MEDBENCHAGENT: AN AGENTIC FRAMEWORK FOR AUTOMATED MEDICAL BENCHMARK CONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The growing use of large language models (LLMs) in the medical domain drives demand for benchmarks to evaluate their capabilities. Constructing these benchmarks manually is labor-intensive, motivating the use of LLM-based agents to automate the process. However, general-purpose agents often lack sufficient medical expertise for benchmark construction, while existing medical benchmark agents typically target specific tasks or modalities and struggle to consistently produce high-quality benchmarks, limiting both generalizability and reliability. To address these limitations, we introduce MedBenchAgent, an agentic framework for reliable medical benchmark construction across multiple tasks and modalities. MedBenchAgent follows a two-stage workflow supported by an extensible medical tool library. The benchmark design stage decomposes the user request into subtasks, integrates subtask responses into a benchmark specification, and uses this specification to generate template items. To improve design reliability, a field-level quality model guides specification refinement. The data construction stage retrieves relevant medical samples for each template item, performs fine-grained annotation on the retrieved samples, and instantiates final benchmark items from the templates and annotated data. To improve annotation reliability, an evidence-guided Thought–Tool–Observation loop verifies and refines annotations. The medical tool library combines diverse medical resources and specialized tools with dynamic tool expansion to support both stages and improve generalizability. Experiments on five benchmark construction tasks spanning CT, pathology, medical video, electrocardiography (ECG), and electronic health records (EHR) show that MedBenchAgent outperforms existing methods overall across LLM jury evaluation, downstream fine-tuning, and human evaluation.

1 INTRODUCTION

Recent advances in large language models (LLMs) have expanded their use across a range of medical applications, including diagnosis, treatment recommendation, clinical decision-making, and medical report generation (Shi et al., 2026a; Zhou et al., 2024; Lu et al., 2024a; Tian et al., 2025; Yang et al., 2024; Albaqshi et al., 2025). These applications involve diverse data modalities and require different capabilities from LLMs, making systematic evaluation of their performance increasingly important. Medical benchmarks provide a standardized basis for such evaluation, enabling meaningful comparisons between LLMs and helping identify their strengths and limitations.

Traditional medical benchmark construction relies heavily on expert involvement and manual curation, making it labor-intensive and difficult to scale. Recent studies use LLM agents to automate benchmark construction, improving efficiency and scalability (Butt et al., 2025; Yoa et al., 2026; Jiang et al., 2026; Li et al., 2025). However, general-purpose agents are designed for broad, domain-agnostic use and often lack sufficient medical expertise, which can limit their effectiveness in medical benchmark construction (Butt et al., 2025; Li et al., 2025; Guo et al., 2026). Specialized medical benchmark agents incorporate domain-specific knowledge and designs to better support medical benchmark construction (Yan et al., 2026; Trinh et al., 2026; Wang et al., 2026b). Nevertheless, they still face two key limitations. *1) Generalizability.* Existing specialized agents are typically tailored to a particular modality or task family, limiting the range of medical benchmarks they can construct. This limited generalizability is reflected in Fig. 1(a), where the five medical benchmark construction

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

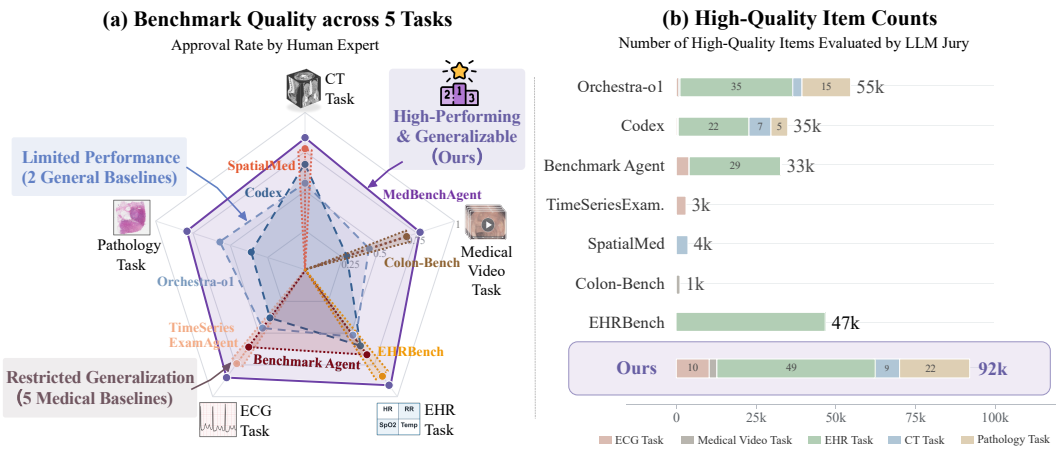

Figure 1: (a) Comparison of benchmark quality. Each axis represents one of five medical tasks, and values for each method indicate the proportion of sampled items approved by both human experts. (b) Comparison of high-quality item counts. Each method generates items from the same source data and aligned benchmark requests within each task; only items whose mean scores across the three judges exceed 8 in all four LLM evaluation dimensions are counted.

baselines each cover only one or two tasks in our evaluation. *2) Reliability.* Existing specialized agents still face challenges in reliably maintaining the quality of generated benchmark items. This reliability challenge is illustrated in Fig. 1(b), where these agents produce only a limited number of high-quality items from the same source data and aligned benchmark requests within each task.

To address these challenges, we introduce MedBenchAgent, an agentic framework for reliable medical benchmark construction across diverse medical modalities and tasks. The framework follows a two-stage workflow comprising Benchmark Design and Data Construction, supported by an extensible Medical Tool Library. First, Benchmark Design decomposes high-level user requests into focused subtasks to better capture the underlying intent, generates corresponding responses, and integrates them into benchmark specifications, which then guide the generation of template items. To improve design reliability, we introduce a field-level quality model trained with supervision derived from general and medical benchmark papers to assess individual specification fields and guide progressive refinement. Second, Data Construction uses multiple distinct queries for each template item to retrieve diverse, relevant medical samples. It performs fine-grained annotation on the retrieved samples and instantiates final benchmark items from the templates and annotated data. To improve annotation reliability, we employ a Thought–Tool–Observation loop to verify and refine annotations against criteria defined by each item's medical evidence. Supporting both stages, the Medical Tool Library provides access to diverse medical resources and specialized tools, incorporating new tools when existing capabilities are insufficient. Its broad resource coverage and dynamic tool expansion jointly improve generalizability across medical tasks and modalities.

Our contributions are as follows:

1) We introduce MedBenchAgent, an agentic framework for reliable medical benchmark construction across diverse tasks and modalities.

2) We improve reliability at both stages: the benchmark design stage uses a field-level quality model to guide specification refinement, while the data construction stage combines evidence guidance with a Thought–Tool–Observation loop to improve annotation quality.

3) We improve generalizability through a Medical Tool Library that combines diverse medical resources and specialized tools with dynamic tool expansion.

4) Experiments on five medical benchmark construction tasks spanning CT, pathology, medical video, ECG, and EHR show that MedBenchAgent outperforms general-purpose and specialized medical agents overall in LLM jury evaluation, fine-tuning effectiveness, and human evaluation.

2 Related Work

Medical Benchmarks. Medical benchmarks for LLM evaluation span diverse data modalities and clinical tasks. PathVQA (He et al., 2020) and PathMMU (Sun et al., 2024) assess pathology understanding. 3D-RAD (Gai et al., 2025), CT-Bench (Zhu et al., 2026), and DeepTumorVQA (Chen et al., 2026) support CT-based visual question answering (VQA). ECG-QA (Oh et al., 2023) and ECGBench (Liu et al., 2024b) cover signal-based ECG question answering (QA) and image-based interpretation, respectively. Medical video benchmarks include MedVidBench (Su et al., 2026) and SurgCoT (Wang et al., 2026a), alongside SurgeryVideoQA and MIMICEchoQA (Thapa et al., 2025). EHRNoteQA (Kweon et al., 2024) and EHRStruct (Yang et al., 2026) support EHR-based QA for clinical information retrieval and reasoning. However, constructing these medical benchmarks still relies substantially on expert-guided design and validation, limiting scalability.

Agents for Benchmark Construction. To reduce manual effort, recent work has explored using LLM agents to automate benchmark construction. BenchAgents (Butt et al., 2025) coordinates agents for benchmark planning, generation, verification, and evaluation. AutoBencher (Li et al., 2025) uses LLM-guided search to construct benchmarks according to specified criteria. AutoConverter (Zhang et al., 2025b) uses agents to convert open-ended VQA into multiple-choice questions. TRACE (Guo et al., 2026) generates harder task variants through agent exploration and trajectory validation. LLM agents have also been applied to domain-specific benchmark construction (Vergopoulos et al., 2025; Lee et al., 2025; Xie et al., 2026a; Qiang et al., 2026). However, these approaches are generally domain-agnostic or tailored to non-medical settings, leaving the specific needs of medical benchmark construction insufficiently addressed.

Agents for Medical Benchmark Construction. Recent studies have begun to explore LLM agents for medical benchmark construction. For text-based medical benchmarks, agents curate and validate real-world clinical cases (Yan et al., 2026), generate guideline-grounded benchmarks (Chen et al., 2025; Unell et al., 2025), and construct counterfactual medical QA for bias evaluation (Xiao et al., 2025). For multimodal medical benchmarks, agents generate spatial QA from 3D CT scans (Trinh et al., 2026), annotate colonoscopy videos (Hamdi et al., 2026), and construct clinical evaluation tasks from case reports (Wang et al., 2026b) and consultations (Shi et al., 2026b). Related approaches include EHRBench (Xie et al., 2026b), which uses an LLM-based pipeline to generate QA from EHRs. Cross-domain agent frameworks, including Benchmark Agent (Xiong et al., 2026) and TimeSeriesExamAgent (Gwiazda et al., 2026), have also been applied to benchmark construction in selected medical settings. Existing methods largely target specific modalities or tasks, limiting generalizability, while ensuring benchmark reliability remains challenging.

3 Methods

3.1 Problem Setup

We study the automated construction of medical benchmarks from user requests and medical data to evaluate LLMs' medical capabilities.

Input. The inputs are a natural-language benchmark request and medical data. (1) The request describes the intended benchmark at a high level or in detail and may specify its evaluation scope, task types, and target item count. (2) The medical data comprises source samples, together with any available annotations and data descriptions.

Output. The output is a benchmark aligned with the request, consisting of multiple-choice or open-ended short-answer items. Each item includes relevant medical data, a question, and a reference answer. Multiple-choice items additionally include answer options.

3.2 MedBenchAgent Framework

MedBenchAgent coordinates specialized agents in a two-stage workflow comprising Benchmark Design and Data Construction, supported by a shared Medical Tool Library.

