysis and interpretation at the slide level. To train SlideChat, we first construct SlideInstruction, the largest pathology WSI instruction-following dataset so far (Fig. 1a). Curated from WSI–report pairs¹⁹ from The Cancer Genome Atlas (TCGA) cohort²⁰, SlideInstruction comprises 274,233 instruction samples covering 24 anatomies, 31 cancer types and diverse formats (Fig. 1b). We then develop an effective MLLM-based architecture for SlideChat (Fig. 1c), integrating fine-grained patch-level histological features, global slide-level context and clinical questions in natural language, enabling interactive, context-aware reasoning for WSI analysis. SlideChat also exhibits question-guided interpretability, assigning high attention weights only to question-relevant regions in the input gigapixel-scale WSIs (Fig. 1c, right).

To evaluate an MLLM's capability on WSI analysis, we further introduce SlideBench, an expert-validated benchmark encompassing 27 pathological tasks across 33 cancer types and five cohorts, covering a wide range of real-world pathological scenarios in clinical practice. These tasks are in multiple question formats, including closed-ended visual question answering (VQA), open-ended VQA and report generation, with both in-domain and out-of-domain cases. SlideChat outperforms state-of-the-art MLLMs such as GPT-4o²¹, MedDr²², Quilt-LLaVA¹⁵ and LLaVA-Med²³, improving the overall accuracy by 19.1% on closed-ended questions and delivering the highest proportion of expert-preferred responses on open-ended questions with superior accuracy, relevance, completeness and clinical applicability. Compared with models specialized in report generation such as HistoGPT²⁴ and PRISM²⁵, SlideChat produces more coherent and higher-quality pathology reports, with a 7.7% improvement in the Metric for Evaluation of Translation with Explicit Ordering (METEOR) score. These results show SlideChat's effectiveness in facilitating the complete diagnostic process from initial assessment to final reporting at the whole-slide level.

Results

Overview of SlideChat

SlideChat consists of four key designs: a patch-level vision encoder, a slide-level vision encoder, a connector and an LLM (Fig. 1c; Methods). The workflow begins by tiling each WSI into non-overlapping patches, each encoded by the patch-level vision encoder⁸, distilling fine-grained cellular and tissue semantics into patch-level tokens. These tokens are then fed into a slide-level vision encoder²⁶, a transformer model with an efficient attention mechanism that captures dependencies across tokens and models global histological context without prohibitive computational cost. Following slide-level encoding, the connector maps these visual tokens into the input space of the LLM²⁷, ensuring seamless integration of visual and textual data. Concurrently, SlideChat processes natural language queries from the user (for example, 'What is the type of tumor in the image?'), converts them into textual tokens and integrates them with visual tokens from WSIs, providing accurate and contextually relevant diagnostic responses.

Using the SlideInstruction dataset (Fig. 1a,b; Methods), SlideChat is trained in two stages²⁸: cross-domain alignment and visual instruction learning. The first stage aligns visual features with the language space, enabling the LLM to interpret visual tokens. In this stage, SlideChat is

trained on 8,583 WSI–caption pairs, with the patch-level encoder and LLM frozen, updating only the slide-level encoder and connector. The second stage unlocks a domain-specific instruction following across diverse pathology tasks, such as describing the extent of tumor invasion or assessing the degree of cellular differentiation. This is achieved by further tuning the slide-level encoder, connector and LLM on 265,650 WSI question–answer (QA) pairs (Supplementary Tables 1 and 2). Examples of training data are shown in Extended Data Fig. 1a,b.

We evaluate SlideChat on the proposed SlideBench benchmark (Supplementary Table 3) across three question formats: closed-ended VQA, open-ended VQA and pathology report generation. These evaluations assess in-domain and out-of-domain diagnostic accuracy, nuanced reasoning and capacity for coherent and end-to-end clinical reporting. Together, SlideBench shows SlideChat's ability to deliver accurate, interpretable and clinically meaningful slide-level analysis, enabling integration of global context, which is inherently limited in existing patch-based MLLMs.

Evaluation on closed-ended question answering

We evaluated the foundational capability of SlideChat in pathology understanding using closed-ended questions via SlideBench-Closed, a multiple-choice VQA benchmark curated from five cohorts: TCGA²⁰, Early Breast Cancer Core-Needle Biopsy WSI (BCNB)²⁹, Clinical Proteomic Tumor Analysis Consortium (CPTAC)³⁰, HISTAI³¹ and an in-house Multi-Organ Pathology (MOP) dataset. TCGA questions assess in-domain performance (Supplementary Table 4), while BCNB, CPTAC, HISTAI and MOP evaluate out-of-domain generalization (Supplementary Tables 5 and 6). For each cohort, we used an LLM-assisted pipeline to extract and revise multiple-choice QA pairs from source pathology reports and clinical notes, which were then reviewed and verified by three board-certified pathologists (Fig. 2a; Methods).

For the TCGA cohort, cases cover 13 pathology tasks across 31 cancer types in three functional groups: 'microscopy' (for example, histopathological changes, tissue architecture), 'diagnosis' (for example, disease classification, staging) and 'clinical' (for example, treatment guidance, biomarker analysis). The BCNB cohort, originally designed for WSI-based classification tasks, was reformulated into multiple-choice questions on the perception and comprehension of estrogen receptor (ER) statuses, progesterone receptor (PR) statuses, human epidermal growth factor receptor 2 (HER2) statuses and tumor subtypes. The CPTAC cohort included cases for tumor subtype classification: uterine corpus endometrial carcinoma (UCEC), lung squamous cell carcinoma (LSCC), lung adenocarcinoma (LUAD) and cutaneous melanoma (CM). The HISTAI cohort focused on disease classification for breast, skin and colorectal tissues, distinguishing malignant from benign and other non-malignant conditions. The MOP cohort covered malignant-versus-benign disease classification across 11 routine diagnostic organs, representing a broad and diverse set of routine diagnostic cases.

Because existing MLLMs cannot directly process gigapixel slides, we compared SlideChat with four representative patch-level baseline models: GPT-4o²¹, a strong general-purpose vision–language model; MedDr²² and LLaVA-Med²³, two generalist medical MLLMs; and Quilt-LLaVA¹⁵, a pathology-focused MLLM. To adapt these MLLMs to the slide-level task, we applied clustering to all non-background patches in each WSI and selected 30 representative patches. These patches were individually fed into each MLLM, giving the final answer determined by majority voting across patch-level predictions.

Across 13 tasks in SlideBench-Closed (TCGA) (Fig. 2b), SlideChat significantly outperformed baseline models with an overall accuracy of 74.12% ($P < 0.0001$). In terms of absolute performance gains, SlideChat outperformed GPT-4o by +18.44%, MedDr by +25.15%, LLaVA-Med by +44.35% and Quilt-LLaVA by +48.50%, with all comparisons reaching statistical significance ($P < 0.0001$). These gains were particularly notable in tasks requiring fine-grained morphological understanding, for example, tissue architecture (SlideChat 83.60% versus next best 61.41%,

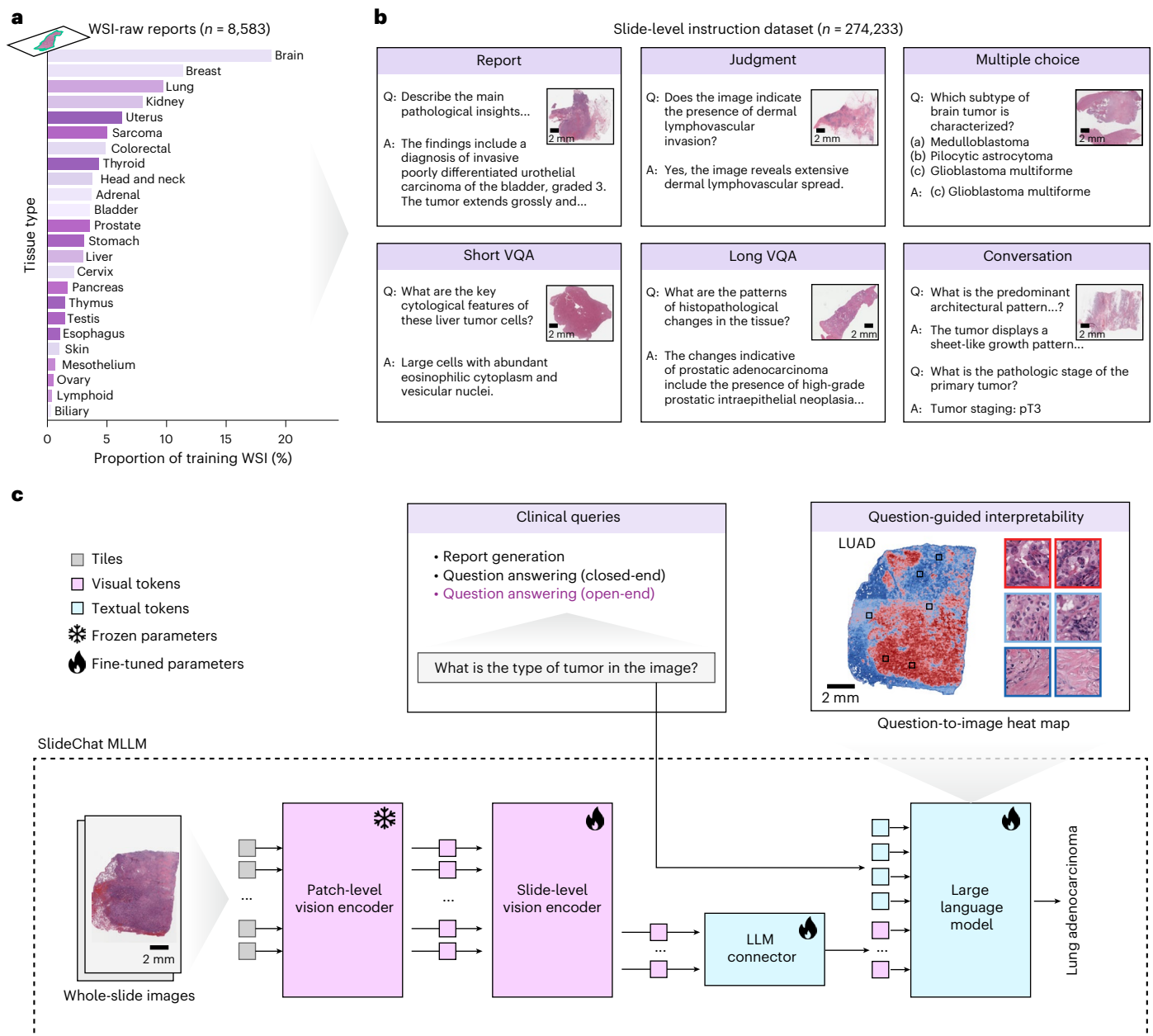


Fig. 1 | Overview of the SlideInstruction dataset and SlideChat architecture.

a, Bar chart of 8,583 WSIs paired with pathology reports used to curate the SlideInstruction dataset, grouped by 24 anatomical sites to illustrate dataset composition. **b**, Representative question formats drawn from the 274,233 samples in the SlideInstruction dataset, including pathological report generation, binary judgment, multiple-choice questions, short-form VQA, long-form VQA and multiturn conversation. **c**, SlideChat architecture with question-guided interpretability. A WSI is tiled and encoded at both patch and

slide levels, projected into the input space of the language model and integrated with a clinical query for multimodal slide-level understanding, responding to a broad range of queries in clinical workflows. Heat maps are computed based on question-to-image attention in SlideChat, revealing that clinical queries guide the model to attend to relevant image regions. This indicates that SlideChat acquires an implicit spatial understanding of WSI content, which supports transparency and interpretability in clinical applications.

+22.19) and histopathological changes (85.91% versus 53.34%, +32.57). SlideChat also excelled in integrative diagnostic tasks such as disease classification (78.12% versus 54.33%, +23.79) and staging (62.45% versus 57.53%, +4.92). For clinical tasks, SlideChat maintained clear advantage in biomarker analysis (73.84% versus 59.05%, +14.79) and prognostic assessment (69.46% versus 61.40%, +8.06). These results highlight the importance of modeling whole-slide context, which patch-level MLLMs often fail to capture. Moreover, across the microscopy, diagnosis and clinical groups, SlideChat achieved the highest accuracy in the majority of cancer types. Per-cancer performance comparisons are provided in Supplementary Tables 7–9.

These advantages also extend to out-of-domain cohorts (Fig. 2c). On BCNB, SlideChat achieved an average accuracy of 75.41% in molecular status prediction tasks (that is, ER, PR and HER2), outperforming the best baseline by +7.46. For tumor type classification, SlideChat achieved 89.42% (versus next best 56.04%, +33.38). On CPTAC, SlideChat led all four subtyping tasks: CM (48.67% versus 41.81%, +6.86), LSCC (68.05% versus 49.99%, +18.07), UCEC (48.20% versus 36.37%, +11.82) and LUAD (41.78% versus 37.00%, +4.775). On HISTAI, SlideChat achieved 79.53%, 73.58% and 69.53% for breast, skin and colorectal tissue disease classification, respectively, surpassing all baseline models. On the more challenging MOP cohort, SlideChat delivered consistent gains across

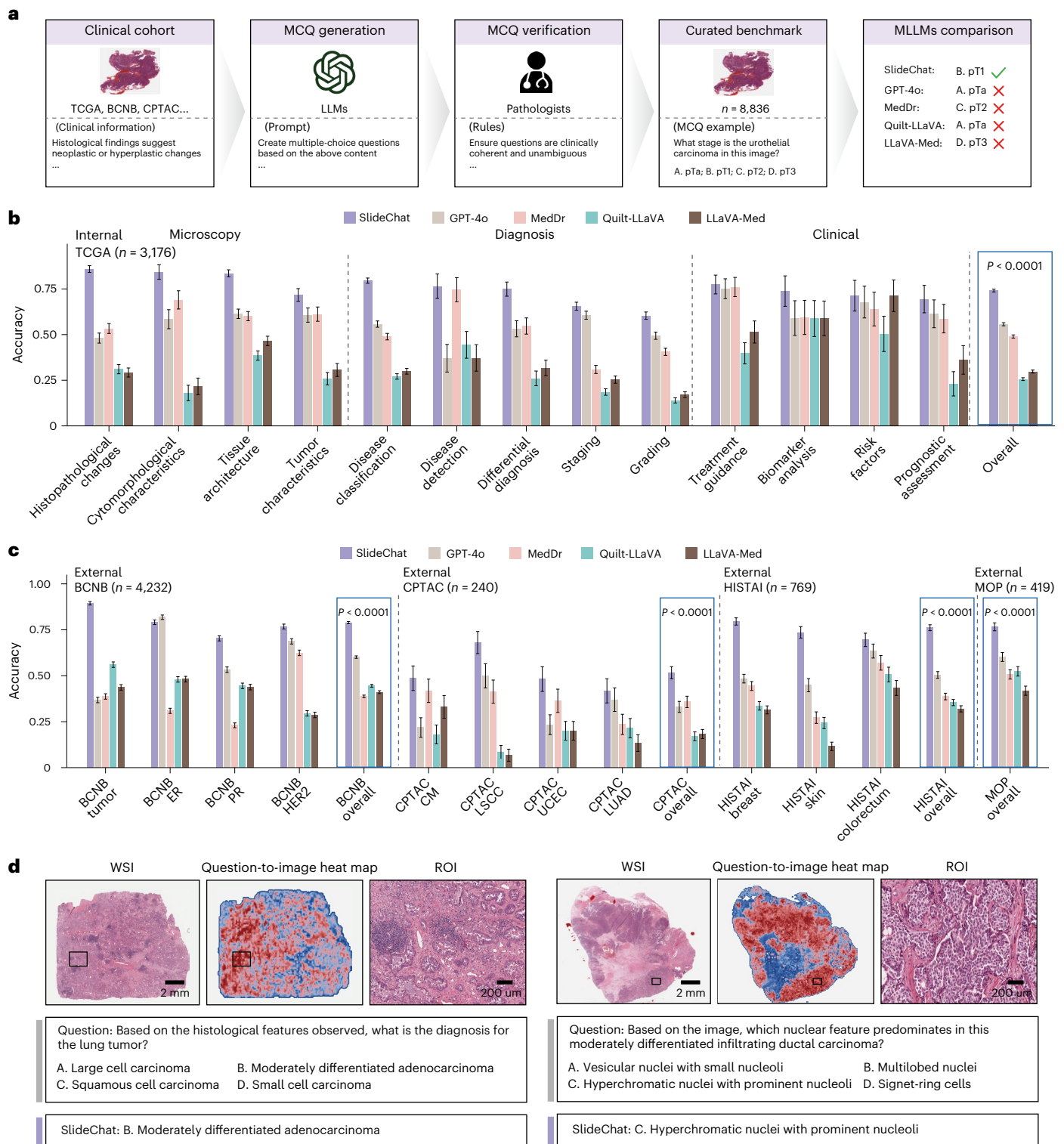


Fig. 2 | Evaluation of SlideChat on closed-ended questions. a, Pipeline for benchmark curation. Across multiple cohorts, candidate multiple-choice questions (MCQs) and answers were generated by LLMs using WSIs and associated annotations. These were then verified by board-certified pathologists to create the final SlideBench-Closed benchmark. Performance was compared across SlideChat and four MLLMs. **b**, In-domain evaluation on the TCGA cohort ($n = 3,176$ samples). SlideChat consistently outperformed the baselines on 13 different pathological tasks. Data are presented as mean values $\pm$ s.d. across 1,000 bootstrap replicates. For the overall comparison, statistical significance between SlideChat and the best-performing baseline was assessed using two-sided Wilcoxon signed-rank tests with Benjamini–Hochberg FDR correction;

