MEDXS: ARE MEDICAL MULTIMODAL AGENTS REASONING ACROSS SOURCES, OR MERELY LANDING ON THE RIGHT ANSWER?

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate clinical decisions rarely rest on a single image. They interleave visual inspection, retrieval over case specific structured records, and corroboration against external medical knowledge. First generation medical agent benchmarks judge outcome alone, and a second adds process aware scoring of tool calls, yet both share one blind spot. They confine tools to a single evidence modality, never treating the structured record as co equal with imaging, and their process metrics check whether calls are well formed rather than whether cross source evidence is used, so a trajectory grounded on one source still scores fully. We introduce MedXS, a process verified benchmark that instantiates imaging, structured record, and medical knowledge as three co equal tool families over $2{,}000$ clinician annotated tasks, with eight question types stratified into three levels by the number of sources fused and grounded on $6{,}000+$ evidence checkpoints, scored by an unweighted, tier increasing suite of six metrics spanning outcome, process, and cross source integration. Across frontier agents, the strongest reaches only 68.2% average result yet a 97.4% workflow score, while its Level 3 coverage collapses to 28.8% and knowledge tools drop to 2.8% of all calls. This exposes *tool tunnel vision*, a persistent gap between correct answers, valid tool ordering, and real cross source integration that stays invisible to both outcome and call level scoring.

1 INTRODUCTION

Medical diagnosis is intrinsically a multi source reasoning process. A radiologist facing a suspicious lesion does not assign a label from a single glance, but magnifies the region of interest, cross references prior studies and laboratory trends, and consults established guidelines before issuing a report (Goh et al., 2024; Tu et al., 2024). This progressive, evidence gated behavior is what *agentic* multimodal models set out to emulate, and a recent wave of systems equips large multimodal models with external tools to inspect images, query records, and retrieve knowledge on demand (Li et al., 2023; Tang et al., 2024; Fallahpour et al., 2025).

Benchmarking has evolved through two generations. The first scores an agent from its *outcome* alone, comparing the final diagnosis against a reference label while discarding the trajectory (Jiang et al., 2025; Schmidgall et al., 2024; Fast et al., 2024). The second adds *process aware* scoring: MedCTA verifies tool selection, parameter legality, and trajectory fidelity (Ashraf et al., 2026); DeepTumorVQA supplies ground truth trajectories over a staged evidence hierarchy (Chen et al., 2026); and MedChain frames workflows as sequential decision making (Liu et al., 2026). Process is thus no longer invisible, and as Figure 1 shows, the decisive difference lies not in whether a tool was called, but whether the returned evidence contributed to the conclusion.

We argue that even this second generation still leaves two structural gaps unfilled. *First*, its tool interactions are usually confined to a *single* evidence modality: MedCTA and DeepTumorVQA couple imaging with its own derived measurements and a single knowledge query, yet never expose case specific structured evidence as a family co equal with imaging and knowledge, while record centric benchmarks conversely never expose imaging (Jiang et al., 2025; Lee et al., 2025). The three families are thus rarely available at once, and few benchmarks test whether an agent picks the right *family* when all three compete. *Second*, process metrics check whether each call is well formed and

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.

Table 1: **Comparison with representative medical benchmarks. Eval. Target** indicates what each benchmark judges, namely a single response, a reasoning chain, or a full trajectory; **Capability Axes** indicate whether it verifies the *process* rather than the answer, along four dimensions: **SL** step level checkpoints, **EU** evidence use, **MS** multi source synergy, and **TF** three co equal tool families ($\mathcal{F}_{\text{img}}, \mathcal{F}_{\text{rec}}, \mathcal{F}_{\text{know}}$). MedXS is the only benchmark supporting all four; its #Samples column reports task count (2,000 tasks, 6,073 checkpoints) to remain comparable with prior work. ✓/◐/✗: full / partial / no support.

Benchmark	Eval. Target	#Dom.	#Samples	Capability Axes				Annotation	
				SL	EU	MS	TF	Annotator	Verdict
Medical reward, judge, and rubric benchmarks.									
Med-RewardBench (Ding et al., 2025)	Model response	13	1,026	✗	✗	✗	✗	Physician	6-dim score
HealthBench (Arora et al., 2025)	Model response	7	5,000	◐	✗	✗	✗	Physician	Rubric
MedHELM (Bedi et al., 2025)	Model response	5	121	✗	✗	✗	✗	Clinic+LLM	Score
K-QA (Manes et al., 2024)	Model response	1	1,212	✗	◐	✗	✗	Physician	Comp/Hall
Medical reasoning, agent, and trajectory benchmarks.									
MedXpertQA (Zuo et al., 2025)	Reasoning steps	17	4,460	◐	✗	✗	✗	Expert	MC / CoT
MedReason (Wu et al., 2025b)	Reasoning steps	7	32,682	◐	◐	✗	✗	KG+Model	Step-valid
MedHallu (Pandit et al., 2025)	Model response	–	10,000	✗	◐	✗	✗	Synthetic	Halluc.
MedAgentBench (Jiang et al., 2025)	Agent trajectory	10	300	◐	✗	✗	◐	Physician	Success
FHIR-AgentBench (Lee et al., 2025)	Agent trajectory	–	2,931	◐	◐	✗	◐	Clinician	Correct
AgentClinic (Schmidgall et al., 2024)	Agent trajectory	9	534	◐	✗	✗	◐	Mixed	Dx-acc
MedChain (Liu et al., 2026)	Agent trajectory	19	12,163	◐	◐	◐	◐	Clinician	Stage-wise
ChestAgentBench (Fallahpour et al., 2025)	Agent trajectory	7	2,500	◐	◐	✗	✗	Clinician	6-choice
MedXS (Ours)	**Medical agent traj.**	**8**	**2,000**	✓	✓	✓	✓	**Model+Clinician**	**6-metric**

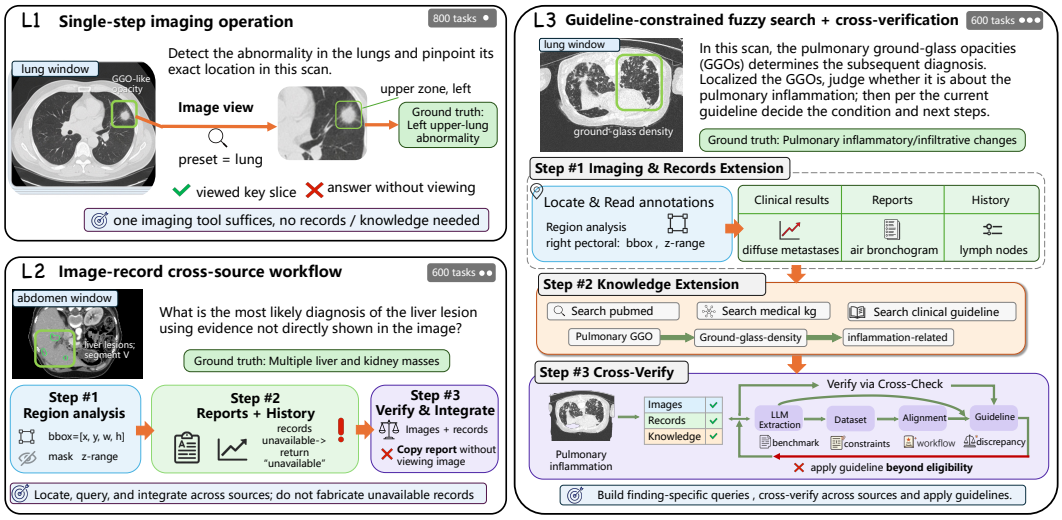


Figure 1: **Tiered task design of MedXS.** Three cases of increasing complexity span the three tiers: Level 1, a single step imaging operation (localizing a pulmonary lesion); Level 2, an imaging plus record workflow (diagnosing abdominal nodules); Level 3, cross source verification of pulmonary opacities under a guideline. Each tier shows the image, question, ground truth, and tool calls, exposing the progression from single source, to imaging plus record, to three source fusion.

ordered, but not whether cross source evidence is *used*: an agent producing a plausible label from imaging alone earns the same trajectory score as one reconciling both. Table 1 makes this clear: process verification is now shared, but a co equal record family and an explicit multi source score remain unoccupied.

