MTBBench: A Multimodal Sequential Clinical Decision-Making Benchmark in Oncology

Kiril Vasilev[1*] **Alexandre Misrahi**[2*] **Eeshaan Jain**[2*†]
Phil Cheng[3] **Petros Liakopoulos**[3] **Olivier Michielin**[3]
Michael Moor[1‡] **Charlotte Bunne**[2‡]
[1]ETH Zürich [2]EPFL [3]HUG

github.com/bunnelab/MTBBench
huggingface.co/datasets/EeshaanJain/MTBBench

Abstract

Multimodal Large Language Models (LLMs) hold promise for biomedical reasoning, but current benchmarks fail to capture the complexity of real-world clinical workflows. Existing evaluations primarily assess unimodal, decontextualized question-answering, overlooking multi-agent decision-making environments such as Molecular Tumor Boards (MTBs). MTBs bring together diverse experts in oncology, where diagnostic and prognostic tasks require integrating heterogeneous data and evolving insights over time. Current benchmarks lack this longitudinal and multimodal complexity. We introduce **MTBBench**, an agentic benchmark simulating MTB-style decision-making through clinically challenging, multimodal, and longitudinal oncology questions. Ground truth annotations are validated by clinicians via a co-developed app, ensuring clinical relevance. We benchmark multiple open and closed-source LLMs and show that, even at scale, they lack reliability—frequently hallucinating, struggling with reasoning from time-resolved data, and failing to reconcile conflicting evidence or different modalities. To address these limitations, MTBBench goes beyond benchmarking by providing an agentic framework with foundation model-based tools that enhance multi-modal and longitudinal reasoning, leading to task-level performance gains of up to 9.0% and 11.2%, respectively. Overall, MTBBench offers a challenging and realistic testbed for advancing multimodal LLM reasoning, reliability, and tool-use with a focus on MTB environments in precision oncology.

1 Introduction

Recent advances in large multi-modal and language models have opened the door to general-purpose clinical agents capable of reasoning across diverse biomedical tasks (Moor et al., 2023). Vision-language models can describe pathology images (Lu et al., 2024a; Dai et al., 2025; Lu et al., 2024b), LLMs can summarize clinical notes (Choudhuri et al., 2025; Yang et al., 2024), and medical agents are increasingly able to query tools, retrieve knowledge, and even hold multi-turn clinical conversations (Schmidgall et al., 2024; Wang et al., 2025a). These developments have prompted growing interest in using agents to support complex workflows (Wang et al., 2024a, 2025b; Gao et al., 2024; Lee et al., 2024; Yue et al., 2024; Fallahpour et al., 2025) like those seen in *molecular tumor boards*

*These authors contributed equally.
†Correspondence to: eeshaan.jain@epfl.ch
‡Co-last authorship.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

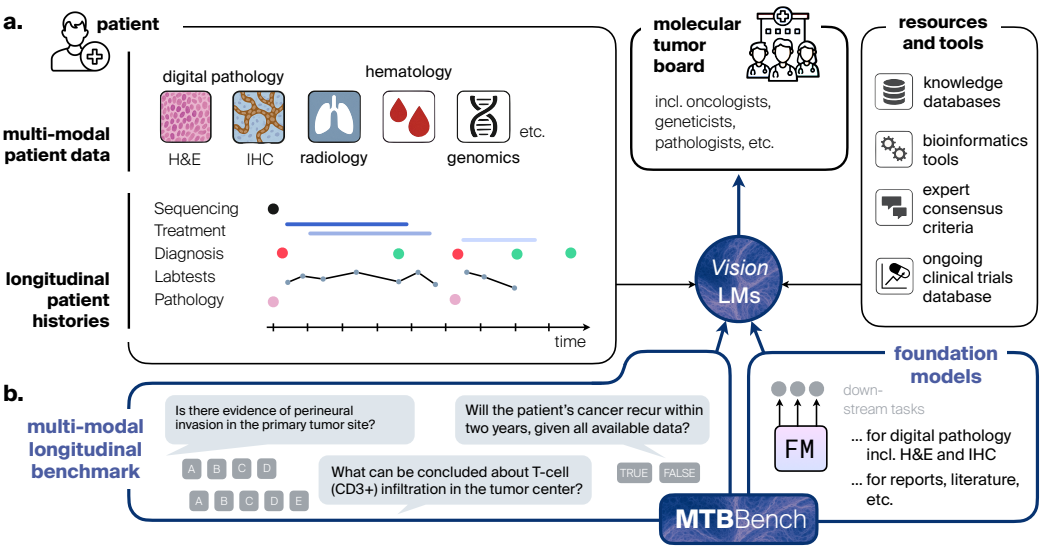


Figure 1: **The MTBBench benchmark and agent framework. a.** MTBBench simulates molecular tumor board workflows, presenting agents with longitudinal, multi-modal patient data (H&E, IHC, hematology, and genomics) along with temporally distributed clinical events. Agents are tasked with integrating this information to support complex decision-making. **b.** MTBBench allows benchmarking agents on their ability to reason across modalities and time in order to accurately tackle clinical questions concerning diagnosis, prognosis, and biomarker interpretation. Lastly, we introduce an agentic framework that enables querying both external tools and pretrained foundation models, allowing agents to more effectively reason over complex, multi-modal and temporally resolved clinical information.

(MTBs) (Tsimberidou et al., 2023), where oncologists, radiologists, pathologists, and geneticists jointly analyze a patient's evolving case (Fig. 4).

However, the evaluation of such agents remains underdeveloped. Existing benchmarks typically frame tasks as *static, uni-modal, single-turn question-answering problems*, where the model is given all necessary inputs at once and evaluated on its ability to predict a discrete answer. This setup diverges sharply from how clinical decisions are made in practice. Real-world oncology reasoning is *interactive, temporal, and multi-modal*: physicians accumulate information over time, integrate findings from multiple data types (e.g., hematoxylin and eosin (H&E) staining, immunohistochemistry (IHC) staining, radiology, blood values, genomics), and make provisional decisions that are updated as new evidence emerges (Fig. 1a). To be useful in these settings, AI agents must not only understand each modality, but also *query, contextualize, and reconcile* information across modalities and time—capabilities rarely assessed in current evaluations.

Recent works such as `MedAgentBench` (Jiang et al., 2025), `MediQ` (Li et al., 2024), and `MedJourney` (Wu et al., 2024) take steps toward interactive or longitudinal evaluation, but typically in limited or uni-modal contexts (e.g., textual EHRs) (Kweon et al., 2024) (Table 1). Likewise, emerging studies on multi-modal agents demonstrate strong promise but lack standardized evaluation across *longitudinal patient trajectories* (Li et al., 2024). Most importantly, these agents are not tested under the cognitive demands of tasks that mirror MTB decision-making: questions involving partial data, sequential updates, conflicting information, highly heterogeneous and different modalities, and high-stakes outcomes.

To address this gap, we introduce **MTBBench**, an oncology-focused benchmark for evaluating AI agents in *longitudinal, multi-modal clinical reasoning*. Inspired by the structure and decision flow of real molecular tumor boards, MTBBench simulates patient case reviews where agents must process heterogeneous patient data across time—including pathology slides, lab data, pathological, surgical and genomic reports—and answer clinically meaningful questions at each step. Questions span diverse task types, including diagnostic classification, spatial biomarker interpretation, and outcome, progression, or recurrence prediction (Fig. 1b). Importantly, the benchmark is *validated by clinicians* using a custom-built expert annotation platform (Fig. 5, for details see Appendix C.1), ensuring both the clinical plausibility of the data and the correctness of model evaluation.

Table 1: Comparison of MTBBench with existing clinical and biomedical benchmarks.

Benchmark	Multi-Modal	Longitudinal	Interactive / Multi-Agent	Clinician-Annotated	Domain
MC-BEC (Chen et al., 2023)	✓	✓	✗	✗	Emergency Medicine
Asclepius (Wang et al., 2024b)	✓	✗	✗	✗	General Medicine
MedJourney (Wu et al., 2024)	✗	✓	✗	✗	General Medicine
EHRNoteQA (Kweon et al., 2024)	✗	✓	✗	✓	General Medicine
MediQ (Li et al., 2024)	✗	✗	✓	✗	General Medicine
ClinicBench (Liu et al., 2024)	✗	✓	✗	✗	General Medicine
HEST-1k (Jaume et al., 2024)	✓	✗	✗	✓	Pathology
BixBench (Mitchener et al., 2025)	✗	✗	✗	✗	Bioinformatics
MTBBench (ours)	✓	✓	✓	✓	**Precision Oncology**

Beyond benchmark construction, we also introduce a modular *agentic framework* designed to interface with different tools as well as **pretrained foundation models** (Fig. 1b). These include models trained on large-scale digital pathology datasets, reports, literature, and other domain-specific modalities. Agents can query these foundation models as part of their reasoning process—invoking them when needed to interpret image regions, extract genomic signatures, or cross-reference trial data—thus mirroring how expert clinicians rely on specialized resources in practice. This framework enables flexible, multi-step decision-making and substantially enhances the agent's ability to synthesize information across modalities and time.

Concretely, our **main contributions** are:

i. **A benchmark for longitudinal, multi-modal clinical reasoning.** MTBBench simulates MTB-style decision-making with temporally evolving patient data across modalities—H&E, IHC, hematology, and genomics—and includes complex, expert-curated questions reflecting real tumor board workflows.
ii. **Systematic evaluation of vision-language models.** We benchmark diverse open and closed-source models and find that performance improves with exposure to more modalities, emphasizing the importance of integrated multi-modal context.
iii. **An agent framework for tool and foundation model integration.** Our agent framework enables dynamic access to external tools and pretrained modality-specific foundation models, significantly boosting reasoning accuracy across tasks.
iv. **Expert-validated data and reproducible tools.** We release curated benchmark data, expert-reviewed annotations, agent logs, and tools to support rigorous and reproducible evaluation of clinical agents.

2 MTBBench: A Multimodal Sequential Clinical Decision-Making Benchmark in Oncology

2.1 Motivation and Positioning

LLMs have shown increasing promise across medical domains, but current benchmarks remain disconnected from the realities of clinical workflows. Ferber et al. (2024) introduce an agent for clinical decision making, however it is limited to general tool frameworks and to single-image inputs. Others evaluate unimodal, static question-answering tasks without requiring the model to gather information, reconcile conflicting inputs, or reason over time (Li et al., 2025). In contrast, real clinical decision-making—especially in molecular tumor boards—is inherently multimodal, interactive, and longitudinal.

MTBBench is a benchmark designed to close this gap. It evaluates how well AI agents can simulate an MTB-style setting, where patient cases evolve across time and require integration of imaging, lab, pathology, genomic, and textual information. MTBBench captures three essential dimensions lacking in prior work:

Multimodality: Clinical data spans digital pathology (H&E, IHC), hematology, radiology, and genomics.
Longitudinality: Patient histories unfold over multiple timepoints, with temporally ordered updates.
Agent workflow: Models must actively request relevant information, access tools, and answer in multi-turn settings.

As shown in Table 1, MTBBench is the first benchmark to jointly address these three dimensions. All questions are expert-validated through a co-developed application, which we further describe below. Together, these components establish MTBBench as a framework for evaluating the capabilities of AI agents in clinical settings that require multi-modal and longitudinal understanding. For further related works, see Appendix B.

2.2 A Benchmark for Molecular Tumor Boards

MTBBench comprises both multiple-choice and true/false questions designed to be answered within a clinically grounded, agent-based framework. Questions span multimodal and longitudinal patient data, requiring the agent to retrieve, interpret, and reason over diverse evidence sources under realistic constraints that mirror the sequential and evolving nature of clinical decision-making.

Companion application for expert validation. To support expert review of the benchmark, we developed a web-based interface allowing clinicians to inspect the clinical context, browse pathology and IHC images, and annotate feedback for each question. The interface presents structured case descriptions alongside image thumbnails of different modalities that open full-resolution slides on demand (Fig. 5). Clinicians can view grouped images by region and marker (e.g., CD3, CD163) and provide detailed assessments directly linked to individual Q&A items. This tool enabled efficient, structured validation of questions and answers by domain experts.