$P = 2.28 \times 10^{-60}$. **c**, Out-of-domain generalization evaluation on closed-ended questions. SlideChat achieved significantly higher overall accuracies on external cohorts, including BCNB ($n = 4,232$ samples, $P = 1.26 \times 10^{-70}$), CPTAC ($n = 240$ samples, $P = 2.02 \times 10^{-5}$), HISTAI ($n = 769$ samples, $P = 6.8 \times 10^{-60}$) and MOP ($n = 419$ samples, $P = 1.24 \times 10^{-54}$) cohorts. Data are presented as mean values $\pm$ s.d. across 1,000 bootstrap replicates. **d**, Question-guided interpretability analysis. Two case studies showed that SlideChat attended to diagnostically salient regions when inferring the correct answer, highlighting invasive glands in lung adenocarcinoma and nuclear atypia in breast ductal carcinoma. These maps represent the spatial weighting of features during inference. ROI, region of interest.

categories and reached an overall accuracy of 76.63% (versus 60.23% for GPT-4o and 50.71% for MedDr). Despite the heterogeneity of tissue types and staining protocols, SlideChat maintained robustness and generalization when applied to these unselected, real-world WSIs (Supplementary Table 10). Our analysis also reveals that adapting existing patch-level models to WSI input via downsampling or multipatch aggregation compromises either cellular detail or global context and that a native slide-level model like SlideChat addresses this key trade-off in WSI analysis by better preserving both (Extended Data Fig. 2).

To further assess model interpretability, we visualized SlideChat's attention heat maps over WSIs when addressing specific queries (Fig. 2d; Methods). We showed two cases illustrating complementary reasoning pathways: one requiring diagnosis from observed features and the other requiring feature identification given a known diagnosis. In the first case, SlideChat correctly inferred 'moderately differentiated adenocarcinoma', focusing on the irregular glandular structures and cytologic atypia. In the second case, SlideChat attended to tumor nests and identified 'hyperchromatic nuclei with prominent nucleoli' as the defining feature for moderately differentiated infiltrating ductal carcinoma. In these examples, SlideChat not only arrived at the correct answers but also did so by attending to diagnostically meaningful regions of interest, underscoring the value of transparent, pathology-aligned whole-slide modeling.

Although attention heat maps reflect token weighting rather than strict causality, our further qualitative analysis shows that SlideChat's attention is not a static map, showing context-aware reasoning. For the same WSI, SlideChat adaptively shifts its focus with queries (Extended Data Fig. 3a–f). For example, in the breast carcinoma case, when asked about the 'architectural pattern' (Extended Data Fig. 3d), SlideChat focuses broadly on the irregular glandular formations. However, when the question shifts to 'changes in the surrounding stroma' (Extended Data Fig. 3e), the heat map re-centers on the dense fibrous regions separating the tumor glands.

Evaluation on open-ended question answering

Building on the closed-ended evaluation, we next examined SlideChat's capacity to handle more nuanced, complex diagnostic scenarios through open-ended questions. To support this, we developed SlideBench-Open, a benchmark comprising 129 pathologist-curated questions requiring detailed free-text interpretations that reflect real-world decision-making and require deep pathological understanding.

For each question, SlideChat and baseline models generated free-form textual responses, independently scored by a panel of three experienced pathologists (Fig. 3a). To adapt baseline MLLMs to the whole-slide level, we extracted 30 representative patches per WSI, fed them individually into each patch-level MLLM and synthesized the outputs into a single textual answer using GPT-4 (ref. 32). Evaluation was performed across five criteria (Supplementary Table 11): accuracy (factual correctness), relevance (alignment with the question), completeness (coverage of key diagnostic information), output clarity (readability and structured presentation) and applicability (potential to support clinical decision-making), scored on a five-point scale from 'insufficient' (1) to 'excellent' (5). Model identities were blinded to the pathologists to reduce bias.

SlideChat showed consistent advantages across all criteria (Fig. 3b and Supplementary Table 12), with the highest proportion of top scores ('excellent', rated 5). As a positive control in the blinded evaluation, reference answers scored 4.68, assessed by the same panel of pathologists under identical criteria. SlideChat achieved an average score of 4.26, approaching the reference (4.68) closely and outperforming GPT-4o (3.75), Quilt-LLaVA (3.55), LLaVA-Med (3.44) and MedDr (3.19). Specifically, SlideChat excelled in accuracy, relevance, output clarity and applicability, surpassing the strongest baseline, GPT-4o. In terms of completeness, SlideChat achieved notably greater number of 'strong' and

'excellent' ratings (300) than the GPT-4o (268). These results are further supported by average expert scores (Extended Data Fig. 4a), where SlideChat exceeded 4.0 on most criteria, while other models remained below this threshold. Interexpert analysis (Fig. 3c) showed moderate-to-good consistency (intraclass correlation coefficient) among the pathologists, with a quadratic-weighted Fleiss' kappa of 0.747. Furthermore, the pairwise quadratic-weighted Cohen's kappa values (expert 1 versus expert 2, 0.638; expert 1 versus expert 3, 0.604; expert 2 versus expert 3, 0.623) fell predominantly in the range of 'substantial agreement'³³, validating the consistency and robustness of the scoring protocol.

Two representative cases illustrate SlideChat's performance (Fig. 3d). In the first case of structural analysis in tumor-infiltrated tissues, SlideChat accurately described disrupted adipose and fibrous tissues, desmoplastic stroma, focal necrosis and relevant cellular features. Baseline models, however, offered less precise or incomplete responses. For instance, GPT-4o provided a broader description mentioning infiltration into fibroadipose tissue without noting the specific stromal or cellular details; MedDr described general disorganization without mentioning stromal or necrotic components; Quilt-LLaVA also missed cellular details; and LLaVA-Med inaccurately described minimal alteration. In the second case of cancer detection, SlideChat correctly identified clear cell renal cell carcinoma (CCRCC) with cellular features, while baseline models like MedDr denied any evidence of malignancy. These results highlighted SlideChat's ability to deliver accurate, comprehensive, clear and clinically relevant responses to open-ended pathology questions, underlining its potential for effective integration into diagnostic and clinical decision-making workflows. These observations are further supported by representative case comparison and question-guided attention visualizations (Extended Data Fig. 4b,c).

Evaluation on report generation

Extending beyond question answering, we further evaluated SlideChat's ability to generate comprehensive pathology reports from WSIs, directly assisting in routine diagnostic documentation. To this end, we prepared 3,149 high-quality reference reports from the TCGA, CPTAC, HISTAI and MOP cohorts (Supplementary Tables 6 and 13). Reference reports from the TCGA and HISTAI cohorts were constructed by extracting and standardizing diagnostically relevant content from their public annotations. CPTAC reports were authored in collaboration with experienced pathologists, ensuring detailed microscopic observations and diagnostic clarity. MOP reports were derived from routine clinical documentation.

We compared SlideChat against two open-source pathology report-generation models, HistoGPT²⁴ and PRISM²⁵, using the following metrics: METEOR³⁴, Recall-Oriented Understudy for Gisting Evaluation (ROUGE)³⁵, Bilingual Evaluation Understudy (BLEU)³⁶ and GPT score. GPT score assesses semantic similarity between generated and reference reports on a scale from 0 to 1 using GPT-4, with higher scores indicating greater similarity and clinical relevance. As each metric captures a different aspect of report quality, all four should be jointly considered when evaluating these generative models, especially GPT score, which more directly reflects factual diagnosis accuracy in pathology narratives. Across the in-domain TCGA and out-of-domain CPTAC, HISTAI and MOP cohorts, SlideChat consistently outperformed baselines in METEOR, ROUGE and GPT score (Fig. 4a and Supplementary Table 14). SlideChat achieved superior METEOR scores across diverse cancer types (Fig. 4b and Supplementary Table 15), especially for cancers requiring intricate morphological descriptions. Consistent improvements in ROUGE, BLEU and GPT score (Supplementary Tables 16–18) further indicated SlideChat's improved phrasing, content coverage and accuracy. We further report the results on two skin-specific datasets (TCGA-SKCM and HISTAI-Skin), enabling a direct comparison with the dermatology-specialized model HistoGPT (Fig. 4c).

To assess clinical utility, we conducted a prospective reader study involving two junior and two senior pathologists, with less than and



Fig. 3 | Evaluation of SlideChat on open-ended questions. **a**, Overview of SlideBench-Open evaluation protocol ($n = 129$ samples). Responses from the MLLMs were anonymized, shuffled and independently rated by three expert pathologists on a five-point scale across five criteria: accuracy, relevance, completeness, output clarity and clinical applicability. **b**, Distribution of scores for each model across five criteria. SlideChat delivered the most 'excellent' (score 5) responses and the fewest 'insufficient' (score 1) responses. **c**, ICC(2,1) (intraclass correlation coefficient, two-way random effects, single rater)

confirms the interpathologist agreement on the scoring. **d**, Representative case examples for qualitative performance comparison. Clinically correct details are highlighted in cyan, inaccurate findings are marked in magenta, and reference answers are highlighted in gray. SlideChat provided detailed and clinically grounded answers closely matching expert references, while other models frequently produced incomplete, ambiguous or less accurate responses. Illustration of pathologists in panel **a** created in BioRender; Chen, Y. <https://biorender.com/fhfegvb> (2026).

more than 5 years of experience, respectively. The study used a paired design across 36 WSIs from five representative cancer types (breast, colorectal, lung, prostate and esophageal), comparing diagnostic

workflows with and without SlideChat assistance. First, SlideChat consistently improved the overall quality of diagnostic summaries (Fig. 4d). GPT scores of pathologist-written summaries increased for

both junior (from 5.94–6.33 to 6.33–7.06) and senior pathologists (from 5.67–6.67 to 6.39–7.56). Second, we performed structured evaluation by extracting key diagnostic entities from written summaries and comparing them with reference annotations. This revealed increased entity-level correctness under model guidance (Fig. 4e), confirming that SlideChat aids in identifying clinically meaningful morphological features. Third, SlideChat improved workflow efficiency, in particular among less experienced readers. Junior pathologists reduced average review time by approximately 18% (from 307.28 to 248.83 s per slide) without compromising rigor (Fig. 4f). Finally, SlideChat acted as a standardizing force and enhanced interreader consistency. Model guidance raised the median interreader agreement from 0.72 to 0.82 (Fig. 4g). These results show SlideChat's potential to enhance diagnostic fidelity, efficiency and consensus in routine practice.

In a representative case (Fig. 4h), SlideChat accurately captured key pathological features, including tumor classification, margin status, invasion patterns and contextual changes. By contrast, HistoGPT, which was developed for dermatopathology, exhibited limited generalizability and frequently generated inaccurate responses for non-skin cancers, despite plausibly high BLEU scores on CPTAC. PRISM produced overly brief and underinformative reports, omitting critical diagnostic details. These comparisons underscored SlideChat's strength in synthesizing visual information into coherent and detailed reports, showing its potential for general-purpose pathology report generation.

Demonstration on diverse use cases

To illustrate SlideChat's potential in real-world practice, we constructed multiturn dialogues that emulate the sequential steps of clinical slide interpretation, including visual assessment, diagnostic classification, prognostic evaluation and therapeutic planning. These cases (Fig. 5) showcased how SlideChat generates coherent, clinically grounded responses at each step. In a cervical biopsy (Fig. 5a), SlideChat began by summarizing the overall tissue architecture, identifying irregular epithelial nests with desmoplastic stroma, along with atypical mitoses and keratinization, capturing key morphological features required for initiating diagnostic consideration. Next, in a liver resection case (Fig. 5b), SlideChat correctly diagnosed hepatocellular carcinoma and supported its conclusion by describing classical histological patterns such as thickened trabeculae, pseudo-acinar formations and centrally located nuclei with prominent nucleoli. In a lung cancer case (Fig. 5c), SlideChat translated morphological findings into prognostic assessment, noting the absence of vascular

invasion and inferring a less aggressive phenotype. In an adrenal tumor case (Fig. 5d), SlideChat integrated morphologic and clinical context, recognized the localized pheochromocytoma and recommended complete laparoscopic adrenalectomy for treatment. Additionally, SlideChat outlined key perioperative strategies to mitigate cardiovascular risks. Together, these cases highlighted SlideChat's end-to-end interpretative workflow, underscoring its promise as an assistant in routine pathology practice.

Beyond qualitative cases, we evaluated SlideChat in multiturn conversations along three key dimensions: contextual accuracy gain, cross-turn inconsistency and self-contradictory hallucinations. We constructed a 'microscopic–diagnosis–clinical' multiturn test set with 66 conversations and 346 questions from SlideBench-Closed (TCGA), using a hybrid query format: 173 questions were maintained as closed-ended questions for accuracy benchmarking, while the rest were converted into open-ended queries to evaluate generation quality. SlideChat leveraged conversational context to improve accuracy from 72.83% (single-turn) to 74.57% (multiturn) (Extended Data Fig. 5a). It maintained logical consistency across turns in 83.27% of cases (Extended Data Fig. 5b), and 87.87% of responses were free from self-contradictory hallucinations (Extended Data Fig. 5c). The remaining errors mainly reflected difficulties in tracking fine-grained attributes, such as tumor–node–metastasis staging, over extended conversational contexts (Extended Data Fig. 5d), as well as in handling complex descriptive tasks involving mutually exclusive pathological features (Extended Data Fig. 5e). These findings show effective context-aware reasoning while indicating areas for future improvement in long-context consistency.