To fill both gaps, we introduce MedXS, a cross source process verified benchmark for medical multimodal agents built on 2,000 clinician annotated tasks from public datasets, normalized into a unified case and instantiated into eight clinical question types. The benchmark exposes three tool families through a common interface: the imaging family renders a study and interrogates spatial regions; the record family exposes source native or derived fields such as reports, annotations, metadata, and measurements; and the knowledge family covers knowledge graph, guideline, and literature retrieval. Because all three compete on every multi source task, the benchmark measures

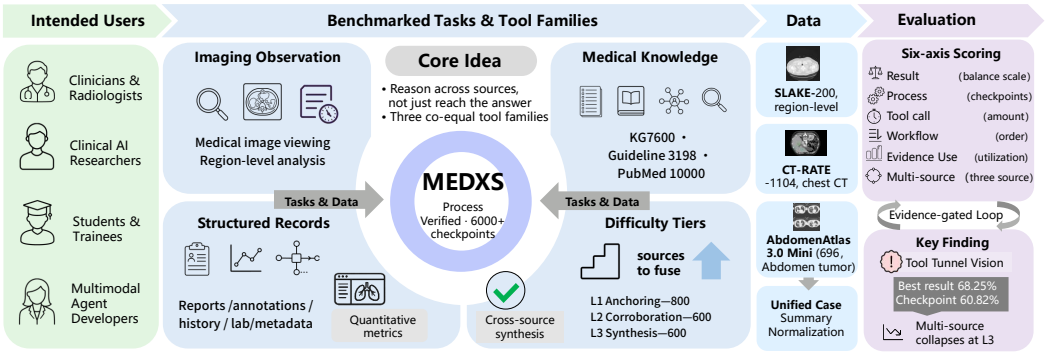


Figure 2: **Overview of MedXS.** The nine tools of the imaging $\mathcal{F}_{\text{img}}$, record $\mathcal{F}_{\text{rec}}$, and knowledge $\mathcal{F}_{\text{know}}$ families are exposed over L1/L2/L3 tiered tasks built from a unified Case normalized from three original datasets, and scored by six evaluation axes (result, checkpoint, successful tool calls, evidence use, workflow, multi source coverage) that expose *tool tunnel vision*.

tool *family* selection rather than isolated manipulation. Tasks are stratified into three tiers by the sources the answer requires: Level 1 is recoverable from imaging alone; Level 2 reconciles at least two families; and Level 3 fuses imaging, record, and knowledge.

The defining feature of MedXS is that it evaluates whether the required evidence families are actually retrieved and reflected in the conclusion, not merely whether calls are well formed. For each task we annotate a reference evidence graph $\mathcal{R}$ and derive $6{,}000+$ stepwise checkpoints from it, on which we report an *unweighted*, tier increasing suite of six metrics (result r, checkpoint c, successful tool calls n, evidence use v, workflow f, multi source coverage m) rather than a single weighted sum, making disagreements between answer correctness, process compliance, and cross source evidence use directly observable.

Our contributions are as follows.

- ◇ We identify two blind spots left open by process aware medical benchmarks: jointly instantiating imaging, records, and knowledge as competing families, and testing whether a model uses the families annotated as necessary rather than merely calling tools correctly.
- ◇ We build MedXS, a benchmark of $2{,}000$ clinician annotated tasks with nine tools across three co equal families, eight question types, three tiers, and $6{,}000+$ checkpoints.
- ◇ We propose the multi source coverage metric m, which scores integration across families rather than answer agreement or trajectory well formedness, embedded in an unweighted, tier increasing suite.
- ◇ We evaluate frontier models as tool-using agents and find that even the strongest averages only 68.2% result at a 97.4% workflow score, with Level 3 coverage collapsing to 28.8%, revealing cross source integration as a distinct, unsolved axis we term *tool tunnel vision*.

2 The MedXS Benchmark

2.1 Problem Formalization

We formalize medical multimodal agent evaluation as an *evidence graph referenced, tier increasing process verification* problem rather than weighted scalar maximization. Given a case $\mathcal{C}$, a task is a quadruple $\mathcal{T} = \langle Q,\ \mathcal{C},\ \mathcal{F}^*,\ \mathcal{R}\rangle$, where Q is a question of one of eight clinical types, $\mathcal{F}^* \subseteq \{\mathcal{F}_{\text{img}}, \mathcal{F}_{\text{rec}}, \mathcal{F}_{\text{know}}\}$ is the set of tool families required to answer, and $\mathcal{R}$ is an annotated reference evidence dependency graph. Under a nine tool interface the agent produces a trajectory $\tau = (a_1, \ldots, a_T)$, where each action $a_t = \langle \phi_t, \theta_t, o_t\rangle$ carries the invoked tool ϕ_t, arguments θ_t, and observation o_t, and commits a final answer $\hat{a}$ before exhausting a turn budget B.

Six unweighted metrics. The defining choice of MedXS is *not* to collapse evaluation into a weighted sum $\sum_k w_k \sigma_k$, because any fixed weighting introduces an artificial exchange rate that

lets an agent compensate a failed axis with a strong one. We instead report six metrics, enabled tier by tier. Let $\mathbb{1}[\cdot]$ denote the indicator and $\mathcal{K}$ the checkpoints derived from $\mathcal{R}$; the metrics are

$$r = \mathbb{1}\big[\text{Judge}(\hat{a},\, y^*) = \text{correct}\big] \in \{0,1\}, \tag{1}$$

$$c = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \mathbb{1}\big[k \text{ effectively hit}\big] \in [0,1], \tag{2}$$

$$n = \min\big(n_{\text{succ}}(\tau),\, B\big), \tag{3}$$

$$v = \frac{\big|\mathcal{E}_{\text{used}}(\tau)\big|}{\big|\mathcal{E}_{\text{ret}}(\tau)\big|} \in [0,1], \tag{4}$$

where r is set by a judge comparing $\hat{a}$ against the gold standard y^*; c counts the fraction of reference checkpoints effectively hit; n is a *purely statistical* count of successful tool calls capped by B; and v measures the fraction of returned evidence $\mathcal{E}_{\text{ret}}$ that the conclusion actually invokes ($\mathcal{E}_{\text{used}}$). For Levels 2 and 3 we additionally enable the workflow score

$$f = \mathbb{1}\big[\text{order obeys ask} \to \text{inspect} \to \text{diagnose}\big] \in \{0,1\}, \tag{5}$$

requiring the trajectory to respect the clinical dependency of retrieving before interpreting before diagnosing. For Level 3 we further enable the multi source coverage score

$$m = \prod_{\mathcal{F} \in \{\mathcal{F}_{\text{img}}, \mathcal{F}_{\text{rec}}, \mathcal{F}_{\text{know}}\}} \mathbb{1}[\mathcal{F}\,\text{inv}]\,\mathbb{1}[\mathcal{F}\,\text{used}] \in \{0,1\}, \tag{6}$$

where $\mathbb{1}[\mathcal{F}\,\text{inv}]$ and $\mathbb{1}[\mathcal{F}\,\text{used}]$ indicate that family $\mathcal{F}$ is *invoked* and that its returned evidence is *marked as used* in the conclusion, so $m = 1$ only when all three required families are marked as used. The tiers thus correspond to an increasing metric set:

$$\mathcal{M}_1 = \{r, c, n, v\},\ \mathcal{M}_2 = \mathcal{M}_1 \cup \{f\},\ \mathcal{M}_3 = \mathcal{M}_2 \cup \{m\}. \tag{7}$$

Because no metric beyond r fires on the final label, an agent cannot inflate c, v, f, m by guessing, while invalid actions, irrelevant calls, unused evidence, and imbalanced call counts are diagnosed on c, v, and n. Process and outcome signals thus do not contaminate each other, and *tool tunnel vision*, invisible to r alone, is exposed explicitly by v and m.

3 Benchmark Construction

Building MedXS means manufacturing tasks whose correctness is *forced* by cross source evidence rather than surface actions, and certifying at every step whether the reference evidence is actually reached. Figure 2 sketches the pipeline: heterogeneous public sources are normalized into a unified *Case*, tasks are instantiated and annotated with a reference evidence graph $\mathcal{R}$, and the resulting trajectories are scored under one nine tool interface and one six metric protocol.