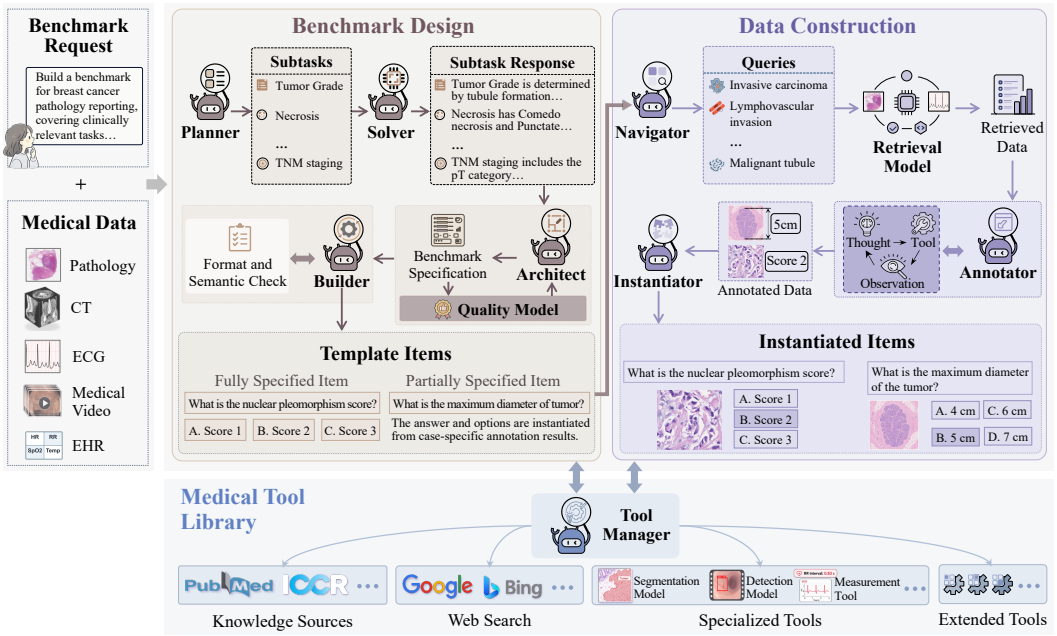

Figure 2: Overview of MedBenchAgent. Given a benchmark request and medical data, MedBenchAgent constructs benchmarks in two stages. (1) Benchmark Design: the Planner decomposes the request into subtasks, and the Solver generates corresponding responses. The Architect integrates these responses into a specification and refines it using feedback from the quality model. The Builder generates template items from the specification. (2) Data Construction: for each template item, the Navigator generates queries and selects a retrieval model. These are used to retrieve candidates. The Annotator annotates them via an evidence-guided Thought–Tool–Observation loop, and the Instantiator combines templates and annotated data into final items. The Tool Manager provides both stages with external knowledge and specialized capabilities via the shared Medical Tool Library.

Benchmark Design. This stage translates the benchmark request into a specification that guides template item generation. As shown in Fig. 2, the Planner agent first decomposes the benchmark request into complementary subtasks to capture user intent more fully and identify the information needed for benchmark design. The Solver agent then generates a detailed response to each subtask using information obtained from tools. The Architect agent integrates these responses into a benchmark specification defining the task taxonomy, item formats, evaluation protocol, and other construction requirements. A quality model evaluates the resulting specification and provides field-level scores to the Architect for progressive refinement. Based on the refined specification, the Builder agent generates template items that may be either fully specified, with all required components such as the question and answer options predefined, or partially specified, with some elements left to be completed during subsequent annotation and instantiation. Inspired by Gwiazda et al. (2026), we check and revise template items for format and semantic errors, as detailed in Appendix A.3.

Data Construction. This stage converts template items into final benchmark items through data retrieval, annotation, and instantiation. As depicted in Fig. 2, the Navigator agent first generates multiple distinct queries for each template item, covering different medical concepts to support varied and comprehensive retrieval results. It also autonomously selects a retrieval model appropriate for the task and modality, such as TITAN (Ding et al., 2025) for the breast pathology reporting task and M3D-CLIP (Bai et al., 2024) for the 3D CT spatial task. The model encodes the queries and medical data into a shared embedding space for retrieval. See Appendix A.4 for retrieval details. The Annotator then processes and annotates the retrieved candidates through an evidence-guided Thought–Tool–Observation loop, repeatedly evaluating tool outputs and refining annotations until they are sufficiently supported by medical evidence. Finally, the Instantiator assigns the corresponding annotated data to each template item and uses the annotations to revise or complete the question, reference answer, and any answer options as needed. The completed items form the final benchmark.

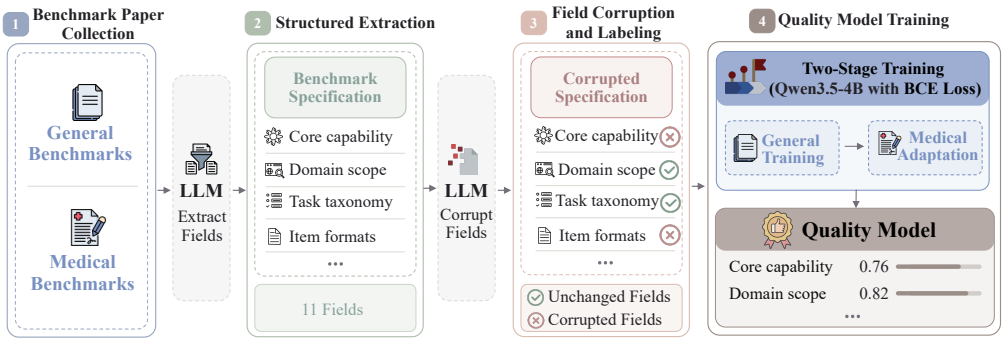

Figure 3: Training pipeline for the field-level quality model. An LLM extracts benchmark specifications with 11 fields from general and medical benchmark papers. Randomly selected fields are corrupted by an LLM and labeled negative, while unchanged fields are labeled positive. Using the resulting specifications and field-level labels, Qwen3.5-4B is trained with BCE loss, first on general-domain specifications and then medical-domain specifications, to predict per-field quality scores.

Medical Tool Library.

Throughout both stages, the shared Medical Tool Library supports the workflow agents with external knowledge and specialized capabilities. The library comprises tool resources and a dedicated Tool Manager agent. As shown in Fig. 4, the workflow sends each agent's predefined instructions to the Tool Manager before the agent executes. Based on these instructions, the Tool Manager selects and invokes suitable tools from the library's tool resources, then passes the results to the agent. The agent then executes its predefined instructions using the returned tool results. This interaction repeats for subsequent agents, except the Architect, ending with the Instantiator receiving tool results and producing the final benchmark.

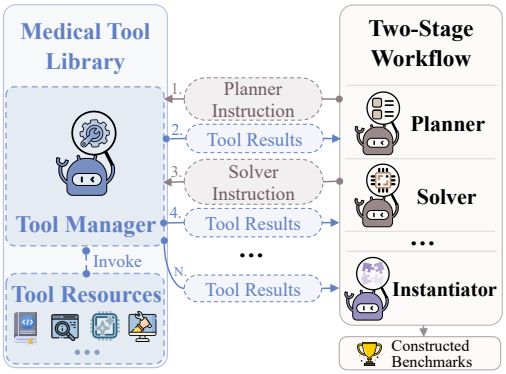

Figure 4: Medical Tool Library support for the two-stage workflow. The Tool Manager selects and invokes tools based on agent instructions, then supplies the results to support agent execution.

3.3 Key Design Components

Tool Resources and Dynamic Expansion. To support generalization across medical tasks and modalities, the library provides a broad range of existing tool resources and supports dynamic expansion as needed. Existing resources include medical knowledge sources such as PubMed and WHO publications, web search via Google and Bing, and specialized medical tools such as SAM-Path (Zhang et al., 2023) for pathology image segmentation, YOLO-SAM2 (Mansoori et al., 2025) for polyp detection and segmentation, and NeuroKit2 (Makowski et al., 2021) for ECG signal analysis. When existing tools cannot meet an agent's requirements, the Tool Manager dynamically expands the tool resources. It searches Hugging Face and GitHub for suitable models or code repositories, wraps selected resources into callable tools with standardized interfaces, and tests their execution. Tools that pass these checks are registered in the library for subsequent reuse.

Field-Level Quality Model. To improve the reliability of benchmark design, we use supervision derived from benchmark papers to train a quality model that scores individual specification fields. As depicted in Fig. 3, we first collect benchmark papers from general and medical domains. An LLM extracts relevant information from each paper and organizes it into a structured benchmark specification consisting of 11 predefined fields, including core capability, domain scope, task taxonomy, and item formats. We randomly select a subset of fields in each specification and prompt an LLM to introduce errors into their content, for example by changing the domain scope from pathology

reporting to radiology reporting. Corrupted fields are labeled negative, while unchanged fields are labeled positive. The resulting specifications and their field-level labels form the supervision data. We replace the language modeling head of Qwen3.5-4B (Qwen Team, 2026a) with a shared linear scoring head followed by a sigmoid activation to produce one quality score per specification field. We train the model with binary cross-entropy (BCE) loss over the field-level labels in two stages: first on general-domain data, then on medical-domain data. Finally, the trained quality model provides field-level scores to guide specification refinement in the MedBenchAgent framework. Further details on the quality model are provided in Appendix A.2.

Evidence-Guided Iterative Annotation. To improve annotation reliability, we introduce an evidence-guided annotation process following an iterative Thought–Tool–Observation loop. Each template item includes medical evidence provided by the Builder, such as medical references or domain rules that support item design and define annotation criteria. For example, for an item on lymph-node metastasis in breast cancer, the evidence may cite an international histopathology reporting guide (Cserni et al., 2021) that specifies classification criteria based on tumour deposit size and tumour-cell count. For each data sample, the Annotator first performs a *Thought* step to determine whether the available information is sufficient to complete the annotation according to the criteria defined by the item's medical evidence. If not, the Annotator specifies the required operation to the Tool Manager, which selects and executes an appropriate medical *Tool*. In the same lymph-node metastasis example, the Tool Manager may invoke tumour-cell segmentation models and size-measurement tools to obtain the tumour-cell count and tumour deposit size. The tool outputs are returned to the Annotator as an *Observation*. The Annotator then returns to *Thought* to reassess the available information, including the latest *Observation*, against the item's medical evidence. If required information is missing or uncertainty remains, the Annotator requests further tool operations; otherwise, the annotation is finalized. By combining medical evidence with the iterative Thought–Tool–Observation loop, this process improves annotation reliability.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on five medical benchmark-construction tasks. For each task, we use source data for benchmark construction and a separate held-out test set to evaluate LLMs fine-tuned on the generated items. 1) ECG Time Series Understanding. Following TimeSeriesExamAgent, we use PTB-XL (Wagner et al., 2020) as the source data for benchmark construction. We test on 12,000 QA items randomly sampled from the MIMIC-IV-ECG-based test split of ECG-QA (Oh et al., 2023). 2) 3D CT Spatial Analysis. Following SpatialMed (Trinh et al., 2026), we use annotated CT scans from AMOS22 (Ji et al., 2022), KiTS23 (Heller et al., 2023), and six MSD (Antonelli et al., 2022) subsets (Liver, Pancreas, Hepatic Vessel, Lung, Spleen, and Colon) as source data. We select 2,082 multiple-choice items from DeepTumorVQA 2.0 (Chen et al., 2026) relevant to SpatialMed's five spatial reasoning tasks for testing. CT cases appearing in the test set are removed from the source data to prevent data leakage. 3) Colonoscopy Video VQA. We map ColonBench (Hamdi et al., 2026) VQA items to their REAL-Colon (Biffi et al., 2024) source videos, using 40 of the 60 videos as source data and the remaining 20, corresponding to 1,216 VQA items, for testing. 4) EHR Clinical Reasoning. Following EHRBench (Xie et al., 2026b), we use MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023) as source data. We test on all 1,440 public EHRBench demo items. Before benchmark construction, we identify the patients associated with the MIMIC-derived test items and exclude all of their records from the source data, ensuring patient-level separation between construction and test data. 5) Breast Pathology Reporting VQA. We use 700 whole-slide images (WSIs) as source data: 400 breast immunohistochemistry (IHC) slides from ACROBAT (Weitz et al., 2024) (100 each for ER, PR, HER2, and Ki67) and 300 H&E slides from TCGA-BRCA (The Cancer Genome Atlas Network, 2012). We evaluate on a private test set of 1,133 multiple-choice items that is patient-disjoint from the source data.