2.2.1 MTBBench-Multimodal

Dataset. We curated a subset of 26 patient cases from the HANCOCK dataset (CC BY 4.0) (Dörrich et al., 2024), a multimodal repository of head and neck cancer patients that includes demographic, pathological, hematological, surgical, and histological data. For each selected patient, an average of 40 modality-specific files are available, including 1.2 H&E slides, 26.2 IHC images, and one hematology report. Among these cases, 32 include a primary tumor H&E slide, and 17 contain at least one lymph node slide. Notably, 27 of the primary tumor slides are accompanied by annotated regions of interest.

Q&A design. We generate 390 multimodal question-answer pairs (15 per patient) using GPT-4o, with select questions verified through expert-in-the-loop review through the companion app (Fig. 5, Appendix C.1). The questions span a range of modalities and clinical reasoning tasks: 2 H&E-based, 4 IHC-based, 3 joint H&E+IHC, 4 hematology-based, and one question each related to clinical outcome and cancer recurrence.

Tasks. The multi-modal track unfolds in three clinically grounded stages. The first focuses on pathological image interpretation (e.g., of H&E, IHC images): agents must infer histologic subtypes and assess spatial patterns of immune infiltration (e.g., $CD3^+$, $CD8^+$ T cells, $CD68^+$, $CD163^+$ macrophages) across tumor subregions such as the invasion front and tumor center. Intermediate questions probe immune correlates of pathology, such as whether lymphovascular or perineural invasion associates with distinct immune profiles. The second stage evaluates hematologic reasoning in a preoperative context—models analyze lab parameters (e.g., CRP, MPV, leukocyte subtypes, creatinine) to infer infection risk, bleeding tendency, renal impairment, and thromboembolic predisposition. Finally, in the post-surgical stage, agents must integrate pathology and lab findings to predict high-level outcomes, including 5-year survival and 2-year recurrence, simulating tumor board-style prognostic deliberation.

2.2.2 MTBBench-Longitudinal

Dataset. We curated a subset of 40 patient cases from the MSK-CHORD dataset (CC BY-NC-ND 4.0) (Jee et al., 2024), a clinicogenomic resource of cancer patients linking genomic profiles with structured clinical timelines, each with an average of five associated files, including copy-number alterations, somatic mutations, specimen pathology reports, and clinical timelines. The timelines capture key clinical events such as diagnostic procedures and treatment transitions, and are segmented into decision-relevant timepoints to support temporally grounded evaluation.

Q&A design. We manually construct 183 question-answer pairs (*i.e.*, on average 4.6 questions per patient), with clinical feedback, targeting outcome prediction, recurrence risk, and treatment progression across clinically actionable stages.

Tasks. The longitudinal track challenges agents to reason over temporally structured patient data segmented into decision-relevant timepoints. Initial questions assess diagnosis and disease trajectory, followed by outcome prediction (e.g., survival), recurrence forecasting, and treatment progression mapping. Genomic data—such as somatic mutations and copy—number alterations—are introduced at key stages, enabling reasoning about resistance patterns or post-treatment stratification. Agents must align treatment regimens with outcomes and integrate evolving context (e.g., updated timelines, new genomic tests, surgical pathology) to justify predictions. This setup mirrors the longitudinal deliberations of MTBs, where clinicians revise hypotheses in light of new events and cumulative history.

2.3 Agent System

Current LLM-based systems struggle to reason across multiple modalities and timepoints (Hager et al., 2024; AlSaad et al., 2024)—a critical requirement in real-world clinical decision-making. In particular, tasks encountered in molecular tumor boards involve dynamic access to evolving patient data, integration of heterogeneous sources such as pathology, lab tests, and genomics, and the ability to contextualize findings over time. Static, single-shot prompting falls short in such settings.

To overcome these limitations, MTBBench implements an agentic framework that enables interactive, multi-turn decision-making. Agents must actively select which files to access, manage evolving memory across turns. A key novelty of our setup is the integration of domain-specific foundation models (FMs) as callable tools besides structured biomedical resources used as tool (e.g., PubMed, DrugBank). These models—trained on large corpora of pathology slides, IHC images, or clinical texts—offer rich, pretrained representations that complement the LLM's general capabilities. Rather than evaluating FMs in isolation, MTBBench enables agents to selectively invoke them as part of a decision-making process, simulating how clinicians consult expert resources. This design of an agentic framework reflects how expert clinicians reason iteratively and selectively, and allows us to benchmark not only factual accuracy but also the agent's ability to gather and use evidence in a realistic clinical workflow.

Agentic workflow. In MTBBench, the agent engages in a multi-turn decision-making process over a temporally evolving patient trajectory. At each turn t, the agent receives a clinical query q_t along with access to a set of modality-specific files $\mathcal{F}_t = \{f_t^1, f_t^2, \ldots, f_t^k\}$, which may include digital pathology images, lab results, clinical notes, or structured genomic and temporal data. The agent may issue a request $\mathcal{R}_t \subseteq \mathcal{F}_t$ to retrieve any subset of these files, which remain accessible only within the current turn. Namely, they do not persist across turns. However, any file from $\mathcal{F}_t$ may be re-requested at a future turn $t' > t$, simulating realistic constraints in clinical workflows where information must be actively re-accessed. The agent's internal memory consists of its reasoning history h_t and a record of previously accessed files $\mathcal{R}_{\leq t}$, forming the basis for answering downstream queries. In the longitudinal track, clinical context is further enriched by an evolving timeline $\mathcal{T}_t = \bigcup_{i=1}^{t} \tau_i$, incrementally summarizing patient history. This setup enforces non-persistent access patterns while encouraging deliberate information gathering and reasoning over temporally non-stationary data. An extensive overview of this workflow is provided in Appendix D.1.

Overview of models. We select a wide range of models with varying sizes. For the multimodal part of our benchmark, we evaluate the vision-text models (including some models with reasoning capabilities): gemma-3-12b, gemma-3-27b, gpt4o, o4-mini (reasoning), internvl3-38b, internvl3-78b, llama90b, mistralsmall, qwen25-7b, and qwen25-32b. For the longitudinal part, we evaluate a mix of text-only and vision-text models: gemma-3-12b, gemma-3-27b, gpt4o, o4-mini (reasoning), llama31-8b, llama33-70b, qwen3-8b (reasoning), and qwen3-32b (reasoning).

2.3.1 Foundation Model-based Tools

While large language models excel at reasoning over textual inputs, they exhibit well-known limitations in visual understanding—especially when interpreting high-resolution biomedical imagery such as histopathology slides (Lu et al., 2024c). In clinical contexts like MTBs, however, the ability to analyze and contextualize pathology images is essential. At the same time, recent advances in

Table 2: Mean accuracy and 95% confidence intervals of various LLMs by task, estimated via bootstrap resampling. Each cell reports the model's mean accuracy, with confidence intervals computed by resampling (with replacement) 1,000 times from the set of patient–question pairs within each task.

Multi-Modal Analysis	**Digital Pathology**	**Hematology**	**Outcome** and **Recurrence**	**Overall**
gemma-3-12b	55.9 ± 6.4	74.9 ± 8.7	53.6 ± 13.5	61.5 ± 10.1
gemma-3-27b	51.8 ± 6.4	76.9 ± 8.2	42.1 ± 13.5	56.9 ± 16.5
gpt4o	63.2 ± 6.0	76.9 ± 7.7	59.9 ± 13.5	66.7 ± 8.1
o4-mini	59.5 ± 6.4	77.8 ± 8.2	55.7 ± 14.4	64.3 ± 10.5
internvl3-38b	54.7 ± 6.4	79.7 ± 8.2	55.9 ± 13.5	63.5 ± 11.9
internvl3-78b	62.0 ± 6.4	79.7 ± 7.7	65.6 ± 11.5	69.1 ± 8.4
llama90b	54.6 ± 6.2	82.8 ± 7.2	51.7 ± 13.5	63.0 ± 14.8
mistralsmall	62.4 ± 6.2	75.8 ± 8.7	51.7 ± 13.5	63.3 ± 11.5
qwen25-7b	42.3 ± 6.2	61.1 ± 9.1	53.9 ± 12.5	52.4 ± 9.0
qwen25-32b	53.3 ± 6.2	73.0 ± 8.7	63.6 ± 12.5	63.3 ± 9.3
Longitudinal Analysis	**Outcome**	**Progression**	**Recurrence**	**Overall**
gemma-3-12b	63.3 ± 11.3	55.9 ± 13.2	54.6 ± 12.8	58.0 ± 4.1
gemma-3-27b	57.7 ± 11.3	50.7 ± 14.0	47.4 ± 13.6	51.9 ± 4.9
gpt4o	72.9 ± 10.6	64.8 ± 13.2	54.8 ± 13.6	64.2 ± 8.6
o4-mini	66.0 ± 10.6	63.1 ± 12.3	51.1 ± 13.6	60.0 ± 7.1
llama31-8b	60.4 ± 11.3	49.0 ± 12.3	45.5 ± 13.6	51.6 ± 7.1
llama33-70b	73.2 ± 9.9	68.2 ± 13.2	56.7 ± 13.6	66.0 ± 7.8
qwen3-8b	63.1 ± 11.3	57.6 ± 13.2	47.4 ± 12.7	56.0 ± 7.5
qwen3-32b	83.0 ± 9.2	63.3 ± 12.3	54.6 ± 13.6	67.0 ± 13.5

vision-language foundation models have shown that pretrained models trained on large-scale medical imaging corpora can capture powerful, domain-specific visual representations (Vaidya et al., 2025a). To harness these capabilities, MTBBench integrates foundation models as external tools (taking inspiration from (Schick et al., 2023; Yao et al., 2023)) that LLM agents can call on-demand. These models are not used passively; instead, agents learn to query them selectively as part of a broader reasoning process. This setup reflects real-world clinical workflows, where specialists consult diagnostic systems or reference image atlases to refine decisions. By exposing FMs as callable components, MTBBench enables systematic evaluation of how agents can leverage visual expertise to complement their textual reasoning.

2.3.2 Digital Pathology Foundation Models

For **H&E images**, we integrate CONCH (Lu et al., 2024d), a vision-language model pretrained on over 1.17 million H&E image–caption pairs. CONCH generates dual visual and textual embeddings, allowing image–text similarity computations. We expose this capability to the LLM by framing it as a tool: given an image and a list of candidate textual descriptors, the tool returns the one with highest embedding similarity to the image, based on dot product in the shared representation space.

For **IHC images**, we develop a custom tool that combines foundation model embeddings with weakly supervised learning to support quantification of marker-specific staining. Tissue regions are segmented and tiled into fixed-size patches (256×256), each embedded using the UNI2 foundation model (Chen et al., 2024a) to produce 1536-dimensional representations. These embeddings are aggregated using an attention-based multiple instance learning (ABMIL) (Ilse et al., 2018) model trained to regress the percentage of positively stained cells. The ABMIL model is trained on a manually curated dataset of IHC images annotated via QuPath (see Appendix E.1), providing marker-level supervision without requiring single-cell labels.

2.3.3 Analysis and Knowledge Database Tools

To support reasoning over temporal sequences of clinical events, we introduce two tools that provide external biomedical knowledge for answering longitudinal questions more accurately: a literature search tool and a pharmacological knowledge base.