3.1 Task Construction

We draw material from three complementary public sources whose modalities dictate the tiers they populate. SLAKE (Liu et al., 2021) contributes 200 single view 2D cases that, lacking volumetric or longitudinal context, seed only Level 1; CT-RATE (Hamamci et al., 2024) (1,104 chest CT) and AbdomenAtlas 3.0 Mini (Bassi et al., 2025) (696 abdominal CT) supply the volumetric structure and records needed for all three tiers (Figure 3), yielding 800 Level 1, 600 Level 2, and 600 Level 3 tasks (2,000 total), with the two CT sources contributing 368 and 232 per tier. Each source is normalized into a unified *Case* holding a raw image handle and *only* the evidence that source provides, every field tagged by provenance as source native (reports, masks, metadata), derived (measurements from masks), or synthetic (history and laboratory proxies), so a task never rewards evidence its origin cannot support. This record family is dense and heterogeneous, linking volume (1,824), image (200), report (1,200), history (1,200), annotation (1,200), and laboratory (464) fields across the six source types of Figure 3.

From each Case we instantiate one or more *Task* objects $\mathcal{T} = \langle Q, \mathcal{C}, \mathcal{F}^*, \mathcal{R} \rangle$ through a two stage *Model+Clinician* procedure: a model drafts candidate questions and evidence chains, and clinicians

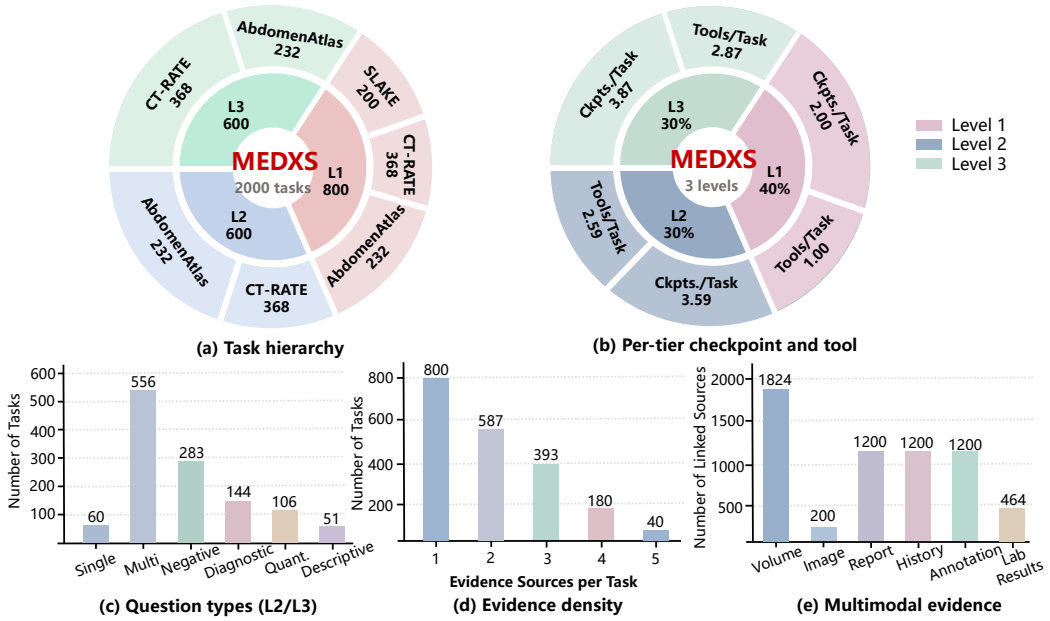


Figure 3: **Data provenance and unified Case construction.** SLAKE, CT-RATE, and ABDOMENATLAS 3.0 MINI are normalized into a unified Case exposing per case structured evidence as source native, derived, or synthetic. Panels report per tier composition, checkpoint and tool budgets, question type mix, evidence density, and the six data source types.

then verify clinical validity and lock the reference graph $\mathcal{R}$, whose nodes record evidence units and whose edges encode their acquisition dependencies. To keep family selection meaningful rather than incidental, all three families are exposed on every higher tier task, and 283 negation or absent finding questions block a default positive shortcut. The eight question types comprise six populating Levels 2 and 3, single source (60), multi source (556), negation (283), diagnostic (144), quantitative (106), and descriptive (51), together with two single step imaging types, spatial localization and existence, that populate Level 1.

3.2 CASE KNOWLEDGE CONSTRUCTION

The knowledge family draws on an external evidence store of three snapshots normalized into one retrievable index with deterministic lexical ranking and a default top 5 cutoff, so retrieval stays reproducible and ranking noise cannot confound the process metrics. An OPENKG medical graph merged with native SLAKE entries yields a KG snapshot of 7,600 deduplicated head relation tail triples (28 OPENKG relations, 2,600 SLAKE entries, 4,972 lexicon expansion records); a guideline snapshot of 3,198 records covers 82 diseases across diagnosis, imaging assessment, TNM staging, and biomarker testing from ESMO, NCCN, and CSCO; and a PUBMED snapshot of 10,000 records, drawn through 47 seed queries and predominantly covering literature from 2023 onward, is deduplicated by PMID, DOI, or title year key. Because retrieval requires non-memorized answers, we audit source identifiers, queries, and overlapping text against gold answers to control leakage, subject to upstream licences.

3.3 TOOL SUITE

The nine tools span the three families a clinician actually consults, and unlike imaging centric benchmarks the record family $\mathcal{F}_{\text{rec}}$ is a first class citizen rather than a byproduct of imaging. The imaging family $\mathcal{F}_{\text{img}}$ renders a study slice, inspects a spatial region of interest, and overlays a segmentation; the record family $\mathcal{F}_{\text{rec}}$ exposes report fields, acquisition and demographic metadata, and quantitative measurements as separately queryable endpoints; and the knowledge family $\mathcal{F}_{\text{know}}$ provides knowledge graph lookup, guideline retrieval, and biomedical literature retrieval. Because all three

are available on every higher tier task, MedXS measures not merely whether an agent uses *a* tool but whether, when several plausible families compete, it selects the *right* family.

3.4 Process and Synergy Verification

Given a trajectory $\tau = (a_1, \ldots, a_T)$, the process side metrics aggregate a stepwise judgment $s_t \in \{0, 1\}$ over three criteria: tool appropriateness under the evidence available before step t, evidence gating (no step consumes evidence not yet retrieved), and consistency of the action order with $\mathcal{R}$; these are carried by the checkpoint score c and the workflow score f. The multi source coverage score m departs from prior trajectory fidelity work: it inspects the evidence attribution of the final conclusion and fires only when all three families are simultaneously invoked *and* marked as contributing, while v penalizes evidence that is returned yet never used. Because c, v, f, m never target the final label, an agent cannot inflate them by guessing, and process and outcome scores are kept from contaminating each other.

4 Experiments

We evaluate a representative set of open and closed multimodal agents under one unified tool interface, organizing the study around three questions: *i)* how far the strongest agent remains from reliable cross source reasoning across the result, checkpoint, workflow, and coverage granularities; *ii)* how difficulty varies across tiers and source domains; and *iii)* how errors decompose across the three tool families, in particular whether agents degenerate into *tool tunnel vision*.

4.1 Experimental Setup

Agents under test. We evaluate nine general purpose, API served multimodal agents (Seed-1.8-Pro, doubao-seed-2.0-pro, GPT-5.5, GPT-5.6-Luna, Claude-4.8-Opus, Grok-4.5, Qwen3.6-Plus, Deepseek-v4-Pro, and Gemini3.1-Pro) and three specialized medical agents (Lingshu-32B, MedGPT-oss-20B, and HuatuoGPT-Vision-34B). All use the same zero shot ReAct style system prompt (Yao et al., 2022), which requires strict JSON tool calls, restricts actions to the available tools, mandates image inspection when an imaging tool is available, and requires a final Conclusion/Evidence response (full prompt in the supplement). The shared namespace contains nine tools but each task exposes only its applicable subset, and every model is evaluated once over the full 2,000 task suite under a per task budget $B = 50$, the maximum tool call rounds before an answer must be committed.