Discussion

The emergence of MLLMs, exemplified by generalist AI models like ChatGPT²¹ and its successors, has fundamentally transformed how humans interact with AI across scientific domains. These models use natural language as a unified interface for both receiving flexible user instructions and generating outputs of varying complexity across diverse analytical tasks. In the pathology domain, this technological advance unlocks transformative potential from medical education and scientific discovery to clinical decision support. While recent pathology foundation models (for example, CONCH⁸, PathAlign⁴, TITAN³⁷ and CHIEF¹⁰) have advanced visual representation learning for downstream tasks, they lack the capability for direct natural language interaction. Emerging MLLMs (for example, PathChat¹⁴ and Quilt-LLaVA¹⁵) enable

Fig. 4 | Evaluation of SlideChat on report generation. **a**, Quantitative comparison of SlideChat against pathology report-generation models (HistoGPT and PRISM) using METEOR, ROUGE, BLEU and GPT score. SlideChat consistently achieved higher scores across the TCGA ($n = 1,547$ samples), CPTAC ($n = 38$ samples), HISTAI ($n = 1,145$ samples) and MOP ($n = 419$ samples) cohorts. For GPT score, P values from two-sided Wilcoxon signed-rank tests with Benjamini–Hochberg FDR correction were TCGA, $P = 3.86 \times 10^{-198}$; CPTAC, $P = 0.0118$; HISTAI, $P = 1.36 \times 10^{-56}$; and MOP, $P = 4.14 \times 10^{-54}$. **b**, Report-generation comparison of different models using METEOR scores across diverse cancer types. Each axis represents a cancer type and is independently normalized to the performance range for visualization. SKCM, skin cutaneous melanoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; LUSC, lung squamous cell carcinoma; MESO, mesothelioma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SARC, sarcoma; STAD, stomach adenocarcinoma; TGCT, testicular germ cell tumors; THCA, thyroid carcinoma; THYM, thymoma; UCS, uterine carcinosarcoma; ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; DLBC, lymphoid neoplasm diffuse large B-cell lymphoma; ESCA, esophageal

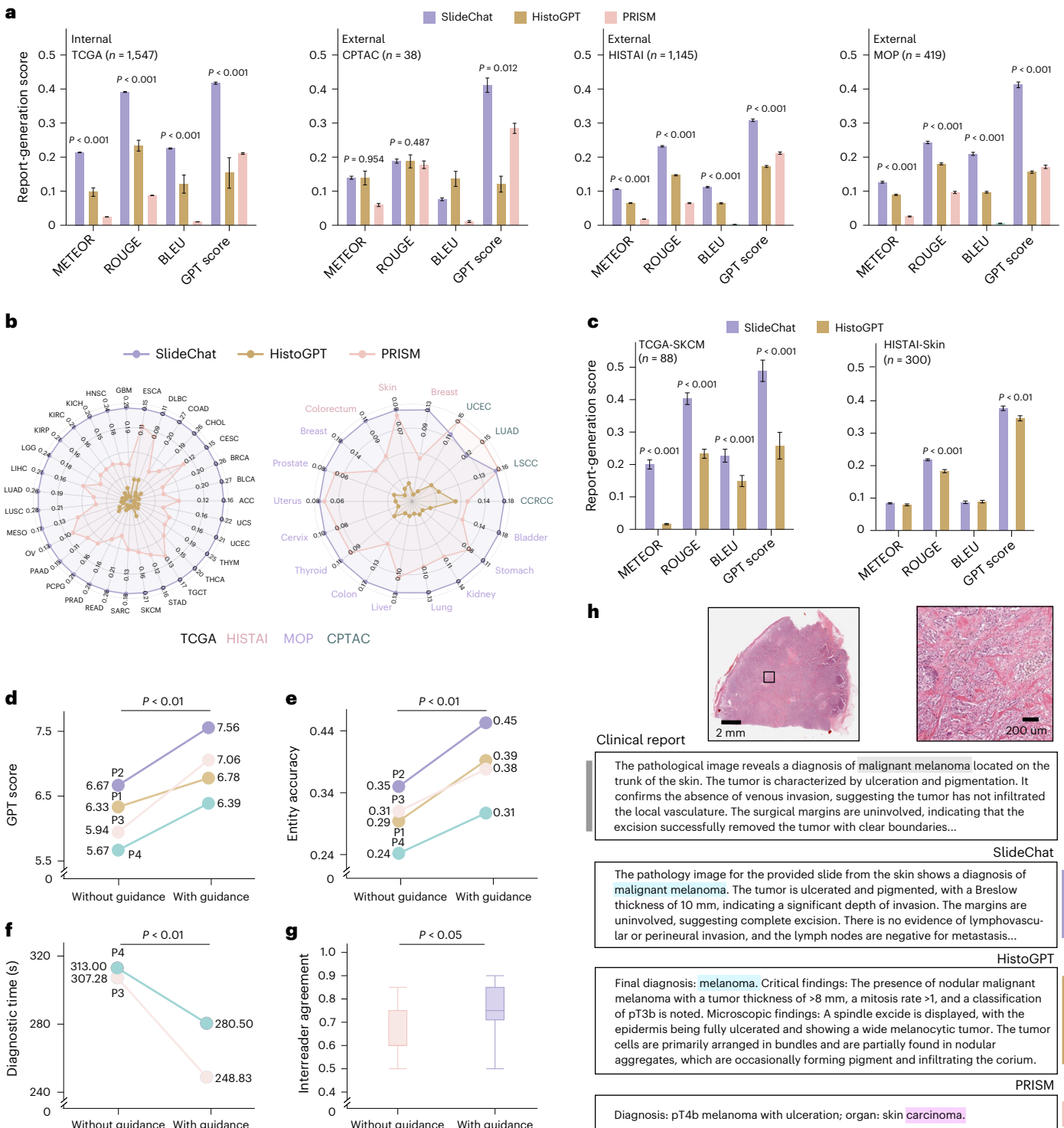
carcinoma. **c**, Report-generation comparison between SlideChat and HistoGPT on skin-related datasets. Evaluation was performed on the TCGA-SKCM ($n = 88$ samples) and HISTAI-Skin ($n = 300$ samples) datasets. For GPT score, P values from two-sided Wilcoxon signed-rank tests were TCGA, $P = 5.23 \times 10^{-4}$, and HISTAI, $P = 0.0039$. Data in **a** and **c** are presented as mean values $\pm$ s.d. across 1,000 bootstrap replicates. **d–g**, Four pathologists reviewed 18 cases with and without SlideChat-generated diagnostic reports in a prospective reader study. Each color in **d–f** represents a different individual. **d**, Diagnostic summaries written with SlideChat guidance achieved higher GPT scores across both junior and senior pathologists compared to unguided review. Statistical significance was assessed using two-sided Mann–Whitney U test; $P = 0.0002$. The same test applies to **e–g**. **e**, The match accuracy of diagnostic entities extracted from clinician-written reports also improved significantly when clinicians had access to SlideChat-generated drafts ($P = 0.006$). **f**, Junior pathologists demonstrated increased efficiency under SlideChat guidance ($P = 0.005$). **g**, Interreader agreement improved from a median of 0.72 to 0.82 with SlideChat assistance, indicating more consistent diagnostic interpretation among pathologists. Box plots show the median center line, first and third quartile box limits and minimum-to-maximum whiskers, with 18 independent measurements per group ($P = 0.045$). **h**, Representative examples comparing reports generated by SlideChat, HistoGPT and PRISM. Correct details are highlighted in cyan, inaccurate findings are marked in magenta, and reference answers are highlighted in gray. SlideChat better captured key morphological and diagnostic features.

multimodal dialogue but remain restricted to patch-level inputs, requiring manual selection of regions of interest.

In this work, we provide a preliminary demonstration of the AI assistant for whole-slide pathology analysis, showing that MLLM can now adapt at the gigapixel scale and complexity required for clinical practice. We present what is, to our knowledge, the most comprehensive evaluation of multimodal AI capabilities for computational pathology, comparing SlideChat against both leading general-purpose and specialized biomedical models across an unprecedented range of pathological tasks. Our evaluation encompasses diverse clinical scenarios, from diagnostic classification to comprehensive pathology report generation, validated through expert pathologist review

and assessed using multiple question formats to ensure broad clinical relevance and robustness. Across these evaluations, SlideChat showed superior performance over existing state-of-the-art solutions, achieving improvements in diagnostic accuracy and providing more clinically pertinent and comprehensive responses. Furthermore, its ability to engage in multiturn diagnostic dialogues with traceable decision process shown by question-guided attention heat maps further underscores its potential to reliably support complex workflows in clinical pathology with transparent AI assistance.

As MLLMs become increasingly integrated into digital pathology workflows, the broader impacts of a whole-slide conversational assistant like SlideChat may be substantial. For clinical practice, it



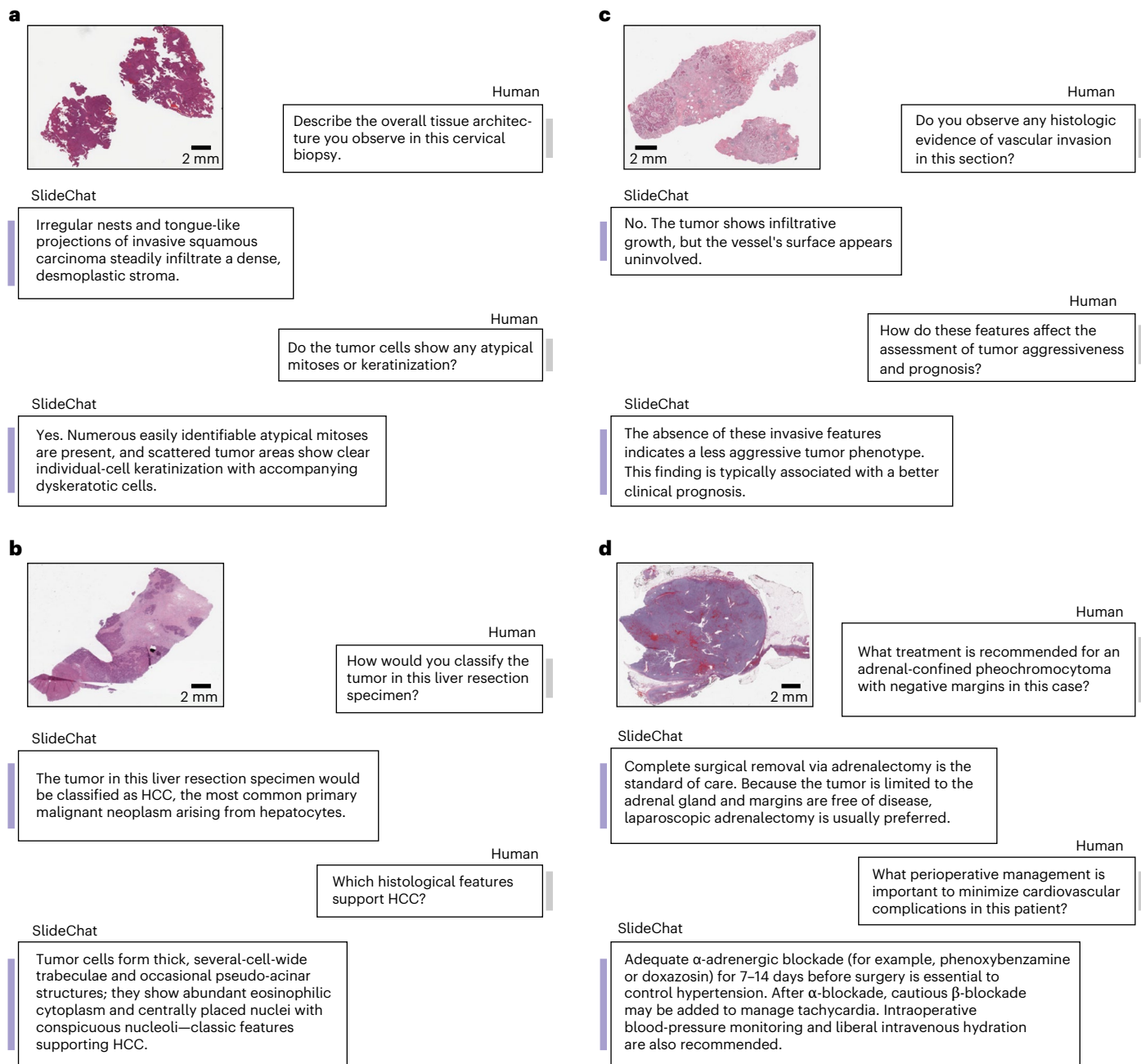


Fig. 5 | Exploring use cases of SlideChat. **a–d**, These cases showcase SlideChat's ability to follow core pathology workflows, progressing from morphological assessment and diagnostic classification to prognostic evaluation and clinical recommendation. **a**, SlideChat analyzes a cervical biopsy and accurately identifies invasive squamous carcinoma with ulceration, describing both architectural and cytologic features such as atypical mitoses and keratinization. **b**, When presented with a liver resection specimen, SlideChat correctly classifies the tumor as

hepatocellular carcinoma (HCC) and highlights histological hallmarks including trabecular architecture, eosinophilic cytoplasm and prominent nucleoli. **c**, SlideChat recognizes the absence of vascular invasion in a lung cancer case and appropriately links these features to a less aggressive tumor phenotype and better prognosis. **d**, For an adrenal-confined pheochromocytoma, SlideChat recommends standard laparoscopic adrenalectomy and details perioperative management strategies to reduce cardiovascular risk.

represents a step toward an interactive 'co-pilot'¹⁴ for pathologists, accelerating diagnostic workflows, reducing interobserver variability, aiding in the formulation of second opinions and standardizing the quality of pathology reports. Its question-answering capabilities could provide rapid verification for WSI features or classifications. For education, SlideChat offers a dynamic learning tool where trainees can explore slides interactively, ask specific questions and receive instant, context-rich feedback that adapts to their learning needs. For research, it opens new possibilities for large-scale automated annotation of histopathology archives, hypothesis generation

from morphological patterns and the accelerated discovery of new pathological markers.

Despite these advances, several limitations remain. SlideChat's performance is inherently influenced by the diversity and representativeness of its training data. Potential biases in SlideInstruction, such as institution-specific diagnostic practices or population demographics, may limit universal generalizability across different clinical settings and patient populations. While current results are promising, further prospective validation in clinical settings is needed to fully establish trust and safety for patient care. Moreover, mitigating hallucinations

in generative AI systems, where model outputs may appear plausible but are factually incorrect, remains a priority for future refinement.

Looking forward, several directions can further enhance the utility of SlideChat. A primary direction is enabling multislide reasoning, which would allow comprehensive analysis of multiple WSIs from a single case, facilitating crucial applications such as differentiating between primary tumor and metastatic lesions and tracking treatment responses over time by comparing pre- and posttreatment specimens. Furthermore, incorporating reinforcement learning from pathologist feedback³⁸ will align model outputs more closely with clinical expertise, bolstering clinical trust and accuracy. This can be complemented by integrating up-to-date knowledge from clinical guidelines and institutional protocols via retrieval-augmented generation³⁹, ensuring its outputs remain current and aligned with best practices. Finally, enhancing SlideChat with tool-using capabilities to invoke specialized AI models will broaden its application, such as enabling precise cell counting, segmentation or object localization in quantitative pathology. These advancements will steer the evolution of SlideChat from a slide-level assistant toward a comprehensive, case-level partner in pathology diagnostics and patient care.

Methods

Ethics statement

All research complied with relevant ethical regulations. The use of the retrospective MOP cohort data was approved by the Medical Ethics Committee of Eastern Hepatobiliary Surgery Hospital, Naval Medical University (approval number EHBHXY2025-K042-P001). The requirement for informed consent for the MOP cohort was waived by the Medical Ethics Committee of Eastern Hepatobiliary Surgery Hospital because the cohort comprised retrospectively collected de-identified data. The remaining datasets (TCGA, CPTAC, BCNB and HISTAI) are publicly available and were used in accordance with their access and usage policies. Sex and gender information were not collected or used in model training or evaluation; therefore, no sex- or gender-based analyses were performed.

