EpiBench: Benchmarking Multi-turn Research Workflows for Multimodal Agents

Xuan Dong*, Huanyang Zheng*, Tianhao Niu*, Zhe Han, Pengzhan Li, Bofei Liu, Zhengyang Liu, Guancheng Li, Qingfu Zhu†, and Wanxiang Che

Harbin Institute of Technology
Harbin, Heilongjiang, China
{xd,huanyzheng,thniu,qfzhu,car}@ir.hit.edu.cn
*Equal contribution
†Corresponding author

Abstract. Scientific research follows multi-turn, multi-step workflows that require proactively searching the literature, consulting figures and tables, and integrating evidence across papers to align experimental settings and support reproducible conclusions. This joint capability is not systematically assessed in existing benchmarks, which largely underevaluate proactive search, multi-evidence integration and sustained evidence use over time. In this work, we introduce EpiBench, an episodic multi-turn multimodal benchmark that instantiates short research workflows. Given a research task, agents must navigate across papers over multiple turns, align evidence from figures and tables, and use the accumulated evidence in the memory to answer objective questions that require cross-paper comparisons and multi-figure integration. EpiBench introduces a process-level evaluation framework for fine-grained testing and diagnosis of research agents. Our experiments show that even the leading model achieves an accuracy of only 29.23% on the hard split, indicating substantial room for improvement in multi-turn, multi-evidence research workflows. Code and data are available at https://github.com/RetroDnix/EPIBench.

Keywords: Multi-turn Research · Proactive Search · Multimodal Integration · Evidence Memory

1 Introduction

The accelerating expansion of the scientific literature has made literature review, verification, and experimental reproduction increasingly costly and time consuming, creating a pressing demand for automated research assistance [15, 30]. This demand has fueled growing interest in research agents for literature review, verification, and experimental reproduction. Recent advances in large multimodal models and tool-using agents make this direction increasingly practical, enabling systems to browse the web, read PDFs, invoke external tools, and produce citation-supported answers with intermediate traces [19,22,36]. As a result, agents are now expected to locate relevant articles, interpret figures and tables,

Table 1: Comparison of agentic reasoning benchmarks by key capabilities. ✓ indicates supported, and × indicates not supported.

Benchmark	Cross-paper Integration	Multimodal Integration	Proactive Search	Evidence Memory	Multi-turn Dialogue
Non-paper benchmarks					
MMCR [35]	×	×	×	×	✓
Vision-DeepResearch [37]	×	✓	✓	×	×
MMDU [13]	×	✓	×	×	✓
VisChainBench [14]	×	✓	×	×	×
MMDeepResearch [8]	✓	✓	✓	×	×
MMSearch-Plus [26]	×	✓	×	×	×
Paper-centric benchmarks					
LiveXiv [23]	×	×	×	×	×
CharXiV [33]	×	×	×	×	×
SPIQA [18]	×	×	×	×	×
SciVQA [16]	×	×	×	×	×
SCIGraphQA [11]	×	×	×	×	×
SIN-Bench [20]	×	✓	×	×	×
PaperArena [32]	✓	✓	×	×	×
EpiBench (Ours)	✓	✓	✓	✓	✓

align experimental settings across sources, and deliver conclusions that can be checked and reproduced.

Despite these advances, current evaluations still fall short of real research workflows, especially in paper-related settings. Table 1 suggests two fundamental mismatches. First, **benchmark design** is often not workflow-faithful. Many tasks begin from an explicitly specified target paper or direct identifiers, so agents are not required to perform *proactive search* from citation cues or indirect hints. When the target is already known, questions are frequently answerable from a single figure or table, and *cross-paper multimodal integration* over *multiple figures and tables* is seldom necessary [20, 33]. In addition, many benchmarks provide limited coverage of *multi-turn* research episodes where evidence must be accumulated, refined, and reused over time. Second, the **evaluation protocol** is incomplete. Because repeated re-browsing is typically allowed, benchmarks rarely measure whether an agent *reuses previously acquired evidence*, and they seldom check whether intermediate evidence is correctly grounded, which makes evidence attribution errors and integration failures hard to diagnose [18, 23]. These two mismatches naturally lead to two core gaps. *(Problem 1)* existing benchmarks do not comprehensively assess human-like workflows that require *proactive search*, *multi-turn interaction*, and *cross-paper multimodal evidence fusion* over time. *(Problem 2)* existing protocols lack process-level metrics for *evidence reuse* and evidence correctness needed for fine-grained diagnosis.

To address *Problem 1*, we introduce EpiBench, a benchmark that evaluates research workflows at the process level. Each instance is an episodic multi-turn task that simulates a short research process. In our formulation, the task description does not provide a target paper or direct identifiers, mirroring realistic user queries that rarely include precise citations. This setting requires agents