Evaluation protocol. For each task the agent produces a trajectory τ and a final answer $\hat{a}$, and scoring is deliberately hybrid so that outcome and process signals stay independent. Result correctness r is assigned by a fixed Seed judge at temperature 0, comparing $\hat{a}$ against the gold standard y^* by medical meaning rather than exact wording; because the top ranked system shares a family with the judge, we cross validate a random subset with an independent judge to rule out self preference. Evidence use v is produced by the same judge from the answer's evidence text and the recorded tool results; crucially, this judge never sees $\mathcal{R}$. Checkpoint c, successful tool calls n, and workflow f are rule based against $\mathcal{R}$, and multi source coverage m is a deterministic conjunction of source family indicators derived after the evidence judgment. Following Eq. 7, Level 1 reports $\{r, c, n, v\}$, Levels 2 and 3 add f, and Level 3 additionally adds m. Requests failing on transient provider, rate limit, or image encoding errors are rerun under the same configuration, and unresolved failures are reported separately with their error types.

4.2 Evaluation Metrics

A single aggregate would mask the very failure we study, since an agent can look strong on one axis yet fail to integrate evidence when sources compete. We therefore report the *unweighted*, tier increasing ladder of Eq. 7, interpreted as follows. ***i) Result*** $r \in \{0, 1\}$ marks whether the judge deems the answer medically correct; as the loosest metric it can overstate ability, since an answer plausible from imaging alone still earns credit. ***ii) Checkpoint*** $c \in [0, 1]$ is the fraction of reference checkpoints hit (a tool checkpoint on a matching `tool_result`, an answer checkpoint on a case insensitive match; $c = 1$ when none are defined). ***iii) Successful tool calls*** $n = \min(n_{\text{succ}}, B)$

Table 2: Main results on MedXS (2025–2026 frontier agents), grouped into general purpose and specialized systems, same family models ordered older to newer. Result r, Checkpoint c, Successful Tool Calls n report per tier (L1/L2/L3) and average scores; Workflow f on L2–3 and Multi-Source m on L3. Values are percentages except n (mean successful tool calls, a diagnostic count with no direction). The clinician row is a reference under the same judge. Best and second per column are **bold** and underlined.

Model Name	Result $r\uparrow$				Checkpoint $c\uparrow$				Successful Tool Calls n				$f\uparrow$	$m\uparrow$
	L1	L2	L3	Avg	L1	L2	L3	Avg	L1	L2	L3	Avg	L2–3	L3
General-purpose														
Seed-1.8-Pro	31.3	80.5	90.8	63.9	57.8	58.9	61.1	59.2	1.10	3.31	5.03	2.94	95.2	**41.5**
doubao-seed-2.0-pro	**38.8**	82.7	93.2	**68.2**	59.4	**60.9**	**62.7**	**60.8**	1.24	4.49	5.99	3.64	**97.4**	28.8
GPT-5.5	36.1	78.6	**95.8**	66.8	58.3	59.7	59.4	59.0	1.95	3.28	3.38	2.78	32.9	1.83
GPT-5.6-Luna	38.5	28.0	31.0	33.1	56.6	35.8	34.3	43.7	1.17	2.36	2.16	1.83	31.3	0.17
Claude-4.8-Opus	35.8	79.8	91.7	65.7	58.8	54.3	60.6	57.9	1.53	3.22	4.10	2.80	32.9	2.73
Grok-4.5	32.3	73.8	80.5	59.2	54.2	59.8	58.0	57.0	1.25	3.54	3.87	2.72	61.4	15.7
Qwen3.6-Plus	33.0	71.5	76.3	57.5	**59.5**	57.3	53.4	57.0	1.18	3.44	3.65	2.60	86.2	0.33
Deepseek-v4-Pro	6.5	**84.8**	**95.8**	56.8	54.3	60.0	62.2	58.4	2.22	4.65	5.63	3.97	84.8	23.0
Gemini3.1-Pro	31.5	52.3	60.8	46.5	**59.5**	47.5	48.7	52.7	1.84	3.70	3.75	2.97	44.0	0.50
Specialized														
Lingshu-32B	32.1	24.1	19.8	26.0	56.1	39.3	35.5	44.9	1.68	3.86	3.54	2.89	61.5	29.3
MedGPT-oss-20B	30.2	25.8	16.7	24.8	54.8	41.0	37.2	45.3	1.70	3.85	3.76	2.96	62.8	34.2
HuatuoGPT-Vision-34B	28.4	13.8	11.4	18.9	49.1	34.4	31.1	39.3	16.3	37.5	34.4	28.1	29.3	9.33
Direct answer (no tool)	9.8	5.1	2.0	5.6	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Clinician reference	92.4	90.1	87.3	89.9	95.2	93.6	92.1	93.6	3.46	5.13	6.82	4.96	95.1	88.7

counts `tool_result` events and exposes imbalanced budgets. ***iv) Evidence use*** $v \in [0,1]$ is the fraction of usable returned evidence reflected in the conclusion, the first probe of ***tool tunnel vision*** since retrieved but discarded evidence lowers it. ***v) Workflow*** $f \in \{0,1\}$ (Levels 2–3) checks the order image $\rightarrow$ inspect $\rightarrow$ diagnose, and ***vi) multi source coverage*** $m \in \{0,1\}$ (Level 3) fires only when $\mathcal{F}_{\text{img}}, \mathcal{F}_{\text{rec}}, \mathcal{F}_{\text{know}}$ are all marked as used. Because the agent never sees $\mathcal{R}$ and observes only the returned observations o_t, checker and judge score process and outcome independently, and the tier enabled set $\mathcal{M}_\ell$ aggregates from the bottom up, keeping scoring explicit and auditable.

4.3 Benchmark Statistics

The benchmark comprises 2,000 tasks with visual inputs, each normalized from three complementary public sources spanning six data source types and eight clinical question types, and stratified into 800 Level 1, 600 Level 2, and 600 Level 3 tasks according to their cross source reasoning demand (Figure 3). It specifies 6,073 reference checkpoints in total, with the mean checkpoints per task rising steadily from 2.00 at Level 1 through 3.59 at Level 2 to 3.87 at Level 3, and the mean reference tool calls rising in parallel from 1.00 through 2.59 to 2.87 (amounting to more than 4,000 reference tool calls overall). Because tier difficulty is set by evidence depth rather than prompt length, the tasks share a compact mean prompt length of only 47.61 words while drawing on 3.04 checkpoints on average, and 1,200 tasks (60.0%) require evidence from at least two distinct families, with evidence density spanning one to five linked sources.

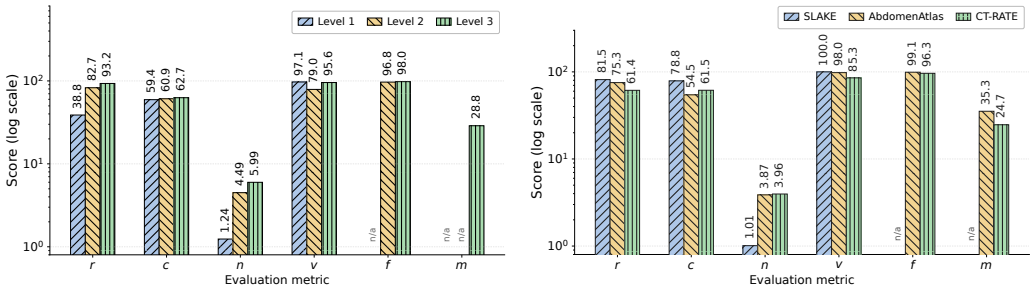
4.4 Overall Results

Table 2 reports the six metrics for frontier agents from 2025–2026, grouped into general purpose and specialized medical systems. A systematic gap separates the loose and tight ends of the ladder: the strongest reaches $\mathbf{62.7}\%$ on the Level 3 checkpoint yet an average result of only $\mathbf{68.2}\%$, a separation from "able to call" to "able to integrate" invisible to outcome only or call level scoring. The no tool *Direct answer* baseline scores only 2.0% at Level 3, confirming these tasks cannot be solved from memorization and require genuine retrieval.

Finding 1: accuracy does not track reasoning complexity. The strongest system averages only $\mathbf{68.2}\%$ on result, yet *rises* from 38.8% at Level 1 to 82.7% at Level 2 and 93.2% at Level 3. Because r is graded by medical meaning, its scale is not directly comparable across tiers, so the rise reflects looser answer matching at higher tiers rather than easier reasoning; outcome and process must therefore be reported separately.

Table 3: Tool family invocation and use. $\mathcal{F}_{\text{img}}$, $\mathcal{F}_{\text{rec}}$, $\mathcal{F}_{\text{know}}$ denote the imaging, record, and knowledge families. Inv. and Use are percentages over tasks requiring that family; All requires every required family used.