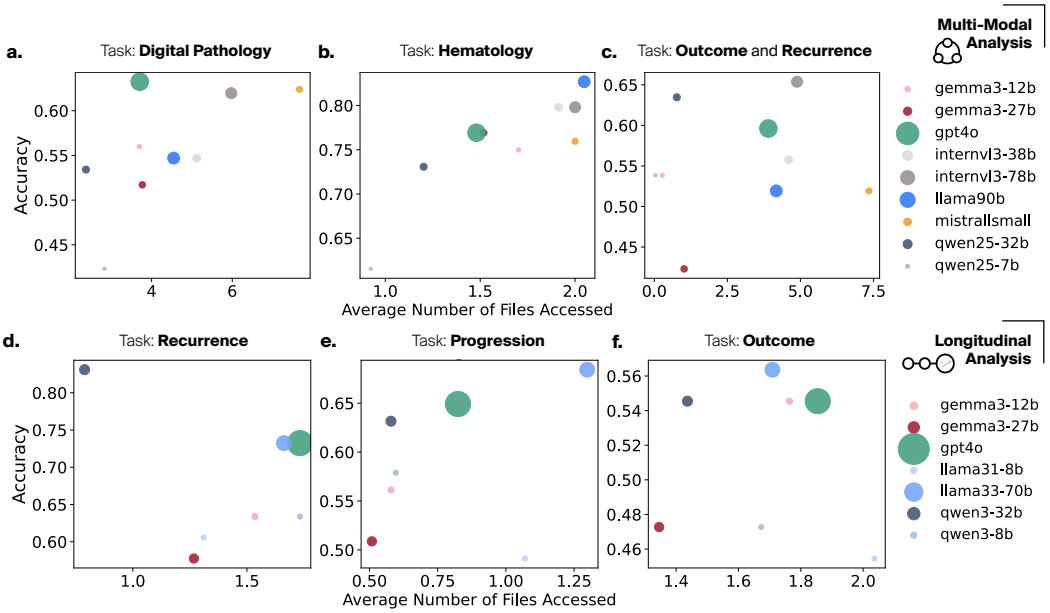


Figure 2: **Accuracy vs. average number of files accessed per question.** Analyzed across tasks for multi-modal understanding (**a–c**) and longitudinal reasoning (**d–f**). Each point represents a model evaluated on a specific task across all patients. Dots indicate model sizes (gpt-4o's size has been reduced for visibility). Higher file access generally correlates with increased accuracy, highlighting the importance of cross-modality and temporal integration for performance.

PubMed. The tool enables the LLM to issue natural language queries to retrieve biomedical literature relevant to a patient's clinical trajectory. The LLM issues natural language queries, which are used to retrieve the top 30 PubMed articles. These are reranked using the `BAAI-bge-reranker-v2-m3 model` (Li et al., 2023), and the top 3 abstracts are returned to the LLM, supporting evidence-grounded reasoning for questions involving treatment effectiveness, sequencing, or disease progression.

DrugBank. To augment drug-related knowledge, we integrate information from DrugBank (Wishart, 2006; Knox et al., 2023). When processing a patient's clinical timeline, drug mentions are automatically linked to corresponding DrugBank entries. Relevant metadata such as therapeutic indications, mechanisms of action, and known drug interactions is incorporated into the model's context. This enrichment enables the language model to reason about treatment sequences with greater specificity, especially in scenarios involving therapeutic decision making and longitudinal disease management.

3 Empirical Evaluation

We evaluate models on both the multi-modal and longitudinal tracks of MTBBench under two conditions: (i) baseline inference without tool support and (ii) augmented inference with access to domain-specific tools, across several metrics, including their accuracy, analysis on multi-modal understanding and ability to reason across temporally-resolved data. Each model acts as an agent interacting with the benchmark via multi-turn dialogues, selectively retrieving and reasoning over available patient files to answer clinical questions. Models are provided only with patient metadata and a list of modality-specific files at each turn. They must request specific files and construct their answers from retrieved content.

3.1 Results on MTBBench without Tools

Settings. We evaluate a diverse set of LLMs and VLMs across all tasks in both the multimodal and longitudinal tracks of MTBBench. To simulate realistic tumor board conditions, we adjust the context and available files for each question type, ensuring that only data plausibly accessible at the corresponding clinical stage is provided. No tools or external resources are available in this setting. Models receive only demographic details, pathology reports, imaging references, and structured

clinical information. The core task remains multiple-choice question answering, but success requires multimodal reasoning, data retrieval, and longitudinal inference rather than simple pattern recognition.

Evaluation metrics. We report mean accuracy per model and task. To quantify uncertainty, we estimate 95% confidence intervals using bootstrap resampling with 1000 iterations. For each task, we sample with replacement from the set of question outcomes per model, compute the mean accuracy per sample, and extract the 2.5th and 97.5th percentiles to define the confidence interval. To assess their ability to incorporate findings from several modalities, we analyze the number of modalities queried compared to the resulting achieved accuracy. The analysis is conducted across 26 patients for MTBBench-Multimodal and 40 for MTBBench-Longitudinal. For details, see Appendix E.

Results for MTBBench-Multimodal. Accuracy across all models (of different parameter sizes) are displayed in Table 2. Model performance varies substantially across the multimodal tasks. Digital pathology, despite being visually complex, does not show a consistent benefit from model size—for instance, gemma-3-12b outperforms its larger counterpart gemma-3-27b. Hematology emerges as the most approachable task, likely due to its structured and interpretable inputs. In contrast, outcome and recurrence prediction remain the most difficult, with accuracies near random (50%), even for leading models. The strongest overall performance is achieved by internvl3-78b at 69.1% accuracy, outperforming closed-source baselines like gpt4o by 2.5%. Nevertheless, large performance gaps persist: up to 36.7% in digital pathology, 17.2% in hematology, and 34.6% in outcome and recurrence prediction—highlighting the need for more robust multimodal reasoning.

Instead of model size, a stronger signal emerges in the relationship between performance and the number of files accessed (Fig. 2a-c), in both MTBBench-Multimodal and MTBBench-Longitudinal. This suggests that effective information gathering, rather than raw scale, is a key determinant of accuracy. This is also demonstrated in Example 1: Compared to qwen25-7b, gpt4o accesses more modalities including higher resolution H&E regions-of-interest, resulting in the correct cancer subtype identification. In Example 2, mistralsmall requests file access to more IHCs and the H&E slide compared to gemma-3-27b, resulting in correct cancer subtype identification. However, this trend does not hold for outcome and recurrence tasks, where high error rates persist across models. We hypothesize that these questions require contextual grounding and biomarker interpretation beyond the current capabilities of uni-modal or tool-free agents. For further results, see Appendix G and Figs. 6).

Results for MTBBench-Longitudinal. The longitudinal track reveals persistent weaknesses in baseline LLMs. While outcome prediction shows some promise—qwen3-32b reaches 83.1% accuracy—recurrence and progression tasks remain near chance (Table 2). Similarly, querying multiple modalities improves model performance (Fig. 2d-f). This is also demonstrated in Example 3: compared to gemma-3-27b, qwen3-32b re-accesses part of the patient timeline of events as well as pathological data, resulting in better cancer progression prediction. This disparity suggests models can detect coarse survival signals but struggle with more nuanced temporal reasoning, reflecting varied evidence complexity across tasks.

3.2 Results on MTBBench with Tools

Settings. In this setup, we augment the baseline LLMs with access to external tools, including foundation model-based modules (UNI, Chen et al. (2024a) and CONCH, Lu et al. (2024b)) and classical biomedical resources (PubMed and DrugBank). These tools are accessible via API-style interfaces, allowing agents to retrieve structured outputs during multi-turn reasoning. Visual tools are primarily used in multimodal tasks, while longitudinal tools support reasoning over treatment history, drug interactions, and literature-based evidence.

Evaluation metrics. We use the same accuracy metrics described in Section 3.1. Results are reported per task and model, and improvements are measured relative to tool-free baselines.

Results for MTBBench-Multimodal. As shown in Figure 3a–c (exact numbers in Fig. 8), access to visual foundation model tools significantly improves performance on all multimodal tasks. Digital pathology, in particular, benefits from integration of the FMs, with models like gemma-3-12b and qwen25-7b showing improvements of up to 9%. Notably, tool augmentation also improves performance on hematology tasks, despite the absence of dedicated tools for lab analysis. We attribute this to better contextual grounding: agents are more effective at integrating diverse file types

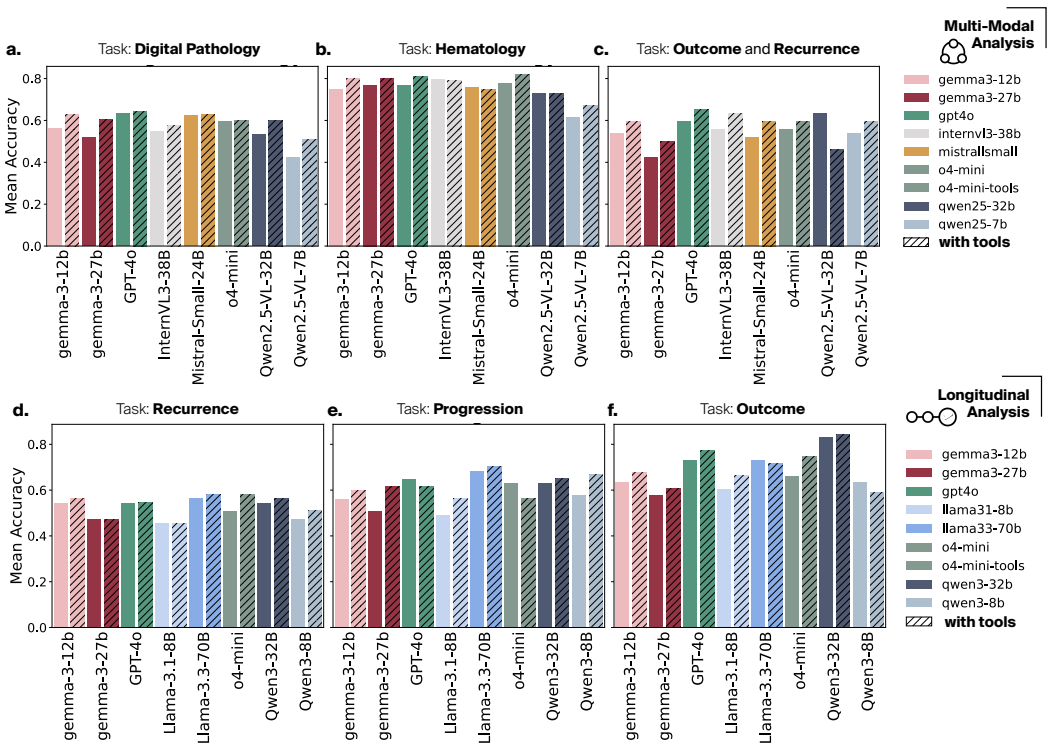


Figure 3: **Accuracy across models and tasks for naive and tool-augmented agents.** For multi-modal (**a.–c.**) and longitudinal (**d.–f.**) evaluation. Models equipped with tool access (hatched bars) generally show improved accuracy, highlighting the benefit of querying external resources in both multi-modal and temporal settings.

when given richer information from related modalities. Outcome and recurrence tasks, which showed low baseline performance, also benefit modestly from the improved visual reasoning capabilities.

This behavior is also demonstrated through an example: in Example 4, mistralsmall + TOOLS in contrast to mistralsmall without tools accesses a digital pathology FM for IHC analysis. Despite both models accessing the same number of modalities, mistralsmall + TOOLS is able to properly address the question, i.e., which marker is critical in a given pathological analysis.

Results for MTBBench-Longitudinal. Figures 3d–f show that tool access improves performance across most longitudinal tasks, though gains are generally modest. This is expected, as no specialized foundation model currently exists for longitudinal clinical reasoning—a key limitation in this track. Instead, agents rely on general-purpose tools like DrugBank and PubMed, which still provide some benefit by enriching context and supporting evidence-based decisions. For example, progression and recurrence predictions improve by over 5% in selected models. Outcome prediction—which already exhibited strong baseline performance—also sees incremental gains, underscoring the potential of even generic tools to enhance structured clinical reasoning.

4 Conclusion

MTBBench introduces a benchmark and agentic framework for evaluating AI agents in longitudinal, multi-modal oncology workflows, modeled on the structure of real molecular tumor boards. By combining temporally evolving patient data, expert-validated clinical questions, and access to external tools and pretrained foundation models, MTBBench enables a rigorous assessment of agents' ability to reason across modalities and time. Our evaluation shows that agents perform significantly better when equipped to query diverse modalities and leverage domain-specific models, underscoring the need for flexible, tool-augmented reasoning in clinical AI. MTBBench shifts the field from static, uni-modal evaluation toward dynamic, decision-centric assessment grounded in clinical complexity. While MTBBench simulates realistic decision-making, it remains a controlled offline benchmark—agents are not yet tested in interactive, real-world clinical workflows or exposed to ambiguous or incomplete inputs requiring clarification or adaptive strategies. Looking forward,

future work will explore extending MTBBench to more diverse clinical domains and incorporating interactive elements—paving the way for evaluating agents not only as reasoning tools but as potential collaborators in real-world precision oncology. A particularly promising direction we would like to explore involves the integration of medical foundation models with capabilities for analyzing complex longitudinal data, enabling deeper temporal reasoning and personalized decision support.