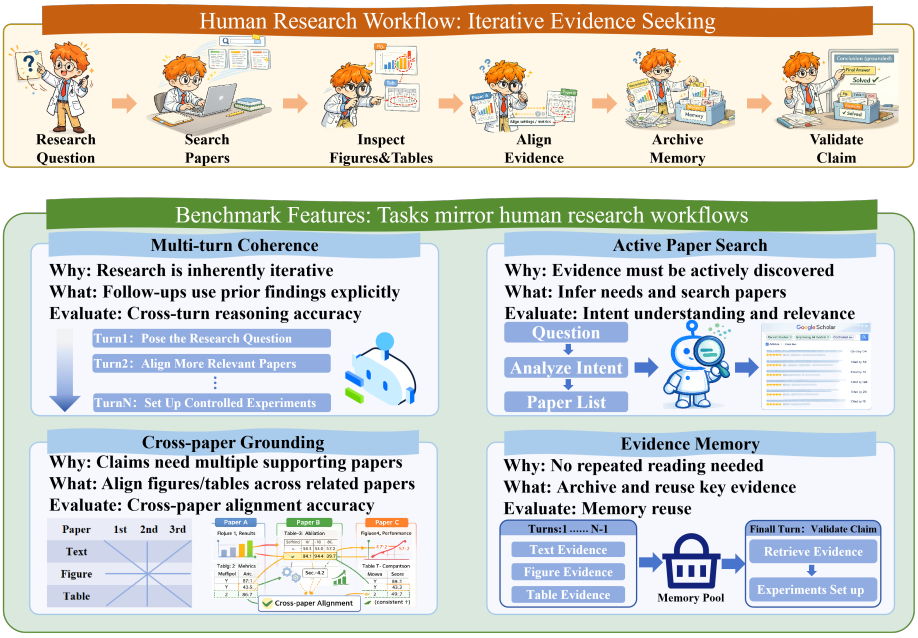


Fig. 1: EpiBench motivation and key features. Top: A human research workflow proceeds iteratively from posing a question to searching papers, inspecting figures and tables, aligning evidence across papers, archiving findings, and validating the final claim. **Bottom:** EpiBench mirrors this workflow with four benchmark features: multi-turn coherence, active paper search, cross-paper grounding over figures and tables, and evidence memory that supports reuse in later turns.

to perform **proactive search**, using **citation cues** and **indirect hints** to discover **multiple papers**, extract structured evidence from specific **figure and table regions**, and repeatedly integrate evidence across turns to produce the final answer. The questions are objective and enforce cross-paper multi-figure and multi-table alignment through comparisons, selection rules, and set-style constraints grounded in visual legends and plotted values. This design reduces the possibility of answering from prior world knowledge or incomplete visual context, and it directly tests an agent's ability to accumulate and integrate multimodal evidence over multi-turn interaction.

To address *Problem 2*, EpiBench goes beyond task design and introduces an access-budget evaluation setting that quantifies evidence reuse and efficiency from agent interaction logs. As illustrated in Fig. 1, we focus on human-like research workflows where evidence is actively collected, aligned, and reused across turns. To enforce memory-based reuse, we restrict the agent to answering questions exclusively from memory without invoking external tools when reuse is required. To measure reuse correctness and evidence grounding, we annotate required evidence units at the granularity of paper identifiers and figure or table identifiers, enabling automatic provenance-based scoring. Using these traces, we

evaluate not only final correctness but also evidence reuse and tool efficiency, and we diagnose failures such as retrieval errors, perception errors, stale or inconsistent reuse, and reasoning errors.

We conduct extensive experiments with representative open source and closed source Large Multimodal Model (LMM) based agents and observe that research workflow performance remains limited. Beyond the final accuracy, intermediate traces reveal brittle behavior. Agents often rely on incomplete visual evidence, link answers to incorrect figures or tables, and resort to erroneous memory or prior knowledge instead of seeking information that truly supports the evidence. These failure modes are difficult to diagnose with accuracy alone, but are exposed by EpiBench through process level evaluation of evidence grounding, memory reuse, and tool efficiency. We expect EpiBench to support the development of more reliable research agents and to accelerate progress in multimodal literature understanding and evidence based scientific reasoning.

Our main contributions are summarized as:

- We introduce EpiBench, an episodic workflow-level benchmark for research agents, filling a key gap in prior evaluations that largely lack process-level research-workflow tasks. It requires proactive search across papers and joint use of figures, tables, and text, with each retrieved paper contributing essential evidence to the final answer.
- We propose a budgeted evaluation protocol with annotations and diagnostics of the evidence unit to measure episode success, evidence grounding, evidence reuse, and tool efficiency under explicit access constraints of paper and figures.
- We identify robustness weaknesses of current LMM agents on research-workflow tasks: they often hallucinate answers based on prior knowledge or incomplete visual evidence, and fail to reliably retrieve, ground, and reuse evidence correctly across papers and figures over multiple turns.

2 Related Work

2.1 Benchmarks on Scientific Literature

Recent benchmarks have explored scientific document understanding, particularly focusing on figures and charts in scholarly papers. Datasets such as SciVQA [16], CharXiv [33], and SPIQA [18] study question answering over scientific figures, while LiveXiv [23] extends this paradigm by generating large-scale and continuously updated VQA tasks from academic papers. More recent work has focused on multimodal reasoning across figures and textual context. For example, SIN-Bench [20] simulates multi-step workflows such as evidence discovery and grounded question answering in long multimodal documents. PaperArena [32] is a closely related benchmark that evaluates tool-augmented multi-hop navigation across academic papers. However, its tasks rarely require jointly integrating information from multiple figures and tables to derive an answer.

However, these benchmarks remain largely limited to single-document settings. Most tasks involve answering isolated questions about a single figure, table, or text span within one paper, so they rarely require integrating multimodal evidence across multiple figures and tables or aligning evidence across multiple papers, and they seldom require reusing information obtained from earlier questions. In contrast, EpiBench evaluates agents in a setting that requires cross-paper reasoning, multi-figure and multi-table evidence integration, and iterative evidence accumulation and reuse across turns.

2.2 Multi-Turn Benchmarks and Long-Horizon Memory

Several recent benchmarks investigate multi-turn reasoning in multimodal settings. MMDU [13] introduces multi-image, multi-turn dialogues derived from Wikipedia content. MULTIVERSE [10] further expands this direction by constructing a dataset with greater scale and depth, while MMCR emphasizes contextual reasoning across dialogue turns to improve model performance.

The development of multi-turn benchmarks has extended the reasoning horizon required to solve a question, necessitating that agents reuse information accumulated during earlier reasoning steps. However, existing studies (e.g., Memory-QA [9] and EMP [4]) primarily focus on retrieving answers to individual queries from previously stored memories. Moreover, they are generally not paper-centric and rarely require cross-paper multimodal evidence integration over figures and tables. In contrast, EpiBench targets research workflows by requiring agents to iteratively explore multiple papers, integrate multimodal evidence, and build upon prior findings across turns to answer objective, evidence-grounded scientific questions.

3 EpiBench: The Benchmark

In this section, we introduce EpiBench and describe its benchmark design. We first formalize the episodic task setting and the corresponding evidence and process annotations, followed by our benchmark construction pipeline. Furthermore, we summarize the key features that distinguish EpiBench from existing benchmarks. Finally, we define the evaluation metrics used throughout the paper.