Model Name	Imaging		Record		Knowledge		All required	
	Inv.	Use	Inv.	Use	Inv.	Use	All-Inv.	All-Use
General-purpose								
Seed-1.8-Pro	100.0	99.2	99.3	99.3	34.8	34.8	92.9	92.4
doubao-seed-2.0-pro	100.0	94.8	97.7	92.8	19.1	19.1	90.2	86.5
GPT-5.5	97.6	95.7	99.8	97.3	0.95	0.95	87.6	85.8
GPT-5.6-Luna	98.3	88.8	47.3	40.3	0.00	0.00	62.9	59.0
Claude-4.8-Opus	96.5	95.1	98.0	96.9	1.95	1.95	85.5	84.5
Grok-4.5	97.1	95.7	97.8	96.6	11.2	11.2	85.9	84.9
Qwen3.6-Plus	99.9	99.2	94.8	94.5	0.00	0.00	86.8	86.5
Deepseek-v4-Pro	100.0	98.6	99.0	98.6	24.3	24.3	91.6	90.7
Gemini3.1-Pro	93.3	92.0	99.0	98.4	0.48	0.48	83.1	82.2
Specialized								
Lingshu-32B	98.8	86.5	54.3	47.4	52.9	49.5	70.3	64.3
MedGPT-oss-20B	80.8	70.9	58.0	51.7	43.0	40.3	70.8	65.5
HuatuoGPT-Vision-34B	96.9	87.2	98.5	96.5	14.7	14.7	88.3	86.5



(a) (b)

Figure 4: **Per-tier and per-source-domain breakdown of the six metrics. (a)** Results across Levels 1–3 on a log scale (percentages except n). Result r increases substantially ($38.8 \rightarrow 93.2$), whereas checkpoint c changes only slightly ($59.4 \rightarrow 62.7$). Coverage m is available only at Level 3 (28.8), while f applies only to Levels 2–3. **(b)** Results across SLAKE, AbdomenAtlas, and CT-RATE. SLAKE leads in r (81.5), c (78.8), and v (100.0); however, f and m are inapplicable to its single-source design. Coverage m remains low for AbdomenAtlas (35.3) and CT-RATE (24.7). Inapplicable results are denoted by n/a.

Finding 2: a high process score does not imply a high task score. The same system leading on average checkpoint ($\mathbf{60.8}\%$) and workflow ($\mathbf{97.4}\%$) still fails most tasks: agents learn to *call* tools competently yet cannot *integrate* their outputs into a correct final conclusion.

4.5 Per Tier and Per Source Analysis

Finding 3: agents exhibit tool tunnel vision. Usage is sharply imbalanced: imaging $\mathcal{F}_{\text{img}}$ accounts for $\mathbf{57.0}\%$ of calls and record $\mathcal{F}_{\text{rec}}$ for 40.2%, whereas knowledge $\mathcal{F}_{\text{know}}$ reaches only $\mathbf{2.8}\%$, falling below 1% for several systems even on tasks that require it (Table 3). This inflates r and c: at Level 3, coverage m ranges from 0.17% to 41.5%, and high workflow agents collapse (Qwen3.6-Plus at 86.2% workflow but $\mathbf{0.33}\%$ coverage). Following a plausible action order thus does not imply the required families reach the conclusion, as Table 4 quantifies through $r{>}m$ and $f{>}m$ disagreements.

Invocation versus use. Table 3 separates a family's invocation from its actual use in the submitted conclusion. All percentages use the tasks requiring that family as their denominator. A successful

return is necessary but does not by itself establish use. The knowledge family shows the sharpest deficit: several systems rarely invoke it even when the task requires that source. Imaging and record tools can be invoked reliably while the knowledge indicator remains zero. Qwen3.6 Plus, for example, reaches 99.9% imaging invocation and 94.8% record invocation with no knowledge invocations in the reported subset. The all required columns apply each task's required set, which varies with the tier; they should not be confused with coverage restricted to Level 3.

Figure 4 traces the per tier and per source behavior of the strongest system, and two regularities stand out. First, result climbs from 38.8% to 93.2% across tiers while checkpoint barely moves ($59.4\% \rightarrow 62.7\%$) and n grows steadily ($1.24 \rightarrow 5.99$); outcome accuracy alone therefore does not validate tier difficulty and must be read with the process measures. Second, the chest 3D CT domain CT-RATE is hardest ($r = 61.4\%$, $m = 24.7\%$), demanding cross evidence verification, whereas SLAKE is easiest ($r = 81.5\%$, $v = 100.0\%$), consistent with its single source design.

4.6 Process Metric Validation

Three board certified physicians independently re scored 300 stratified trajectories, with 100 per tier, blind to the automatic scores. They assessed whether conclusions used returned evidence, whether every required family contributed, whether tool calls respected prerequisites, and whether the final answer was correct. Agreement with the majority clinician verdict reaches $\kappa \geq 0.81$ for all four judged metrics, including $\kappa = 0.86$ for coverage. Residual evidence use disagreements concentrate on paraphrased measurements, which the automatic matcher occasionally misses. This validation supports interpreting the outcome versus coverage gap as a behavioral distinction. The full protocol and agreement table appear in the supplementary material.

Independent judge check. A different model family re grades 400 randomly sampled tasks to check the shared lineage of the leading agent and the primary judge. The judges agree on 95.2% of items with $\kappa = 0.90$; the top system's result changes from 68.2 to 67.6. This small change suggests that judge lineage is not the dominant source of the result versus coverage gap.

Oracle retrieval control. The supplementary oracle experiment returns the gold knowledge evidence through the standard knowledge tool interface on Level 3 tasks. The return counts as one successful invocation, but use is credited only if the agent incorporates the evidence into its submitted Evidence field. Coverage increases by 17.3 percentage points while evidence use changes from 95.6% to 96.1%. Retrieval therefore explains part of the deficit, yet more than half of the trajectories still fail to invoke and use all three families with the relevant knowledge returned directly. Improving retrieval alone does not ensure that the final conclusion integrates every required source. Together, these controls indicate that the observed result–coverage gap cannot be explained solely by judge lineage or retrieval availability, but persists at the stage of cross source evidence integration.

5 Conclusion

We introduced MedXS, a process verified benchmark that evaluates medical multimodal agents not only by whether they reach the correct final answer or call tools in the expected order, but also by whether they actually use the tool families annotated as necessary for each individual task. This task-level requirement turns evidence use from an implicit assumption into an observable and auditable evaluation criterion. By jointly exposing imaging, case specific structured evidence, and medical knowledge through three co equal tool families, and by reporting six unweighted outcome and process measurements against a reference evidence graph, MedXS reveals substantial disagreements between answer correctness, trajectory compliance, and genuine multi source coverage. The resulting metric profile distinguishes procedural compliance from the successful integration of all evidence required by the clinical question. Our results identify a pronounced tool family imbalance, which we consistently term *tool tunnel vision*, while further source ablation and external clinical validation are still needed to determine how directly this pattern reflects semantic cross source reasoning in real world clinical settings.

AI USE STATEMENT

The evaluation uses generative language models as agents and evaluators as described in the methodology and supplementary material. An AI assistant was also used to adapt the LaTeX source to the ICLR 2027 template and check compilation.

Author completion required before submission: disclose any additional AI assistance in conceptual development, hypotheses, experimental design, implementation, data preparation, translation, and interpretation of results; describe the human verification performed and confirm author responsibility for the final manuscript and artifacts.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Tajamul Ashraf, Hyewon Jeong, Fida Mohammad Thoker, and Bernard Ghanem. Medcta: A benchmark for clinical tool agents. *arXiv preprint arXiv:2606.11702*, 2026.

Pedro RAS Bassi, Mehmet Can Yavuz, Ibrahim Ethem Hamamci, Sezgin Er, Xiaoxi Chen, Wenxuan Li, Bjoern Menze, Sergio Decherchi, Andrea Cavalli, Kang Wang, et al. Radgpt: Constructing 3d image-text tumor datasets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23720–23730, 2025.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*, 2025.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

Yixiong Chen, Wenjie Xiao, Pedro RAS Bassi, Xinze Zhou, Sezgin Er, Ibrahim Ethem Hamamci, Zongwei Zhou, and Alan Yuille. Are vision language models ready for clinical diagnosis? a 3d medical benchmark for tumor-centric visual question answering. *arXiv preprint arXiv:2505.18915*, 2025.