Baselines. We use Codex CLI (OpenAI, 2025) and Orchestra-o1 (Zhang et al., 2026) as general-purpose baselines across all five tasks. For the first four tasks in Table 1, the specialized medical baselines are TimeSeriesExamAgent (Gwiazda et al., 2026), SpatialMed (Trinh et al., 2026), ColonBench (Hamdi et al., 2026), and EHRBench (Xie et al., 2026b), respectively. We additionally evaluate Benchmark Agent (Xiong et al., 2026) as a medical baseline on the ECG and EHR tasks. For

Table 1: Comparison of benchmark construction methods across five medical tasks. LLM evaluation scores (1–10) measure specificity (Spec.), unambiguity (Unamb.), domain relevance (Rel.), and answerability (Answer.). FT Accuracy reports Qwen3-VL-4B-Instruct accuracy on the corresponding test set before and after fine-tuning with the constructed benchmark. Human Approval denotes the proportion of constructed benchmark items approved by both human experts.

Task	Scope	Method	LLM Jury Evaluation (↑)				FT Accuracy (↑)	Human Approval (↑)
			Spec.	Unamb.	Rel.	Answer.		
ECG Time Series Understanding	General	Codex CLI	7.18	6.32	9.26	7.97	0.480→0.556	38%
		Orchestra-o1	6.40	7.20	9.41	9.07	0.480→0.528	46%
	Medical	Benchmark Agent	7.53	7.78	9.40	8.86	0.480→0.579	61%
		TimeSeriesExamAgent	7.81	7.59	**9.62**	9.09	0.480→0.616	74%
		Ours	**8.28**	**8.74**	9.51	**9.22**	**0.480→0.672**	**85%**
3D CT Spatial Analysis	General	Codex CLI	8.66	8.59	9.45	8.93	0.382→0.440	67%
		Orchestra-o1	8.27	7.19	9.38	8.21	0.382→0.401	55%
	Medical	SpatialMed	8.82	8.02	8.85	9.25	0.382→0.467	77%
		Ours	**9.18**	**9.05**	**9.64**	**9.63**	**0.382→0.527**	**84%**
Colonoscopy Video VQA	General	Codex CLI	6.94	7.03	9.08	6.99	0.423→0.485	28%
		Orchestra-o1	8.02	6.92	8.32	8.24	0.423→0.547	43%
	Medical	Colon-Bench	8.17	7.29	8.53	9.03	0.423→0.605	68%
		Ours	**8.48**	**8.15**	**9.88**	**9.36**	**0.423→0.680**	**77%**
EHR Clinical Reasoning	General	Codex CLI	9.22	8.83	7.88	9.25	0.487→0.574	60%
		Orchestra-o1	9.26	8.34	9.22	9.23	0.487→0.621	52%
	Medical	Benchmark Agent	9.04	8.61	8.57	8.63	0.487→0.680	67%
		EHRBench	**9.27**	**9.09**	9.24	9.30	0.487→0.745	85%
		Ours	9.08	8.84	**9.47**	**9.59**	**0.487→0.776**	**91%**
Breast Pathology Reporting VQA	General	Codex CLI	8.59	8.24	9.51	6.05	0.510→0.561	36%
		Orchestra-o1	9.00	8.48	9.39	8.47	0.510→0.607	58%
	Medical	**Ours**	**9.22**	**9.14**	**9.84**	**9.46**	**0.510→0.724**	**79%**

a fair comparison, all methods within each task use the same source data and receive benchmark requests aligned in terms of task descriptions, item counts, and question formats, with minor adaptations to accommodate method-specific capabilities. For the four tasks with task-specific baselines, benchmark requests follow the construction requirements of the corresponding baselines. For the remaining task, breast pathology reporting VQA, we define a benchmark request organized around core report components. Appendix A.1 provides an overview of all five tasks.

Metrics. We assess benchmark quality using LLM jury evaluation, fine-tuning accuracy, and human approval. 1) LLM Jury Evaluation. Following TimeSeriesExamAgent, we score each item on a 1–10 scale across four dimensions: specificity, unambiguity, domain relevance, and answerability. We use Gemini-3.1-Pro-Preview (Google DeepMind, 2026a), GPT-5.6 Sol (OpenAI, 2026a), and Qwen3.5-397B-A17B (Qwen Team, 2026a) as jury models. For each dimension, we report the mean of the three judges' scores across all generated items. 2) FT Accuracy. Following TimeSeriesExamAgent, we separately fine-tune Qwen3-VL-4B-Instruct (Bai et al., 2025) on each method's generated benchmark using low-rank adaptation (LoRA). We report accuracy on the corresponding test set before and after fine-tuning. 3) Human Approval. Following TrustGen (TrustGen Team, 2026), we randomly sample 100 items per method per task for independent review by two medical experts. Experts assess benchmark design and data annotation, considering request alignment, clinical relevance, clarity, and annotation correctness. Human Approval is the percentage of sampled items approved by both experts. See Appendix A.5 for metric definitions and calculation details.

Implementation Details. For a fair comparison, we use Qwen3.8-27B (Qwen Team, 2026c) as the shared backbone for agents across all methods. The Architect refines the benchmark specification for up to three rounds using field-level scores from the quality model, stopping early once all scores

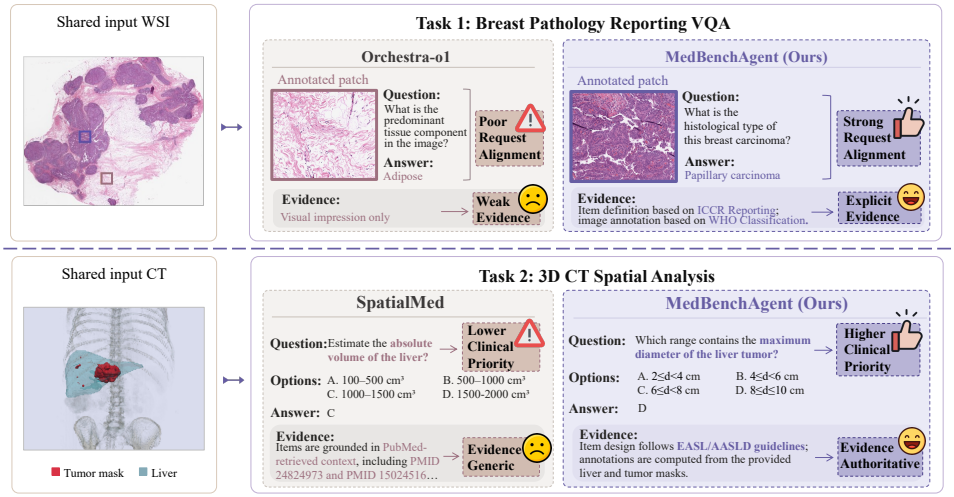

Figure 5: Qualitative comparison of benchmark items for two representative medical tasks. For each task, MedBenchAgent and the competing baseline use the same input. MedBenchAgent produces items with stronger request alignment, clinical relevance, and evidence support.

are > 0.8. For each template item, the Builder performs up to three rounds of format and semantic checking and revision; items that remain invalid are discarded. For annotation, the Thought–Tool–Observation loop runs for up to 6 iterations per data sample. The loop stops early when sufficient information is available to complete the annotation according to the item's medical evidence, or when the sample is deemed unsuitable. The Navigator generates up to five retrieval queries for each template item and adaptively determines the retrieval size for each query. When the item count is not specified by the user, the Architect determines it based on data availability and construction feasibility.

4.2 Quantitative Results

Table 1 compares MedBenchAgent with general-purpose and medical baselines across five medical benchmark construction tasks. MedBenchAgent achieves the highest fine-tuning accuracy and human approval rate on all tasks and ranks first in 17 of the 20 LLM jury evaluations. Notably, MedBenchAgent leads on all metrics in three tasks: 3D CT spatial analysis, colonoscopy video VQA, and breast pathology reporting VQA. The largest gains over the strongest baseline occur in breast pathology reporting VQA, where MedBenchAgent improves fine-tuning accuracy from 0.607 to 0.724 and human approval from 58% to 79%. On ECG time series understanding, MedBenchAgent leads on all metrics except domain relevance, where it performs comparably to the strongest baseline. On EHR clinical reasoning, MedBenchAgent leads on four of six metrics, while scoring slightly below EHRBench in specificity and unambiguity. Taken together, these results support MedBenchAgent's effectiveness in constructing high-quality medical benchmarks across the evaluated tasks and modalities. We report LLM performance on our benchmarks (Appendix A.6), medical-input ablations (Table 13), and non-approval reasons in human evaluation (Table 14).

4.3 Qualitative Results

Figure 5 provides a qualitative comparison between MedBenchAgent and strong baselines on two representative tasks, using the same medical input for each comparison. For Task 1, Orchestra-o1 focuses on adipose tissue in the annotated patch and question, with limited alignment to the breast pathology reporting request and weak evidence. MedBenchAgent instead annotates a tumor-containing patch and generates a reporting-oriented question aligned with the request and grounded in medical evidence. In Task 2, SpatialMed constructs an item on absolute liver volume, which has limited clinical relevance in this setting, drawing on generic PubMed-retrieved literature. MedBenchAgent instead targets the more clinically meaningful maximum tumor diameter, guided by authoritative medical guidelines. See Appendix Figure 6 for additional results.

4.4 Ablation Study

We conduct three ablation studies on the Breast Pathology Reporting VQA task. Results on the other four tasks are reported in Appendix Tables 9, 10, and 11.

Impact of Field-Level Quality Model. We compare four specification refinement settings in Table 2. Passing the initial specification directly to the Builder without refinement yields the lowest overall performance, while single-score feedback, obtained by averaging the 11 scores from our trained quality model, improves fine-tuning accuracy to 0.681. Direct LLM field-level scoring gives comparable overall performance, achieving 0.702 fine-tuning accuracy but 66% human approval. Our field-level quality model performs best across all metrics, reaching 0.724 fine-tuning accuracy and 79% human approval, demonstrating its effectiveness in improving benchmark quality.

Table 2: Ablation study of the quality model. *No Refinement* performs no specification refinement. *Learned Single Score* uses the mean score from our trained quality model. *LLM Field Scores* uses field-level scores produced directly by the LLM Qwen3.8-27B. *Learned Field Scores* uses field-level scores from quality model.

Refinement Strategy	LLM Jury Evaluation (↑)				FT Accuracy (↑)	Human Approval (↑)
	Spec.	Unamb.	Rel.	Answer.		
No Refinement	8.74	8.67	9.23	8.62	0.510→0.625	63%
Learned Single Score	8.89	8.80	9.49	9.02	0.510→0.681	70%
LLM Field Scores	8.81	8.82	9.66	9.10	0.510→0.702	66%
Learned Field Scores	**9.22**	**9.14**	**9.84**	**9.46**	**0.510→0.724**	**79%**

Effect of Evidence-Guided Iterative Annotation. We evaluate two core mechanisms of annotation strategy: evidence guidance and iterative Thought–Tool–Observation refinement. Table 3 reports ablation results for these mechanisms individually and jointly. Evidence guidance alone improves most metrics, increasing specificity from 8.87 to 9.23 and fine-tuning accuracy from 0.643 to 0.704. Iterative refinement alone improves all four LLM jury dimensions and fine-tuning accuracy. Combining both achieves the strongest overall results, with domain relevance of 9.84 and answerability of 9.46. These results support their joint use for more reliable medical benchmark construction.

Table 3: Ablation of annotation strategies. *Evidence* denotes medical evidence guidance. *Iterative* denotes refinement through the Thought–Tool–Observation loop.

Annotation Strategy		LLM Jury Evaluation (↑)				FT Accuracy (↑)	Human Approval (↑)
Evidence	Iterative	Spec.	Unamb.	Rel.	Answer.		
×	×	8.87	8.79	9.47	8.77	0.510→0.643	68%
✓	×	**9.23**	8.72	9.67	8.68	0.510→0.704	72%
×	✓	9.07	8.93	9.48	9.10	0.510→0.670	65%
✓	✓	9.22	**9.14**	**9.84**	**9.46**	**0.510→0.724**	**79%**

Contribution of the Tool Library. Table 4 compares three tool-library configurations. Adding a broad range of manually selected tools to the basic set improves all metrics, increasing fine-tuning accuracy from 0.631 to 0.672 and human approval from 58% to 69%. Enabling dynamic expansion on the same initial library further increases answerability from 8.85 to 9.46, fine-tuning accuracy to 0.724, and human approval to 79%. These results demonstrate the benefits of predefined tool resources and dynamic expansion.