3.1 Task Formulation

EpiBench evaluates research agents on episodic multi-turn tasks grounded in multimodal evidence from scientific papers. Each episode is a sequence of turns $\{(q_t, \mathcal{I}_t)\}_{t=1}^{T}$, where q_t is an objective question and $\mathcal{I}_t$ optionally provides a seed cue such as an image snippet or a short bibliographic hint. The agent answers each q_t and may invoke an external toolkit $\mathcal{T}$ for paper discovery and content access, including searching for papers, opening a paper, and extracting figures, tables or text. Figure 2 illustrates a representative episode and the corresponding tool-use workflow.

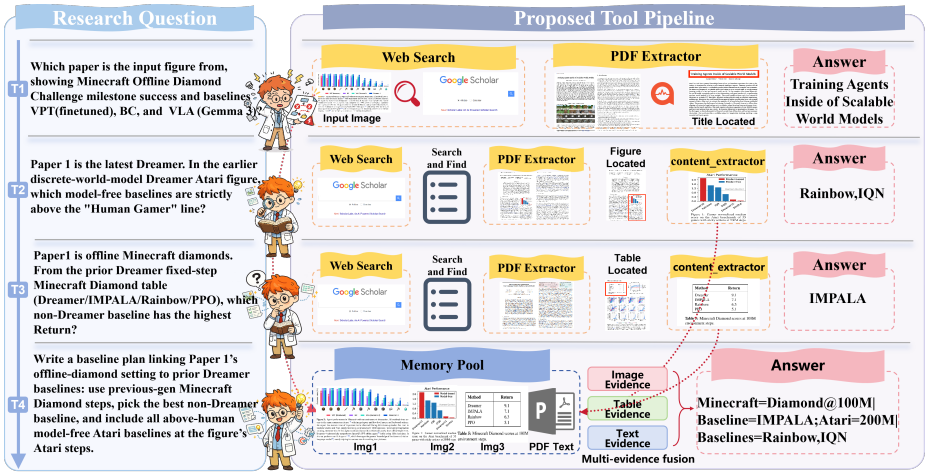


Fig. 2: Example episode and tool-use workflow in EpiBench. A research task is instantiated as a multi-turn question chain. At each turn, the agent uses tools to search for relevant papers, open PDFs, and locate and extract target figures or tables. Retrieved evidence is stored in a memory pool and reused in later turns. The final turn requires multi-evidence fusion across papers, combining visual and textual evidence in the memory pool to produce the structured answer.

3.2 Benchmark Construction

We construct EpiBench through an expert extraction, expansion, and curation pipeline on papers. We first collect a seed set of classic papers from public venues and preprint repositories, including OpenReview [2] and arXiv [1]. The seed set covers six broad areas of computer vision and machine learning and contains 68 representative papers.

Next, we expand the candidate pool using Connected Papers [25], a citation graph exploration tool that discovers relevant papers around each seed paper at scale. After deduplication and basic quality filtering, we obtain a pool of 485 papers as the source corpus for episode construction.

We then use GPT-5.2 [17] to draft episodic question chains. Each draft episode is designed to follow a research workflow. Early turns locate papers and extract evidence from specific figures or tables. Later turns require reusing previously accessed evidence and integrating multiple evidence units across papers. For each turn, we also generate a checklist of tool calls, where each call specifies the target paper and, when applicable, the target figure or table. This supports verification of the intended solution path.

Finally, five master-level candidates in computer science refine, validate, and curate the generated episodes. They remove ambiguous cases, correct evidence targets, and resolve annotation inconsistencies through a final consolidation pass, ensuring that each required paper contributes essential evidence to the final answer. The resulting benchmark contains 102 high-quality episodes, split into

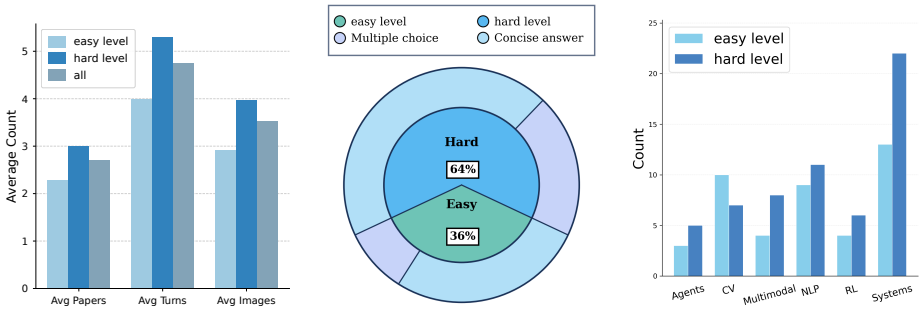


Fig. 3: Dataset overview of EpiBench. **Left:** Statistics of the EpiBench dataset by difficulty level. **Middle:** Data distribution across categories. **Right:** Episode counts across subject areas stratified by difficulty.

Easy (37) and Hard (65). Difficulty is assigned based on the required number of turns, the amount of cross-paper and multimodal evidence integration, and the extent of evidence reuse required in later turns. Figure 3 illustrates key statistics of the benchmark. The left panel shows the average count of papers, turns, and images across different difficulty levels. The middle panel presents the overall distribution across difficulty levels and answer formats. The right panel further breaks down the number of episodes by subject area for the easy and hard splits.

3.3 Key Features of EpiBench

Multi-turn research episodes. EpiBench evaluates agents on short research workflows rather than isolated questions. Each instance is organized as an episode with multiple turns, where later questions depend on evidence acquired earlier. This structure captures common research behaviors such as iterative retrieval, progressive refinement of hypotheses, and cross-turn consistency requirements, and it exposes failure modes that do not appear in single-turn settings.

Multimodal evidence integration across papers. The benchmark emphasizes evidence fusion from figures, tables, and text. Questions are designed so that a correct answer requires aligning information across multiple visual artifacts and, in many cases, across multiple papers linked by citations or indirect cues. This setting stresses multimodal grounding, such as reading plotted values and legends, extracting table entries, and reconciling experimental settings across sources.