Yixiong Chen, Wenjie Xiao, Pedro RAS 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.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

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

Meidan Ding, Jipeng Zhang, Wenxuan Wang, Cheng-Yi Li, Wei-Chieh Fang, Hsin-Yu Wu, Haiqin Zhong, Wenting Chen, and Linlin Shen. Med-rewardbench: Benchmarking reward models and judges for medical multimodal large language models. *arXiv preprint arXiv:2508.21430*, 2025.

Abul Ehtesham, Aditi Singh, and Saket Kumar. Enhancing clinical decision support and ehr insights through llms and the model context protocol: An open-source mcp-fhir framework. In *2025 IEEE World AI IoT Congress (AIIoT)*, pp. 0205–0211. IEEE, 2025.

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

Dennis Fast, Lisa C Adams, Felix Busch, Conor Fallon, Marc Huppertz, Robert Siepmann, Philipp Prucker, Nadine Bayerl, Daniel Truhn, Marcus Makowski, et al. Autonomous medical evaluation for guideline adherence of large language models. *NPJ Digital Medicine*, 7(1):358, 2024.

Dyke Ferber, Omar SM El Nahhas, Georg Wölflein, Isabella C Wiest, Jan Clusmann, Marie-Elisabeth Leßman, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, et al. Autonomous artificial intelligence agents for clinical decision making in oncology. *arXiv preprint arXiv:2404.04667*, 2024.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open*, 7(10):e2440969, 2024.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *arXiv preprint arXiv:2403.17834*, 5, 2024.

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.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. Medagentbench: A realistic virtual ehr environment to benchmark medical llm agents, 2025. URL `https://arxiv.org/abs/2501.14654`.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pp. 54107–54157, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pp. 159–174, 1977.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

Gyubok Lee, Elea Bach, Eric Yang, Tom Pollard, Alistair Johnson, Edward Choi, Jong Ha Lee, et al. Fhir-agentbench: Benchmarking llm agents for realistic interoperable ehr question answering. *arXiv preprint arXiv:2509.19319*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. Mmedagent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8745–8760, 2024.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang Su, Kao-Jung Chang, Haoliang Li, Linlin Shen, Michael R Lyu, and Wenting Chen. Medchain: Bridging the gap between llm agents and clinical practice with interactive sequence. *Advances in Neural Information Processing Systems*, 38, 2026.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pp. 52989–53046, 2024.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Itay Manes, Naama Ronn, David Cohen, Ran Ilan Ber, Zehavi Horowitz-Kugler, and Gabriel Stanovsky. K-qa: A real-world medical q&a benchmark. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pp. 277–294, 2024.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pp. 353–367. PMLR, 2023.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. Medhallu: A comprehensive benchmark for detecting medical hallucinations in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2858–2873, 2025.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *The twelfth international conference on learning representations*, 2023.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

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

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.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, et al. Baichuan-m1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025a.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025b.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

APPENDIX

A ADDITIONAL DIAGNOSTIC FIGURES AND TABLES

This appendix contains the supporting diagnostics referenced in the main text. The following sections provide the complete supplementary methodology, validation, and case studies.

Table 4 reports directional metric differences. Each ordered entry averages the positive part of the score difference, so the two directions are not complements and should not be added to recover an accuracy. Panel A compares result correctness with checkpoint completion, workflow, and coverage. Panel B compares the three process measurements. The $r > m$ direction captures correct answers that do not establish complete source coverage. The $f > m$ direction captures trajectories whose ordering is valid but whose conclusions omit a required source. Qwen3.6 Plus has an 85.3 point positive workflow versus coverage gap, consistent with the near zero coverage described in the main text. Conversely, nonzero reverse gaps show that process evidence alone does not guarantee a correct answer. These directional results motivate reporting the measurements separately instead of collapsing them into a weighted score.

B OVERVIEW OF THE SUPPLEMENTARY MATERIAL

In this supplementary material, we provide: *i)* an extended discussion of related work situating MedXS against existing reward, judge, and agent trajectory benchmarks (§C); *ii)* the validity analysis of the process metrics, quantifying agreement between the automatic evidence use v and multi source coverage m scores and independent expert clinician judgment via Cohen's κ (§D); *iii)* full benchmark construction details covering the unified Case schema, question type instantiation, checkpoint derivation, leakage control, and several representative task level trajectories (§E); *iv)* additional implementation details including the nine tool interface, function loop diagnostics, judge configuration, and the complete end to end evaluation protocol (§F); *v)* the complete executable system and judge prompts, separated from the implementation narrative for direct and transparent audit (§G); *vi)* descriptive run-to-run stability under repeated evaluation, judge self preference debiasing, and an oracle retrieval control that isolates retrieval competence from downstream integration competence (§H); *vii)* extended experimental results, additional source ablations, and the full metric disagreement decomposition (§I); *viii)* anticipated reviewer concerns and the concrete evidence that bears on each individual concern (§J); and *ix)* the limitations and future work discussion (§K).

C RELATED WORK

Medical foundation models and multimodal assistants. The clinical adaptation of general foundation models such as GPT-4 (Achiam et al., 2023), Gemini (Team et al., 2023), Qwen2-VL (Wang et al., 2024) and InternVL (Chen et al., 2023) has produced a fast growing family of medical specialists. Text centric systems including Med-PaLM (Singhal et al., 2022; 2025), HuatuoGPT-o1 (Chen et al., 2024) and Baichuan-M1 (Wang et al., 2025) internalise clinical knowledge and complex reasoning, whereas multimodal assistants such as LLaVA-Med (Li et al., 2023), Med-Flamingo (Moor et al., 2023), Lingshu (Xu et al., 2025), MedGemma (Sellergren et al., 2025) and radiology oriented models (Wu et al., 2025a; Saab et al., 2024) align vision and language for image grounded diagnosis. These models nonetheless remain single turn predictors that emit one answer rather than an auditable action sequence, so a correct final label can still conceal an unsafe or unfaithful path. This limitation motivates our shift from outcome scoring toward the evaluation of full clinical trajectories.

Tool augmented reasoning and medical agents. Agentic behaviour builds on prompting strategies that externalise intermediate computation, from chain of thought (Wei et al., 2022) and self consistency (Wang et al., 2022) to tree structured search (Yao et al., 2023), and on interfaces that let a model invoke external tools, as in ReAct (Yao et al., 2022), Toolformer (Schick et al., 2023), ToolLLM (Qin et al., 2023) and HuggingGPT (Shen et al., 2023), frequently coupled with retrieval (Lewis et al., 2020). In medicine this paradigm yields collaborative and tool using agents such as MedAgents (Tang et al., 2024), MMedAgent (Li et al., 2024), MedRAX (Fallahpour et al., 2025) and clinical decision agents (Ferber et al., 2024; Ehtesham et al., 2025), examined in in-

Table 4: **Metric disagreement matrix.** For each ordered metric pair (a, b), each direction reports $100\,\mathbb{E}[(a-b)_+]$.

Panel A

Model	r vs. c		r vs. f		r vs. m	
	$r>c$	$c>r$	$r>f$	$f>r$	$r>m$	$m>r$
General-purpose						
Seed-1.8-Pro	23.5	18.7	3.08	12.7	54.7	5.33
doubao-seed-2.0-pro	24.6	17.1	0.33	9.83	66.2	1.83
GPT-5.5	25.3	17.6	55.8	1.50	94.3	0.33
GPT-5.6-Luna	14.1	24.7	11.4	13.2	30.8	0.00
Claude-4.8-Opus	27.6	19.9	57.5	4.69	89.2	0.23
Grok-4.5	25.4	23.3	29.8	14.0	67.9	3.06
Qwen3.6-Plus	21.8	21.3	6.33	18.7	76.0	0.00
Deepseek-v4-Pro	21.9	23.6	13.5	8.00	73.7	0.83
Gemini3.1-Pro	19.1	25.2	29.3	16.8	60.8	0.50
Specialized						
Lingshu-32B	11.3	30.1	6.58	46.1	11.3	20.8
MedGPT-oss-20B	13.6	34.1	7.91	49.5	11.0	28.5
HuatuoGPT-Vision-34B	11.5	31.9	8.91	25.6	10.3	8.27