Table 4: Ablation of tool-library configurations. *Basic*: a minimal tool set for pipeline execution. *Predefined*: a broad range of additional tools manually selected before construction. *Dynamic*: on-demand acquisition and integration of new tools during construction.

Tool Library			LLM Jury Evaluation (↑)				FT Accuracy (↑)	Human Approval (↑)
Basic	Predefined	Dynamic	Spec.	Unamb.	Rel.	Answer.		
✓	×	×	8.81	8.50	9.59	8.52	0.510→0.631	58%
✓	✓	×	8.97	8.87	9.66	8.85	0.510→0.672	69%
✓	✓	✓	**9.22**	**9.14**	**9.84**	**9.46**	**0.510→0.724**	**79%**

5 Conclusion

We present MedBenchAgent, an agentic framework for reliable medical benchmark construction across diverse tasks and modalities. Its two-stage workflow decomposes benchmark requests into subtasks to derive specifications and template items, then retrieves and annotates medical data to instantiate final items. A field-level quality model and an evidence-guided Thought–Tool–Observation loop improve design and annotation reliability, respectively. The Medical Tool Library supports both stages and improves generalizability through broad resource coverage and dynamic tool expansion. Experiments show stronger overall LLM jury performance and the highest fine-tuning accuracy and human approval rates on all five tasks. MedBenchAgent offers a practical approach to constructing clinically grounded benchmarks for systematic evaluation and improvement of medical LLMs.

AI USE STATEMENT

Generative AI was used in two ways. First, LLMs are integral components of the proposed benchmark construction and evaluation pipelines, as described in Sections 3 and 4. Second, ChatGPT (OpenAI) was used to assist with translation, language polishing, and literature lookup. AI-assisted suggestions were independently reviewed, verified, and revised by the authors. The authors take full responsibility for the final manuscript, claims, experiments, and artifacts.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the framework and experimental setup to support reproducibility. Section 3 describes the two-stage workflow, field-level quality model, evidence-guided iterative annotation, and medical tool library. Section 4.1 specifies the datasets, baselines, evaluation metrics, and implementation settings. Appendix A.1– A.6 further provides task-specific benchmark construction settings, quality-model training and evaluation, item validation procedures, retrieval models, data preprocessing, and evaluation protocols. Within each task, compared methods use the same source data, aligned benchmark requests, matched fine-tuning data sizes, and consistent evaluation procedures.

ETHICS STATEMENT

This work studies automated medical benchmark construction for research evaluation and is not intended for clinical diagnosis, treatment, or patient care. The experiments use medical data from the sources described in Section 4.1 and Appendix A.1, with source and evaluation data separated at the case or patient level where applicable. Automatically constructed benchmark items may contain errors, ambiguities, or biases inherited from the source data, language models, external knowledge sources, or specialized tools, and should therefore not be treated as clinical ground truth without appropriate expert validation. We mitigate these risks through evidence-guided annotation, automated item validation, and blinded evaluation by domain experts. Remaining benchmark-item failure modes are reported in Table 14. The proposed framework is intended to support research on medical AI evaluation, and the use or release of underlying medical data should follow their respective access conditions and licenses.

REFERENCES

Ahmed Albaqshi, Ji Su Ko, Chong Hyun Suh, Pae Sun Suh, Woo Hyun Shim, Hwon Heo, Chang-Yun Woo, and Hyungjun Park. Evaluating diagnostic accuracy of large language models in neuroradiology cases using image inputs from jama neurology and jama clinical challenges. *Scientific Reports*, 15:43027, 2025. doi: 10.1038/s41598-025-06458-z.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, et al. The medical segmentation decathlon. *Nature Communications*, 13:4128, 2022. doi: 10.1038/s41467-022-30695-9.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q.-H. Meng, and Bo Zhao. M3D: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Carlo Biffi, Giulio Antonelli, Sebastian Bernhofer, Cesare Hassan, Daizen Hirata, Mineo Iwatate, Andreas Maieron, Pietro Salvagnini, and Andrea Cherubini. REAL-Colon: A dataset for developing real-world ai applications in colonoscopy. *Scientific Data*, 11:539, 2024. doi: 10.1038/s41597-024-03359-0.

Natasha Butt, Varun Chandrasekaran, Neel Joshi, Besmira Nushi, and Vidhisha Balachandran. Benchagents: Automated benchmark creation with agent interaction. In *ICLR Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2025.

Xiuyuan Chen, Tao Sun, Dexin Su, Ailing Yu, Junwei Liu, Zhe Chen, Gangzeng Jin, Xin Wang, Jingnan Liu, Hansong Xiao, et al. Gaps: A clinically grounded, automated benchmark for evaluating ai clinicians. *arXiv preprint arXiv:2510.13734*, 2025.

Yixiong Chen, Wenjie Xiao, Pedro R. A. S. Bassi, Boyan Wang, Liang He, Xinze Zhou, Sezgin Er, Ibrahim Ethem Hamamci, Zongwei Zhou, and Alan Yuille. DeepTumorVQA: A hierarchical 3d ct benchmark for stage-wise evaluation of medical vlms and tool-augmented agents. *arXiv preprint arXiv:2605.09679*, 2026.

G. Cserni, E. Brogi, H. S. Cody, III, R. Deb, G. Farshid, S. O'Toole, E. Provenzano, C. M. Quinn, A. Sahin, F. Schmitt, D. Weaver, R. Yamaguchi, and P. H. Tan. *Surgically Removed Lymph Nodes for Breast Tumours Histopathology Reporting Guide*. International Collaboration on Cancer Reporting, Sydney, Australia, 1st edition, 2021. URL https://www.iccr-cancer.org/datasets/published-datasets/breast/surgically-removed-lymph-nodes-for-breast-tumours/. Version 1.2.

Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31:3749–3761, 2025. doi: 10.1038/s41591-025-03982-3.

FreedomIntelligence. HuatuoGPT-Vision-7B-Qwen2.5VL. Hugging Face model card, 2025. URL https://huggingface.co/FreedomIntelligence/HuatuoGPT-Vision-7B-Qwen2.5VL.

Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks. In *Advances in Neural Information Processing Systems*, 2025.

Google DeepMind. Gemini 3.1 Pro Model Card. Model Card, February 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Google DeepMind. Gemini 3.6 Flash model card, July 2026b. URL https://deepmind.google/models/model-cards/gemini-3-6-flash/.

Google DeepMind. Gemini 3.7 Flash model card, August 2026c. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/.

Dadi Guo, Tianyi Zhou, Dongrui Liu, Chen Qian, Qihan Ren, Shuai Shao, Zhiyuan Fan, Yi R. Fung, Kun Wang, Linfeng Zhang, and Jing Shao. Towards self-evolving agent benchmarks: Validatable agent trajectory via test-time exploration. In *The Fourteenth International Conference on Learning Representations*, 2026.

Malgorzata Gwiazda, Yifu Cai, Mononito Goswami, Arjun Choudhry, and Artur Dubrawski. TimeSeriesExamAgent: Creating time series reasoning benchmarks at scale. In *International Conference on Learning Representations*, 2026.

Abdullah Hamdi, Changchun Yang, and Xin Gao. Colon-Bench: An agentic workflow for scalable dense lesion annotation in full-procedure colonoscopy videos. In *Medical Image Computing and Computer Assisted Intervention – MICCAI*, 2026.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Nicholas Heller, Fabian Isensee, Resha Tejpaul, Andrew W. Wood, Nikolaos Papanikolopoulos, and Christopher Weight. 2023 kidney and kidney tumor segmentation challenge, 2023. URL https://doi.org/10.5281/zenodo.7840134.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, and Ping Luo. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Baoyang Jiang, Fengchun Zhang, Leyuan Wang, Haotian Li, Yida Wang, Zhe Ji, Jinshan Lai, Xi Ren, Jianwei Hu, and Qiang Ma. Embodied-benchclaw: An autonomous multi-agent system for embodied spatial intelligence benchmark construction. *arXiv preprint arXiv:2606.11909*, 2026.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-Wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x.

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Hwiwon Lee, Ziqi Zhang, Hanxiao Lu, and Lingming Zhang. SEC-bench: Automated benchmarking of LLM agents on real-world software security tasks. In *Advances in Neural Information Processing Systems*, 2025.

Xiang Li, Farzaan Kaiyom, Evan Liu, Yifan Mai, Percy Liang, and Tatsunori Hashimoto. Autobencher: Towards declarative benchmark construction. In *International Conference on Learning Representations*, 2025.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31949–31963. PMLR, 2024a.

Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. Teach multimodal llms to comprehend electrocardiographic images. *arXiv preprint arXiv:2410.19008*, 2024b.

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, 2024a. doi: 10.1038/s41586-024-07618-3.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024b. URL https://openreview.net/forum?id=KUNzEQMWU7.

Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, 2021. doi: 10.3758/s13428-020-01516-y.

Mobina Mansoori, Sajjad Shahabodini, Jamshid Abouei, Konstantinos N. Plataniotis, and Arash Mohammadi. Self-prompting polyp segmentation in colonoscopy using hybrid YOLO-SAM2 model. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10887638.

Jungwoo Oh, Gyubok Lee, Seongsu Bae, Joon-myoung Kwon, and Edward Choi. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

OpenAI. Introducing codex. OpenAI, May 2025. URL https://openai.com/index/introducing-codex/.

OpenAI. Gpt-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, July 2026a. Accessed: 2026-08-26.

OpenAI. GPT-6 Astra system card, September 2026b. URL https://deploymentsafety.openai.com/gpt-6-astra.

Alejandra Perez, Chinedu Nwoye, Ramtin Raji Kermani, Omid Mohareri, and Muhammad Abdullah Jamal. SurgLaVi: Large-scale hierarchical dataset for surgical vision–language representation learning. *Medical Image Analysis*, 110:103982, 2026. doi: 10.1016/j.media.2026.103982.

Rushi Qiang, Yuchen Zhuang, Anikait Singh, Percy Liang, Chao Zhang, Sherry Yang, and Bo Dai. MLE-Smith: Scaling MLE tasks with automated multi-agent pipeline. In *The Fourteenth International Conference on Learning Representations*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.8-27B. Hugging Face model card, 2026c. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Qwen Team. Qwen3.8-Flash. QwenCloud model documentation, 2026d. URL https://www.qwencloud.com/models/qwen3.8-flash.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. URL https://arxiv.org/abs/2507.05201.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. MedGemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026. URL https://arxiv.org/abs/2604.05081.

Miaojing Shi, Zijie Yue, Huihui Sun, Bo Li, Yikun Ma, Tianyu Cen, Ying Chen, Xiaofen Wu, Hanli Wang, Jun Wang, and Shuchang Xu. Bootstrapping multimodal large language model with medical knowledge for automatic esophagogastroduodenoscopy diagnosis and reporting. *Nature Communications*, 2026a. doi: 10.1038/s41467-026-75377-y.

Runhan Shi, Quan Zhou, Yuqian Xu, Shuai Yang, Xin Wu, Zitong Zhou, Hui Liu, Bin Zha, Zheming Wang, Liya Li, Wei Wei, Jinru Ding, Wenrao Pang, Mouxiao Bian, Haoyuan Hu, Jun Xu, and Jie Xu. MedRealMM: A real-world multimodal benchmark for Chinese online medical consultation. *arXiv preprint arXiv:2607.09142*, 2026b. URL https://arxiv.org/abs/2607.09142v3.