Proactive citation-guided search. EpiBench requires agents to proactively discover relevant papers instead of starting from an explicitly specified target title or identifier. Agents must use citation cues and indirect hints accumulated during the episode to formulate search queries, follow reference trails, and resolve paper identities, which better reflects realistic research settings where users rarely provide precise citations.

Evidence reuse with process supervision. EpiBench explicitly targets memory use by marking evidence units that should be reused in later turns.

In particular, for the final high-difficulty turn that requires fusing all previously acquired multimodal evidence, we disable tool access to enforce memory-based reuse rather than repeated re-browsing. This enables process-level evaluation of whether agents revisit the right evidence and reuse earlier findings correctly when appropriate.

3.4 Evaluation Metrics

We report four metrics that capture success rate, evidence correctness, minimality gap and error analysis.

Success rate. We report episode success rate (ESR), where an episode is counted as correct only if the agent answers all turns correctly. Since the final turn is intentionally harder and requires multi-evidence fusion, we additionally report turn-level accuracy for the final turn ($\mathrm{Acc}_{\mathrm{final}}$) and the average accuracy over all non-final turns ($\mathrm{Acc}_{\mathrm{pre}}$).

Evidence correctness. To verify that correct answers follow the intended reasoning path rather than lucky guesses, we annotate each turn t with a set of required evidence units $\mathcal{E}_{i,t}$, indexed by paper identifier and figure or table identifier, and evaluate whether the agent's accessed evidence $\mathcal{A}_{i,t}$ from the tool trace aligns with $\mathcal{E}_{i,t}$ on correctly answered turns:

$$\mathrm{EC} \;=\; \frac{\sum_{i=1}^{N}\sum_{t=1}^{T_i} y_{i,t}\,|\mathcal{E}_{i,t}\cap\mathcal{A}_{i,t}|}{\sum_{i=1}^{N}\sum_{t=1}^{T_i} y_{i,t}\,|\mathcal{E}_{i,t}|}. \tag{1}$$

This measures the fraction of required evidence units that are actually accessed when the answer is correct.

Minimality gap. Each episode is annotated with a checklist that specifies a minimal viable tool-use path with $c_i^\star$ tool calls. Let c_i be the number of tool calls executed by the agent in episode i, and let $s_i \in \{0,1\}$ indicate episode success. We report the minimality gap on successful episodes:

$$\mathrm{MG}_{\mathrm{succ}} \;=\; \frac{1}{\sum_{i=1}^{N} s_i}\sum_{i=1}^{N} s_i\cdot\frac{c_i}{c_i^\star}. \tag{2}$$

A larger gap indicates more redundant tool usage relative to the minimal viable solution path.

Error analysis. To understand failure modes beyond aggregate accuracy, we perform trace-based error attribution. For each failed episode, we assign a single error label based on the earliest point of failure in the tool trace and evidence alignment, and then report the proportion of episodes in each category.

4 Agent Framework

This section describes the agent implementation used in our experiments. Our goal is to provide a clean and reproducible interface that couples an LMM

with external tools and a persistent memory, and supports detailed logging for process-level evaluation. We build on SMOLAGENTS [21] and adapt it to our episodic paper-centric setting.

4.1 Agentic Workflow

At each turn t, the agent receives a question q_t and optional images. It maintains a memory state M_t that stores all previously observed content in the episode. The agent follows an iterative loop that alternates between tool invocation and answering. In practice, we implement this loop using function-calling. The model decides whether to call a tool and how to parameterize the call based on the current question and memory context.

4.2 Tool Environment

We provide four tools tailored to literature search and paper content access.

- **WEB SEARCH.** Queries a search backend and returns ranked results with titles, snippets, and URLs. We use both Google search and DuckDuckGo search as backends [6, 7]. This tool is used to locate candidate papers, resolve paper identities from indirect cues, and follow citation leads.
- **PDF EXTRACTOR** [31]. Given a paper identifier (for example, title or arXiv ID), this tool downloads the PDF and parses it into a normalized markdown representation.
- **PDF EXTRACTOR RAG.** Performs retrieval over the cached markdown for a given paper. Given a text query, it returns the most relevant spans together with lightweight provenance, such as section context and figure or table references. This tool supports targeted evidence lookup without re-reading the full paper.
- **CONTENT EXTRACTOR.** Extracts a specified figure or table from a cached paper. For figures, it returns the figure image and caption context. For tables, it returns a structured table representation together with surrounding text when available. This tool is used when a question requires visual grounding.

All tools share a unified interface and return structured outputs that can be appended to memory and traced during evaluation.

4.3 Memory Management

Memory is implemented as an episode-level chronological store. We append retrieved PDFs, extracted figures and tables, and tool observations in the order they are accessed. This design keeps the agent state transparent and enables post hoc computation of reuse and efficiency metrics from the interaction trace.

To stress evidence reuse, we run the final turn in a memory-only mode for episodes that require reuse. In this mode, we disable all tools. The agent must answer using evidence already stored in memory rather than re-browsing or re-downloading documents. This setting isolates memory utilization from additional retrieval and makes reuse behavior measurable and comparable across models.

5 Experiment

5.1 Experimental Setup

Models. We evaluate six representative open-source and four closed-source LMMs as agent backbones. Our open-source baselines include the Qwen3.5-9B, Qwen3.5-35B-A3B [28], Qwen3-VL-235B-A22B-Instruct, Qwen3-VL-235B-A22B-Thinking [3], GLM-4.5V [29], and Kimi-K2.5 [27]. Our closed-source baselines include GPT-5-Mini [24], GPT-5.2 [17], Grok-4.1 [34], and Gemini-2.5-Pro [5]. We also include a human baseline from two computer-science master-level candidates, who solve the same episodes under the same tool interface.
Evaluation Metrics. We report three primary outcome metrics in the main table, episode success rate (ESR), accuracy on the final turn ($\mathrm{Acc}_{\mathrm{final}}$), and accuracy on all non-final turns ($\mathrm{Acc}_{\mathrm{pre}}$). Full metric definitions are given in Sec. 3.4. All turn answers are scored with the LLM-as-a-Judge [12], where GPT-5.2 [17] assigns binary correctness labels under a fixed rubric. In addition, we present diagnostic analyses using evidence correctness and minimality gap, together with an error analysis that attributes failures to retrieval, perception, memory reading, or reasoning.
Implementation Details. We employ a ReAct-like [36] workflow, in which the agent first generates a plan outlining its intended actions before producing executable code for tool invocation. For each turn within an episode, the agent is constrained by a limited step budget (10 steps by default). If the agent fails to provide a final answer within the allotted steps, it is granted one additional opportunity to summarize its exploration process and deliver an alternative answer in a single response. For generation, we use the default decoding configuration for all evaluated models and enable their reasoning mode when available.