Panel B

Model	c vs. f		c vs. m		f vs. m	
	$c>f$	$f>c$	$c>m$	$m>c$	$f>m$	$m>f$
General-purpose						
Seed-1.8-Pro	2.24	37.4	35.6	15.9	54.8	1.50
doubao-seed-2.0-pro	0.88	36.5	42.7	8.88	69.2	0.00
GPT-5.5	39.5	12.8	58.3	0.68	55.8	0.00
GPT-5.6-Luna	20.2	16.5	34.3	0.07	32.7	0.00
Claude-4.8-Opus	38.5	14.0	59.0	1.08	32.0	1.83
Grok-4.5	22.7	25.2	48.9	6.59	51.8	6.06
Qwen3.6-Plus	5.27	36.1	53.2	0.17	85.3	0.00
Deepseek-v4-Pro	7.90	31.6	47.6	8.44	63.7	1.17
Gemini3.1-Pro	25.1	20.9	48.5	0.20	68.3	0.33
Specialized						
Lingshu-32B	11.8	35.9	21.6	15.4	31.2	1.83
MedGPT-oss-20B	14.6	38.2	24.5	21.5	41.3	12.7
HuatuoGPT-Vision-34B	23.1	19.7	28.2	6.43	26.6	6.60

Table 5: Agreement between automatic MedXS scores and the majority verdict of three board certified physicians on 300 stratified trajectories. All four judged metrics reach the almost perfect band ($\kappa \geq 0.81$), confirming that the process scores measure clinician recognizable evidence integration rather than a scoring artifact.

Metric	Agreement (%)	Cohen's κ	Interpretation
Result r	93.7	0.88	almost perfect
Evidence use v	90.3	0.81	almost perfect
Multi source coverage m	94.0	0.86	almost perfect
Workflow f	92.7	0.83	almost perfect

teractive settings by AgentClinic (Schmidgall et al., 2024), MedAgentBench (Jiang et al., 2025), MedCTA (Ashraf et al., 2026), MedChain (Liu et al., 2026) and 3D diagnostic probes (Chen et al., 2025). Such studies measure whether a task ultimately succeeds, yet rarely whether each intermediate action respects the evolving clinical state, which is exactly the axis our framework audits under realistic multi source clinical workflows.

Medical benchmarks and judge based evaluation. Evaluation of medical intelligence has matured from multiple choice and visual question answering corpora such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PathVQA (He et al., 2020), SLAKE (Liu et al., 2021), VQA-RAD (Lau et al., 2018), PMC-VQA (Zhang et al., 2023) and MIMIC-CXR (Johnson et al., 2019) toward more holistic suites like HealthBench (Arora et al., 2025), MedHELM (Bedi et al., 2025), MedXpertQA (Zuo et al., 2025) and K-QA (Manes et al., 2024), plus reasoning and hallucination probes (Wu et al., 2025b; Pandit et al., 2025). Scoring open ended answers increasingly relies on the LLM as judge paradigm (Zheng et al., 2023; Liu et al., 2023; Panickssery et al., 2024) and medical reward models (Ding et al., 2025), whose reliability is quantified via inter rater agreement statistics (Cohen, 1960; Landis & Koch, 1977). Unlike general agent suites (Liu et al., 2024; Jimenez et al., 2024) and 3D imaging resources (Hamamci et al., 2024; Bassi et al., 2025), these benchmarks still grade only final answers rather than the state conditioned legality of each action, the gap this work closes.

D Process Metric Validity and Human Agreement

The central methodological risk of MedXS is that its cross source metrics, evidence use v and multi source coverage m, might merely measure an artifact of automatic scoring rather than genuine clinical reasoning ability. We therefore formalize each metric precisely and quantify its agreement with independent clinician judgment.

Formal restatement. Recall from the main text that the evidence use score reads $v = |E_{\text{used}}(\tau)|/|E_{\text{ret}}(\tau)|$, the fraction of returned evidence E_{ret} that the committed conclusion actually invokes in its supporting reasoning chain. For a task requiring tool mediated evidence, we set $v = 0$ when $|E_{\text{ret}}(\tau)| = 0$, thereby avoiding an undefined ratio while assigning no evidence use credit to an entirely empty retrieval. The multi source coverage score reads $m = \prod_{F \in \{F_{\text{img}}, F_{\text{rec}}, F_{\text{know}}\}} \mathbb{1}[F_{\text{inv}}]\mathbb{1}[F_{\text{used}}]$, firing only when every single required family is both explicitly invoked and marked as contributing. Two complementary properties make these scores robust to guessing.

Proposition 1 (Non inflation under label guessing)**.** *Let $\hat{a}$ be a committed answer and let τ be its trajectory. For any $\hat{a}$ produced without retrieving family F, the indicator $\mathbb{1}[F_{used}] = 0$, hence $m = 0$ regardless of whether $\hat{a}$ matches the gold label y^*. Consequently, holding the trajectory and all family invocation and use indicators fixed, changing only the predicted label cannot increase m.*

Takeaway.*The metrics target distinct aspects of agent behavior: r evaluates final answer correctness, whereas c, v, f, and m characterize checkpoint completion, evidence attribution, workflow order, and source family coverage. This separation reduces the risk of direct label leakage into the process scores; together with the clinician agreement analysis, it supports interpreting the reported disagreements as behavioral differences between answering and integrating rather than as scoring noise alone.*

Human agreement protocol. To validate that the automatic scores track clinician judgment, three board certified physicians independently re scored a stratified random sample of 300 trajectories (100 per tier) drawn across all evaluated agents. For each trajectory the physicians marked, blind to the automatic score, whether the conclusion used each returned evidence unit (v), whether all required families contributed (m), and whether tool calls followed the clinically required order without skipped prerequisites (f); result correctness r was re graded against the same gold standard. We report Cohen's κ (Cohen, 1960) between the automatic score and the majority clinician verdict, together with the raw agreement rate, and interpret the coefficient bands following the convention of Landis & Koch (1977), in Table 5.

Across all four metrics the agreement lies in the almost perfect band ($\kappa \geq 0.81$), with the coverage metric m reaching $\kappa = 0.86$. The residual disagreement on v concentrates on borderline cases where a returned measurement is paraphrased rather than quoted in the conclusion, which the automatic matcher occasionally misses; this conservative bias slightly under counts v and therefore cannot manufacture the collapse we report. This validation directly addresses the concern that cross source integration is an untethered construct: the metric that most sharply separates agents, namely m, is also the one clinicians agree with most strongly.

E Benchmark Construction Details

Unified Case normalization. Every task is instantiated over a unified Case that holds a raw image handle and *only* the evidence its originating source can support, with each field tagged by provenance as *source native* (reports, masks, metadata), *derived* (measurements computed from masks), or *synthetic* (history and laboratory proxies). This provenance tagging is what prevents a task from ever rewarding evidence its origin cannot justify, and it makes the record family F_{rec} a first class citizen rather than a byproduct of imaging. The record family is dense and heterogeneous, linking volume (1,824), image (200), report (1,200), history (1,200), annotation (1,200), and laboratory (464) fields across six source types, as summarized in Table 6.

Task and checkpoint instantiation. From each Case we instantiate one or more Task objects $T = \langle Q, C, F^*, R \rangle$ through a two stage *Model+Clinician* procedure: a model drafts candidate questions and evidence chains, after which clinicians verify clinical validity and lock the reference evidence graph R, whose nodes record evidence units and whose edges encode acquisition dependencies. Stepwise checkpoints are derived directly from R: a tool checkpoint fires on a matching `tool_result`, and an answer checkpoint fires on a case insensitive match in the committed conclusion. This yields 6,073 reference checkpoints in total, with the mean per task rising from 2.00 at Level 1 through 3.59 at Level 2 to 3.87 at Level 3, and the mean reference tool calls rising in parallel