Curation of SlideInstruction

To develop SlideChat, we curated SlideInstruction, a large-scale instruction-following dataset derived from WSI-report associations in the TCGA²⁰ cohort. These include (1) one report linked to multiple WSIs and (2) one-to-one WSI-report pairs. For the former, we retained all WSIs to introduce controlled label noise, improving robustness under real-world variability. For the latter, 80% were randomly selected for SlideInstruction and 20% reserved for benchmarking in SlideBench. In total, 8,583 WSIs with reports were included. Raw reports were cleaned using GPT-4 to remove noise, administrative metadata and redundancy, producing concise, diagnostically focused text to construct WSI-caption and instruction data in SlideInstruction.

For WSI-caption data, we leveraged GPT-4 to distill each cleaned report into a concise, clear and clinically informative paragraph capturing key findings. These summaries provide high-quality slide-level textual representations for training SlideChat (Extended Data Fig. 1a). For instruction data, GPT-4 generated diverse QA pairs per WSI to emulate clinical workflows, grouped into 'microscopy', 'diagnosis' and 'clinical' (Extended Data Fig. 1b). The microscopy group covers histopathological changes, cytomorphology, tissue architecture and tumor characteristics, emphasizing morphological recognition. The diagnosis group includes disease detection, classification, grading, differential diagnosis and staging, requiring integrative reasoning. The clinical group comprises biomarker analysis, prognostic assessment, treatment guidance and risk factors, linking image features to clinical decisions.

Before full-scale curation, we conducted a human-in-the-loop validation. First, we prompted GPT-4 to generate an initial batch of approximately 5,000 instruction pairs, which were then audited by board-certified pathologists, revealing a 36% rejection rate owing to

insufficient visual dependency or incorrect task categorization. We then iteratively refined prompts by incorporating subcategory definitions (Supplementary Table 1) and enforcing visual understanding, reducing the rejection rate to less than 5% and ensuring that the final prompts (Extended Data Fig. 6) yielded clinically relevant instructions for full-scale data curation. To enhance topical diversity and reduce redundancy, multiple QA pairs were generated per WSI, followed by semantic deduplication using Sentence-BERT⁴⁰ embeddings with a cosine similarity threshold of 0.7. Consequently, SlideInstruction contains 8,583 WSI-caption pairs and 265,650 instruction-following samples, totaling 274,233 training instances and providing rich supervision for SlideChat training.

Curation of SlideBench

To evaluate SlideChat, we curated SlideBench, an expert-reviewed benchmark involving five cohorts, comprising SlideBench-Closed, SlideBench-Open and SlideBench-Report. Data from the TCGA cohort were drawn from the held-out 20% split, ensuring no overlap with SlideInstruction. Annotation and validation were performed by three board-certified pathologists (with 3, 6 and 8 years of experience). We provide the prompts for curating SlideInstruction and SlideBench via LLM in Extended Data Fig. 6.

SlideBench-Closed comprises 8,836 expert-reviewed multiple-choice questions from five cohorts to assess the closed-ended VQA. For TCGA, GPT-4 generated candidate questions, options and answers across 13 task categories. To ensure visual dependency, we applied LLM-based filtering⁴¹, excluding questions answerable by at least three of four text-only LLMs (that is, GPT-4, InternLM2-Chat-7B⁴², Qwen-7B-Chat⁴³ and DeepSeek-7B-Chat⁴⁴) without images. The remaining pairs were independently reviewed by pathologists. Examples of accepted and rejected items are shown in Extended Data Fig. 7. Each QA pair was verified to require visual inspection, ensure clinical clarity and correct task categorization; invalid items were revised or removed. This yielded 3,176 TCGA QA pairs.

The remaining cohorts were incorporated to test out-of-domain generalization. For BCNB, 4,232 QA pairs were derived from four established classification tasks: ER status, PR status, HER2 status and tumor subtype classification. CPTAC included 240 QA pairs on subtype identification across four cancer types: UCEC, LSCC, LUAD and CM. HISTAI contributed 769 QA pairs across breast, skin and colorectum, tasked with diagnosing distinct diseases. MOP provided 419 QA pairs derived from routine diagnostic cases in 11 organ systems, testing the ability to differentiate between malignant and benign organ-specific conditions.

SlideBench-Open was built from the TCGA cohort for open-ended VQA evaluation. Specifically, we first generated candidate questions from reports using GPT-4. Each candidate was reviewed by the pathologist panel to ensure that the question was clinically relevant, clearly worded and grounded in the slide. Irrelevant or ambiguous questions were discarded or revised for clarity. The final set contains 129 questions spanning the same 13 task categories, grouped into broader domains of microscopy, diagnosis and clinical decision-making.

SlideBench-Report evaluates free-text report generation. It includes 3,149 reports from TCGA ($n = 1,547$), CPTAC ($n = 38$), HISTAI ($n = 1,145$) and MOP ($n = 419$). TCGA reports were cleaned to retain diagnostic content. CPTAC reports were newly authored by pathologists based on WSI review. HISTAI reports were constructed from structured metadata, while MOP reports were derived from de-identified clinical records. These reports reflect expert-level diagnostic language, with emphasis on morphological specificity, terminological accuracy and clinical coherence, thus providing high-quality targets for evaluation in real-world settings.

SlideChat model architecture

SlideChat adopts a hierarchical vision-language architecture with four modules: (1) a patch-level vision encoder, (2) a slide-level vision encoder, (3) a connector and (4) an LLM.

The patch-level vision encoder is initialized from CONCH⁸, a ViT-B/16 (ref. 45) model pretrained on large-scale pathology image-text pairs using a CoCa-style loss⁴⁶. Given a WSI, we tile it into non-overlapping 224×224 patches, $X = [x_1, \dots, x_M]$, where M is the total number of patches, typically ranging within $[10^4, 10^5]$. Each patch is mapped by the patch-level encoder E_{patch} to an embedding with dimension $d_{\text{vis}} = 768$

$$\mathbf{p}_i = E_{\text{patch}}(x_i) \in \mathbb{R}^{d_{\text{vis}}}, \quad (1)$$

yielding a long sequence of patch-level tokens $P = [\mathbf{p}_1, \dots, \mathbf{p}_M] \in \mathbb{R}^{M \times d_{\text{vis}}}$. E_{patch} is frozen throughout the training. To capture global context without prohibitive computational cost, we follow LongNet²⁶ to implement our slide-level vision encoder E_{slide} , which consists of two transformer layers with dilated global attention. Token sequence P is projected to form the queries, keys and values; partitioned into segments; and processed in parallel via dilated attention. Outputs are concatenated to produce slide-level tokens with near-linear complexity:

$$S = E_{\text{slide}}(P) \in \mathbb{R}^{M \times d_{\text{vis}}}, \quad (2)$$

with $S = [\mathbf{s}_1, \dots, \mathbf{s}_M]$. Please refer to LongNet²⁶ for details. After the slide-level processing, we use a lightweight connector E_{conn} (that is, two multilayer perceptrons with Gaussian Error Linear Unit activation⁴⁷) to map each slide-level visual token $\mathbf{s}_i$ to language space, producing a language-aligned visual token:

$$\mathbf{v}_i = E_{\text{conn}}(\mathbf{s}_i) \in \mathbb{R}^{d_{\text{lang}}}, \quad (3)$$

forming a visual token sequence $V = [\mathbf{v}_1, \dots, \mathbf{v}_M] \in \mathbb{R}^{M \times d_{\text{lang}}}$. The LLM is Qwen2.5-7B-Instruct⁴⁸ with hidden dimension $d_{\text{lang}} = 4,096$.

During inference, visual tokens V are concatenated with textual tokens T to form the input multimodal token sequence for LLM, where $T = [\mathbf{t}_1, \dots, \mathbf{t}_L] \in \mathbb{R}^{L \times d_{\text{lang}}}$ contains L tokens embedded from the user instruction via byte-pair encoding tokenization⁴⁹. The decoder-only LLM autoregressively generates response tokens, with cross-modal fusion handled natively in its transformer layers.

SlideChat model training

SlideChat is trained in two stages: cross-domain alignment and visual instruction learning. The first stage aligns WSI-derived visual tokens with the LLM input space via a WSI-captioning task, enabling interpretation of slide-level representations, while the second stage focuses on visual question-answering task and enables accurate responses to domain-specific WSI queries. Both stages share a unified maximum-likelihood objective. Let f_{backbone} and f_{head} denote the LLM backbone and head, θ the trainable parameters in a specific stage and Y the full token sequence:

$$Y = \left[\underbrace{\mathbf{t}_1, \dots, \mathbf{t}_L}_{\text{instruction tokens}}, \underbrace{\mathbf{v}_1, \dots, \mathbf{v}_M}_{\text{visual tokens}}, \underbrace{\mathbf{a}_1, \dots, \mathbf{a}_K}_{\text{response tokens}} \right]. \quad (4)$$

Then, the training objective is given by:

$$\mathcal{L}(\theta) = - \sum_{k \in \mathcal{K}} \log p_{\theta}(y_k | y_{1:k-1}), \quad (5)$$

$$p_{\theta}(y_k | y_{1:k-1}) = \text{softmax}(f_{\text{head}}(\mathbf{h}_k)), \quad (6)$$

$$\mathbf{h}_k = f_{\text{backbone}}(Y_{1:k-1}), \quad (7)$$

where $\mathcal{K}$ is a set of textual token indices, and $\mathbf{h}_k$ denotes the LLM hidden states at the k th step. The full response token sequence, after byte-pair encoding de-tokenization, reconstructs the final human-readable response.

In the first stage, we use 8,583 WSI-caption pairs in SlideInstruction, where the user queries a summary of the WSI. Only the slide-level encoder and connector are updated. This stage runs for 3 epochs on eight A100 GPUs (80 GB) with a batch size of 1 per GPU, taking approximately 3 h, using AdamW⁵⁰ optimizer with a learning rate of 1×10^{-3} and no weight decay. In the second stage, the slide-level encoder, the connector and the LLM are jointly optimized using 265,650 QA pairs in SlideInstruction. This stage runs for 1 epoch with the same batch size and optimizer, using a learning rate of 2×10^{-5} and no weight decay. Total training time is approximately 24 h on the same hardware setup (Supplementary Table 19).

Evaluation of closed-ended VQA

We evaluated closed-ended VQA on SlideBench-Closed. Four strong MLLM baselines were included for comparison: GPT-4o²¹, MedDr²², Quilt-LLaVA¹⁵ and LLaVA-Med²³. As these models cannot directly process gigapixel WSIs, we adapted them using an independent patch voting strategy. Specifically, for each WSI, we applied K -means clustering to tissue patches and selected 30 representatives. Each patch-question pair was processed independently, producing 30 predictions, which were aggregated via majority voting into a slide-level answer. SlideChat outperformed these patch-level models (Extended Data Fig. 8a). We also evaluated two alternative adaptation strategies for patch-level baseline models: slide thumbnail input (resizing a WSI to a 512×512 thumbnail; Extended Data Fig. 8b) and joint multipatch input (feeding multiple patches jointly into the model; Extended Data Fig. 8c). Across all settings, patch-level models underperformed compared with SlideChat, which directly processes full WSIs and achieves higher accuracy. We further assessed the robustness of the patch voting strategy via sensitivity analysis over different patch numbers (10, 20, 30, 40) and selection methods (Extended Data Fig. 8d). Performance remained stable across patch counts, and K -means clustering-based selection consistently outperformed random sampling.

During evaluation, model outputs were matched to reference answers. For free-text responses, GPT-4 was used as an adjudicator to assess semantic consistency, ensuring fair comparison across formats. We also conducted five independent runs using different sampling seeds for SlideBench-Closed. The mean and 95% confidence intervals across runs are summarized in Supplementary Tables 20–24 and Extended Data Fig. 8e. The consistently narrow intervals show SlideChat's stable predictions.

Evaluation of open-ended VQA

We evaluated SlideChat's free-form diagnostic response on SlideBench-Open. For patch-level baselines (GPT-4o, MedDr, Quilt-LLaVA and LLaVA-Med), we follow the aforementioned independent patch voting strategy to adapt them to WSI input. By contrast, SlideChat jointly encodes the entire slide and question, allowing for more whole-slide-informed and coherent outputs (Extended Data Fig. 4b). All responses were anonymized, randomized and then evaluated by three board-certified pathologists independent of dataset construction. Each response was independently scored across five criteria (that is, accuracy, relevance, completeness, clarity and clinical applicability) on a five-point scale ranging from 1 (insufficient) to 5 (excellent). Criteria are defined in Supplementary Table 11, and interexpert agreement was quantified to ensure reliability.

Evaluation of pathology report generation

We used SlideBench-Report to evaluate SlideChat's capability to generate diagnostic reports. SlideChat was compared against HistoGPT²⁴ and PRISM²⁵, two specialized report-generation models. All models received identical WSIs and generated slide-level reports (Extended Data Fig. 9). Evaluation used standard natural language processing metrics, including BLEU³⁶, ROUGE-L³⁵ and METEOR³⁴. These metrics measure n-gram overlap but may not capture semantic quality. We further used GPT-4

as a judge to assess accuracy, completeness, clinical relevance and professional language, providing a more comprehensive evaluation. We additionally evaluated SlideChat over five independent runs using distinct sampling seeds for SlideBench-Report. Mean performance and 95% confidence intervals are reported in Supplementary Tables 25–28.

Ablation studies

Ablation studies on SlideBench-Closed (Extended Data Fig. 10a) validate key components. We first evaluated the necessity of slide-level encoder; removing it resulted in convergence failure during stage 1. When we tested an alternative variant without slide-level encoder using only stage 2 (row 1 in Extended Data Fig. 10a), the accuracy dropped significantly across the TCGA, BCNB and CPTAC cohorts compared with our default version (row 3), confirming the critical role of slide-level encoder in modeling gigapixel-scale context. We then assessed the training efficacy by comparing our two-stage strategy against joint single-stage training. The results showed that the two-stage approach (row 3) consistently outperformed the single-stage counterpart (row 2), showing that progressive alignment is critical for bridging visual and language modalities.

Computational efficiency

We report the computational resources required for SlideChat inference by measuring GPU memory and end-to-end runtime across different patch-token counts (Extended Data Fig. 10b,c). Both scale approximately linearly with input size. For WSIs with up to 20,000 patches, inference completes in 10 s using less than 20 GB memory, showing SlideChat's practical feasibility on standard hardware.

Analysis of potential performance bias

We implemented a class-balanced sampling strategy during SlideChat training to mitigate cancer type imbalance. To further examine whether such imbalance induces systematic performance bias, we analyzed the correlation between training sample size and accuracy across 31 cancer types in SlideBench-Closed. No significant correlation was observed (Spearman $\rho = 0.035$, $P = 0.824$) (Extended Data Fig. 10d), suggesting that SlideChat does not systematically favor high-prevalence cancer types.

Question-guided interpretability

To better understand how SlideChat uses visual information when answering specific clinical queries, we designed a question-guided interpretability method based on token-level attention analysis. Specifically, we identify WSI-derived visual tokens and aggregate their attention weights across all attention heads and layers in the LLM. Averaging these yields a single attention score per token, quantifying its contribution to the generated response. These scores are then mapped back to their spatial locations on the WSI, producing a heat map that highlights which regions SlideChat attended to most strongly in the context of the given question (Extended Data Figs. 3 and 4c). This provides intuitive insight into SlideChat's decision process and visual evidence supporting its responses.