Yuhao Su, Anwesa Choudhuri, Zhongpai Gao, Benjamin Planche, Van Nguyen Nguyen, Meng Zheng, Yuhan Shen, Arun Innanje, Terrence Chen, Ehsan Elhamifar, and Ziyan Wu. Medgrpo: Multi-task reinforcement learning for heterogeneous medical video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Dan Wan, Xiaoxiao Lan, Mengyue Zheng, Jingxiong Li, Xinheng Lyu, Tao Lin, and Lin Yang. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, pp. 56–73. Springer, 2024.

Rahul Thapa, Andrew Li, Qingyang Wu, Bryan He, Yuki Sahashi, Christina Binder, Angela Zhang, Ben Athiwaratkun, Shuaiwen Leon Song, David Ouyang, and James Zou. How well can general vision-language models learn medicine by watching public educational videos? *arXiv preprint arXiv:2504.14391*, 2025.

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70, 2012. doi: 10.1038/nature11412.

Weiwei Tian, Xinyu Huang, Tianhao Cheng, Wen He, Jinwu Fang, Rui Feng, Daoying Geng, and Xiaobo Zhang. A medical multimodal large language model for pediatric pneumonia. *IEEE Journal of Biomedical and Health Informatics*, 29(9):6869–6882, 2025. doi: 10.1109/JBHI.2025.3569361.

Quoc-Huy Trinh, Xi Ding, Yang Liu, Zhenyue Qin, Xingjian Li, Gorkem Durak, Halil Ertugrul Aktas, Elif Keles, Ulas Bagci, and Min Xu. Beyond medical diagnostics: How medical multimodal large language models think in space. *arXiv preprint arXiv:2603.13800*, 2026. URL https://arxiv.org/abs/2603.13800v1.

TrustGen Team. TrustGen: A platform of dynamic benchmarking on the trustworthiness of generative foundation models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Fcf5fLmaeG.

Alyssa Unell, Noel C. F. Codella, Sam Preston, Peniel Argaw, Wen-wai Yim, Zelalem Gero, Cliff Wong, Rajesh Jena, Eric Horvitz, Amanda K. Hall, Ruican Rachel Zhong, Jiachen Li, Shrey Jain, Mu Wei, Matthew Lungren, and Hoifung Poon. CancerGUIDE: Cancer guideline understanding via internal disagreement estimation. *arXiv preprint arXiv:2509.07325*, 2025.

Konstantinos Vergopoulos, Mark Niklas Mueller, and Martin Vechev. Automated benchmark generation for repository-level coding tasks. In *International Conference on Machine Learning*, 2025.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7:154, 2020. doi: 10.1038/s41597-020-0495-6.

Gui Wang, YongSong Zhou, Kaijun Deng, Wooi Ping Cheah, Rong Qu, Jianfeng Ren, and Linlin Shen. Surgcot: Advancing spatiotemporal reasoning in surgical videos through a chain-of-thought benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

Xidong Wang, Shuqi Guo, Yue Shen, Junying Chen, Jian Wang, Jinjie Gu, Ping Zhang, Lei Liu, and Benyou Wang. LiveClin: A live clinical benchmark without leakage. In *International Conference on Learning Representations*, 2026b.

Philippe Weitz, Masi Valkonen, Leslie Solorzano, Circe Carr, Kimmo Kartasalo, Constance Boissin, Sonja Koivukoski, Aino Kuusela, Dusan Rasic, Yanbo Feng, et al. The ACROBAT 2022 challenge: Automatic registration of breast cancer tissue. *Medical Image Analysis*, 97:103257, 2024. doi: 10.1016/j.media.2024.103257.

Ying Xiao, Jie Huang, Ruijuan He, Jing Xiao, Mohammad Reza Mousavi, Yepang Liu, Kezhi Li, Zhenpeng Chen, and Jie M. Zhang. FairMedQA: Benchmarking bias in large language models for medical question answering. *arXiv preprint arXiv:2505.19562*, 2025.

Jingxu Xie, Dylan Xu, Xuandong Zhao, and Dawn Song. AgentSynth: Scalable task generation for generalist computer-use agents. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Yuzhang Xie, Keqi Han, Yunpeng Xiao, Hejie Cui, Guanchen Wu, Ziyang Zhang, Kai Shu, Jiaying Lu, Xiao Hu, and Carl Yang. Ehrbench: An automated and reliable ehr-based benchmark for clinical decision making with llms. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026b.

Shiyun Xiong, Dongming Wu, Peiwen Sun, Yuang Ai, Bokang Yang, Wencheng Han, Xiao-Hui Li, and Xiangyu Yue. Benchmark everything everywhere all at once. *arXiv preprint arXiv:2606.06462*, 2026.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025. URL https://arxiv.org/abs/2506.07044.

Zhiling Yan, Dingjie Song, Zhe Fang, Yisheng Ji, Xiang Li, Quanzheng Li, and Lichao Sun. Livemedbench: A contamination-limited medical benchmark for llms with automated rubric evaluation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 10162–10173, 2026.

Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, Eric Wang, Ellery Wulczyn, Fayaz Jamil, Theo Guidroz, Chuck Lau, Siyuan Qiao, Yun Liu, Akshay Goel, Kendall Park, Arnav Agharwal, Nick George, Yang Wang, Ryutaro Tanno, David G. T. Barrett, Wei-Hung Weng, S. Sara Mahdavi, Khaled Saab, Tao Tu, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, Jorge Cuadros, Gregory Sorensen, Yossi Matias, Katherine Chou, Greg Corrado, Joelle Barral, Shravya Shetty, David Fleet, S. M. Ali Eslami, Daniel Tse, Shruthi Prabhakara, Cory McLean, Dave Steiner, Rory Pilgrim, Christopher Kelly, Shekoofeh Azizi, and Daniel Golden. Advancing multimodal medical capabilities of Gemini. *arXiv preprint arXiv:2405.03162*, 2024.

Xiao Yang, Xuejiao Zhao, and Zhiqi Shen. Ehrstruct: A comprehensive benchmark framework for evaluating large language models on structured electronic health record tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34340–34348, 2026.

Seungdong Yoa, Sanghyu Yoon, Suhee Yoon, Dongmin Kim, Ye Seul Sim, Junhyun Lee, and Woohyung Lim. From static benchmarks to dynamic protocol: Agent-centric text anomaly detection for evaluating LLM reasoning. In *International Conference on Learning Representations*, 2026.

Fan Zhang, Vireo Zhang, Shengju Qian, Haoxuan Li, Hao Wu, Jinyang Wu, Donghao Zhou, Zhihong Zhu, Zheng Lian, Xin Wang, and Pheng-Ann Heng. Orchestra-o1: Omnimodal agent orchestration, 2026. URL https://arxiv.org/abs/2606.13707.

Jingwei Zhang, Ke Ma, Saarthak Kapse, Joel Saltz, Maria Vakalopoulou, Prateek Prasanna, and Dimitris Samaras. Sam-path: A segment anything model for semantic segmentation in digital pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 161–170. Springer, 2023.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025a.

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, Ludwig Schmidt, and Serena Yeung-Levy. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, and Xin Gao. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nature Communications*, 15:5649, 2024. doi: 10.1038/s41467-024-50043-3.

Qingqing Zhu, Qiao Jin, Tejas S. Mathai, Yin Fang, Zhizheng Wang, Yifan Yang, Maame Sarfo-Gyamfi, Benjamin Hou, Ran Gu, Praveen T. S. Balamuralikrishna, Kenneth C. Wang, Ronald M. Summers, and Zhiyong Lu. Ct-bench: A benchmark for multimodal lesion understanding in computed tomography. *arXiv preprint arXiv:2602.14879*, 2026.

A Appendix

A.1 Benchmark Construction Tasks

Within each task, all methods use the same source data and share the same overall construction objectives. Their benchmark requests are aligned in evaluation scope, target item count, and question formats, with only minor adaptations to accommodate method-specific capabilities. Below, we summarize the source data and intended benchmark for each task.

ECG Time Series Understanding. Following the ECG benchmark-construction setting of TimeSeriesExamAgent (Gwiazda et al., 2026), we aim to construct 15,000 four-option multiple-choice items for evaluating ECG interpretation and reasoning. (1) The source data consists of 10-second, 12-lead ECG recordings from PTB-XL, together with SCP-ECG annotations and associated metadata. (2) The task focuses on interpreting ECG recordings, for example, to identify rhythm abnormalities, assess waveform morphology, or recognize conduction abnormalities. Each question should target a clearly defined ECG finding or clinical interpretation, require analysis of the accompanying recordings, and have an unambiguous answer. Reference answers are derived from dataset annotations or calculations on the ECG signals, depending on the item. Each item contains one or more ECG recordings, a question, four answer options, and a reference answer.

3D CT Spatial Analysis. Following the task formulation of SpatialMed (Trinh et al., 2026), we aim to construct 10,000 four-option multiple-choice VQA items to evaluate spatial reasoning over anatomical structures and tumors in volumetric CT scans. (1) The source data consists of CT volumes from AMOS22, KiTS23, and selected MSD subsets, together with corresponding segmentation masks, label definitions, and voxel-spacing metadata. (2) The task covers five types of four-option multiple-choice questions: directional reasoning (DIR), distance reasoning (DIST), extent/size/shape reasoning (EXT), volume magnitude reasoning (VOL), and comparative reasoning (COMP). Reference answers are derived from the segmentation masks and associated spatial information, depending on the question type. Each item contains one or more CT volumes, a question, four answer options, and a reference answer.

Colonoscopy Video VQA. Following the VQA task formulation of Colon-Bench (Hamdi et al., 2026), this task aims to construct 3,000 five-option multiple-choice VQA items for evaluating colonoscopy video understanding. (1) The source data consists of 40 selected full-procedure colonoscopy videos from REAL-Colon. (2) The task covers three question categories: lesion identification, clinical characteristics, and temporal reasoning, assessing lesion recognition, interpretation of features such as morphology, color, and anatomical location, and reasoning about lesion-related events within video clips. Construction involves locating relevant temporal segments and annotating their lesion content and associated clinical and temporal attributes to generate questions grounded in the video content. Each item contains a video clip, a question, five answer options comprising one correct answer and four plausible distractors, and the reference answer.

EHR Clinical Reasoning. Following the clinical task definitions of EHRBench (Xie et al., 2026b), we aim to construct 60,000 QA items in multiple-choice (4, 5, or 6 options) and open-ended short-answer formats for evaluating diagnosis, treatment, and prognosis reasoning. (1) The source data consists of encounter-level records from MIMIC-III and MIMIC-IV, preprocessed following EHRBench's procedure. All methods receive the same preprocessed records as input, with encounter order preserved within each patient. (2) The task covers diagnosis, treatment, and prognosis. Diagnosis and treatment items use single encounter-level records, whereas prognosis items use consecutive record pairs from the same patient. Diagnosis items assess clinical relationships among diagnoses within an encounter. Treatment items assess relationships between observed diagnoses and recorded prescriptions or procedures within the same encounter. Prognosis items assess how current diagnoses and treatments relate to diagnoses at the next encounter. Each item includes a natural-language clinical context drawn from selected events, a question, and a reference answer. Multiple-choice items include one correct option and three to five plausible distractors, with the target answer excluded from the context. Open-ended items require short answers.

Breast Pathology Reporting VQA. This task aims to construct 25,000 multiple-choice VQA items for standardized breast pathology reporting. (1) We use H&E WSIs from TCGA-BRCA and ER, PR, HER2, and Ki67 IHC WSIs from ACROBAT as source data. (2) We define a task covering the diagnostic assessments needed to determine core elements of breast pathology reports. Items based on H&E images cover diagnostic assessments in areas such as breast cancer subtyping, tumor grading, and lymphovascular invasion. Items based on IHC images address biomarker staining patterns and interpretation. Annotation involves locating and delineating WSI regions that match predefined answer labels, followed by patch extraction. Each item contains an image, a question, answer options, and a reference answer.