Societal impacts. MTBBench offers clear benefits by promoting clinically relevant evaluation of AI agents, potentially advancing more trustworthy decision support in oncology. It encourages models to reason over multimodal, longitudinal data, closer to real-world needs. However, risks include misuse of the benchmark to suggest clinical readiness, and reduced transparency when agents rely heavily on tools. We emphasize that MTBBench is for research only, not clinical deployment, and should be used with proper oversight and ethical safeguards.

5 Acknowledgements

This work was supported by the Swiss AI Initiative through a Swiss National Supercomputing Centre (CSCS) allocation under the SWISS AI Large Grant No. 46 ("Virtual Patient Platform") on Alps. The authors would like to thank Stavros Pantelakos, Martha Nifora, Jan Rüschoff, and Claudio de Vito for their valuable feedback for validation of the benchmark.

References

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Melissa Zhao, Aaron K. Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, Amr Soliman, Chengkuan Chen, Tong Ding, Judy J. Wang, Georg Gerber, Ivy Liang, Long Phi Le, Anil V. Parwani, Luca L. Weishaupt, and Faisal Mahmood. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033): 466–473, Oct 2024a. ISSN 1476-4687. doi: 10.1038/s41586-024-07618-3.

Dawei Dai, Yuanhui Zhang, Qianlan Yang, Long Xu, Xiaojing Shen, Shuyin Xia, and Guoyin Wang. Pathologyvlm: a large vision-language model for pathology image understanding. *Artificial Intelligence Review*, 58(6):186, Mar 2025. ISSN 1573-7462. doi: 10.1007/s10462-025-11190-1.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, Mar 2024b. ISSN 1546-170X. doi: 10.1038/s41591-024-02856-4.

Akash Choudhuri, Philip Polgreen, Alberto Segre, and Bijaya Adhikari. Summarizing clinical notes using llms for icu bounceback and length-of-stay prediction. *medRxiv*, 2025. doi: 10.1101/2025.01.19.25320797.

Zhichao Yang, Avijit Mitra, Sunjae Kwon, and Hong Yu. Clinicalmamba: A generative clinical language model on longitudinal clinical notes. *arXiv preprint arXiv:2403.05795*, 2024.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Wenting Chen, Xiang Li, and Yixuan Yuan. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*, 2025a.

Zifeng Wang, Hanyin Wang, Benjamin Danek, Ying Li, Christina Mack, Hoifung Poon, Yajuan Wang, Pranav Rajpurkar, and Jimeng Sun. A perspective for adapting generalist ai to specialized medical ai applications and their challenges. *arXiv preprint arXiv:2411.00024*, 2024a.

Eric Wang, Samuel Schmidgall, Paul F. Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. Txgemma: Efficient and agentic llms for therapeutics. *arXiv preprint arXiv:2504.06196*, 2025b.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *arXiv preprint arXiv:2404.02831*, 2024.

Yongju Lee, Dyke Ferber, Jennifer E. Rood, Aviv Regev, and Jakob Nikolas Kather. How ai agents will change cancer research and oncology. *Nature Cancer*, 5(12):1765–1767, Dec 2024. ISSN 2662-1347. doi: 10.1038/s43018-024-00861-7.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning. *arXiv preprint arXiv:2404.14777*, 2024.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673*, 2025.

Apostolia M Tsimberidou, Michael Kahle, Henry Hiep Vo, Mehmet A Baysal, Amber Johnson, and Funda Meric-Bernstam. Molecular tumour boards—current and future considerations for precision oncology. *Nature Reviews Clinical Oncology*, 20(12):843–863, 2023.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. Medagentbench: A realistic virtual ehr environment to benchmark medical llm agents. *arXiv preprint arXiv:2501.14654*, 2025.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2024.

Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin Wang, Zhenxi Lin, Jie Yang, Shuang Zhao, and Yefeng Zheng. Medjourney: Benchmark and evaluation of large language models over patient clinical journey. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 87621–87646. Curran Associates, Inc., 2024.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. *arXiv preprint arXiv:2402.16040*, 2024.

Emma Chen, Aman Kansal, Julie Chen, Boyang Tom Jin, Julia Rachel Reisler, David A. Kim, and Pranav Rajpurkar. Multimodal Clinical Benchmark for Emergency Care (MC-BEC): A Comprehensive Benchmark for Evaluating Foundation Models in Emergency Medicine. *arXiv preprint arXiv:2311.04937*, 2023.

Wenxuan Wang, Yihang Su, Jingyuan Huan, Jie Liu, Wenting Chen, Yudi Zhang, Cheng-Yi Li, Kao-Jung Chang, Xiaohan Xin, Linlin Shen, et al. Asclepius: A spectrum evaluation benchmark for medical multi-modal large language models. *arXiv preprint arXiv:2402.11217*, 2024b.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. Large language models are poor clinical decision-makers: A comprehensive benchmark. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Perez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, and Faisal Mahmood. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems*, December 2024.

Ludovico Mitchener, Jon M Laurent, Benjamin Tenmann, Siddharth Narayanan, Geemi P Wellawatte, Andrew White, Lorenzo Sani, and Samuel G Rodriques. Bixbench: a comprehensive benchmark for llm-based agents in computational biology. *arXiv preprint arXiv:2503.00096*, 2025.

Dyke Ferber, Omar S. M. El Nahhas, Georg Wölflein, Isabella C. Wiest, Jan Clusmann, Marie-Elisabeth Leßman, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, Manuel Salto-Tellez, Nikolaus Schultz, Daniel Truhn, and Jakob Nikolas Kather. Autonomous artificial intelligence agents for clinical decision making in oncology. *arXiv preprint arXiv:2404.04667*, 2024.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2025.

Marion Dörrich, Matthias Balk, Tatjana Heusinger, Sandra Beyer, Hassan Kanso, Christian Matek, Arndt Hartmann, Heinrich Iro, Markus Eckstein, Antoniu-Oreste Gostian, and Andreas M. Kist. A multimodal dataset for precision oncology in head and neck cancer. *medRxiv*, 2024. doi: 10.1101/2024.05.29.24308141.

Justin Jee, Christopher Fong, Karl Pichotta, Thinh Ngoc Tran, Anisha Luthra, Michele Waters, Chenlian Fu, Mirella Altoe, Si-Yang Liu, Steven B Maron, et al. Automated real-world data integration improves cancer outcome prediction. *Nature*, pages 1–9, 2024.

Paul Hager, Friederike Jungmann, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Robbie Holland, Rickmer Braren, Marcus Makowski, Georgios Kaisis, and Daniel Rueckert. Evaluating and mitigating limitations of large language models in clinical decision making. *medRxiv*, 2024. doi: 10.1101/2024.01.26.24301810.

Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. Multimodal large language models in healthcare: Applications, challenges, and future outlook. *Journal of Medical Internet Research*, 26:e59505, August 2024. doi: 10.2196/59505.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Melissa Zhao, Aaron K. Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, Amr Soliman, Chengkuan Chen, Tong Ding, Judy J. Wang, Georg Gerber, Ivy Liang, Long Phi Le, Anil V. Parwani, Luca L. Weishaupt, and Faisal Mahmood. A multimodal generative AI copilot for human pathology. *Nature*, 634 (8033):466–473, October 2024c. ISSN 1476-4687. doi: 10.1038/s41586-024-07618-3. Publisher: Nature Publishing Group.

Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H. Song, Tong Ding, Sophia J. Wagner, Ming Y. Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, Richard J. Chen, Dina ElHarouni, Georges Ayoub, Connor Bossi, Keith L. Ligon, Georg Gerber, Long Phi Le, and Faisal Mahmood. H&E, DNA, scRNA-seq: Molecular-driven Foundation Model for Oncologic Pathology. *arXiv preprint arXiv:2501.16652*, January 2025a.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024d.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024a.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 10–15 Jul 2018.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. Making large language models a better foundation for dense retrieval. *arXiv preprint arXiv:2312.15503*, 2023.

D. S. Wishart. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34(90001):D668–D672, January 2006. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkj067.

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, Marysol Garcia-Patino, Ray Kruger, Aadhavya Sivakumaran, Selena Sanford, Rahil Doshi, Nitya Khetarpal, Omolola Fatokun, Daphnee Doucet, Ashley Zubkowski, Dorsa Yahya Rayat, Hayley Jackson, Karxena Harford, Afia Anjum, Mahi Zakir, Fei Wang, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, November 2023. doi: 10.1093/nar/gkad976.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J Seibel, Junjun He, and Yu Qiao. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *arXiv preprint arXiv:2408.03361*, 2024b.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and Yuyin Zhou. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024.

Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F. Steiner, Hester Van Boven, Robert Vink, Christina Hulsbergen-Van De Kaa, Jeroen Van Der Laak, Mahul B. Amin, Andrew J. Evans, Theodorus Van Der Kwast, Robert Allan, Peter A. Humphrey, Henrik Grönberg, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature Medicine*, 28(1):154–163, January 2022. doi: 10.1038/s41591-021-01620-2.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*, 2025c.

Qiaoyu Zheng, Chaoyi Wu, Pengcheng Qiu, Lisong Dai, Ya Zhang, Yanfeng Wang, and Weidi Xie. How well can modern llms act as agent cores in radiology environments? *arXiv preprint arXiv:2412.09529*, 2025.

Fatemeh Ghezloo, Mehmet Saygin Seyfioglu, Rustin Soraki, Wisdom O Ikezogwo, Beibin Li, Tejoram Vivekanandan, Joann G Elmore, Ranjay Krishna, and Linda Shapiro. Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. *arXiv preprint arXiv:2502.08916*, 2025.

Ziyue Wang, Junde Wu, Chang Han Low, and Yueming Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. 2025d.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 09 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac409. URL https://doi.org/10.1093/bib/bbac409.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. Biomedlm: A 2.7b parameter language model trained on biomedical text, 2024. URL https://arxiv.org/abs/2403.18421.

Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0, 2024. URL https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0.

Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction, 2024. URL https://arxiv.org/abs/2409.09173.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Ellen Yang, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan H. Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Hannah Wen, Juan A. Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David S. Klimstra, Brandon Rothrock, Siqi Liu, and Thomas J. Fuchs. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, Oct 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-03141-0. URL https://doi.org/10.1038/s41591-024-03141-0.

Fang Yan, Jianfeng Wu, Jiawen Li, Wei Wang, Jiaxuan Lu, Wen Chen, Zizhao Gao, Jianan Li, Hong Yan, Jiabo Ma, Minda Chen, Yang Lu, Qing Chen, Yizhi Wang, Xitong Ling, Xuenian Wang, Zihan Wang, Qiang Huang, Shengyi Hua, Mianxin Liu, Lei Ma, Tian Shen, Xiaofan Zhang, Yonghong He, Hao Chen, Shaoting Zhang, and Zhe Wang. Pathorchestra: A comprehensive foundation model for computational pathology with over 100 diverse clinical-grade tasks, 2025. URL https://arxiv.org/abs/2503.24345.

Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, Kun-Hsing Yu, Sierra Willens, Francesca Maria Olguin, Jeffrey J. Nirschl, Joel Neal, Maximilian Diehn, Sen Yang, and Ruijiang Li. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778, Feb 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08378-w. URL https://doi.org/10.1038/s41586-024-08378-w.

Jacqueline Lammert, Tobias Dreyer, Sonja Mathes, Leonid Kuligin, Kai J Borm, Ulrich A Schatz, Marion Kiechle, Alisa M Lörsch, Johannes Jung, Sebastian Lange, et al. Expert-guided large language models for clinical decision support in precision oncology. *JCO precision oncology*, 8: e2400478, 2024.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5):378–382, 1971.

Ted Byrt, Janet Bishop, and John B. Carlin. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5):423–429, 1993.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

Klaus Krippendorff. Computing krippendorff's alpha-reliability. https://repository.upenn.edu/asc_papers/43, 2011. University of Pennsylvania ScholarlyCommons.

Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al. Qupath: Open source software for digital pathology image analysis. *Scientific reports*, 7(1):1–7, 2017.

Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H Song, Tong Ding, Sophia J Wagner, Ming Y Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, et al. Molecular-driven foundation model for oncologic pathology. *arXiv preprint arXiv:2501.16652*, 2025b.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Appendix

A Background on Molecular Tumor Boards

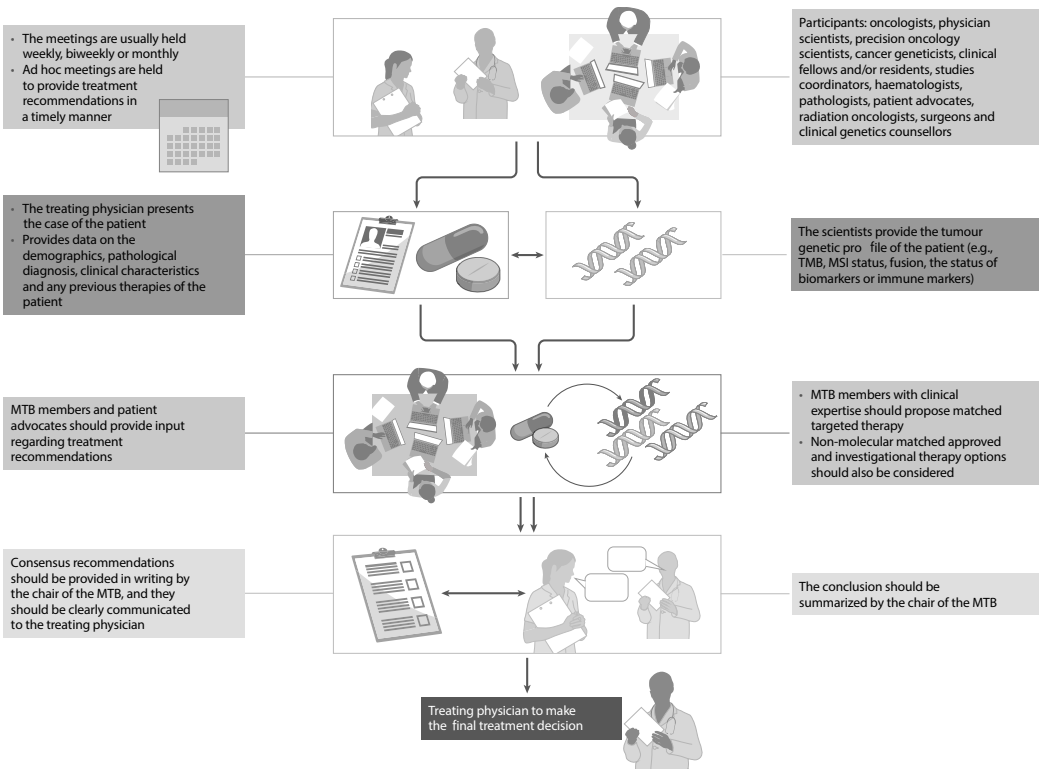


Figure 4: **Overview of the Molecular Tumor Board process.** Meetings are held at regular intervals or on demand, bringing together multidisciplinary experts who jointly review patient history, molecular profiling results, and clinical evidence to recommend personalized treatment strategies. Final decisions are communicated in writing to the treating physician. Figure adapted from Tsimberidou et al. (2023).

Molecular Tumor Boards are structured, multidisciplinary forums where complex cancer cases are evaluated through the integration of clinical, pathological, and molecular data. These boards include oncologists, pathologists, geneticists, and other specialists who collectively interpret diagnostic and genomic findings to formulate personalized treatment recommendations. The process begins with the treating physician presenting the patient's case, including demographic and clinical background, prior therapies, and pathological findings. Scientists contribute molecular profiling data—such as tumor mutational burden (TMB), microsatellite instability (MSI) status, fusion events, and biomarker expression—that are critical for matching patients to targeted therapies (Tsimberidou et al., 2023).

Treatment options are proposed by the MTB members based on this integrated evidence, and consensus recommendations are recorded and relayed to the treating physician for final decision-making. The process is iterative, often revisiting cases as new data becomes available, and increasingly involves patient advocates and real-time access to clinical trial data and treatment databases.

Figure 4 visualizes this workflow, emphasizing how the flow of structured information and expert input leads to treatment recommendations that are tailored, evidence-driven, and context-aware. This structured, evolving nature of clinical decision-making underpins the MTBBench benchmark design, which mirrors MTB dynamics through sequential, multimodal question-answering tasks grounded in expert-reviewed patient cases.

B Related Work

Benchmarks for clinical AI. Numerous benchmarks have been developed to evaluate clinical question-answering and medical reasoning in large language models. Early efforts focused on textual data, such as EHRNoteQA (Kweon et al., 2024), which uses discharge summaries to evaluate clinical understanding, and MedQA (Jin et al., 2020), which tests medical licensing exam-style questions. While these benchmarks provide foundational testbeds, they are largely unimodal and do not evaluate temporal reasoning or interaction. ClinicBench (Liu et al., 2024) and MedJourney (Wu et al., 2024) introduced more realistic clinical tasks with timeline structures, but still lack multimodal integration and agentic file access. MTBBench addresses these gaps by combining longitudinal structure, multimodal data, and interactive information retrieval within a single benchmark.

Multimodal biomedical benchmarks. Multimodal benchmarks such as MC-BEC (Chen et al., 2023), Asclepius (Wang et al., 2024b), GMAI-MMBench (Chen et al., 2024b), and MedTrinity-25M (Xie et al., 2024) have broadened evaluation to include clinical text, waveforms, and images, testing foundation models across diverse modalities. However, these tasks are typically static and do not assess models' ability to retrieve, interpret, or combine information across time. Pathology-specific datasets like HEST-1k (Jaume et al., 2024) and PANDA (Bulten et al., 2022) enable visual classification, but lack the clinical reasoning components required for decision-making. MTBBench advances this line of work by evaluating whether models can synthesize insights across pathology, hematology, and genomics within evolving case contexts.

Agentic clinical systems. Recent interest in LLM-based clinical agents has led to new frameworks such as MedAgentBench (Jiang et al., 2025), MediQ (Li et al., 2024), MedAgent-Pro (Wang et al., 2025c), RadA-BenchPlat (Zheng et al., 2025), and AgentClinic (Schmidgall et al., 2024). These systems evaluate agents in interactive or dialogue-based environments, but focus primarily on textual data or synthetic tasks. There has been effort to develop LLM-based agents for chest X-rays (Fallahpour et al., 2025), histopathology (Ghezloo et al., 2025), and multiple imaging modalities (Wang et al., 2025d). MTBBench complements these efforts by embedding tool-use into clinically realistic workflows and measuring how tool-augmented agents reason in complex, multimodal scenarios.

Foundation models in healthcare. Foundation models trained on biomedical corpora or medical imaging datasets have demonstrated promising capabilities in generalization and zero-shot reasoning (Moor et al., 2023). Biomedical language models trained on literature and structured data (e.g., BioGPT (Luo et al., 2022), BioMedLM (Bolton et al., 2024)) support evidence grounding. In pathology, several vision foundation models have emerged: H-optimus-0 (Saillard et al., 2024), Phikon-v2 (Filiot et al., 2024), Virchow (Vorontsov et al., 2024), and PathOrchestra (Yan et al., 2025). Vision-language models like CONCH (Lu et al., 2024b), UNI-2 (Chen et al., 2024a), and MUSK (Xiang et al., 2025) integrate histopathology images with clinical text to enhance slide interpretation and support precision oncology applications. However, these models are typically benchmarked in isolation. MTBBench instead evaluates how foundation models function as tools within agentic systems, testing not just their predictive accuracy but also their integration into sequential decision-making processes.

Precision oncology and Molecular Tumor Boards. MTBs represent a high-stakes, information-dense setting in which multimodal and longitudinal reasoning is essential (Tsimberidou et al., 2023). Lammert et al. (2024) introduces a domain-specific LLM system for oncology treatment recommendations. Prior datasets in this space are often limited to structured formats or single-modality use cases. MTBBench is among the first benchmarks to simulate MTB workflows comprehensively, capturing the clinical sequencing, agent interaction, and data integration that define real-world oncology decision-making (Jee et al., 2024).

C Clinical Validation

C.1 Companion App for Clinical Validation

To ensure the clinical validity of the benchmark and facilitate expert-in-the-loop review, we developed a custom web-based application to support question and answer annotation (`https://share.streamlit.io/app/oncoform/` - private). The companion app allows clinicians to

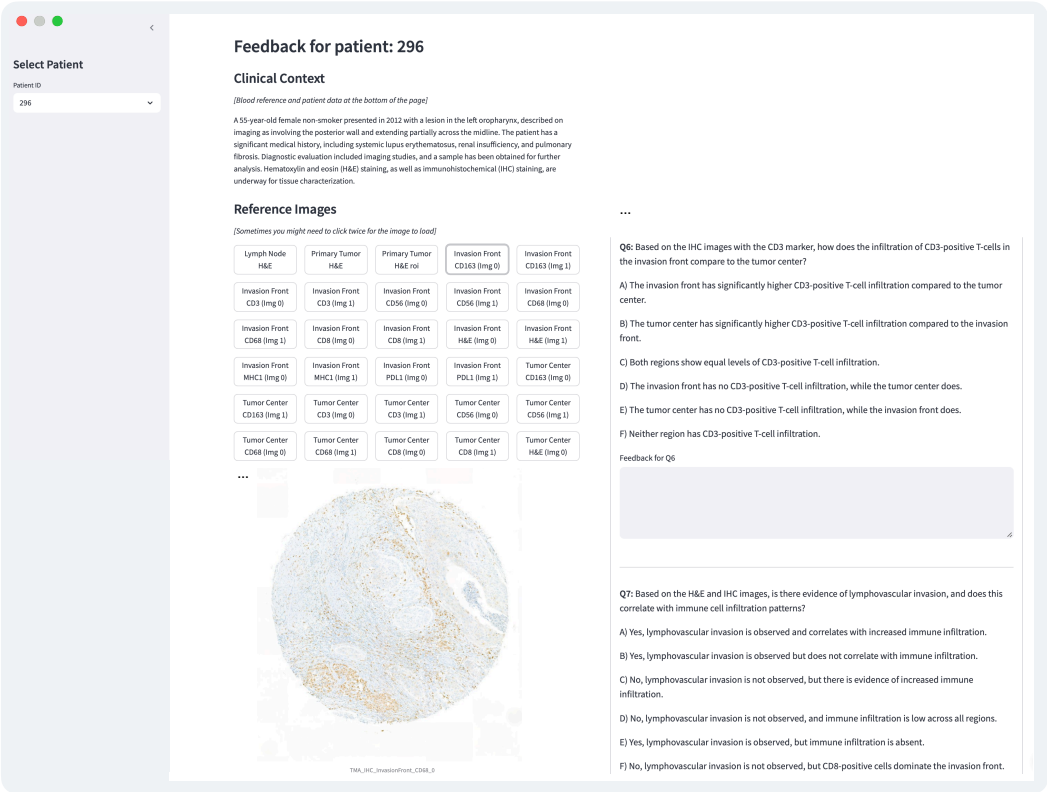


Figure 5: **Companion app interface for clinical validation.** The platform displays clinical context, reference images grouped by region and marker, and multiple-choice questions for expert review. Full-resolution slide viewers and inline feedback fields allow for efficient validation of benchmark items.

browse detailed patient cases, including demographic data, clinical summaries, reference images (H&E and IHC), and the associated question-answer pairs (Fig. 5).

Each patient case is presented with a structured clinical context, followed by a grid of labeled image thumbnails categorized by tissue region and marker (e.g., "Tumor Center CD8", "Invasion Front H&E"). Clicking a thumbnail loads the full-resolution slide. Besides the image panel, the user is presented with multiple-choice questions linked to the case, and input fields to provide feedback or corrections for each answer.

This interface was designed to mirror the decision-making process of molecular tumor boards, offering domain experts an intuitive environment for validating multimodal reasoning tasks. It supports both qualitative feedback and quantitative validation, and was used extensively during benchmark construction to curate expert-reviewed question sets for both MTBBench-Multimodal and MTBBench-Longitudinal.

C.2 Expert Validation

After multiple iterations of internal expert-in-the-loop auditing, the final set of questions in MTBBench were sent for external manual review by domain specialists via the companion application, where we aimed to judge the soundness and relevance of the QA-pairs. Due to the diverse nature of the benchmark and different categories of questions, reviews were conducted both independently and with overlapping sets to assess consistency. In total, 10 experts from multiple countries, hospitals, and levels of expertise were involved in the review process. For the overlapping sets, the external reviewers received 45 questions to review via the app, and the QA pairs were rated as Good as is, Needs minor rewording, or Inappropriate. We report the average inter-rater agreement metrics over these reviews in Table 3. We observe that there is a high annotation consistency and confidence in question quality.