Beyond attention maps, we conducted a patch-level perturbation experiment to validate SlideChat's dependence on meaningful visual evidence. For each WSI, we corrupted patches with the top 25% and bottom 25% attention scores using Gaussian noise, respectively, and test SlideChat's output. We observed that the former led to a significantly sharper performance decline compared to the latter (Extended Data Fig. 10e), confirming that SlideChat relies on informative visual regions when generating responses.

Comparison with representational foundation models

We distinguish SlideChat from current pathology foundation models such as CONCH⁸, TITAN³⁷, CHIEF¹⁰ and Prov-GigaPath⁶. These models function as visual encoders that require task-specific fine-tuning (for

example, training separate classifiers) to perform diagnostic tasks. By contrast, SlideChat is an interactive MLLM that interprets WSIs through natural language, addressing queries without task-specific adaptation. We compared SlideChat against these models on four BCNB classification tasks (Extended Data Fig. 10f). Foundation models were fine-tuned using fivefold cross-validation, while SlideChat was evaluated in a zero-shot conversational setting yet achieved comparable performance.

Analysis of the clinical fidelity of SlideInstruction data

To assess the clinical fidelity and potential bias in our SlideInstruction dataset, we curated two clinician-authored datasets for comparative benchmarking: SlideInstruction+ and SlideInstruction++. SlideInstruction+ comprises 436 QA pairs manually authored by four board-certified pathologists across 11 tumor types in TCGA, capturing real-world reasoning patterns. SlideInstruction++ scales this to 10,386 QA pairs by extracting ten high-frequency diagnostic categories (for example, morphology, subtyping and staging) from SlideInstruction+ and using clinician-written seed exemplars to guide LLM-based data curation, followed by manual review by experts to ensure factual accuracy and alignment with standard diagnostic criteria.

We compared SlideChat variants trained on these datasets against the default model trained on the original SlideInstruction dataset (Extended Data Fig. 10g). Across the SlideBench-Closed cohorts, adding clinician-authored instructions or clinician-verified synthetic expansions yielded consistent but modest improvements. This suggests that the original SlideInstruction dataset captures sufficient clinical signals for developing strong pathology models without the prohibitive cost of large-scale manual curation.

Statistics and reproducibility

This study was primarily based on retrospective and publicly available datasets; no randomization of patients, specimens or slide-level data pairs was performed. Randomization was applied only in the prospective reader study, in which the order of cases was randomized and the assisted and unassisted reading conditions were counterbalanced across readers. Blinding was used only in expert evaluation experiments: pathologists were blinded to model identity in the open-ended QA evaluation and the reader study. For dataset curation, model development and automated benchmark evaluation, blinding was not applicable or was not performed. No additional post hoc exclusions were applied when benchmarking. We filtered out tiles with low tissue occupancy as background during model development. Data distribution was assumed to be normal, but this was not formally tested. No statistical methods were used to predetermine sample sizes, but our sample sizes were similar to those reported in previous publications^{8,14,24}.

We estimate 95% confidence intervals using non-parametric bootstrapping⁵¹ with 1,000 replicates and assess performance differences between models using a two-sided Wilcoxon signed-rank test⁵², reporting the corresponding P values for statistical significance. Unless stated otherwise, we applied the Benjamini–Hochberg procedure to adjust P values for the false discovery rate (FDR) to control for multiple comparisons. Specifically, corrections were applied across task-level comparisons for the SlideBench-Closed and across organ-level comparisons for the SlideBench-Report benchmark.

Computing hardware and software

All experiments were implemented using Python (v3.10.15) and PyTorch (v2.4.0) with CUDA 12.1 support. The experiments are fully reproducible using the open-source libraries listed below. WSI preprocessing was performed using OpenSlide (v3.4.1) and openslide-python (v1.4.1), with each slide tiled into non-overlapping 224×224 patches at $\times 20$ magnification. Our model was developed based on the Xtuner framework (<https://github.com/InternLM/xtuner>). Training was conducted on eight NVIDIA A100 GPUs, and inference was carried out using a single

NVIDIA A100 GPU. To accelerate training, we used Flash Attention (v2.5.8) and DeepSpeed (v0.16.1). For baseline comparison, we used Faiss (v1.8.0) to perform K -means clustering for selecting representative patches. Pandas (v2.2.3) was used to perform random sampling during the non-parametric bootstrapping procedure (1,000 iterations) to estimate 95% confidence intervals for all reported metrics. Visualization of model outputs and evaluation results was performed using matplotlib (v3.9.3) and seaborn (v0.13.2).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The TCGA and CPTAC datasets, including whole-slide images and associated labels, are publicly available through the NIH Genomic Data Commons (<https://portal.gdc.cancer.gov>) and the Proteomic Data Commons (<https://proteomic.datacommons.cancer.gov>), respectively. The extracted TCGA pathology reports are available at <https://github.com/tatonetti-lab/tcga-path-reports>. The BCNB dataset, including whole-slide images and diagnostic labels, is available at <https://bcnb.grand-challenge.org/>. The HISTAI dataset, including metadata and whole-slide images, is available at <https://github.com/HistAI/HISTAI>. The SlideChat training and evaluation benchmark data are released at <https://huggingface.co/datasets/General-Medical-AI/SlideChat>. The MOP cohort was collected under a retrospective study approved by the Eastern Hepatobiliary Surgery Hospital (approval number EHBHXY2025-K042-P001). The MOP cohort is subject to controlled access owing to patient privacy obligations and institutional data governance requirements. Requests for access may be submitted to the corresponding author with a brief research plan and are subject to institutional approval and completion of a data use agreement. We aim to respond to all data access requests in 4 weeks. Data usage is restricted to non-commercial academic research purposes. Source data are provided with this paper.

Code availability

The source code for model training and inference is publicly available at <https://github.com/uni-medical/SlideChat>. The model weights are available at https://huggingface.co/General-Medical-AI/SlideChat_Weight.

References

- Bahadir, C. D. et al. Artificial intelligence applications in histopathology. *Nat. Rev. Electr. Eng.* **1**, 93–108 (2024).
- Van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nat. Med.* **27**, 775–784 (2021).
- Shmatko, A., Ghaffari Laleh, N., Gerstung, M. & Kather, J. N. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat. Cancer* **3**, 1026–1038 (2022).
- Ahmed, F. et al. PathAlign: a vision–language model for whole slide images in histopathology. In *Proc. MICCAI Workshop on Computational Pathology* (eds Ciompi, F. et al.) 72–108 (PMLR, 2024).
- Wang, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559 (2022).
- Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
- Vorontsov, E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* **30**, 2924–2935 (2024).
- Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).
- Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat. Med.* **29**, 2307–2316 (2023).
- Wang, X. et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 970–978 (2024).
- Xiang, J. et al. A vision–language foundation model for precision oncology. *Nature* **638**, 769–778 (2025).
- Bilal, M. et al. Foundation models in computational pathology: a review of challenges, opportunities, and impact. Preprint at <https://doi.org/10.48550/arXiv.2502.08333> (2025).
- Marra, A. et al. Artificial intelligence entering the pathology arena in oncology: current applications and future perspectives. *Ann. Oncol.* **36**, 712–725 (2025).
- Lu, M. Y. et al. A multimodal generative AI copilot for human pathology. *Nature* **634**, 466–473 (2024).
- Seyfioglu, M. S., Ikezogwo, W. O., Ghezloo, F., Krishna, R. & Shapiro, L. Quilt-LLaVA: visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (ed. Mortensen, E.) 13183–13192 (IEEE, 2024).
- Sun, Y. et al. PathAsst: a generative foundation AI assistant towards artificial general intelligence of pathology. In *Proc. Thirty-Eighth AAAI Conference on Artificial Intelligence* (eds Wooldridge, M. et al.) 5034–5042 (AAAI Press, 2024).
- Li, X. et al. A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. *Artif. Intell. Rev.* **55**, 4809–4878 (2022).
- Song, A. H. et al. Artificial intelligence for digital and computational pathology. *Nat. Rev. Bioeng.* **1**, 930–949 (2023).
- Kefeli, J. & Tatonetti, N. TCGA-Reports: a machine-readable pathology report resource for benchmarking text-based AI models. *Patterns* **5**, 100933 (2024).
- Hutter, C. & Zenklusen, J. C. The cancer genome atlas: creating lasting value beyond its data. *Cell* **173**, 283–285 (2018).
- Hurst, A. et al. GPT-4o system card. Preprint at <https://doi.org/10.48550/arXiv.2410.21276> (2024).
- He, S. et al. Towards generalizable AI in medicine via Generalist-Specialist Collaboration. *Nat. Biomed. Eng.* <https://doi.org/10.1038/s41551-026-01653-3> (2026).
- Li, C. et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. In *Proc. 37th International Conference on Neural Information Processing Systems* (eds Oh, A. et al.) 28541–28564 (Curran Associates, Inc., 2023).
- Tran, M. et al. Generating dermatopathology reports from gigapixel whole slide images with HistoGPT. *Nat. Commun.* **16**, 4886 (2025).
- Shaikovski, G. et al. PRISM: a multi-modal generative foundation model for slide-level histopathology. Preprint at <https://doi.org/10.48550/arXiv.2405.10254> (2024).
- Ding, J. et al. LongNet: scaling transformers to 1,000,000,000 tokens. Preprint at <https://doi.org/10.48550/arXiv.2307.02486> (2023).
- Yang, A. et al. Qwen2 technical report. Preprint at <https://doi.org/10.48550/arXiv.2407.10671> (2024).
- Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. In *Proc. 37th International Conference on Neural Information Processing Systems* (eds Oh, A. et al.) 34892–34916 (Curran Associates, Inc., 2023).
- Xu, F. et al. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Front. Oncol.* **11**, 759007 (2021).
- Edwards, N. J. et al. The CPTAC Data Portal: a resource for cancer proteomics research. *J. Proteome Res.* **14**, 2707–2713 (2015).

31. Nechaev, D., Pchel'nikov, A. & Ivanova, E. HISTAI: an open-source, large-scale whole slide image dataset for computational pathology. Preprint at <https://doi.org/10.48550/arXiv.2505.12120> (2025).
32. Achiam, J. et al. GPT-4 technical report. Preprint at <https://doi.org/10.48550/arXiv.2303.08774> (2023).
33. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).
34. Banerjee, S. & Lavie, A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (eds Goldstein, J. et al.) 65–72 (Association for Computational Linguistics, 2005).
35. Lin, C.-Y. ROUGE: a package for automatic evaluation of summaries. In *Text Summarization Branches Out* 74–81 (Association for Computational Linguistics, 2004).
36. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics* (eds Isabelle, P. et al.) 311–318 (Association for Computational Linguistics, 2002).
37. Ding, T. et al. A multimodal whole-slide foundation model for pathology. *Nat. Med.* **31**, 3749–3761 (2024).
38. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
39. Gao, Y. et al. Retrieval-augmented generation for large language models: a survey. Preprint at <https://doi.org/10.48550/arXiv.2312.10997> (2023).
40. Reimers, N. & Gurevych, I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing* (eds Inui, K. et al.) 3982–3992 (Association for Computational Linguistics, 2019).
41. Sun, Y. et al. PathMMU: a massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision* (eds Leonardi, A. et al.) 56–73 (Springer, 2024).
42. Cai, Z. et al. InternLM2 technical report. Preprint at <https://doi.org/10.48550/arXiv.2403.17297> (2024).
43. Bai, J. et al. Qwen technical report. Preprint at <https://doi.org/10.48550/arXiv.2309.16609> (2023).
44. Bi, X. et al. DeepSeek LLM: scaling open-source language models with longtermism. Preprint at <https://doi.org/10.48550/arXiv.2401.02954> (2024).
45. Dosovitskiy, A. et al. An image is worth 16 × 16 words: transformers for image recognition at scale. In *Proc. 9th International Conference on Learning Representations* 1–21 (ICLR, 2021); <https://openreview.net/forum?id=YicbFdNTTy>
46. Yu, J. et al. CoCa: contrastive captioners are image-text foundation models. Preprint at <https://doi.org/10.48550/arXiv.2205.01917> (2022).
47. Hendrycks, D. & Gimpel, K. Gaussian error linear units (GELUs). Preprint at <https://doi.org/10.48550/arXiv.1606.08415> (2016).
48. Yang, A. et al. Qwen2.5 technical report. Preprint at <https://doi.org/10.48550/arXiv.2412.15115> (2024).
49. Gage, P. A new algorithm for data compression. *C Users J.* **12**, 23–38 (1994).
50. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *Proc. 7th International Conference on Learning Representations* 1–18 (ICLR, 2019); <https://openreview.net/forum?id=Bkg6RiCqY7>
51. Efron, B., Tibshirani, R. J. *An Introduction to the Bootstrap* 1st edn (Chapman and Hall, 1993).
52. Wilcoxon, F. in *Breakthroughs in Statistics: Methodology and Distribution* (eds Kotz, S. & Johnson, N. L.) Ch. 16 (Springer, 1992).

Acknowledgements

We thank the pathologists who contributed to pathology review, reader studies and the curation of a clinician-authored dataset. In particular, we acknowledge X. Fang (Shanghai Pudong Hospital); S. Huang and X. Yang (International Peace Maternity and Child Health Hospital); R. Ding (Huai'an Second People's Hospital); and Y. Zhang, X. He, C. Ding and D. Xia (Eastern Hepatobiliary Surgery Hospital, Naval Medical University) for their valuable clinical expertise and support.

Author contributions

Y.J., J.H., R.Y. and Ying Chen conceived and designed the study. Ying Chen and Y.J. developed the model architecture. Ying Chen, Q.L., T.L. and Y. Li were responsible for data collection and processing. Ying Chen, C.M., Q.L. and G.W. performed experimental analysis and interpreted the results. Q.L., H.D., Xiang Wang, X.X., Y.Z., X.Z. and L.Z. implemented quality control of the data; pathology review, annotation and validation were performed by pathologists among the authors. Ying Chen, C.M. and Y.J. prepared the first draft of the paper. F.Y., S.Y., Xiyue Wang, Yirong Chen, T.L., J.Y., M.H., Y. Lin, H.X., R.Y. and Y.Q. provided feedback on the analysis. All authors contributed to the preparation of the paper and approved the submission for publication. R.Y., J.H. and Y.J. supervised the research.

Funding

This work was supported by the National Natural Science Foundation of China (62272450, to J.H. and Y.Q.) and AI for Science Program, Shanghai Municipal Commission of Economy and Informatization (2025-GZL-RGZN-BTBX-02023 to Ying Chen and J.H.).

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s43018-026-01220-4>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s43018-026-01220-4>.

Correspondence and requests for materials should be addressed to Rongshan Yu, Junjun He or Yuanfeng Ji.