A.2 Field-Level Quality Model

We describe the benchmark specification and detail the quality model's data, architecture, and training.

Benchmark specification. Each benchmark specification contains the 11 fields listed in Table 5. Each field is treated as one scoring unit, including all nested content. For example, `task_taxonomy` receives a single score based on all the task information it contains.

Table 5: Benchmark specification fields.

Field	Description
`benchmark_name`	A concise and descriptive benchmark name.
`motivation`	The need for the benchmark and the evaluation gap it addresses.
`evaluation_target`	The types of LLMs the benchmark is designed to evaluate, such as text-only or multimodal LLMs.
`core_capability`	The primary capability assessed by the benchmark.
`medical_domain_scope`	The diseases, data modalities, clinical context, and evaluation boundaries covered.
`task_taxonomy`	A list of task types, with each entry specifying the task name, evaluation focus, how reference answers are determined, and target item count.
`item_formats`	The supported item formats and how each format is used.
`evaluation_protocol`	Four aspects: how models are tested, what inputs they receive, what outputs they must produce, and what baselines are used for comparison.
`evaluation_metrics`	The metrics used, what each measures, and how scores are calculated and aggregated.
`analysis_and_reporting`	How results are reported for the whole benchmark and each task category, how models are compared, and what conclusions are drawn, with uncertainty and error analysis where appropriate.
`intended_use`	What the benchmark is intended for, what questions it can answer about model capabilities, and which uses it does not support.

Data Curation. We assembled 15,008 PDF records for the general-domain collection using Crossref, arXiv, Europe PMC, ACL Anthology, CVF proceedings, OpenAlex, and Semantic Scholar for paper discovery and full-text retrieval. Search queries included terms such as "benchmark," "benchmarking," "test suite," and "evaluation suite," as well as combinations of "dataset" or "corpus" with "benchmark" or "evaluation." For the medical-domain collection, we assembled 1,206 PDF records through sources including arXiv, Europe PMC, OpenAlex, and PMLR, supplemented by medical papers selected from our general-domain collection. Medical searches combined medical and biomedical topic terms with "benchmark," "dataset," "corpus," and "challenge." For both collections, we verified that each PDF was readable and that its title matched the corresponding catalog entry. We screened relevant sections of the papers to identify papers providing substantive information about benchmark specifications, data descriptions, and benchmark requests. Deduplication used normalized DOIs, titles, and PDF content, followed by verification of relationships among preprints, published versions, and manuscript revisions. Confirmed duplicate versions were consolidated, and accompanying materials were not counted as separate papers. These procedures yielded 11,699 papers in the general-domain collection after consolidating 32 duplicate or accompanying records

and excluding 3,277 ineligible papers. The medical-domain collection retained 1,050 papers after excluding 156 records, including 14 duplicate versions and five invalid or incomplete documents.

Supervision Data Construction. We construct supervision data from source collections containing 11,699 general-domain and 1,050 medical-domain benchmark papers, following the specification extraction and corruption procedure described in Section 3.3. Papers used in our external benchmark evaluation are excluded from all supervision-data splits. Using Qwen3.8-27B, we extract one original 11-field specification and generate one corrupted variant per retained paper. All fields in the original specification are assigned a label of 1. For each corrupted variant, we randomly select between 1 and 11 fields for corruption, assigning labels of 0 to the selected fields and 1 to the unselected fields. We also use this model to derive a benchmark request and data descriptions from each paper, which are shared by the original and corrupted specifications. Each supervision example consists of the request and data descriptions followed by two identical copies of the selected specification, with a single set of 11 field-level labels as supervision. We partition the supervision data globally by unique source paper into training, validation, and internal test sets at an approximate 80:10:10 ratio. All versions of the same paper are assigned to the same split across the general-domain and medical-domain collections. Papers assigned to validation or internal test are excluded from training in both stages.

Table 6: Human audit of original fields and fields selected for corruption. Accuracies are computed using the reviewers' final consensus labels.

Metric	General domain	Medical domain
Original-field accuracy	0.967	0.938
Corrupted-field accuracy	0.927	0.920

Human Audit of Supervision Labels. We randomly sample 50 general-domain and 50 medical-domain papers from our internal test split, retaining one original and one corrupted specification per paper (200 specifications and 2,200 field positions). Two expert reviewers with extensive experience in medical benchmark design independently evaluate each field. They receive the source paper and the full specification, with all specification versions presented in random order. Reviewers are not informed of version identities, corruption positions, or the assigned field labels. Each field receives a label of 1 if no substantive issue is identified and 0 if a clear substantive error is found. Minor wording differences are not treated as errors. After independent review, the reviewers discuss disagreements until they reach a consensus on the final labels. Table 6 reports the proportion of original fields confirmed as valid positive examples and the proportion of fields selected for corruption confirmed as valid negative examples.

Model Architecture. We initialize the backbone from Qwen3.5-4B (Qwen Team, 2026a) and replace its language modeling head with a randomly initialized shared linear scoring head. We concatenate the benchmark request, available data descriptions, and two copies of the complete serialized specification into a single input sequence. The two copies contain identical field contents in the same order. During preprocessing, we identify one scoring position at the end of each field in the second copy. We retain the backbone's causal computation. Since every scoring position follows a complete copy of the specification, all 11 fields are available as preceding context for each prediction. We apply the shared linear scoring head followed by a sigmoid activation to the final-layer hidden state at each of these 11 positions to obtain one quality score per field.

Model Training. For each training specification, let $s_i \in (0, 1)$ denote the predicted quality score for field i, and let $y_i = 1$ indicate an unchanged field and $y_i = 0$ indicate a corrupted field. We train the quality model using BCE averaged over the 11 fields:

$$\mathcal{L}_{\text{field}} = -\frac{1}{11} \sum_{i=1}^{11} \left[y_i \log s_i + (1 - y_i) \log(1 - s_i) \right]. \tag{1}$$

We minimize the mean of this loss over the training specifications. Only the 11 designated scoring positions in the second copy receive direct supervision, so each field contributes equally regardless of its token length.

Training proceeds in two stages: first on general-domain specifications, then on medical-domain specifications from the first-stage checkpoint, retaining both the backbone and the scoring head. Both stages use the same 11 specification fields and field-level BCE objective. This two-stage training allows the model to first learn general benchmark-design principles and then adapt to the requirements of medical benchmarks, such as accurate medical terminology, clinically relevant task design, and alignment with medical guidelines.

Score-Guided Specification Refinement. Let R and D denote the benchmark request and available data descriptions, respectively. At refinement round t, the quality model evaluates the current specification $S^{(t)}$ and produces 11 field-level scores:

$$\mathbf{s}^{(t)} = Q(R, D, S^{(t)}) \in (0, 1)^{11}. \tag{2}$$

For each call to $Q(R, D, S^{(t)})$, we construct a single input sequence containing R, D, and two copies of the current specification $S^{(t)}$, and read scores from the second copy. Both copies are rebuilt from the current specification after each revision. If all scores exceed 0.8, refinement stops. Otherwise, the Architect receives the complete specification and its field-level scores, together with the benchmark request, data descriptions, and Solver outputs. It reviews all fields, identifies issues, and revises the specification accordingly. The quality model supplies numerical scores; the Architect determines the issue and the corresponding revision. Fields above the threshold are retained unless changes are needed to maintain consistency with a revised field. Each revised specification replaces the preceding version, and all 11 fields are rescored before the next decision. We allow at most three rounds, excluding initial specification generation. The latest specification is passed to the Builder when either all scores exceed 0.8 or the revision budget is exhausted. Reaching the budget does not imply that all fields satisfy the threshold.

Table 7: Quality-model evaluation on real Architect outputs. Counts show correct predictions over assessable fields for accuracy, and detected problems over human-identified problematic fields for recall.

Metric	Count	Score
Field-level accuracy	498 / 550	0.905
Error detection recall	209 / 236	0.886

Model Evaluation on Real Architect Outputs. To evaluate the quality model on naturally occurring errors, we prepare 50 held-out medical benchmark requests with distinct requirements. None of these requests is used for quality-model training or tuning. We run the Planner–Solver–Architect workflow and retain the first specification for every request before quality-model feedback, yielding 50 specifications and 550 fields. We apply the fixed quality model to each specification to obtain 11 field-level scores. Using the existing threshold of 0.8, we classify fields with scores > 0.8 as reasonable and those with scores ≤ 0.8 as problematic. Reviewers independently judge each field as reasonable or problematic using the same criteria as in the label audit described above, with access to the request, data descriptions, and complete specification but not the model scores. Table 7 reports field-level accuracy, the proportion of assessable fields whose model predictions match the human labels, and error detection recall, the proportion of human-identified problematic fields assigned scores ≤ 0.8.

A.3 Item Validation

Inspired by the automated template verification in Gwiazda et al. (2026), we validate each template item using two separate procedures: a programmatic format validation and an LLM-based semantic validation. Items that fail either validation are returned to the Builder with the identified errors for revision.

Format Validation. We implement deterministic checks to ensure that each generated item conforms to the predefined output schema. Specifically, we verify: (1) JSON Validity, whether the output can be successfully parsed as JSON; (2) Required Fields, whether all fields required by the corresponding item type are present; (3) Field Types, whether each field has the expected data type; (4) Valid Values, whether fields with predefined value sets contain only allowed values; and (5) Identifier Uniqueness, whether each `item_id` is unique.

Semantic Validation. We further use an LLM to evaluate each item together with the original user request and its associated medical evidence. The validator examines four aspects: (1) User Request Alignment, whether the item remains within the scope and requirements of the user request; (2) Evidence Sufficiency, whether the provided medical evidence sufficiently supports the item and its intended evaluation target; (3) Data Dependence, whether the item requires sample-level medical data or annotations rather than being answerable from the template alone; and (4) Unambiguity, whether the item is clearly defined without allowing multiple reasonable interpretations or answers. The validation prompt is shown below.

> You are an expert validator of medical benchmark template items.
>
> You are given:
> - User request: {`USER_TASK`}
> - Template item: {`ITEM_TEMPLATE`}
> - Medical evidence: {`EVIDENCE`}
>
> Your task is to validate the template item using the following criteria:
>
> 1. User Request Alignment:
> Does the item remain consistent with the provided user request and stay within its specified scope and requirements?
>
> 2. Evidence Sufficiency:
> Does the provided medical evidence sufficiently support the item and justify its intended evaluation target?
>
> 3. Data Dependence:
> Does the item require sample-level medical data or annotations, rather than being answerable from the template alone?
>
> 4. Unambiguity:
> Is the item clearly defined, without ambiguity that could lead to multiple interpretations or answers?
>
> If all criteria are YES or MOSTLY YES, return only `1`. Otherwise, briefly identify the failed criterion(s) and describe the specific issue(s).

A.4 Retrieval Models

For each template item, the Navigator generates up to five queries and autonomously selects a suitable retrieval model. Retrieval models are not manually pre-assigned to individual tasks. We store data embeddings in Milvus, together with their corresponding source identifiers and file paths, to support efficient similarity search. Each query is encoded using the corresponding text encoder to retrieve the top-K source samples by embedding similarity, with K determined by the Navigator for each query. Here, K counts distinct source samples rather than individual patches or clips. Results from different queries are deduplicated by source identifier before annotation. For example, five queries with $K = 10$ yield at most 50 distinct source samples.

(1) For the ECG Time Series Understanding task, the Navigator selects retrieval model MERL (Liu et al., 2024a) to encode source ECG recordings and textual queries into a shared embedding space, enabling retrieval of relevant ECG recordings.