5.2 Main Results

Table 2 summarizes the main results on EpiBench across ten LMMs under our agentic workflow. Across all models, GPT-5.2 achieves the best overall performance, and it is also the strongest model on the Hard split in terms of ESR, reaching 29.23%. Among the newly included Qwen3.5 models, Qwen3.5-35B-A3B substantially outperforms Qwen3.5-9B in overall ESR, but still trails the strongest closed-source model and Kimi-K2.5. However, even for the most capable closed-source systems, high turn-level accuracy on non-final turns does not translate into reliable end-to-end performance. $\mathrm{Acc}_{\mathrm{pre}}$ remains relatively high for several closed-source models, yet $\mathrm{Acc}_{\mathrm{final}}$ and ESR drop sharply, indicating that multi-evidence fusion across papers and figures or tables, together with memory-based evidence reuse, remains a major bottleneck.

Notably, $\mathrm{Acc}_{\mathrm{final}}$ is particularly informative in our setting because the final turn is designed to require reuse of previously retrieved evidence and joint reasoning over multiple evidence units. The persistent gap between $\mathrm{Acc}_{\mathrm{pre}}$ and $\mathrm{Acc}_{\mathrm{final}}$, together with the low ESR on hard episodes, suggests that failures often arise after initial evidence acquisition, when agents must organize, recall,

Table 2: Main results on EpiBench. We report episode success rate (ESR, %), accuracy on the final turn (Acc_{final}, %), and accuracy on all non-final turns (Acc_{pre}, %), stratified by difficulty. Instr. denotes instruction-tuned models and Think. denotes thinking-mode models.

Base LMM	Easy			Hard			Average		
	ESR	Acc_{final}	Acc_{pre}	ESR	Acc_{final}	Acc_{pre}	ESR	Acc_{final}	Acc_{pre}
Open-source models									
Qwen3.5-9B	5.26	55.26	41.11	0.00	36.07	24.07	2.02	43.43	28.33
Qwen3.5-35B-A3B	32.43	77.78	57.61	10.77	46.88	52.30	18.63	58.00	53.60
Qwen3-VL-235B-Instr.	16.98	41.51	70.67	8.16	**51.02**	49.77	12.75	46.08	55.17
Qwen3-VL-235B-Think.	30.00	65.00	76.09	3.23	43.55	43.84	13.73	51.96	51.90
GLM-4.5V	8.11	51.35	38.95	0.00	30.77	29.17	2.94	38.24	31.59
Kimi-K2.5	52.94	73.53	89.77	26.15	49.23	79.51	35.35	57.58	81.91
Closed-source models									
GPT-5-Mini	29.73	62.16	84.21	21.54	50.77	74.65	24.51	54.90	77.02
GPT-5.2	**58.33**	**77.78**	**91.30**	**29.23**	49.23	**86.46**	**39.60**	**59.41**	**87.63**
Grok-4.1	15.38	66.67	63.10	6.35	39.68	51.07	9.80	50.00	53.85
Gemini-2.5-Pro	27.03	64.86	75.79	7.69	46.15	57.29	14.71	52.94	61.88
Human Baseline									
Human	91.43	97.14	96.67	81.36	94.42	95.79	85.11	95.74	96.01

and align evidence under constraints. In contrast, two master-level candidates achieve substantially higher performance, highlighting a large gap between current agents and expert research workflows. To understand where models break down, we next present a process-level analysis of tool use and evidence chains in Sec. 5.3, followed by a focused study of memory reuse behavior and its impact on success in Sec. 5.4.

5.3 Process-Level Diagnostics and Error Analysis

To understand why high turn-level accuracy does not translate into high episode success, we analyze process-level behaviors beyond final answers. Table 3 reports evidence correctness (EC), measuring alignment to required evidence units, and minimality gap (MG), measuring tool-call redundancy relative to the minimal viable checklist. GPT-5.2 and Kimi-K2.5 achieve consistently high EC, suggesting that their correct answers are more often supported by sufficient evidence, whereas other models more frequently rely on prior knowledge or guessing. MG varies widely and is weakly coupled with EC, and redundancy is often higher on Easy than on Hard, consistent with over-exploration on simpler tasks. Compared with master-level candidates, both EC and MG reveal a substantial gap, motivating finer-grained error analysis from interaction traces.

Notably, EC and MG quantify whether agents follow the intended evidence path and *how* efficiently they use tools, but they do not explain *why* episodes fail.

Table 3: Process-level diagnostics on representative models with complete tool traces. We report evidence correctness (EC, %) and minimality gap (MG) by difficulty. Higher EC is better. Lower MG is better. Instr. denotes instruction-tuned models and Think. denotes thinking-mode models.