Peer review information *Nature Cancer* thanks Tian-Gen Chang, Olivier Elemento, Danh Tai Hoang, Eytan Ruppin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

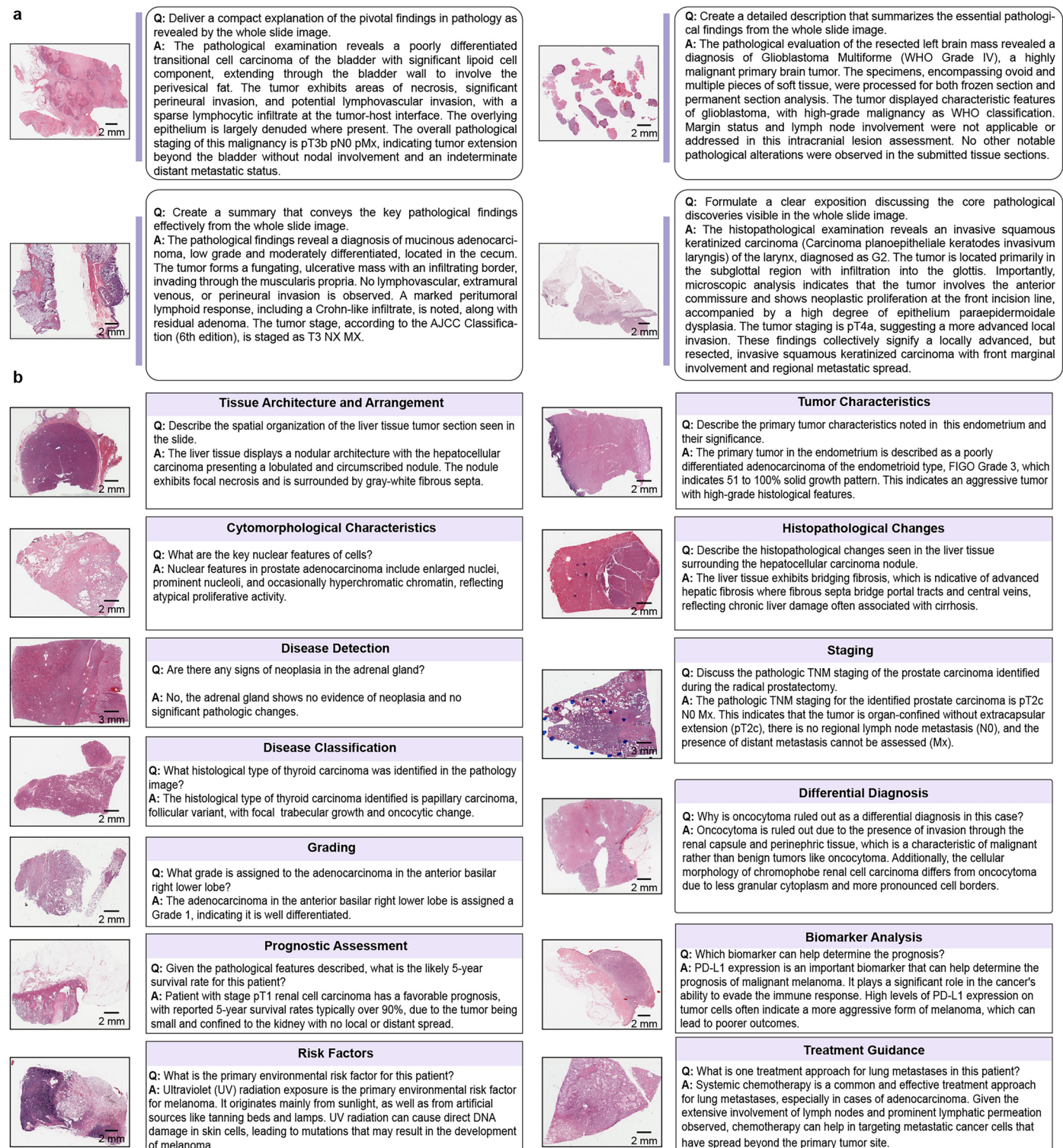
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2026

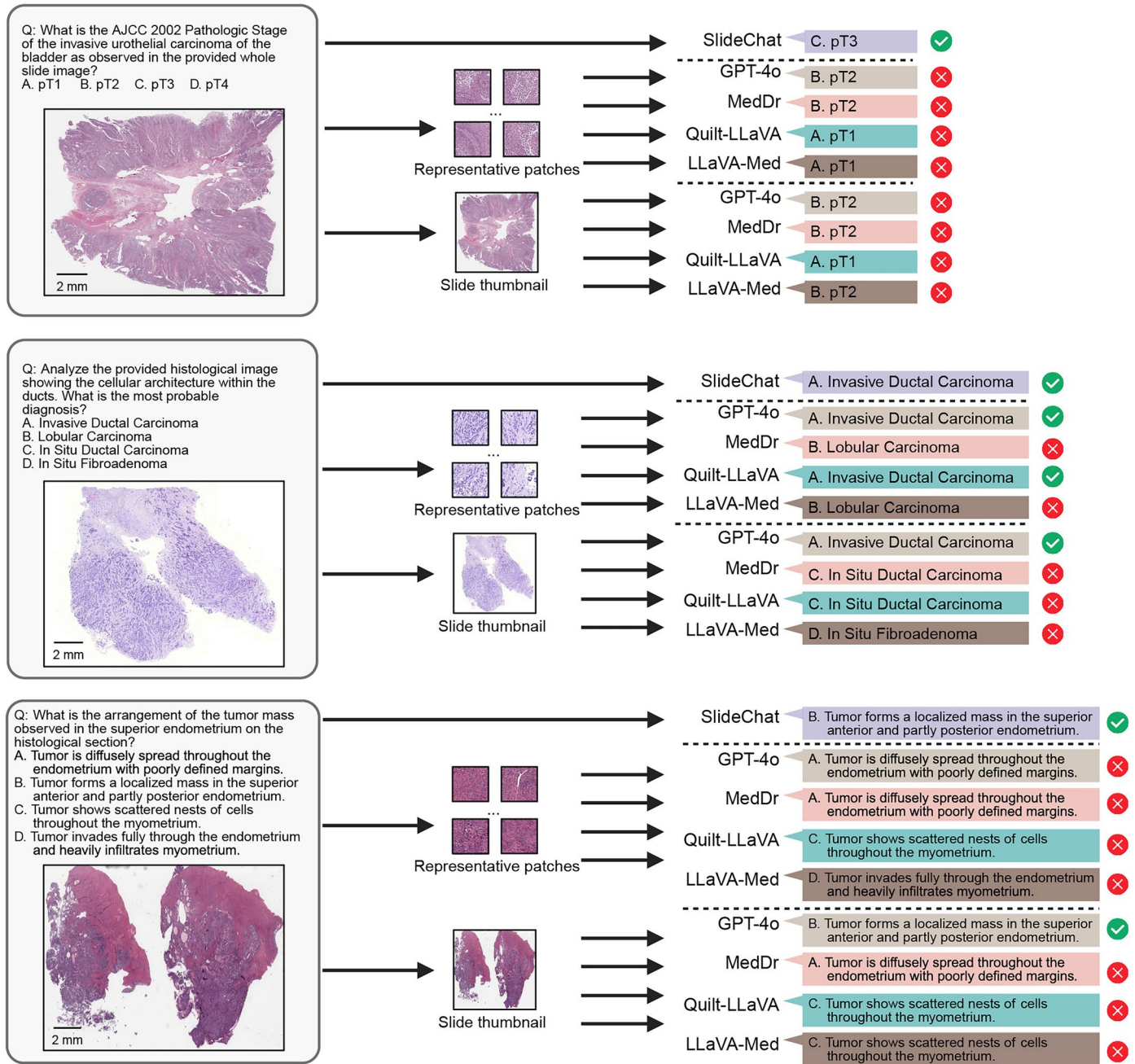
¹Shanghai Artificial Intelligence Laboratory, Shanghai, China. ²Shanghai Innovation Institute, Shanghai, China. ³School of Informatics, Xiamen University, Xiamen, China. ⁴Department of Pathology, Shanghai General Hospital, Shanghai, China. ⁵National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China. ⁶Department of Pathology, Eastern Hepatobiliary Surgery Hospital, Naval Medical University, Shanghai, China. ⁷Second Department of Biliary Surgery, Eastern Hepatobiliary Surgery Hospital, Naval Medical University, Shanghai, China. ⁸Department of Pathology, Huadong Hospital Affiliated to Fudan University, Shanghai, China. ⁹Department of Radiation Oncology, Stanford University School of Medicine, Stanford, CA, USA. ¹⁰Department of Radiology, The First Affiliated Hospital of Jinan University, Guangzhou, China. ¹¹Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. ¹²These authors contributed equally: Ying Chen, Chenglong Ma.

✉ e-mail: rsyu@xmu.edu.cn; hejunjun@sjtu.edu.cn; yfj@stanford.edu



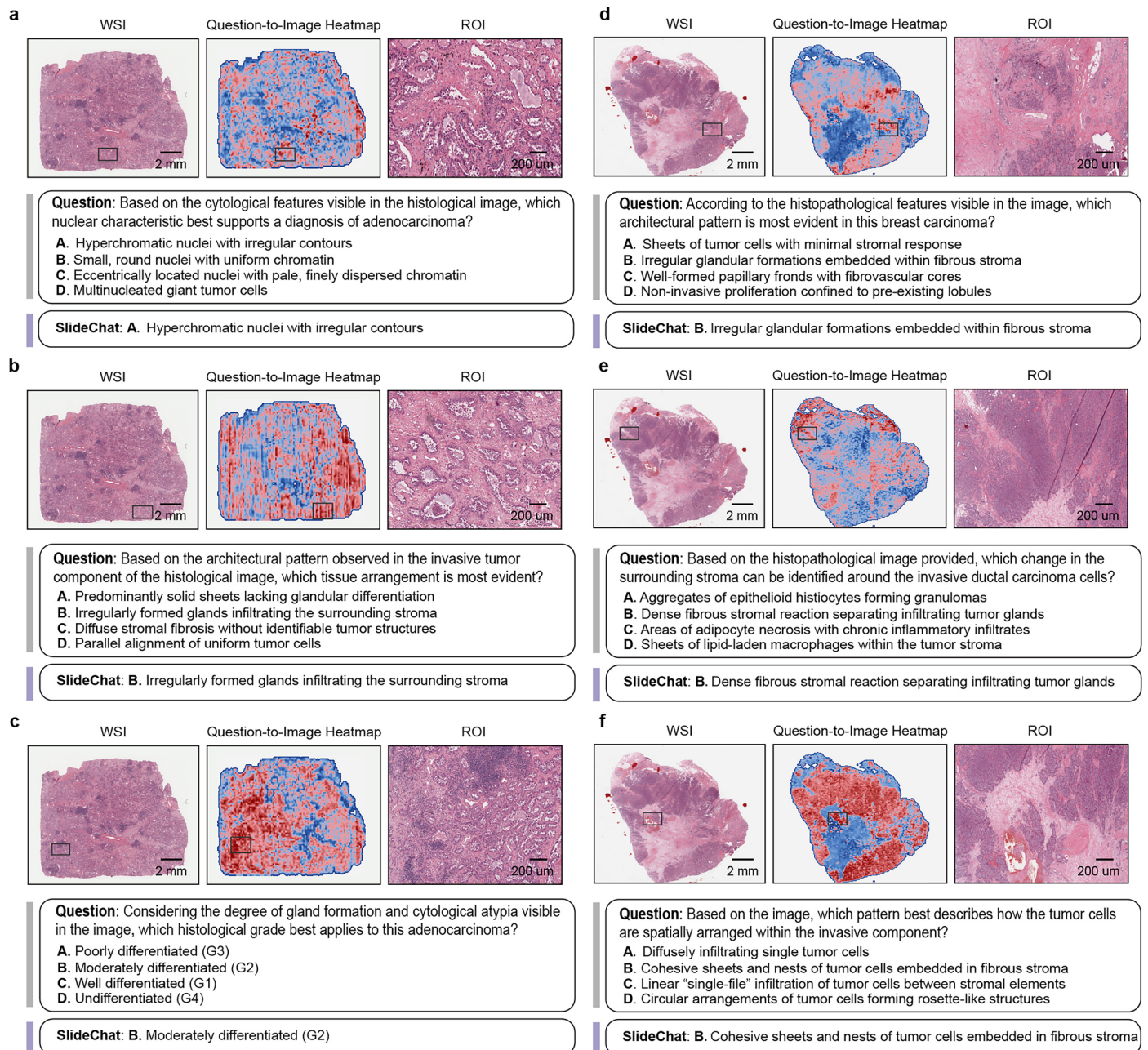
Extended Data Fig. 1 | Examples of the SlideInstruction dataset. a, Examples of WSI-caption pairs. Raw diagnostic reports are parsed to extract concise summaries that highlight key pathological descriptions, forming instruction-style captions for WSI-language alignment during the first training stage of

SlideChat. **b,** Examples of instruction-style question-answer pairs, generated from raw pathology reports across multiple groups to guide the model toward understanding and responding to complex pathology instructions.



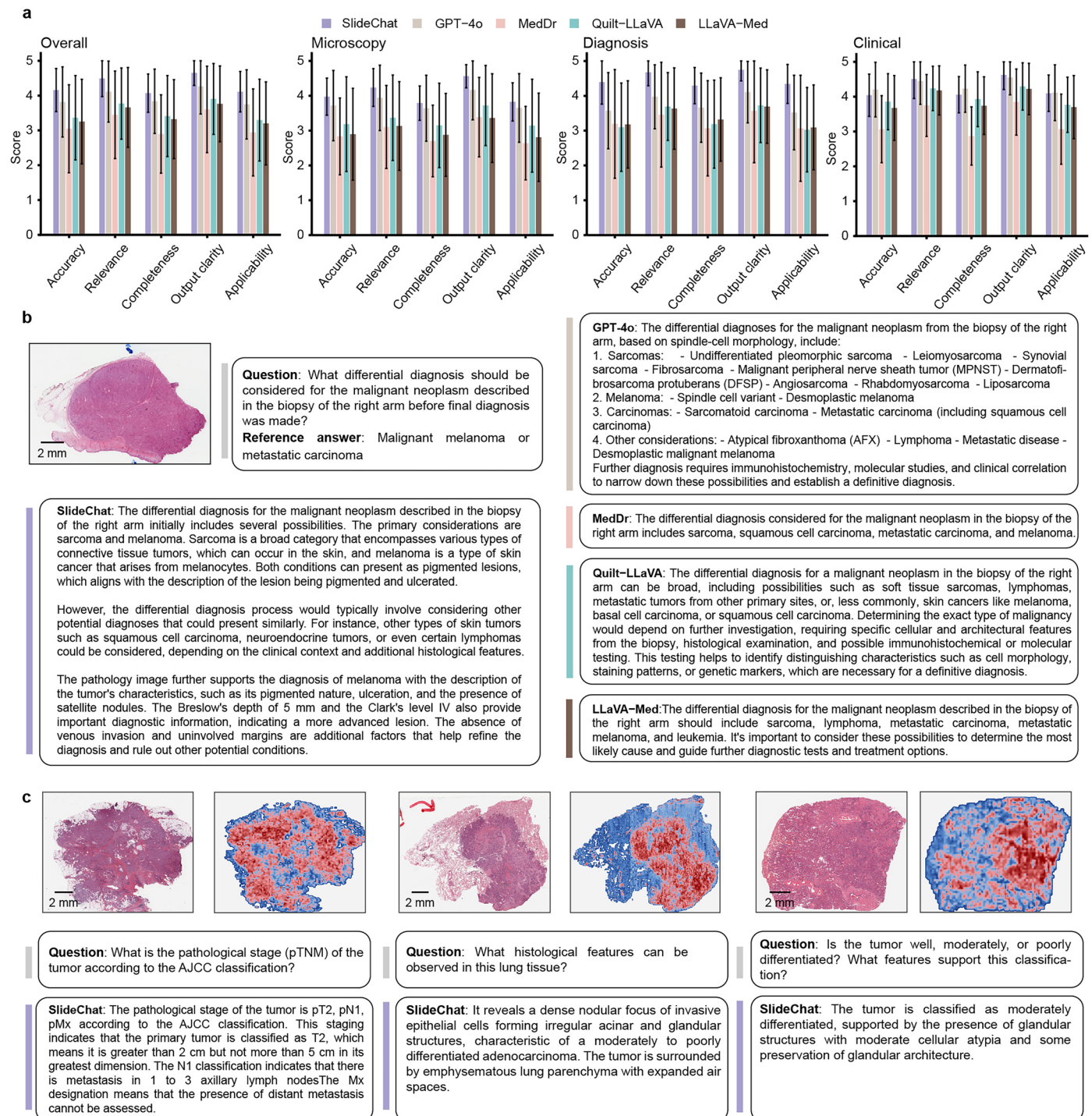
Extended Data Fig. 2 | Comparison between SlideChat and patch-level models on complex examples. Patch-level models can struggle with high-level reasoning tasks such as pathologic staging, tumor subtyping, and describing global tumor arrangement. For instance, in a challenging bladder cancer case,

patch-level models failed to correctly identify the invasive stage (pT3), likely due to their inability to integrate spatial context across the entire slide. In contrast, SlideChat correctly analyzed the global architecture to provide the accurate stage.