(2) For the 3D CT Spatial Analysis task, the Navigator selects retrieval model M3D-CLIP (Bai et al., 2024) to encode CT volumes and textual queries, enabling retrieval at the CT-volume level.

(3) For the Colonoscopy Video VQA task, the Navigator selects SurgCLIP-β (Perez et al., 2026) as the retrieval model for video–text retrieval. We encode each candidate clip using 16 sampled frames. Queries are encoded by the same retrieval model. For each query, a source video's retrieval score is

the maximum similarity over its clips. We then select the top-K distinct source videos, retaining the matched clips' timestamps and source identifiers for subsequent annotation.

(4) For the EHR Clinical Reasoning task, the Navigator selects retrieval model Qwen3-Embedding-4B (Zhang et al., 2025a) to encode the preprocessed encounter-level records and queries for semantic retrieval. Patient identifiers and encounter order are retained to support subsequent construction of items requiring longitudinal records.

(5) For the Breast Pathology Reporting VQA task, the Navigator selects retrieval model TITAN (Ding et al., 2025) to encode source WSIs and textual queries for slide-level retrieval. The retrieved WSIs are subsequently examined by the Annotator to select and annotate relevant regions.

A.5 Evaluation Metrics

Fine-Tuning Evaluation. (1) Fine-tuning setup. For each task, we separately fine-tune Qwen3-VL-4B-Instruct on the benchmark generated by each method. The number of fine-tuning items is matched across methods within each task: 15,000 for ECG Time Series Understanding, 10,000 for 3D CT Spatial Analysis, 3,000 for Colonoscopy Video VQA, 60,000 for EHR Clinical Reasoning, and 25,000 for Breast Pathology Reporting VQA. We report accuracy before and after fine-tuning on the corresponding held-out test sets described in the main text. (2) Input preparation. We convert medical data into model-compatible inputs using the same preprocessing procedures across methods. For the ECG Time Series Understanding task, following TimeSeriesExamAgent, we render each 12-lead recording as a waveform image with lead labels and an ECG grid. For the 3D CT Spatial Analysis task, we adapt CT volumes to the models' 2D image interfaces following SpatialMed. Specifically, segmentation masks guide the selection of three orthogonal slices (axial, coronal, and sagittal), with the full body cross-section retained in each view to preserve surrounding anatomical context. SpatialMed's slice-sampling ablation reports the best average performance for three orthogonal views among the evaluated sampling strategies (Trinh et al., 2026, Appendix A.8, Table 5), supporting our choice of this 2D input representation. We apply the same three-view preprocessing protocol across all compared benchmark construction methods. For colonoscopy videos, following Colon-Bench, we uniformly sample 32 frames over the full duration of each clip and provide them in temporal order as a video input to Qwen3-VL. For the remaining two tasks, EHR Clinical Reasoning and Breast Pathology Reporting VQA, we directly use the textual clinical contexts and associated image patches, respectively, without additional modality-specific conversion. (3) Accuracy calculation. For each task, accuracy is the proportion of correctly answered test items, with equal weight assigned to each item. Multiple-choice answers are evaluated by exact matching between the predicted and reference options. Open-ended answers are evaluated by exact match against the reference answers. Invalid or unparseable responses are counted as incorrect. The same scoring rules are applied to all methods.

LLM Jury Evaluation. For each method and task, we evaluate all generated items using the associated medical input, question, and reference answer, with answer options additionally provided for multiple-choice items. ECG, CT, and video inputs follow the same preparation as in Fine-Tuning Evaluation: ECG waveform images, three orthogonal CT views, and 32 uniformly sampled video frames presented in temporal order. We adapt the evaluation prompts from TimeSeriesExamAgent to assess item quality in the context of the provided medical input. We use a separate prompt for each of the five tasks, tailoring the wording to each task while retaining the same four evaluation dimensions and 1–10 scoring scale. Table 8 summarizes the shared evaluation criteria.

Human Evaluation. For each task, we randomly sample 100 items from each method and assign two experts with relevant domain expertise to independently review all sampled items. Items are presented in random order with the generating method concealed. Reviewers assess benchmark design and data annotation in terms of request alignment, clinical relevance, clarity, and annotation correctness. Reviewers are provided with the benchmark request and data descriptions, along with each item's medical input, question, reference answer, and options where applicable. Reviewers provide independent approve/reject judgments. We report Human Approval as the percentage of sampled items approved by both experts.

Table 8: Definitions of the four dimensions in the LLM jury evaluation.

Dimension	Definition
Specificity	Specificity assesses whether the item has a single, clearly defined assessment focus, considering the medical input, question, and options or reference answer, as applicable. Higher scores indicate a precise assessment target and clear medical terminology. Lower scores indicate that the target remains overly broad or unclear in context, or that the item combines multiple unrelated objectives.
Unambiguity	Unambiguity assesses whether the item can be interpreted consistently and evaluated objectively. The wording and references to the medical input should be clear. For multiple-choice items, options should be clearly distinguishable in context; for open-ended items, the question and reference answer should be clear in meaning and scope. Higher scores indicate clarity and objectivity; lower scores reflect vague or subjective wording, unclear references, or multiple reasonable interpretations.
Domain relevance	Domain relevance assesses whether the item is relevant to the intended medical domain and task, considering the medical input, question, and options or reference answer, as applicable. Higher scores indicate strong relevance, appropriate medical terminology, and suitable medical context. Lower scores reflect weak relevance, inappropriate terminology, or content unrelated to the intended medical domain or task.
Answerability	Answerability assesses whether the item can be answered by analyzing the provided medical input, considering the question and options or reference answer, as applicable. Higher scores indicate that answering requires analysis of the input, that the input provides sufficient information to support an answer, and that the reference answer is correct. Lower scores reflect insufficient information, incorrect reference answers, reliance on case details not provided, or questions answerable without analyzing the medical input.

Table 9: Ablation study of specification refinement strategies across four medical benchmark construction tasks. *No Refinement* performs no specification refinement. *Learned Single Score* uses the mean score from our trained quality model. *LLM Field Scores* uses field-level scores produced directly by the LLM Qwen3.8-27B. *Learned Field Scores (Ours)* uses field-level scores from our trained quality model.

Task	Refinement Strategy	LLM Jury Evaluation (↑)				FT Accuracy (↑)	Human Approval (↑)
		Spec.	Unamb.	Rel.	Answer.		
ECG Time Series Understanding	No Refinement	7.30	7.53	9.24	8.69	0.480→0.588	62%
	Learned Single Score	7.82	8.23	9.26	8.89	0.480→0.647	76%
	LLM Field Scores	7.51	7.82	9.46	8.98	0.480→0.628	74%
	Learned Field Scores (Ours)	**8.28**	**8.74**	**9.51**	**9.22**	**0.480→0.672**	**85%**
3D CT Spatial Analysis	No Refinement	8.52	8.64	9.39	9.08	0.382→0.470	75%
	Learned Single Score	8.87	8.70	9.58	8.90	0.382→0.491	80%
	LLM Field Scores	8.81	9.01	9.60	9.19	0.382→0.496	77%
	Learned Field Scores (Ours)	**9.18**	**9.05**	**9.64**	**9.63**	**0.382→0.527**	**84%**
Colonoscopy Video VQA	No Refinement	7.87	7.53	9.25	8.88	0.423→0.629	63%
	Learned Single Score	8.06	7.85	9.42	8.85	0.423→0.645	71%
	LLM Field Scores	8.18	7.82	9.51	9.10	0.423→0.650	72%
	Learned Field Scores (Ours)	**8.48**	**8.15**	**9.88**	**9.36**	**0.423→0.680**	**77%**
EHR Clinical Reasoning	No Refinement	8.87	8.53	9.04	9.22	0.487→0.704	73%
	Learned Single Score	**9.15**	8.75	9.36	9.27	0.487→0.747	87%
	LLM Field Scores	8.94	8.79	9.22	9.36	0.487→0.723	85%
	Learned Field Scores (Ours)	9.08	**8.84**	**9.47**	**9.59**	**0.487→0.776**	**91%**

Table 10: Ablation study of annotation strategies across four medical benchmark construction tasks. *Evidence* denotes medical evidence guidance. *Iterative* denotes refinement through the Thought–Tool–Observation loop. Enabling both components corresponds to our full annotation strategy.

Task	Annotation Strategy		LLM Jury Evaluation (↑)				FT Accuracy (↑)	Human Approval (↑)
	Evidence	Iterative	Spec.	Unamb.	Rel.	Answer.		
ECG Time Series Understanding	×	×	7.92	8.02	9.26	9.11	0.480→0.620	70%
	✓	×	7.70	8.28	9.46	8.89	0.480→0.631	74%
	×	✓	8.14	8.44	9.11	8.93	0.480→0.625	80%
	✓	✓	**8.28**	**8.74**	**9.51**	**9.22**	**0.480→0.672**	**85%**
3D CT Spatial Analysis	×	×	8.71	8.87	9.34	9.15	0.382→0.496	76%
	✓	×	8.79	8.91	9.51	9.37	0.382→0.506	81%
	×	✓	9.04	8.64	9.40	9.37	0.382→0.501	78%
	✓	✓	**9.18**	**9.05**	**9.64**	**9.63**	**0.382→0.527**	**84%**
Colonoscopy Video VQA	×	×	8.08	7.83	9.48	9.01	0.423→0.613	68%
	✓	×	8.10	7.69	9.72	9.02	0.423→0.664	71%
	×	✓	8.46	8.08	9.52	9.29	0.423→0.647	75%
	✓	✓	**8.48**	**8.15**	**9.88**	**9.36**	**0.423→0.680**	**77%**
EHR Clinical Reasoning	×	×	8.75	8.82	9.08	9.37	0.487→0.725	83%
	✓	×	8.76	8.74	9.42	9.25	0.487→0.738	79%
	×	✓	9.03	**9.03**	9.28	9.31	0.487→0.741	86%
	✓	✓	**9.08**	8.84	**9.47**	**9.59**	**0.487→0.776**	**91%**

Table 11: Ablation of tool-library configurations across four medical benchmark construction tasks. *Basic*: a minimal tool set for pipeline execution. *Predefined*: a broad range of additional tools manually selected before construction. *Dynamic*: on-demand acquisition and integration of new tools during construction.

Task	Tool Library			LLM Jury Evaluation (↑)				FT Accuracy (↑)	Human Approval (↑)
	Basic	Predefined	Dynamic	Spec.	Unamb.	Rel.	Answer.		
ECG Time Series Understanding	✓	×	×	7.72	7.82	9.21	8.61	0.480→0.605	67%
	✓	✓	×	7.90	8.54	9.45	9.02	0.480→0.643	76%
	✓	✓	✓	**8.28**	**8.74**	**9.51**	**9.22**	**0.480→0.672**	**85%**
3D CT Spatial Analysis	✓	×	×	8.86	8.75	9.01	9.43	0.382→0.517	73%
	✓	✓	×	8.98	**9.14**	9.47	9.50	**0.382→0.524**	78%
	✓	✓	✓	**9.18**	9.05	**9.64**	**9.63**	0.382→0.527	**84%**
Colonoscopy Video VQA	✓	×	×	7.79	7.86	9.37	8.60	0.423→0.582	60%
	✓	✓	×	8.20	7.78	9.52	8.92	0.423→0.634	65%
	✓	✓	✓	**8.48**	**8.15**	**9.88**	**9.36**	**0.423→0.680**	**77%**
EHR Clinical Reasoning	✓	×	×	8.87	8.63	9.10	9.11	0.487→0.741	81%
	✓	✓	×	8.81	8.65	9.35	9.50	0.487→0.752	85%
	✓	✓	✓	**9.08**	**8.84**	**9.47**	**9.59**	**0.487→0.776**	**91%**

A.6 Evaluation of LLMs on Constructed Benchmarks

Evaluation Protocol. (1) Evaluation setup. We evaluate general-domain and medical-domain LLMs on the five benchmarks constructed by MedBenchAgent. For each task, we randomly sample 1,000 items, with sample allocation proportional to the capability-category distribution in the corresponding benchmark. All models are evaluated on the same fixed subset in a zero-shot setting. Each model receives the medical input and question, together with options for multiple-choice items. (2) Input preparation. We follow the preprocessing described under Fine-Tuning Evaluation in Appendix A.5, with two adjustments to accommodate input limits of the evaluated inference endpoints. For 3D CT, we combine the three orthogonal views into a single image, arranging views horizontally and using one row per case with consistent view ordering for cross-case questions. For colonoscopy videos, we arrange all 32 uniformly sampled frames chronologically in a single 4×8 grid, from left to right and top to bottom. All models receive the same prepared inputs for each item. (3) Accuracy calculation. We follow the scoring rules described under Fine-Tuning Evaluation in Appendix A.5 and report the proportion of correctly answered items, assigning equal weight to each item and counting invalid or unparseable responses as incorrect.