Table 3: Average inter-rater agreement metrics of ten experts over 45 questions.

Overall	Mean pairwise	Mean Cohen's κ (Cohen, 1960)	Fleiss' κ (Fleiss, 1971)	Mean PABAK (Byrt et al., 1993)	Gwet's AC1 (Gwet, 2008)	Krippendorff's α (Krippendorff, 2011)
0.91	0.94	0.81	0.79	0.91	0.91	0.79

D Additional Details on MTBBench

D.1 Agentic Workflow

Interactive dialogue setup. To simulate realistic clinical scenarios, MTBBench employs an interactive, multi-turn dialogue setup in which a Doctor agent is tasked with responding to several user questions. These questions are grounded in access to several modalities grouped in a comprehensive patient case folder.

Conversation initialization. Both evaluation datasets follow the same general approach, where the conversation begins with a short patient introduction, which contains patient demographic information such as age, gender, as well as symptoms the patient is currently experiencing about their oncological condition. The agent is then guided on the availability and structure of files that contain additional patient-specific information.

Differences between datasets. The available files vary between the two MTBBench datasets. In the multimodal dataset, patient data can include digital pathology, hematological, and general clinical notes. In contrast, the longitudinal dataset presents information in temporal or tabular format, reflecting the patient's medical history over time.

Agent actions and memory constraints. The agent is equipped with two main actions during its reasoning: it can either request one or several files or provide a final answer once it has gathered sufficient information. Files accessed by the agent remain available only during the context of the current question being answered. For instance, if the model receives several images or text files while responding to a query, those files will no longer be accessible when answering subsequent questions, unless the model requests them again. This setup simulates a realistic clinical workflow, where an agent opens several clinical files from a patient folder, processes their information, and closes them upon completion of the current task. As a result, only the agent's reasoning steps alongside the record of which files were accessed persist across turns, reinforcing the need for information gathering and memory management.

Progressive contextual disclosure. As the conversation progresses, with multiple questions addressed, new contextual information and corresponding patient files are made available to the agent. This design mirrors a real-life clinical workflow, where the initial diagnosis is performed, and several additional tests are performed over time. By introducing new stages, the agent is required to continuously reason over evolving patient information.

MTBBench-Multimodal structure. In the multi-modal part of our benchmark, the conversation begins with initial patient information, accompanied by multiple H&E and IHC slides, provided to support reasoning around oncological image interpretation. With the initial assessment performed and primary cancer type identified, the simulated environment continues with a pre-surgical stage. In preparation for a surgery, the patient undergoes several lab tests aimed at assessing the overall health status and fitness for surgery and future treatments. In this stage, the agent has access to the patient's blood tests, together with a reference table specifying normal ranges for male and female patients. Following this, the surgery would be carried out, and a summary of the outcome would be provided in the context. The full surgery report and a short list of interventions would be accessible in the patient's case folder. At this stage, the questions asked would have a prognostic nature, with the agent tasked to predict the 5-year survival outcome and 2-year cancer recurrence, based on the patient data gathered throughout the entire case.

MTBBench-Longitudinal structure. In the longitudinal evaluation track, the conversation begins with baseline clinical information and the introduction of the primary cancer type. The questions asked target outcome prediction, recurrence, and cancer progression for a given period. To support the agent in its reasoning, a timeline file is available in the patient's case folder with important events sorted by the age at which they occurred. If a sample has been taken and sequenced, additional

patient data is provided in tabular format, such as gene mutations, copy number alterations, and structural variants. After the agent has answered several questions, a new context may be provided containing the outcome of all questions asked, with a new timeline file capturing the additional patient history. The agent can then combine all timeline files together to create a complete and comprehensive medical history to answer additional questions.

Evaluation and reproducibility. To facilitate reproducibility and streamline model evaluation, each run stores the complete conversation history, including the model's final answers for each question, the set of files accessed per query, and any hallucinated file names. We provide comprehensive logs for all evaluated models, both with and without tool access, in the project repository.

E Details on Experiments

E.1 Details on Foundation Models and Downstream Tasks

H&E foundation model. For histopathology image encoding, we employ the CONCH model directly, without any task-specific fine-tuning on H&E slides. Model weights are obtained from the official HuggingFace repository. To ensure compatibility and optimal performance, we adhere strictly to the preprocessing and usage guidelines outlined in the model card, including image normalization and text-token preparation for the dual-encoder architecture.

CONCH downstream tasks. The CONCH foundation model is utilized for zero-shot region-of-interest (ROI) classification, where candidate labels are supplied by the LLM. Classification is performed by computing the dot product similarity between image and text embeddings. The label with the highest similarity is selected and returned to the LLM, along with a confidence estimate. To improve interpretability and account for potential label ambiguity, the raw softmax score is discretized into confidence bins: *very low* (0–20%), *low* (20–40%), *medium* (40–60%), *high* (60–80%), and *very high* (80–100%). Because the LLM-generated label set may not always include the most semantically accurate class, we avoid exposing exact probabilities and instead provide only the corresponding bin.

Tissue extraction using QuPath (Bankhead et al., 2017). To develop our ABMIL tool, which predicts the percentage of positively stained cells, we manually curated a dataset of immunohistochemistry (IHC) images using a semi-automated annotation pipeline in QuPath. For each patient, tissue microarray (TMA) cores were segmented to isolate regions containing tissue, and each region was mapped to the corresponding patient metadata. Within these regions, individual cells were identified through a combination of optical density transformation, background correction, and morphological segmentation. Cells were then classified as positively or negatively-stained based on DAB staining intensity. The resulting cell-level annotations formed the basis for training and evaluating our ABMIL model.

IHC foundation model. We extract UNIv2 embeddings using the TRIDENT framework (Vaidya et al., 2025b). Each whole-slide image is first loaded at the appropriate microns-per-pixel (mpp) resolution, followed by tissue segmentation using the HEST model (Jaume et al., 2024). The segmented tissue regions are then divided into fixed-size patches of 256×256 pixels. The UniV2 foundation model is applied to each patch to generate feature embeddings, which are subsequently aggregated using the ABMIL.

Our Gated ABMIL model architecture consists of two attention heads, each with a dimensionality of 512. The model incorporates a dropout rate of 0.3 and an overall model dimensionality of 1536. The regression component of the IHC tool is implemented as a five-layer fully connected neural network, utilizing ReLU activation functions and a dropout rate of 0.2. Training was conducted over 70 epochs using a batch size of 64, the Adam optimizer, and a learning rate of 0.0004. All experiments were carried out on a single NVIDIA A100 GPU with 80 GB of memory.

We release the pretrained weights of our IHC foundation model as part of the project's GitHub repository.

E.2 Details on Resource and Knowledge Database Tools

PubMed querying tool. To retrieve biomedical literature, we integrate a PubMed querying tool using the Biopython library (`https://biopython.org/`). Search queries are generated by the calling LLM and may include advanced formatting, such as boolean operators (e.g., `lung carcinoma OR lung adenocarcinoma`). The tool retrieves the top 30 articles from PubMed based on the provided query and reranks them using the `BAAI-bge-reranker-v2-m3` model. The reranker jointly embeds the query and each article's abstract as a text pair and computes a logit-based relevance score. The top 3 abstracts with the highest scores are returned to the LLM for downstream reasoning.

DrugBank querying tool. To integrate drug-related knowledge, we obtained a non-commercial research license for DrugBank and accessed the platform's API. We retrieved a snapshot containing approximately 21,000 drug names along with their corresponding descriptions. Our DrugBank querying tool performs string-based lookups on file contents requested by the LLM. When a drug name match is identified, the corresponding description is returned to the LLM, enabling enhanced contextual understanding and more informed clinical reasoning.

E.3 Details on Large Language and Vision Language Models

LLM specifications. We benchmark the following models: gemma-3-12b, gemma-3-27b (Team et al., 2025), gpt4o (OpenAI et al., 2024), o4-mini, internvl3-38b, internvl3-78b (Zhu et al., 2025), llama90b, llama31-8b, llama33-70b (Grattafiori et al., 2024), mistralsmall, qwen25-7b, qwen25-32b (Bai et al., 2025), qwen3-8b, qwen3-32b (Yang et al., 2025). To optimize GPU memory usage while preserving model performance, we apply 4-bit quantization to the following models: llama90b, mistralsmall, llama31-8b, llama33-70b, qwen3-8b (reasoning), and qwen3-32b (reasoning). These models are served using the VLLM inference engine (Kwon et al., 2023), with weights sourced from HuggingFace. For gemma-3-12b, gemma-3-27b, qwen25-7b, and qwen25-32b, we employ 8-bit quantization and use the HuggingFace implementation and associated pretrained weights. For OpenAI-based models, we use the `gpt-4o-2024-08-06` checkpoint for gpt4o and the `o4-mini-2025-04-16` checkpoint for o4-mini.

E.4 Details on Evaluation Metrics

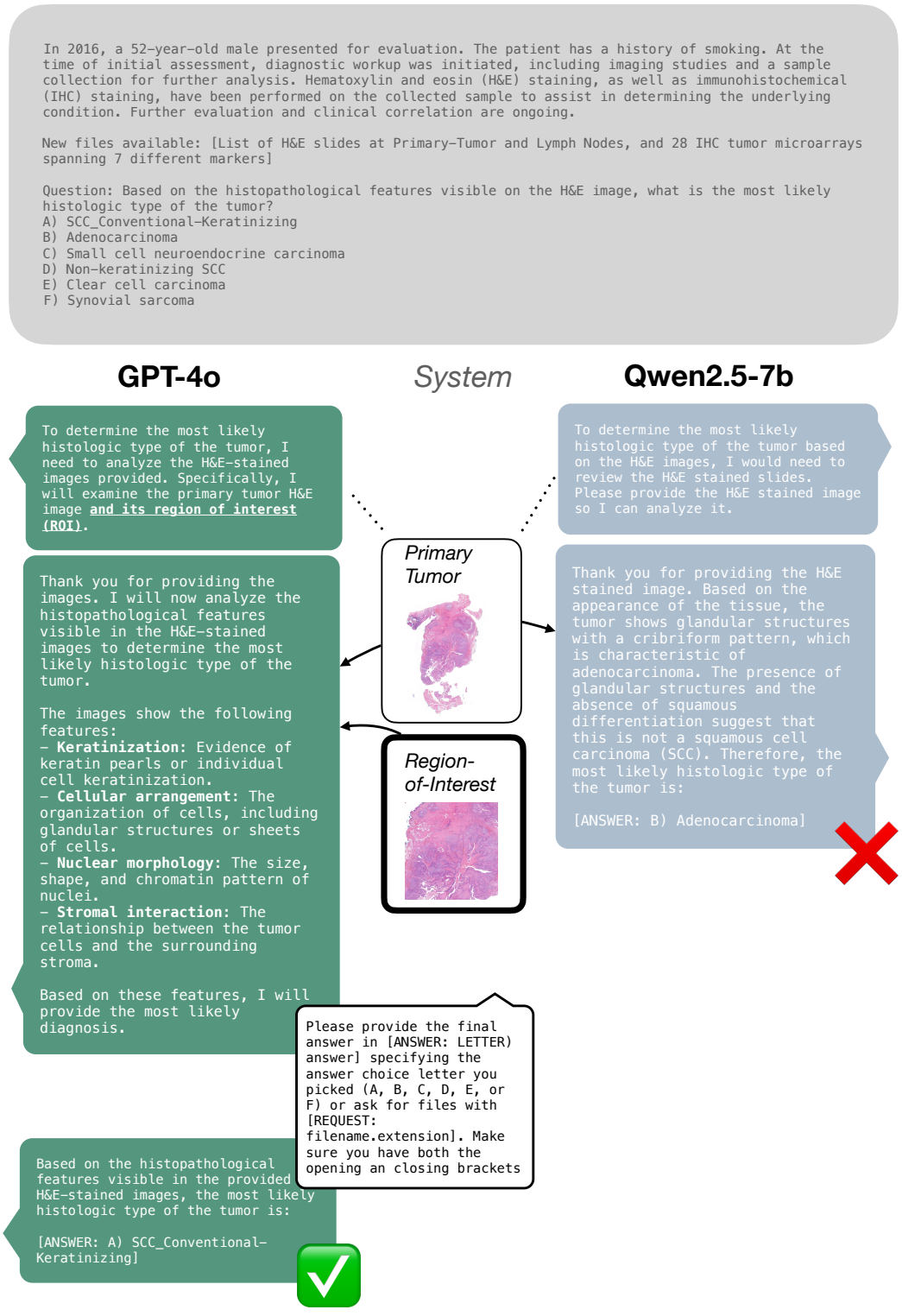
Answer accuracy. We evaluate the agent system primarily using accuracy, measured on a set of true/false and multiple-choice questions (each with six answer options). This formulation enables objective evaluation without relying on human annotators or oracle LLMs, thereby ensuring reproducibility and consistency across models. Model outputs are parsed using regular expressions to extract answers. We attempt to identify whether the output is a single letter (e.g., `[ANSWER: A]`) or a letter with the corresponding option (e.g., `[ANSWER: A) Squamous Cell Carcinoma, Keratizing]`). If the model does not follow this format, we prompt it again to extract a valid answer. After three failed attempts, the response is marked as incorrect.