Base LMM	Easy		Hard		Average	
	EC ↑	MG ↓	EC ↑	MG ↓	EC ↑	MG ↓
Open-source models						
Qwen3-VL-235B-Instr.	64.63	1.41	63.48	1.03	63.85	1.23
Qwen3-VL-235B-Think.	54.55	**1.02**	45.66	**1.00**	48.63	**1.01**
GLM-4.5V	24.49	2.60	31.20	2.60	29.31	2.60
Kimi-K2.5	81.30	1.49	84.91	1.20	84.13	1.35
Closed-source models						
GPT-5-Mini	75.40	3.00	82.43	1.94	80.75	2.32
GPT-5.2	**85.93**	1.34	**86.15**	1.27	**86.10**	1.30
Grok-4.1	85.71	3.33	78.33	2.88	80.12	3.07
Gemini-2.5-Pro	71.30	2.17	68.71	1.41	69.39	1.80
Human Baseline						
Human	83.45	1.00	88.35	1.02	86.88	1.01

We therefore conduct an error-rate analysis by attributing each failed episode to its earliest point of deviation in the tool trace and evidence alignment. Specifically, based on their underlying causes, we categorize the failures into five types: **Retrieve** errors arise when the agent fails to locate the intended papers. **Perception** errors occur when the agent accesses the relevant evidence but misreads or misinterprets figures or tables. **Reasoning** errors refer to incorrect integration or constraint satisfaction despite having accessed the necessary evidence. **Reading Memory** errors capture cases where the required evidence was retrieved earlier but is not correctly reused at later turns, including conflation across sources or missing evidence selection during final fusion. **Others** collects residual failures that do not fit the above categories, such as max-step terminations, runtime errors or pdf-parsing errors.

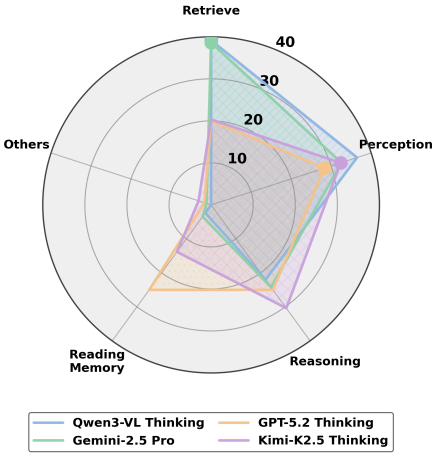
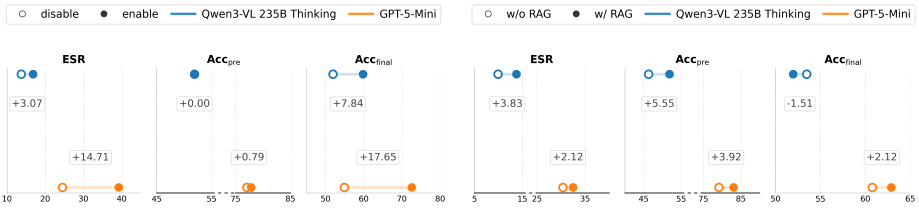


Fig. 4: Failure-type distribution (%) across models.

Figure 4 summarizes failure-type distributions across models. We observe



(a) Final-turn tool access. **(b)** Removing PDF text RAG.

Fig. 5: Ablations on memory-centric constraints. **(a)** Allowing tool use in the final turn yields a large recovery in ESR and $\text{Acc}_{\text{final}}$, indicating that multimodal evidence cached across turns is often insufficient for reliable final fusion under the memory-only protocol. **(b)** Removing PDF text RAG has mixed effects and generally smaller impact, suggesting that the primary bottleneck lies in multimodal evidence selection and alignment rather than text retrieval alone.

a clear shift in where different agents
break down. Gemini-2.5-Pro and Qwen3-VL-Thinking are dominated by **Retrieve** errors, suggesting that they often fail before reaching the stage where multi-evidence fusion becomes decisive. In contrast, stronger reasoning-oriented models shift mass toward later-stage failures. GPT-5.2 and Kimi-K2.5 show larger shares of **Perception** and **Reasoning** errors, indicating that once retrieval succeeds, the remaining bottleneck lies in reliable multimodal evidence alignment and integration. GPT-5.2 also exhibits a higher **Reading Memory** error rate, which is consistent with it reaching memory-reuse turns more often due to stronger early evidence acquisition. This still highlights substantial room for improvement in multimodal evidence reuse from memory.

5.4 Impact of Evidence Memory and Final-Turn Constraints

We probe how memory constraints affect episodic success with two ablations in settings where EpiBench requires multi-evidence fusion across turns. First, we compare the main *memory-only* protocol (tools disabled on the final turn) with a relaxed setting that *enables* tool use on the final turn, which serves as an upper-bound style reference when agents are allowed to re-open and re-browse documents. Second, we remove the PDF text RAG tool to isolate the contribution of within-document text retrieval.

As shown in Fig. 5a, allowing tool use on the final turn substantially improves ESR and $\text{Acc}_{\text{final}}$, suggesting that memory-based reuse and multi-evidence integration remain a primary bottleneck. Figure 5b shows that removing text RAG has smaller and model-dependent effects, suggesting that text retrieval is not the main limiting factor for these episodes. Overall, the results point to multimodal evidence selection, indexing, and alignment in memory as the primary challenge for reliable research-workflow completion.

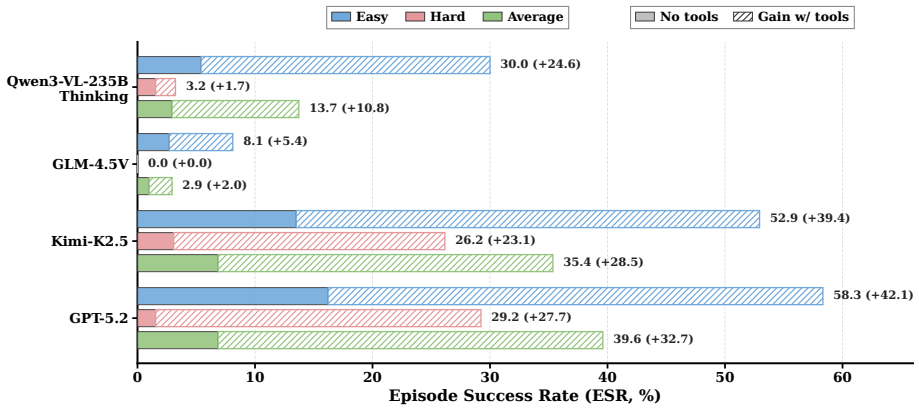


Fig. 6: Closed-book no-tool ablation on EpiBench. Solid bars show no-tool ESR, and hatched bars show the gain from the tool-augmented agent. The sharp drop indicates that EpiBench requires active evidence acquisition and multimodal integration.