Random baseline. Following MathVista (Lu et al., 2024b), we define random chance by selecting one answer uniformly at random from the available options for multiple-choice questions and returning an empty response for open-ended questions. We report the expected accuracy of this strategy. For ECG and 3D CT, all evaluation items are four-option multiple-choice questions, resulting in a random baseline of 0.250. Colonoscopy contains five-option multiple-choice questions, giving a random baseline of 0.200. The EHR benchmark contains four-option, five-option, six-option, and open-ended questions in equal proportions. Since open-ended questions receive an empty response under the MathVista random-chance convention, the resulting expected accuracy is 0.154. For Breast Pathology, the sampled evaluation subset contains 488 two-option, 111 three-option, 379 four-option, and 22 five-option questions; weighting the corresponding random-guess probabilities by their frequencies gives an expected accuracy of 0.380.

Results and Analysis. Table 12 shows that model rankings vary across the constructed benchmarks. GPT-6 Astra achieves the highest accuracy on ECG (0.684) and 3D CT (0.726), whereas Gemini 3.7 Flash leads on colonoscopy (0.852), EHR (0.760), and breast pathology (0.866). Among the evaluated models, a general-domain model achieves the highest accuracy on each of the five benchmarks, while Lingshu-32B remains competitive on EHR (0.725). Medical-domain models therefore do not show a consistent advantage under this evaluation protocol. Even the strongest evaluated models leave room for improvement, particularly on ECG and 3D CT. These results illustrate how the constructed benchmarks reveal differences in model strengths across medical tasks and identify areas for further improvement.

Table 12: Accuracy of general-domain and medical-domain LLMs on the five benchmarks constructed by MedBenchAgent. *Random* denotes the expected accuracy under uniform random guessing for each benchmark. For each benchmark, the highest accuracy across all models is shown in **bold**, and the second-highest is underlined.

		Benchmarks constructed by MedBenchAgent				
Domain	**Model**	ECG Time Series Understanding	3D CT Spatial Analysis	Colonoscopy Video VQA	EHR Clinical Reasoning	Breast Pathology Reporting VQA
	Random	0.250	0.250	0.200	0.154	0.380
General	GPT-6 Astra (OpenAI, 2026b)	**0.684**	**0.726**	0.581	0.718	0.714
	GPT-5.6 Luna (OpenAI, 2026a)	0.469	0.636	0.558	0.657	0.730
	Gemini 3.7 Flash (Google DeepMind, 2026c)	0.480	0.707	**0.852**	**0.760**	**0.866**
	Gemini 3.6 Flash (Google DeepMind, 2026b)	0.522	0.679	0.767	0.704	0.811
	Qwen3.8-Flash (Qwen Team, 2026d)	0.543	0.589	0.670	0.752	0.797
	Qwen3.6-27B (Qwen Team, 2026b)	0.496	0.557	0.602	0.672	0.752
Medical	MedGemma 27B IT (Sellergren et al., 2025)	0.454	0.376	0.294	0.588	0.530
	MedGemma 1.5 4B IT (Sellergren et al., 2026)	0.374	0.350	0.223	0.330	0.545
	Lingshu-32B (Xu et al., 2025)	0.443	0.437	0.463	0.725	0.691
	Lingshu-7B (Xu et al., 2025)	0.404	0.284	0.364	0.669	0.674
	HuatuoGPT-Vision-7B-Qwen2.5VL (FreedomIntelligence, 2025)	0.495	0.269	0.435	0.570	0.572

Table 13: Accuracy of selected LLMs on the five benchmarks constructed by MedBenchAgent under three medical-input conditions. *Random* denotes uniform random guessing. *w/o medical input* removes the case-specific medical input; *Mismatched medical input* replaces the original medical input with an answer-incompatible input from another case within the same benchmark and task category; *Full input* uses the original case-specific medical input. For each model, the three medical-input conditions use the same evaluation items and differ only in the medical input.

Model	**Input Condition**	**ECG**	**3D CT**	**Colonoscopy**	**EHR**	**Breast Pathology**
Random		0.250	0.250	0.200	0.154	0.380
GPT-6 Astra	w/o medical input	0.303	0.286	0.230	0.190	0.384
	Mismatched medical input	0.254	0.359	0.251	0.221	0.419
	Full input	0.684	0.726	0.581	0.718	0.714
Gemini 3.7 Flash	w/o medical input	0.353	0.404	0.305	0.234	0.388
	Mismatched medical input	0.247	0.371	0.247	0.254	0.429
	Full input	0.480	0.707	0.852	0.760	0.866
Qwen3.8-Flash	w/o medical input	0.324	0.305	0.227	0.181	0.392
	Mismatched medical input	0.265	0.321	0.214	0.234	0.406
	Full input	0.543	0.589	0.670	0.752	0.797
Lingshu-32B	w/o medical input	0.273	0.278	0.204	0.193	0.365
	Mismatched medical input	0.234	0.266	0.211	0.217	0.393
	Full input	0.443	0.437	0.463	0.725	0.691

Table 14: Breakdown of non-approval reasons for items constructed by our method across five tasks. The original human-evaluation sample of 100 items per task and the independent approve/reject decisions are retained. Two domain experts review the items not jointly approved using the benchmark request and item materials, independently assign non-exclusive reason labels, and discuss disagreements. Each item is counted once in each reason category confirmed by both experts and may appear in multiple categories. Percentages are calculated over the non-jointly-approved items in each task (first row). The last row counts items with no jointly confirmed rejection reason.

Reason / audit status	ECG	3D CT	Colonoscopy	EHR	Breast Pathology
Not jointly approved	15	16	23	9	21
Incorrect reference answer	3 (20.0%)	2 (12.5%)	7 (30.4%)	2 (22.2%)	5 (23.8%)
Insufficient medical input	4 (26.7%)	1 (6.3%)	3 (13.0%)	3 (33.3%)	6 (28.6%)
Ambiguous question or options	5 (33.3%)	6 (37.5%)	8 (34.8%)	4 (44.4%)	4 (19.0%)
Request misalignment	2 (13.3%)	3 (18.8%)	2 (8.7%)	1 (11.1%)	4 (19.0%)
Limited clinical relevance	1 (6.7%)	4 (25.0%)	1 (4.3%)	0 (0.0%)	2 (9.5%)
Other confirmed issues	1 (6.7%)	1 (6.3%)	3 (13.0%)	2 (22.2%)	0 (0.0%)
Unresolved or unconfirmed rejection reason	0 (0.0%)	1 (6.3%)	2 (8.7%)	1 (11.1%)	3 (14.3%)

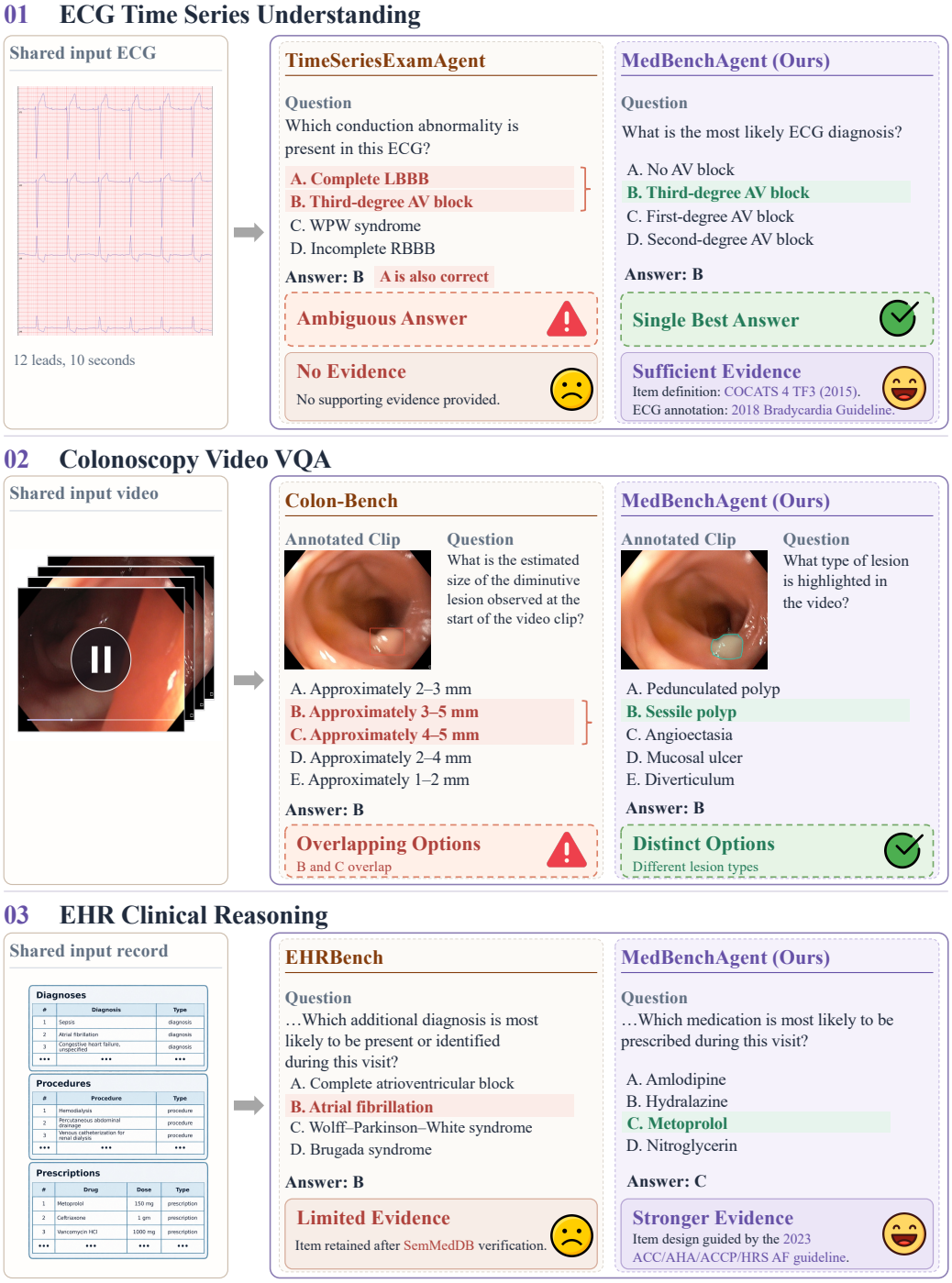

Figure 6: Additional qualitative comparisons of benchmark items for the remaining three medical tasks: ECG time series understanding, colonoscopy video VQA, and EHR clinical reasoning. For each task, MedBenchAgent and the corresponding baseline use the same medical input. MedBenchAgent constructs an ECG item with a single best answer and a colonoscopy video item with distinct answer options. The ECG and EHR examples further illustrate stronger medical evidence support for item construction.