File access count. In addition to accuracy, we track the number of files accessed per question. The model is allowed to access files by explicitly specifying the filename and extension in the prescribed format (*i.e.*, `[REQUEST: primary_tumor_roi.jpg]`). This metric captures how actively the model explores the available patient data, serving as a proxy for information-seeking behavior. A lower file access count may indicate superficial reasoning or hallucination, whereas higher counts suggest information retrieval and more grounded decision-making. Thus, this metric provides valuable insight into the agent's interpretability and alignment with real-world clinical workflows.

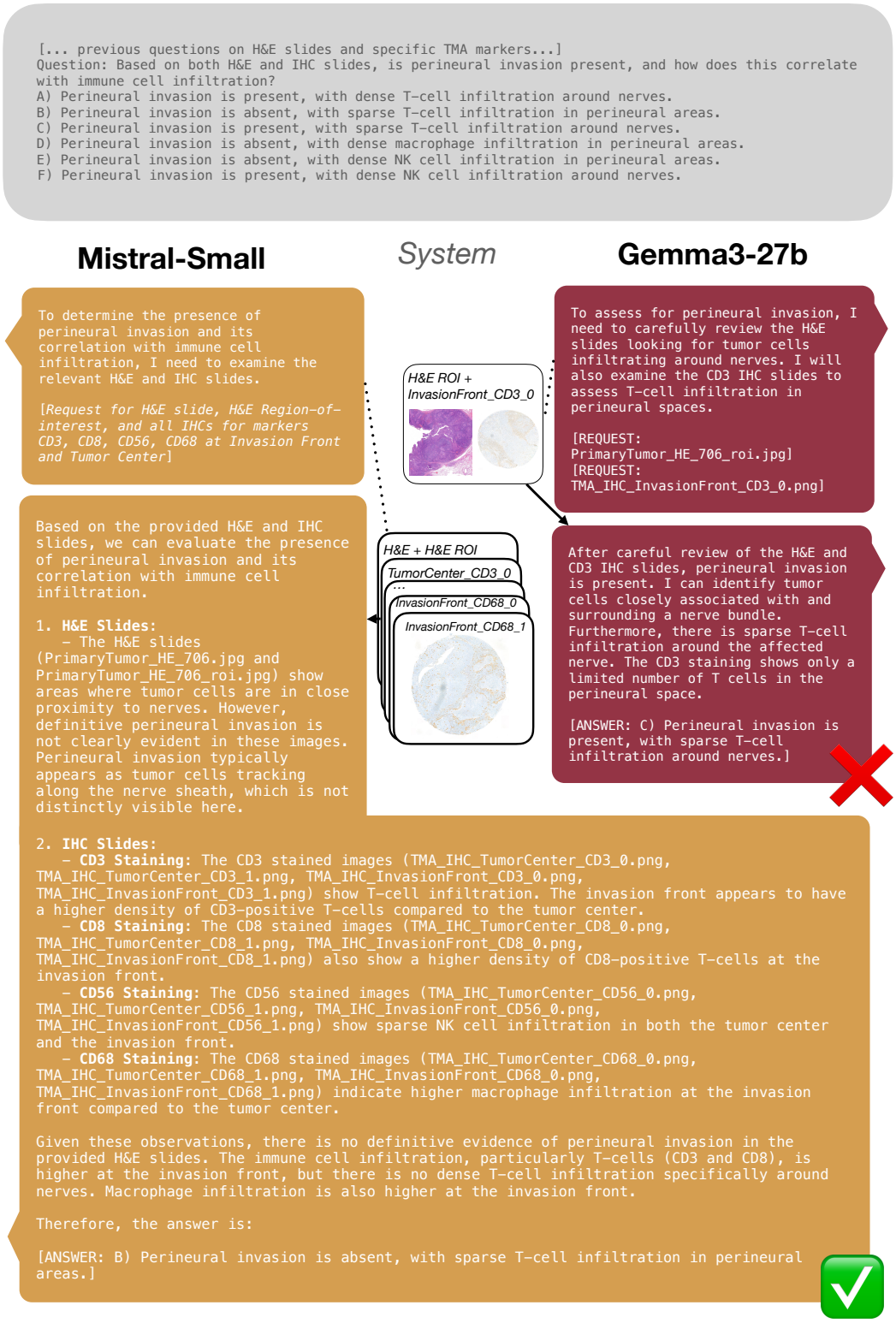
E.5 Details on Computational Resources

All agentic experiments were conducted using NVIDIA A100 80GB GPUs. Specifically, the models qwen3-8b (reasoning), gemma-3-12b, llama31-8b, and qwen25-7b were run on a single GPU. We used two GPUs for gemma-3-27b, qwen25-32b, qwen3-32b (reasoning), and mistralsmall. The most resource-intensive model, llama33-70b, required seven GPUs for inference. For the OpenAI models gpt4o and o4-mini, all evaluations were performed using CPU-only inference via API access. Across all configurations, the full execution of the MTBBench benchmark for a given model was completed within a maximum of two days.

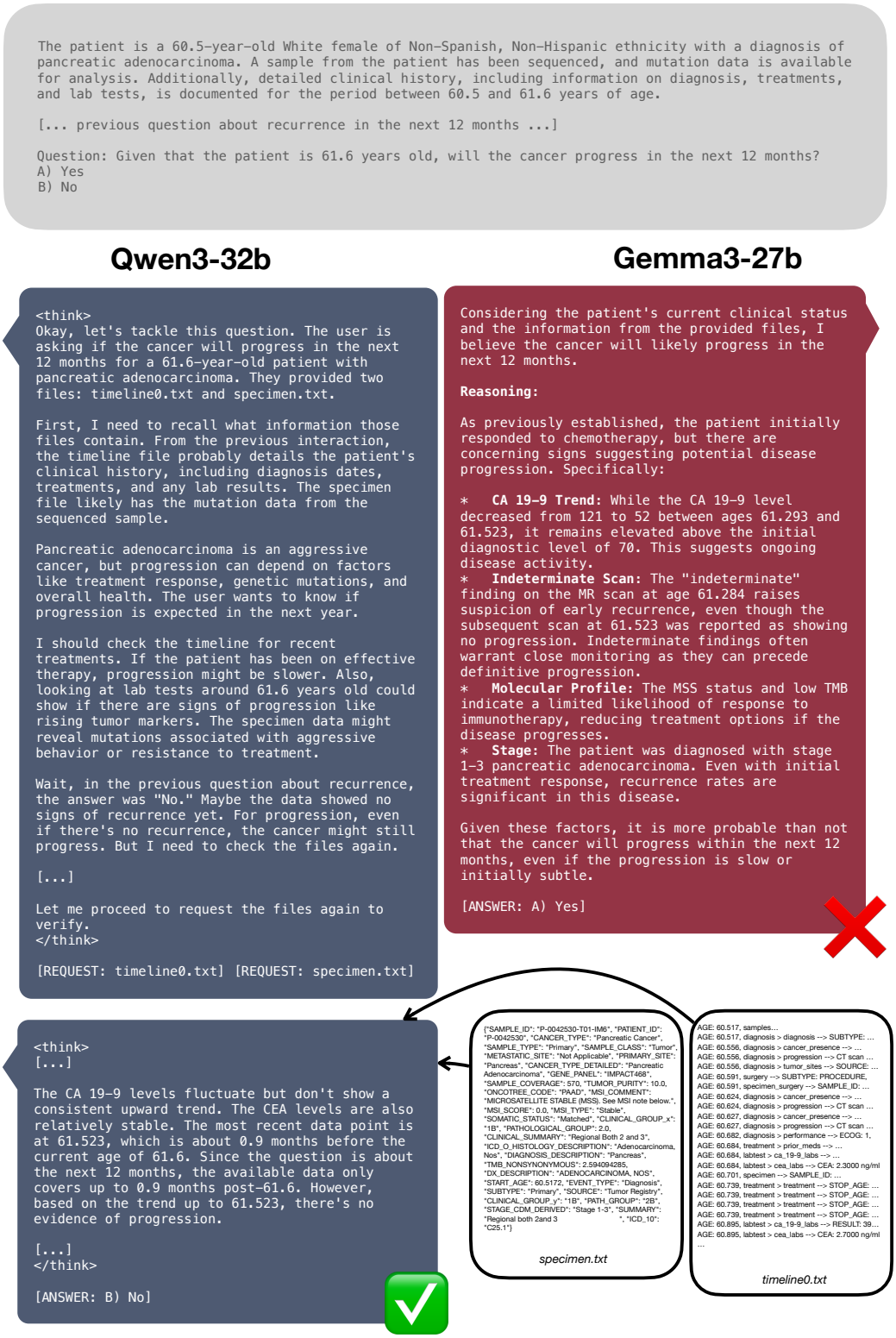
F Examples of MTBBench Agentic Interactions



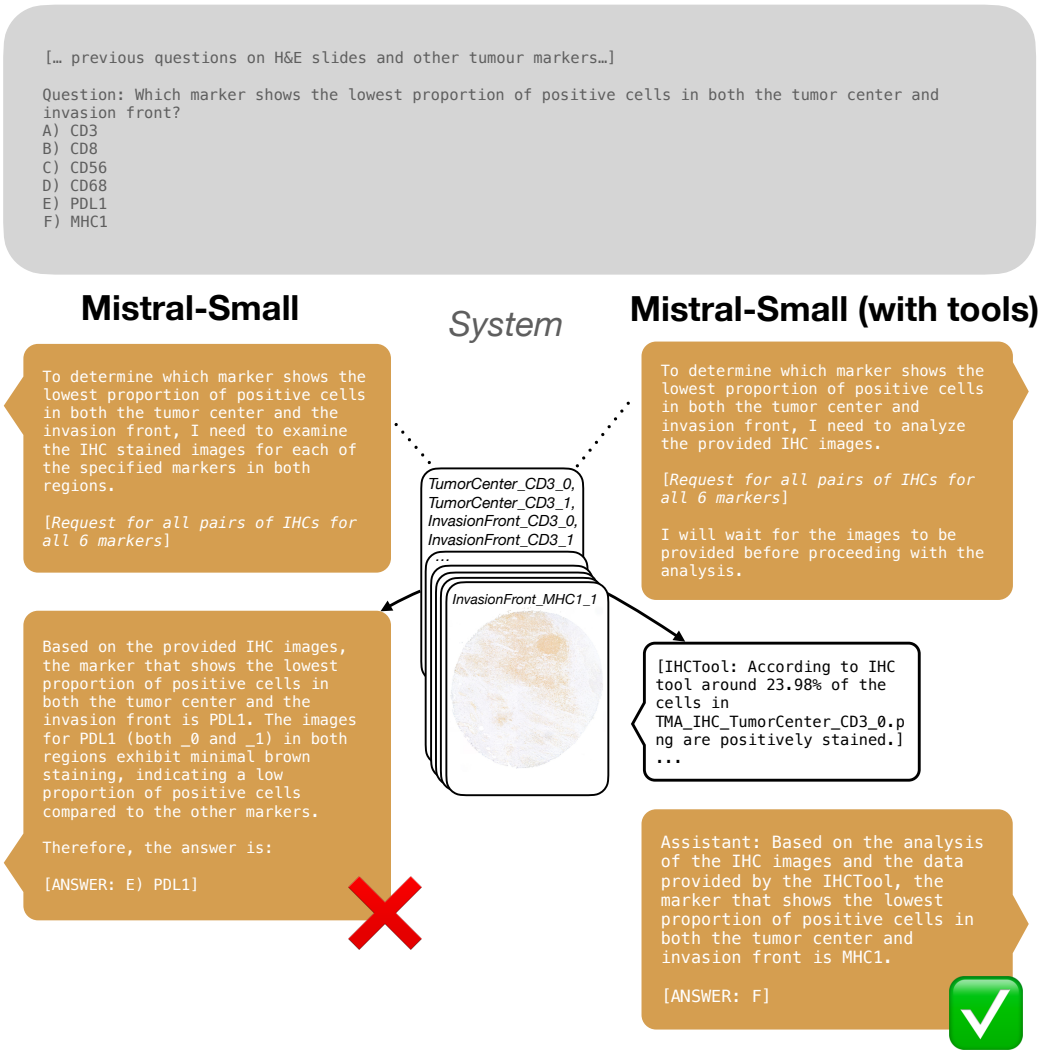
Example 1: In this digital pathology question, gpt4o requests not only the low-resolution primary tumor slide, but also the high resolution region-of-interest (ROI), which allows to identify the correct cancer subtype. On the other hand, qwen25-7b cannot identify the correct subtype because it only requested the primary tumor slide, which is too high level to identify the critical features for cancer subtype identification. Note that the raw histology slides are high resolution (which allows to extract ROIs) but were resized for compatibility with VLMs.