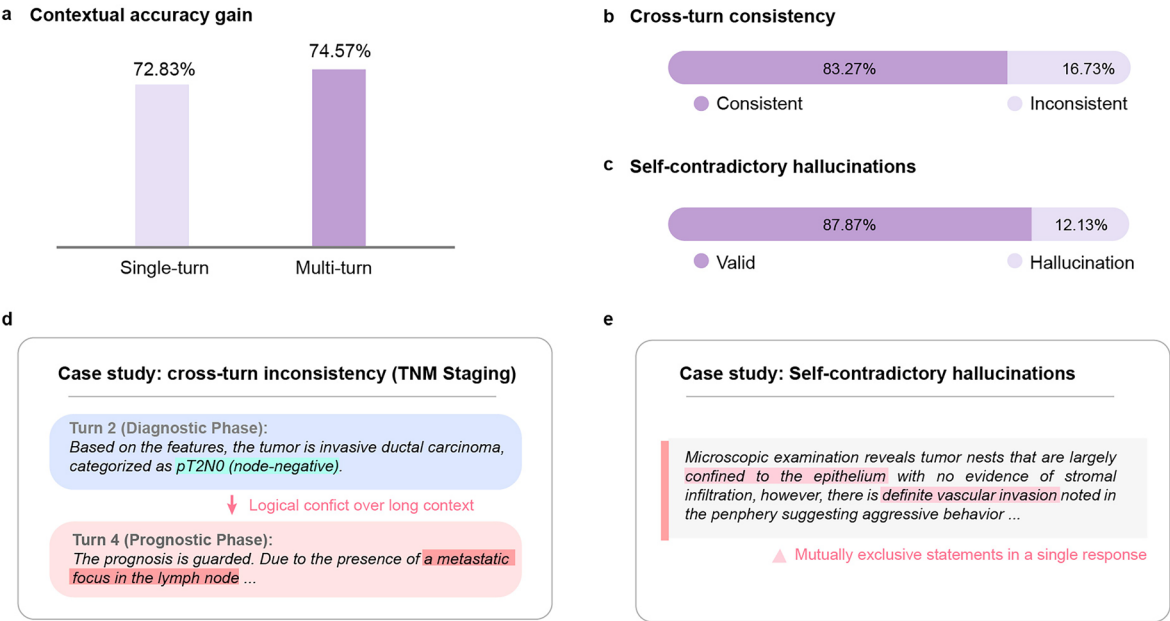
Extended Data Fig. 3 | Question-guided attention heatmaps based on different queries, demonstrating how SlideChat reallocates focus based on the specific diagnostic question. a–c, In an adenocarcinoma case, attention shifts from nuclear details to glandular architecture depending on whether the

prompt queries cytology or differentiation. d–f, In a breast carcinoma case, the model distinctly attends to fibrous stroma versus tumor nests when the inquiry shifts from stromal reaction to cellular arrangement.



Extended Data Fig. 4 | Pathologist evaluation and qualitative comparison of open-ended question answering. **a**, Pathologist evaluation on SlideBench-Open (N = 129 samples). Models were assessed on five criteria across four groups of tasks. SlideChat outperforms all baselines across most dimensions. Data are presented as mean values $\pm$ s.d. **b**, A case from SlideBench-Open assessing responses to a differential diagnosis question for a malignant neoplasm. SlideChat generated the most diagnostically accurate and case-specific answer,

integrating key pathological features to support a refined melanoma diagnosis. GPT-4o produced a broad but weakly prioritized differential, MedDr gave a brief but aligned response, whereas Quilt-LLaVA and LLaVA-Med returned generic answers with limited case-specific reasoning. **c**, Question-guided attention heatmaps. Representative examples showing how SlideChat integrates question context and visual evidence for open-ended reasoning on WSIs.



Extended Data Fig. 5 | Evaluation of SlideChat in multiturn conversational settings. **a**, Comparison of diagnostic accuracy between single-turn and multiturn interactions. **b**, Assessment of cross-turn consistency. **c**, Evaluation of self-contradictory hallucinations. **d**, A representative case study of cross-turn inconsistency, where the model initially identifies a node-negative tumor but

later contradicts this finding during prognostic assessment. **e**, A representative case study of self-contradictory hallucination, illustrating a response that simultaneously claims the tumor is confined to the epithelium while reporting vascular invasion.

[Report Clean]

Prompt: This is the content from the pathology report. Please remove some redundant irrelevant information from the original report, such as technical details of pathology department procedures, Symbols unrelated to the pathological report, specimen handling and processing information, redundant administrative or legal statements, and some repeated information. Show me the cleaned report content.

[Summarized Report Generation] → *SlideInstruction & SlideBench-Report*

Prompt: Based on the above pathological report content, generate a detailed paragraph that summarizes the essential pathological findings. The paragraph should include key information such as the diagnosis, tumor characteristics, margin status, lymph node involvement, and other relevant pathological findings. The summary should not mention the source being a report and should exclude any specific sizes or measurements. The paragraph should be written in a clear and cohesive manner, covering all important points without unnecessary details.

[Question-Answers from Report] → *SlideInstruction & SlideBench-Closed & SlideBench-Open*

System Prompt: You are an AI assistant proficient in digital pathology.

General Prompt: You will receive a pathology report for a whole slide image (WSI). Based on the content of the provided pathological report and the specified narrow task category, your goal is to propose high-quality question-answer pairs (including multiple-choice questions and open-ended questions) that strictly require visual analysis of the image to answer, following the rules below. When generating questions, indicate their broad category and narrow category. Every question must be answerable strictly by observing the visual content of the image. Design medical questions very carefully and only ask questions when you are sure of the answer. Answers should be specific and avoid ambiguity. For multiple-choice questions: (1) Set "question type" to "close-ended"; (2) Provide four options (A, B, C, and D); (3) Ensure that only one option is correct and the other three are plausible but incorrect; (4) Include at least one "hard negative", an option that is incorrect but very similar to the correct one, to increase the difficulty level. For open-ended questions: (1) Set "question type" to "open-ended"; (2) Propose questions with contents different from the multiple-choice questions; (3) Leave the "options" list empty. For formatting: (1) Do not mention/refer to the text report as the source. Phrase question-answer pairs as if looking directly at the WSI. (2) Return a valid JSON object for each question: {"question type": xxx, "question": xxx, "options": [], "answer": xxx, "broad category": xxx, "narrow category": xxx}.

Objective Prompt: <Definition of Broad Category and its corresponding Narrow Categories. >

Prompt: System Prompt + General Prompt + Objective Prompt

[Question-Answers from Label] → *SlideBench-Closed*

Prompt: You will receive the classification result for a whole slide images (WSI). For each specified task and its associated classification labels, your goal is to propose high-quality multiple-choice question-answer pairs that strictly require visual analysis of the image to answer.

Task: <task>

Labels: <label 1> <label 2> <label 3>...

Each question should directly test knowledge based on the labels provided without introducing additional details or characteristics beyond the labels. The number of answer options should match the number of labels provided. Each question-answer pair should include the following:

Question: A concise, visually-oriented question that is strictly based on the labels.

Options: A set of answer choices based only on the given labels, with only one correct label and the rest plausible but incorrect alternatives.

Answer: The correct label from the list.

[LLM Filtering]

Prompt: As an AI assistant, you are analyzing a medical question. <Question> <Option A> <Option B> <Option C> <Option D>

You are responsible to answer the question by selecting only one option. Note that each question includes only one correct answer. Answering with specific option letter. Please avoid answering with any redundant explanations.

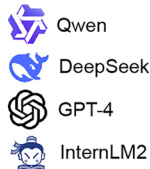
Extended Data Fig. 6 | Prompt templates and dataset construction pipeline.

The workflow illustrates the generation of SlideInstruction and SlideBench. The process includes report cleaning, summarized report generation, question-

answer generation from reports and labels, and LLM-based filtering. LLM-based filtering is applied specifically to SlideBench-Closed to ensure the rigor and validity of the evaluation options.

LLM Verification

Question-Answers
+ LLM Filtering Prompt



Question: What is the observed tissue architecture in the adrenal cortical carcinoma in the microscopic examination?
A. Sheet-like infiltration with necrosis
B. Papillary architecture with fibrovascular cores
C. Glandular arrangement with mucin production
D. Reticular network pattern with normal vasculature
Answer: A
LLM Output: A A A B

✗ Reject

Question: Examine the tissue surrounding invasive squamous carcinoma in the provided slide. What histopathological change is commonly observed?
A. Extensive fibrosis
B. Normal, unaffected tissue
C. Increased mucin production
D. Atrophy of surrounding tissues
Answer: C
LLM Output: B A Undecided A

✓ Pass

Question: What subtype of renal cell carcinoma is identified in this case?
A. Clear cell renal cell carcinoma
B. Papillary renal cell carcinoma
C. Chromophobe renal cell carcinoma
D. Collecting duct carcinoma
Answer: C
LLM Output: Undecided C A B

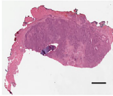
✓ Pass

Pathologists Verification



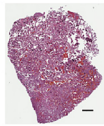
Image Dependency
Question Correctness
Answer Correctness
Task Categorization

Question: Examine the tissue surrounding invasive squamous carcinoma in the provided slide. What histopathological change is commonly observed?
A. Extensive fibrosis
B. Normal, unaffected tissue
C. Increased mucin production
D. Atrophy of surrounding tissues
Answer: C
Category: Microscopy
Subcategory: Histopathological Changes



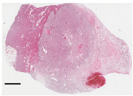
✗ Answer Correctness

Question: What subtype of renal cell carcinoma is identified in this case?
A. Clear cell renal cell carcinoma
B. Papillary renal cell carcinoma
C. Chromophobe renal cell carcinoma
D. Collecting duct carcinoma
Answer: C
Category: Diagnosis
Subcategory: Disease Classification



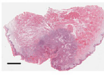
✓ Image Dependency ✓ Question Correctness
✓ Task Categorization ✓ Answer Correctness

Question: What histological growth pattern is associated with the tumor in this case?
A. Trabecular
B. Papillary
C. Acinar
D. Spindle cell
Answer: C
Category: Diagnosis
Subcategory: Disease Classification



✗ Task Categorization

Question: Based on the provided histological image, which subtype of oral squamous cell carcinoma is characterized by prominent keratinization?
A. Basaloid squamous cell carcinoma
B. Verrucous carcinoma
C. Moderately differentiated squamous cell carcinoma
D. Small cell carcinoma
Answer: C
Category: Diagnosis
Subcategory: Disease Classification



✗ Image Dependency

Question: What architectural change is observed in the lymph nodes on the provided slide image which have been replaced by metastatic squamous carcinoma?
A. Normal lymphoid architecture
B. Follicular pattern
C. Complete replacement by tumor cells
D. Increased adipose tissue
Answer: C
Category: Microscopy

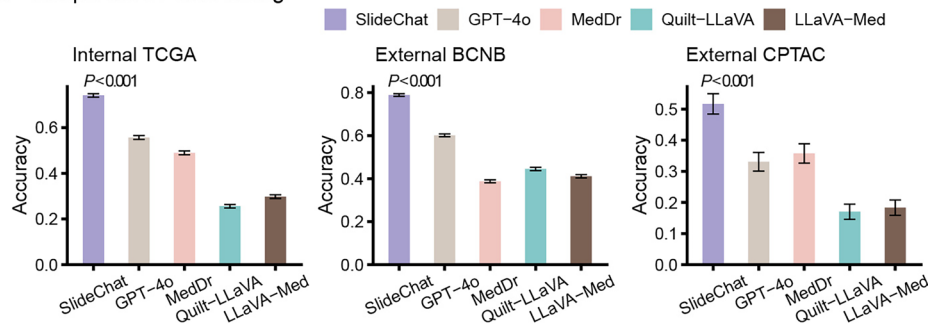


✗ Question Correctness

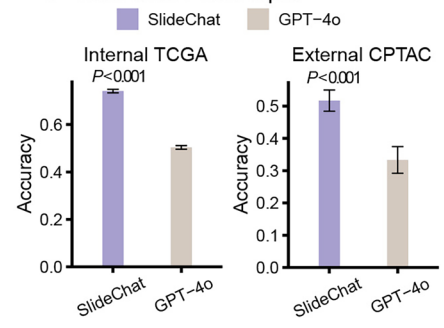
Extended Data Fig. 7 | LLM and pathologist verification workflow for quality control. The LLM verification ensures that the question-answer pairs cannot be answered without the visual inspection of the WSI. Then, pathologists evaluate

the image dependency, task categorization and QA correctness. Examples of accepted and rejected items are shown. Illustrations of pathologists created in BioRender; Chen, Y. <https://biorender.com/fhfeqvb> (2026).