5.5 Closed-book No-tool Ablation

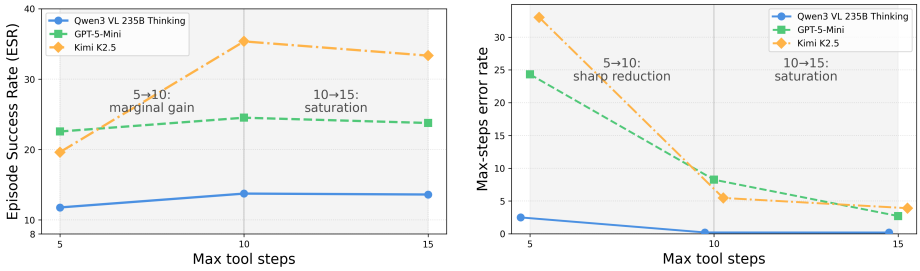
To further examine whether the benchmark can be solved by parametric memory alone, we conduct a closed-book no-tool ablation. In this setting, all external tools are disabled throughout the episode, including web search, PDF extraction, PDF text retrieval, and figure/table extraction. The agent must answer each turn directly from its internal knowledge and the dialogue context.

As shown in Fig. 6, disabling tools causes a substantial performance drop across the models included in this ablation. On the overall split, GPT-5.2 decreases from 39.6% to 6.9% ESR, and Kimi-K2.5 decreases from 35.4% to 6.9%. Qwen3-VL-235B-Thinking also drops from 13.7% to 2.9%, while GLM-4.5V remains low in both settings. The gap is especially clear on the hard split, where no-tool ESR is at most 3.1% for all models in this ablation, compared with 29.2% for GPT-5.2 and 26.2% for Kimi-K2.5 under the full tool-augmented setting. These results suggest that EpiBench cannot be reliably solved by closed-book parametric knowledge alone. Instead, successful episode completion depends on proactive search, evidence access, and cross-paper multimodal integration.

5.6 Effect of Reasoning Step Budget

We analyze how the step budget affects episodic performance by varying the maximum number of tool calls $S_{\max} \in \{5, 10, 15\}$ while keeping all other settings fixed. Figure 7 reports episode success rate (ESR) together with the max-steps error rate, which captures failures caused by exceeding the step limit.

Figure 7a shows a consistent improvement in ESR when increasing the budget from 5 to 10 across the evaluated models, but little to no additional gain when further increasing to 15, indicating a saturation regime. Figure 7b explains part of this behavior: max-steps errors drop sharply from 5 to 10, suggesting that



(a) Episode success rate (ESR). **(b)** Max-steps error rate.

Fig. 7: Effect of step budget $S_{\max}$ on episodic performance. **(a)** ESR increases from 5 to 10 steps but saturates from 10 to 15. **(b)** Max-steps error rate drops sharply from 5 to 10, indicating fewer budget-induced terminations, while additional budget yields diminishing returns.

a tighter budget often truncates episodes before agents complete evidence acquisition and recovery steps. In contrast, the decrease from 10 to 15 is small while ESR remains nearly unchanged, implying that remaining failures are not primarily due to insufficient steps. Instead, beyond a moderate budget, performance is limited by systematic issues in evidence grounding, cross-evidence alignment, and memory-based evidence reuse that additional tool calls do not resolve.

6 Conclusion

We introduced EpiBench, an episodic multi-turn multimodal benchmark that evaluates research agents in short, process-level workflows requiring proactive cross-paper navigation, visual evidence extraction, and memory-based reuse for objective multi-evidence questions. By emphasizing cross-paper multi-figure and multi-table integration under workflow constraints, EpiBench provides fine-grained evaluation and diagnostics beyond single-turn accuracy. Experiments on leading open-source and closed-source agents show that current systems remain far from reliable research assistance, with the best model reaching only 29.23% on the hard split, highlighting persistent bottlenecks in evidence grounding, multi-evidence fusion, and robust reuse of previously acquired evidence. We hope EpiBench will serve as a practical platform for developing more verifiable, efficient, and reproducible research agents.

Acknowledgements

We gratefully acknowledge the support of the National Natural Science Foundation of China (NSFC) via grant 62236004 and 62476073.

References

1. arxiv. https://arxiv.org/, accessed: 2026-02-27
2. Openreview. https://openreview.net/, accessed: 2026-02-27
3. Bai, S., Cai, Y., Chen, R., et al.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
4. Chen, Z., Geng, X., Wang, X., Jiang, Y., Zhang, Z., Xie, P., Tu, K.: Efficient multimodal planning agent for visual question-answering (2026), https://arxiv.org/abs/2601.20676
5. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Helmholz, W.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), https://arxiv.org/abs/2507.06261
6. DuckDuckGo: Duckduckgo search engine. https://duckduckgo.com/ (2026), accessed: 2026-03-03
7. Google for Developers: Custom search json api: Use rest to invoke the api. https://developers.google.com/custom-search/v1/using_rest, accessed 2026-03-02
8. Huang, P., Zhong, Z., Wan, Z., Zhou, D., Alam, S., Wang, X., Li, Z., Dou, Z., Zhu, L., Xiong, J., Tao, C., Xu, Y., Dimitriadis, D., Zhang, T., Zhang, M.: Mmdeepresearch-bench: A benchmark for multimodal deep research agents (2026), https://arxiv.org/abs/2601.12346
9. Jiang, H., Zhang, X., Garg, S., Arora, R., Kuo, S.Z., Xu, J., Bansal, A., Brossman, C., Liu, Y., Colak, A., Aly, A., Kumar, A., Dong, X.L.: Memory-qa: Answering recall questions based on multimodal memories (2025), https://arxiv.org/abs/2509.18436
10. Lee, Y.J., Lee, B.K., Zhang, J., Hwang, Y., Ko, B., Kim, H.G., Yao, D., Rong, X., Joo, E., Han, S.H., Ko, B., Choi, H.J.: Multiverse: A multi-turn conversation benchmark for evaluating large vision and language models (2025), https://arxiv.org/abs/2510.16641
11. Li, S., Tajbakhsh, N.: Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs (2023)
12. Liu, Z., Li, J., Zhuang, Y., Liu, Q., Shen, S., Ouyang, J., Cheng, M., Wang, S.: amelo: A stable framework for arena-based llm evaluation (2025), https://arxiv.org/abs/2505.03475
13. Liu, Z., Chu, T., Zang, Y., Wei, X., Dong, X., Zhang, P., Liang, Z., Xiong, Y., Qiao, Y., Lin, D., Wang, J.: Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms (2024), https://arxiv.org/abs/2406.11833
14. Lyu, W., Du, Y., Zhao, J., Zhen, X., Shao, L.: Vischainbench: A benchmark for multi-turn, multi-image visual reasoning beyond language priors (2025), https://arxiv.org/abs/2512.06759
15. Mikriukov, A., Senokosov, A., Succi, G., Tormasov, A., Plaksin, Y., Trofimova, E., Sitnikov, V.: Ai tools for automating systematic literature reviews. In: Proceedings of the 2025 International Conference on Software Engineering and Computer Applications. p. 25–30. SECA '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3747912.3747962, https://doi.org/10.1145/3747912.3747962
16. Movva, P., Marupaka, N.H.: Enhancing scientific visual question answering through multimodal reasoning and ensemble modeling (2025), https://arxiv.org/abs/2507.06183