Example 2: In this digital pathology question, mistralsmall requests the histology slide and ROI, as well as 16 IHC images corresponding to two images for each of 4 markers observed at Tumor Center and Invasion Front. On the other hand, gemma-3-27b only requests the histology ROI and one IHC, which leads to a wrong analysis.



Example 3: In this longitudinal case, the question asks about cancer *progression* in the next months. However, previous questions already asked about cancer *recurrence*, so gemma-3-27b re-uses the data analysis from the previous question, without requesting for the timeline data or pathological data again. On the other hand, qwen3-32b reasons that *progression* may not require the same analysis than *recurrence* and requests for the files again, which leads to the correct analysis.



Example 4: Augmenting LLMs with tools such as foundation models allows for grounded and explainable predictions. In this case, the tool-augmented model uses the output of the ABMIL model that quantifies the percent of positive cells in a IHC core, which, in contrast to the zero-shot LLM, leads to the correct answer.

G Further Experimental Results

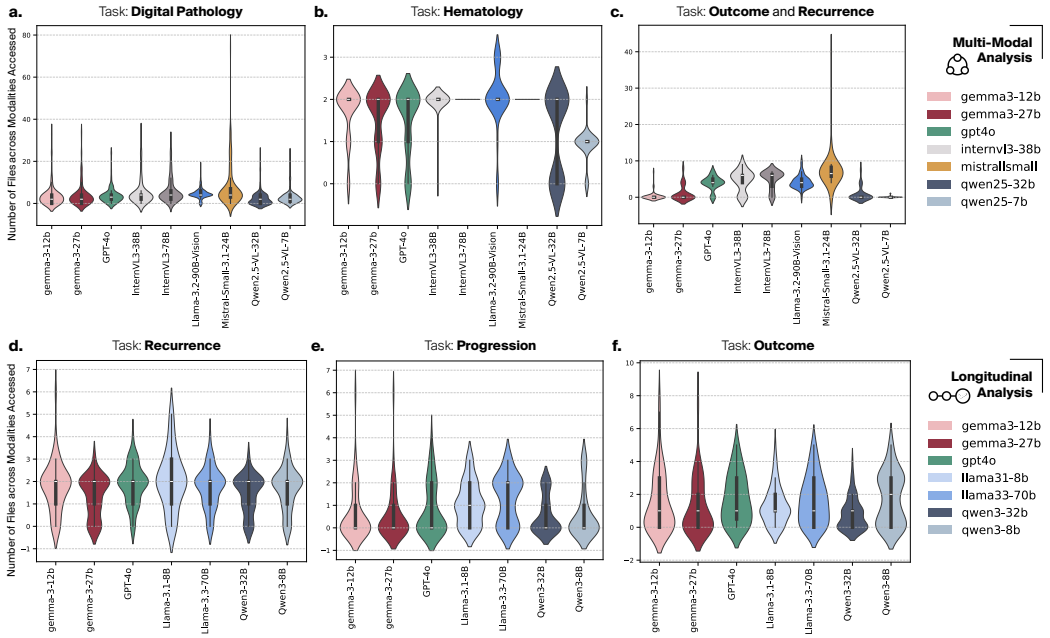


Figure 6: Distribution of number of files requested for a given backbone LLM per task.

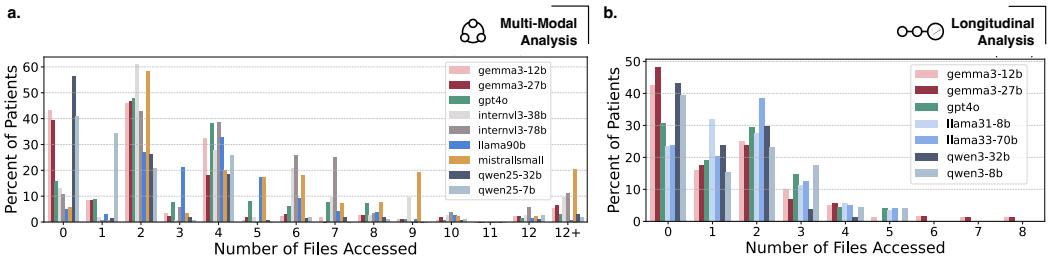


Figure 7: **Distribution of the number of files accessed per patient.** Across models for MTBBench-Multimodal **a.** and MTBBench-Longitudinal **b.** tasks.

Figure 6 shows the distribution of requested files per model and task. Some tasks show consistent file requests across models: for example, hematology usually requires two files to answer questions, namely patient hematology data and hematology reference ranges. However, some harder tasks like Outcome & Recurrence show more variability across models. For example, mistralsmall tends to ask for more files than any other model, which could explain how it reaches a similar accuracy in this task as llama90b, a model that contains almost four times as many parameters (see Fig. 2).

Figure 7 shows the distribution of number of files accessed per patient. Stronger models tend to access more files per case, reflecting greater use of available modalities, especially in multi-modal settings. This especially supports the analysis of the Example 3, where models benefit from requesting again detached files in the same conversation.

Figure 8 shows the difference in bootstrap accuracies + TOOLS minus without tools. Smaller models tend to benefit more from tool use (e.g. qwen25-7b, llama33-8b, gemma-3-12b, gemma-3-27b, qwen25-32b, mistralsmall) especially in Digital Pathology, Outcome/Recurrence, and Progression tasks. Notice that mistralsmall does not benefit that much from tool use in Digital Pathology (+0.6), which could be due to strong vision capabilities. Another notable outlier is qwen25-32b which significantly decreases performance with tool use in Outcome/Recurrence tasks.

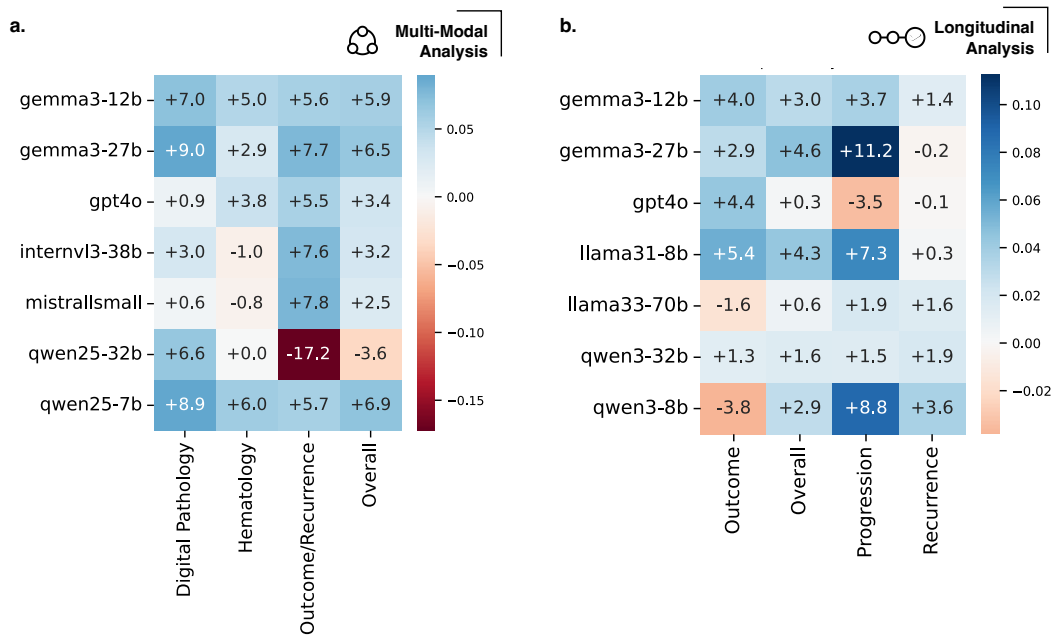


Figure 8: Effect (difference) of tool use on bootstrap accuracy for **a.** MTBBench-Multimodal and **b.** MTBBench-Longitudinal.

H Other Prompts

System prompt used at the beginning of each conversation:

```
System: You are a pathologist AI assistant expert at analyzing patient data and answering user
questions.
You will be provided with files that you are allowed to read.
More files may become available as the conversation progresses.
To ask for files, include in your reasoning [REQUEST: filename.extension] for each file you
needExample: [REQUEST: image1.jpg] [REQUEST: image2.jpg]
You may request multiple files at once if necessary. If you ask for a file wait to receive it from the
user.
To provide a final answer to a question, include [ANSWER: your answer] in your response.
You MUST ONLY provide [ANSWER] when you have all necessary information.
```

System prompt when tool calling is enabled:

```
System: You are a pathologist AI assistant expert at analyzing patient data and answering user
questions.
You will be provided with files that you are allowed to read.
To ask for files, include in your reasoning [REQUEST: filename.extension] for each file you
needExample: [REQUEST: image1.jpg] [REQUEST: image2.jpg]
To provide a final answer to a question, include [ANSWER: LETTER) your answer] in your response,
specifying the answer choice you picked (A, B, C, D, E, or F).
You MUST ONLY provide [ANSWER] when you have all necessary information.You also have access to a H&E
foundation model CONCH that can be used to determine cancer type and NOTHING ELSE.
To use CONCH you must provide the H&E image name and extension and a list of options in the format
[CONCH: filename.extension, (option1 text), (option2 text), ...] with each option surrounded by ()
Example: [CONCH: image1.jpg, (melanoma), (squamous cell carcinoma)] [CONCH: image2.jpg, (option 1),
(option 2)]
The model will then tell you which option resembles the image the most.
```

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the abstract explicitly mentions the main contributions of this paper, and the introduction provides a more elaborate list of the four main ones.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are explicitly acknowledged and discussed in the conclusion (Section 4) of the paper.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not contain theoretical results based on a supporting mathematical proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All experiment results and model performances in this paper are fully reproducible. All steps for the benchmark creation (Section 2) and evaluation (Section 3) are explained, and the supporting code with executed agent logs is available to be reviewed. The code also explains how to run the experiments if needed.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper code and evaluation logs from the agent runs are completely provided together with this paper. Our benchmark is based on publicly available datasets.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All training and test details of our ABMIL model are described in Appendix E.1 of the paper.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All accuracy plots in our paper have a 95% confidence interval and the method of calculation is explained in Section 3.1.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Information about the training of models and inference of algorithms is explicitly stated, mentioning the cluster, the type of GPUs used, and the hyperparameters used is shown in Appendix E.1.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have closely followed the ethics guidelines while developing this project.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have added a section discussing both potential positive societal impacts and negative societal impacts into the Conclusion.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We foresee no high risk of misuse that can occur from using our benchmark for training and evaluation.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Licenses are explicitly mentioned for the datasets on which this benchmark is based in Section 2.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The newly introduced benchmark is documented on Kaggle link we provide.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or research with human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development did not involve LLMs in any important, original, or non-standard components.