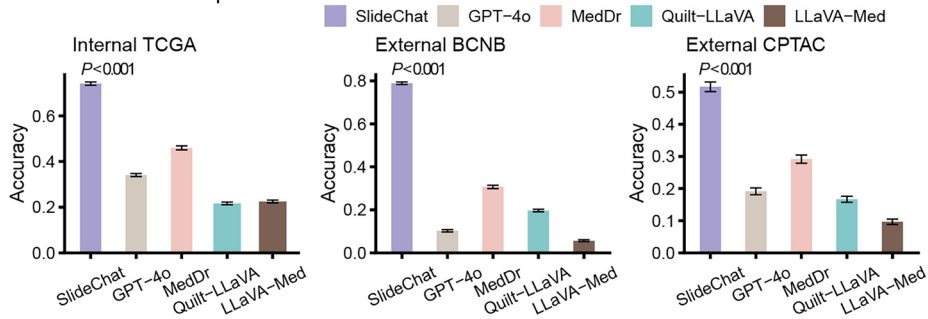
a Independent Patch Voting



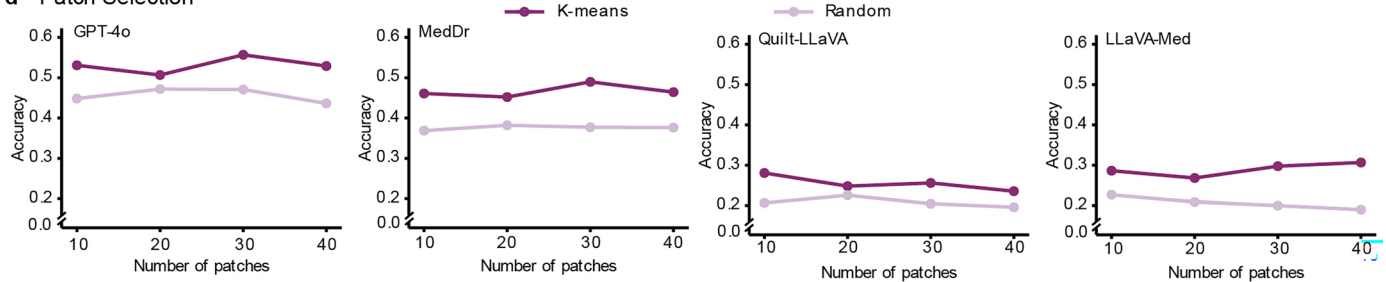
c Joint Multi-Patch Input



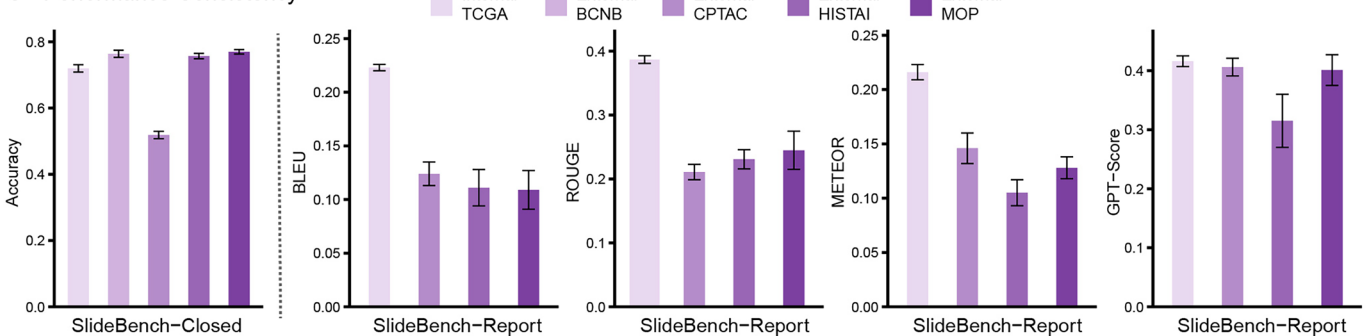
b Slide Thumbnail Input



d Patch Selection



e Performance Consistency



Extended Data Fig. 8 | Patch-level model adaptation and SlideChat

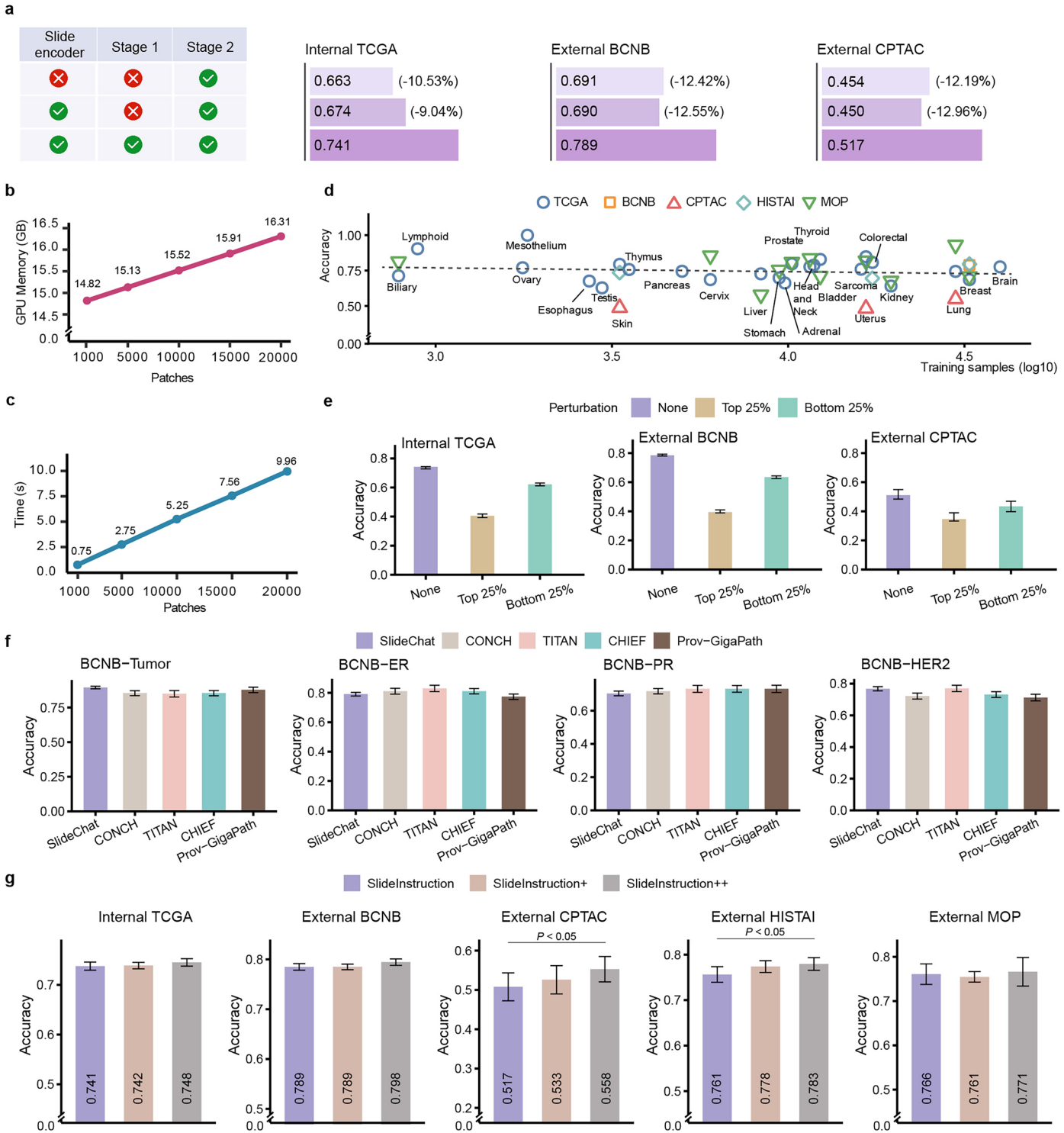
performance consistency analysis. **a–c**, Comparison of WSI-native SlideChat with patch-level baselines under different WSI adaptation strategies. Performance was evaluated on the SlideBench-Closed benchmark across TCGA (N = 3,176 samples), BCNB (N = 4,232 samples), and CPTAC (N = 240 samples). Data are presented as mean values $\pm$ s.d. across 1,000 bootstrap replicates. P values were calculated by comparing SlideChat with the best-performing baseline using two-sided Wilcoxon signed-rank tests with Benjamini–Hochberg FDR correction. **a**, Independent patch voting, where the model independently receives representative patches from K-means clusters and predicts the slide label through majority voting. TCGA: P = 2.28e-60; BCNB: P = 1.26e-70; CPTAC: P = 2.02e-05. **b**, Slide thumbnail input, where the model receives downsampled whole-slide thumbnails. TCGA: P = 7.27e-67; BCNB: P = 5.13e-86; CPTAC: P = 2.54e-16. **c**, Joint multi-patch input, where representative patches from K-means clusters are jointly fed into the model. TCGA: P = 6.35e-65; BCNB: P = 9.11e-53; CPTAC: P = 1.12e-13. **d**, Patch selection sensitivity analysis on SlideBench-Closed: TCGA. Accuracy remains stable when varying input patch counts. Baseline models consistently perform better with K-means patch selection than random sampling. **e**, Performance consistency across five independent runs. Results are shown for SlideBench-Closed (N = 8,836 samples) and SlideBench-Report (N = 3,149 samples). Error bars represent 95% confidence intervals.

P = 2.02e-05. **b**, Slide thumbnail input, where the model receives downsampled whole-slide thumbnails. TCGA: P = 7.27e-67; BCNB: P = 5.13e-86; CPTAC: P = 2.54e-16. **c**, Joint multi-patch input, where representative patches from K-means clusters are jointly fed into the model. TCGA: P = 6.35e-65; BCNB: P = 9.11e-53; CPTAC: P = 1.12e-13. **d**, Patch selection sensitivity analysis on SlideBench-Closed: TCGA. Accuracy remains stable when varying input patch counts. Baseline models consistently perform better with K-means patch selection than random sampling. **e**, Performance consistency across five independent runs. Results are shown for SlideBench-Closed (N = 8,836 samples) and SlideBench-Report (N = 3,149 samples). Error bars represent 95% confidence intervals.



Extended Data Fig. 9 | Comparing model outputs on report generation. a. Two case examples from TCGA demonstrating model performance on free-text pathology report generation across different cancer types (SKCM and BLCA). SlideChat consistently produced accurate and complete reports, correctly capturing tumor type, anatomical location, invasion status, lymphovascular features, TNM staging, and relevant histological details. Its outputs closely matched the reference clinical reports, using structured language and expert-level terminology. In contrast, HistoGPT produced anatomically inconsistent or diagnostically mismatched responses, while PRISM returned brief diagnoses lacking contextual or morphological explanation. **b.** Two additional case

examples from CPTAC on pathology report generation across distinct tumor types: clear cell renal cell carcinoma (CCRCC) (left) and lung squamous cell carcinoma (LSCC) (right). SlideChat generated highly accurate and structured reports, correctly identifying the tumor subtype, anatomical location, and pathological characteristics such as Fuhrman grade, pT2a staging, absence of vascular/perineural invasion, and margin negativity. Its responses also contextualized benign changes (for example, emphysematous alterations in the lung) and avoided irrelevant speculation. In contrast, HistoGPT produced anatomically or diagnostically inconsistent descriptions, while PRISM's outputs were short and often lacked core histological detail.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Additional analysis of SlideChat. **a**, Ablation study of the slide-level encoder and two-stage training, evaluated on SlideBench-Closed across TCGA (N = 3,176 samples), BCNB (N = 4,232 samples), and CPTAC (N = 240 samples). Both components are indispensable for SlideChat performance. **b**, GPU memory usage during SlideChat inference. **c**, SlideChat inference time. **d**, Relationship between organ-level training size in SlideInstruction and SlideChat performance on SlideBench-Closed. Each point shows accuracy for a specific organ in one benchmark dataset; colors and shapes indicate the source dataset. The x-axis shows the number of training samples per organ on a log10 scale, and the y-axis shows SlideChat accuracy for the corresponding organ. The dashed line denotes a linear regression fit, showing no strong monotonic association between training sample size and performance. **e**, Sensitivity of SlideChat to visual corruption. Gaussian noise was introduced to patches with top or bottom 25% attention scores. The observed monotonic drop in accuracy across all datasets confirms that SlideChat relies on meaningful visual content

rather than artifacts. **f**, Comparison between SlideChat and state-of-the-art pathology foundation models (CONCH, TITAN, CHIEF, and Prov-GigaPath) on the BCNB cohort across four classification tasks: Tumor, ER, PR, and HER2 status. Each task includes an independent test set of N = 1,058 samples. SlideChat achieves competitive performance with these specialized models. **g**, Performance of SlideChat trained with different instruction datasets. Accuracy on SlideBench-Closed using the original SlideInstruction dataset, the clinician-authored SlideInstruction+ dataset, and the expanded SlideInstruction++ dataset. The three variants achieve comparable performance across TCGA, BCNB, CPTAC, HISTAI (N = 769 samples) and MOP (N = 419 samples) cohorts. Statistical significance was assessed by comparing the default SlideInstruction-trained SlideChat model against the best-performing enhanced training variant using two-sided Wilcoxon signed-rank tests with Benjamini–Hochberg FDR correction; CPTAC, P = 0.033; HISTAI, P = 0.0036. For panels **e–g**, data are presented as mean values $\pm$ s.d. across 1,000 bootstrap replicates.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Codes for data preprocessing are available at https://github.com/uni-medical/SlideChat . For the pathology reports, we used GPT-4 provided to extract relevant information.
Data analysis	Data analysis was completed using custom software and codes available at https://github.com/uni-medical/SlideChat . Whole-slide image (WSI) preprocessing was performed using OpenSlide (v3.4.1) and openslide-python (v1.4.1). Pandas (v2.2.3) was used to perform random sampling during the nonparametric bootstrapping procedure to estimate 95% confidence intervals for all reported metrics. Visualization of model outputs and evaluation results was performed using matplotlib (v3.9.3) and seaborn (v0.13.2).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The TCGA and CPTAC datasets, including whole-slide images and associated labels, are publicly available through the NIH Genomic Data Commons (<https://portal.gdc.cancer.gov>) and the Proteomic Data Commons (<https://proteomic.datacommons.cancer.gov>), respectively. The extracted TCGA pathology reports can be accessed at <https://github.com/tatonetti-lab/tcga-path-reports>. The BCNB dataset, including whole-slide images and diagnostic labels, is available at <https://bcnb.grand-challenge.org/>. The HISTAI dataset, including metadata and whole-slide images, is available at <https://github.com/HistAI/HISTAI>. The SlideChat training and evaluation benchmark data are released at <https://huggingface.co/datasets/General-Medical-AI/SlideChat>. The MOP cohort was collected under a retrospective study approved by the Eastern Hepatobiliary Surgery Hospital (Approval No. EHBHKY2025-K042-P001). The MOP cohort is subject to controlled access owing to patient privacy obligations and institutional data governance requirements. Requests for access may be submitted to the corresponding author ([yfy@stanford.edu] (<mailto:yfy@stanford.edu>)) with a brief research plan and are subject to institutional approval and completion of a data use agreement. We aim to respond to all data access requests within 4 weeks. Data usage is restricted to non-commercial academic research purposes. Source data are provided with this paper.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\)](#), [and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	The objectives of our study are not related to sex or gender, and these variables were not considered in the analysis.
Reporting on race, ethnicity, or other socially relevant groupings	Our study does not involve any socially constructed or socially relevant categorization variables such as race, ethnicity, or socioeconomic status, and therefore no related analyses were conducted.
Population characteristics	The diagnostic distribution of the training dataset is shown in Figure 1. Sensitive patient information, such as names and dates of birth, was removed during data curation and processing.
Recruitment	No active recruitment was performed.
Ethics oversight	All research complied with relevant ethical regulations. Use of the retrospective Multi-Organ Pathology (MOP) cohort data was approved by the Medical Ethics Committee of Eastern Hepatobiliary Surgery Hospital, Naval Medical University (Approval No. EHBHKY2025-K042-P001). The requirement for informed consent for the MOP cohort was waived by the Medical Ethics Committee of Eastern Hepatobiliary Surgery Hospital because the cohort comprised retrospectively collected de-identified data. The remaining datasets (TCGA, CPTAC, BCNB and HISTAI) are publicly available and were used in accordance with their access and usage policies. Sex and gender information were not collected or used in model training or evaluation; therefore, no sex- or gender-based analyses were performed.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We curated 8,583 raw WSI reports and 274,233 multimodal instruction samples for model training. Evaluation was performed on 8,836 closed-ended questions, 129 open-ended questions, and 3,149 WSI reports collected from five cohorts spanning 31 cancer types. No formal statistical method was used to predetermine sample size. Sample size was determined by data availability after curation and quality control. The performance observed across multiple cohorts and evaluation tasks suggests that the available sample size was adequate for the analyses performed.
Data exclusions	We identified patch images with little or no tissue content as background regions and excluded them from SlideChat training and testing.
Replication	We performed five independent runs with different sampling seeds for the automated evaluations on SlideBench-Closed and SlideBench-Report, and reported mean results with 95% confidence intervals across runs.

Randomization

Sample allocation into experimental groups was generally not applicable because this study was based on retrospective/public datasets rather than interventional group assignments. No randomization of patients, specimens, or slide-level data pairs was performed. For TCGA, one-to-one WSI-report pairs were randomly split, with 80% used for SlideInstruction and 20% held out for SlideBench; the remaining cohorts were predefined external test sets. Because all models were evaluated on the same fixed benchmark datasets, covariate control through group allocation was not relevant. Randomization was applied only in the prospective reader study, where case order was randomized and assisted/unassisted conditions were counterbalanced across readers.

Blinding

Blinding was used in expert evaluation experiments: pathologists were blinded to model identity in the open-ended QA evaluation and the reader study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

NA

Novel plant genotypes

NA

Authentication

NA