17. OpenAI: Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/ (2025), accessed: 2026-03-05
18. Pramanick, S., Chellappa, R., Venugopalan, S.: Spiqa: A dataset for multimodal question answering on scientific papers (2025), https://arxiv.org/abs/2407.09413
19. Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., Sun, M.: Toolllm: Facilitating large language models to master 16000+ real-world apis (2023), https://arxiv.org/abs/2307.16789
20. Ren, Y., Wang, J., Meng, Y., Shi, Y., Lin, Z., Chu, R., Xu, Y., Li, Z., Zhao, Y., Wang, Z., Qiao, Y., Tang, R., Liu, M., Yang, Y.: Sin-bench: Tracing native evidence chains in long-context multimodal scientific interleaved literature (2026), https://arxiv.org/abs/2601.10108
21. Roucher, A., del Moral, A.V., Wolf, T., von Werra, L., Kaunismäki, E.: 'smolagents': a smol library to build great agentic systems. https://github.com/huggingface/smolagents (2025)
22. Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach themselves to use tools. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 68539–68551. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/d842425e4bf79ba039352da0f658a906-Paper-Conference.pdf
23. Shabtay, N., Polo, F.M., Doveh, S., Lin, W., Mirza, M.J., Chosen, L., Yurochkin, M., Sun, Y., Arbelle, A., Karlinsky, L., Giryes, R.: Livexiv – a multi-modal live benchmark based on arxiv papers content (2025), https://arxiv.org/abs/2410.10783
24. Singh, A., Fry, A., Perelman, A., et al.: Openai gpt-5 system card (2025), https://arxiv.org/abs/2601.03267
25. Smolyansky, E.: Announcing connected papers — a visual tool for researchers to find and explore academic papers. https://medium.com/connectedpapers/announcing-connected-papers-a-visual-tool-for-researchers-to-find-and-explore-academic-papers-89146a54c7d4 (Jun 2020), accessed: 2026-02-27
26. Tao, X., Teng, Y., Su, X., Fu, X., Wu, J., Tao, C., Liu, Z., Bai, H., Liu, R., Kong, L.: Mmsearch-plus: Benchmarking provenance-aware search for multimodal browsing agents (2025), https://arxiv.org/abs/2508.21475
27. Team, K., Bai, T., Bai, Y., et al.: Kimi k2.5: Visual agentic intelligence (2026), https://arxiv.org/abs/2602.02276
28. Team, Q.: Qwen3.5: Accelerating productivity with native multimodal agents (February 2026), https://qwen.ai/blog?id=qwen3.5
29. Team, V., Hong, W., Yu, W., et al.: Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning (2026), https://arxiv.org/abs/2507.01006
30. Vishesh, O., Khadilkar, H., Akkil, D.: Aegis: An agent for extraction and geographic identification in scholarly proceedings (2025), https://arxiv.org/abs/2509.09470
31. Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., Zhang, B., Wei, L., Sui, Z., Li, W., Shi, B., Qiao, Y., Lin, D., He, C.: Mineru: An open-source solution for precise document content extraction (2024), https://arxiv.org/abs/2409.18839

32. Wang, D., Cheng, M., Yu, S., Liu, Z., Guo, Z., Li, X., Liu, Q.: Paperarena: An evaluation benchmark for tool-augmented agentic reasoning on scientific literature (2026), https://arxiv.org/abs/2510.10909
33. Wang, Z., Xia, M., He, L., Chen, H., Liu, Y., Zhu, R., Liang, K., Wu, X., Liu, H., Malladi, S., Chevalier, A., Arora, S., Chen, D.: Charxiv: Charting gaps in realistic chart understanding in multimodal llms (2024), https://arxiv.org/abs/2406.18521
34. xAI: Grok 4.1. https://x.ai/news/grok-4-1 (Nov 2025), accessed: 2026-03-05
35. Yan, D., Li, Y., Chen, Q.G., Luo, W., Wang, P., Zhang, H., Shen, C.: Mmcr: Advancing visual language model in multimodal multi-turn contextual reasoning (2025), https://arxiv.org/abs/2503.18533
36. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=WE_vluYUL-X
37. Zeng, Y., Huang, W., Fang, Z., Chen, S., Shen, Y., Cai, Y., Wang, X., Yin, Z., Chen, L., Chen, Z., Huang, S., Zhao, Y., Tang, X., Hu, Y., Torr, P., Ouyang, W., Cao, S.: Vision-deepresearch benchmark: Rethinking visual and textual search for multimodal large language models (2026), https://arxiv.org/abs/2602.02185