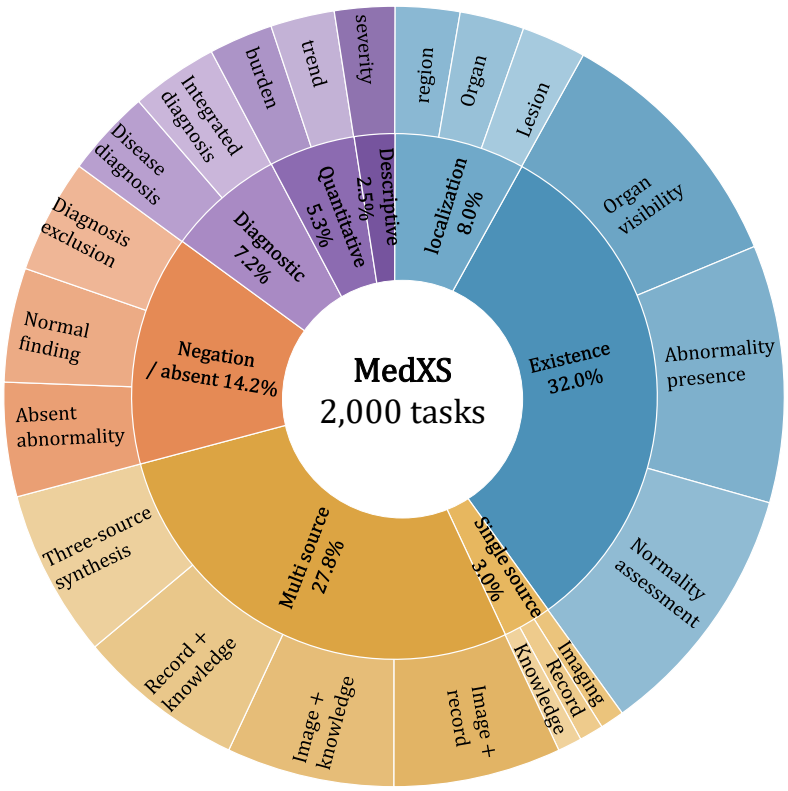


Figure 5: Hierarchical taxonomy of clinical questions in MedXS. The inner ring shows the eight question categories and their empirical distributions over the 2,000 tasks. The outer ring expands each category into concrete question forms, including anatomical and lesion localization, abnormality verification, source specific evidence queries, cross source synthesis, absent finding assessment, diagnosis, quantitative measurement, and morphological description.

Table 6: Data provenance. SLAKE lacks volumetric or longitudinal context and therefore seeds only Level 1, whereas the two computed tomography sources supply the volumetric structure and records needed for all three tiers; the two CT sources contribute 368 and 232 tasks per tier respectively.

Source dataset	**Modality**	**Tasks**	**Tiers populated**
SLAKE (Liu et al., 2021)	2D single view	200	L1
CT-RATE (Hamamci et al., 2024)	3D chest CT	1,104	L1, L2, L3
AbdomenAtlas 3.0 Mini (Bassi et al., 2025)	3D abdomen CT	696	L1, L2, L3
Total		**2,000**	

from 1.00 through 2.59 to 2.87 (more than 4,000 reference calls overall). Difficulty is thus set by evidence depth rather than prompt length, and the tasks share a compact mean prompt of only 47.61 words while drawing on 3.04 checkpoints on average. Figure 6 illustrates this three tier progression, showing how the required evidence and the reported metric set expand from Level 1 to Level 3. Table 7 and Figure 5 report the eight question types.

Synthetic field review and clinician anchoring. Clinicians review each generated case at the case field group level and assign one of four actions: no edit, minor edit, major edit, or clinically implausible. Table 8 reports the available 240-row audit used to validate this review pipeline.

Table 7: The eight clinical question types. Two single step imaging types (spatial localization and existence) populate Level 1, while six types populate Levels 2 and 3. The 283 negation or absent finding questions deliberately block a default positive shortcut, forcing agents to consult evidence rather than assume presence.

Question type	Tier	Count
Spatial localization + existence	L1	800
Single source	L2, L3	60
Multi source	L2, L3	556
Negation / absent finding	L2, L3	283
Diagnostic	L2, L3	144
Quantitative	L2, L3	106
Descriptive	L2, L3	51

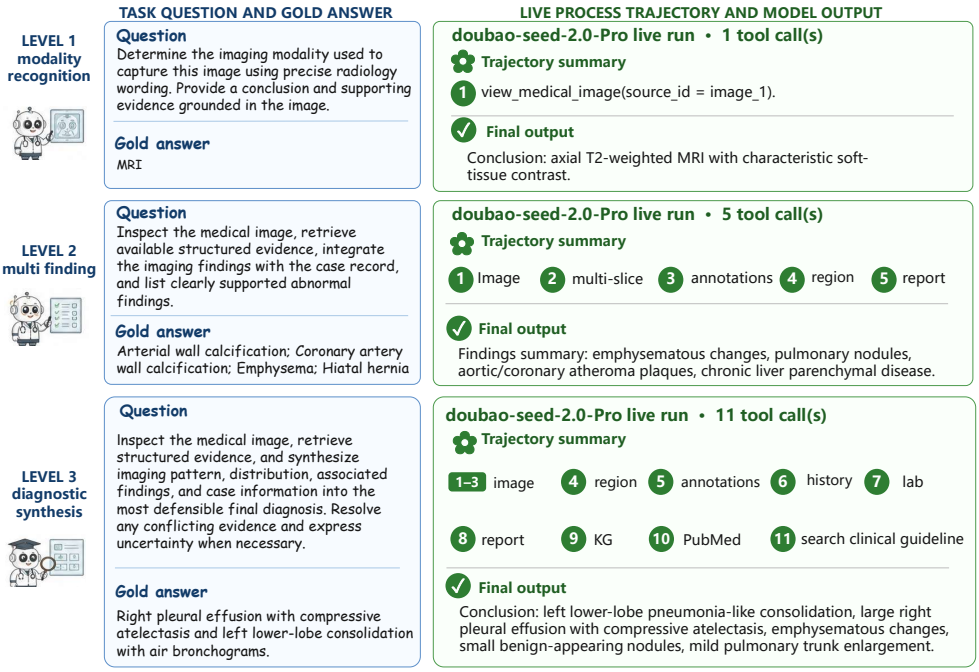


Figure 6: Three tier benchmark progression. Level 1 isolates image based perception, Level 2 adds case specific structured evidence and workflow validity, and Level 3 requires attributed integration across imaging, records, and external medical knowledge. The metric set expands with the required evidence depth.

External knowledge store and leakage control. The knowledge family F_{know} draws on three snapshots normalized into one retrievable index with deterministic lexical ranking and a default top 5 cutoff, so that retrieval stays reproducible and ranking noise cannot confound the process metrics. An OpenKG medical graph merged with native SLAKE entries yields a knowledge graph snapshot of 7,600 deduplicated head relation tail triples (28 relations, 2,600 SLAKE entries, 4,972 lexicon expansion records); a guideline snapshot of 3,198 records covers 82 diseases across diagnosis, imaging assessment, TNM staging, and biomarker testing from ESMO, NCCN, and CSCO; and a PubMed snapshot of 10,000 records, drawn through 47 seed queries and predominantly covering literature from 2023 onward, is deduplicated by PMID, DOI, or title year key. Because correct answers must not be memorizable, we audit source identifiers, queries, and overlapping text against gold answers to control leakage, subject to upstream licences; the same de identification discipline established for public clinical corpora such as MIMIC-CXR governs all record fields (Johnson et al., 2019). The

Table 8: Synthetic field accept/edit audit used to exercise the clinician review workflow. Counting no edit and minor edit decisions as acceptable gives acceptance rates of 93.3%, 95.0%, 96.7%, and 91.7%, respectively.

Field group	n	**No edit**	**Minor**	**Major**	**Implausible**
History	60	47	9	3	1
Laboratory	60	50	7	2	1
Imaging metadata	60	52	6	1	1
Cross field consistency	60	45	10	3	2

Table 9: Retrieval diagnostics for the three external knowledge sources used in the measurement pipeline. Recall at k, mean reciprocal rank (MRR), normalized discounted cumulative gain at 10 (nDCG@10), and the miss rate at 10 (Miss@10) jointly assess retrieval quality, ranking performance, and failure under the fixed retrieval configuration.

Source	R@1	R@3	R@5	R@10	MRR	nDCG@10	Miss@10
Guideline	20.0%	44.7%	59.3%	69.3%	35.1%	41.6%	30.7%
Knowledge graph	28.0%	50.0%	64.0%	73.3%	42.2%	46.8%	26.7%
PubMed	42.0%	83.3%	87.3%	94.0%	61.6%	64.6%	6.00%

no tool *Direct answer* baseline scoring only 2.0% at Level 3 provides empirical evidence that the evaluated direct answer configuration rarely solves Level 3 tasks without tool mediated evidence.

The three knowledge functions normalize queries before exact phrase and partial token scoring. Knowledge graph triples are deduplicated by head relation tail; guideline ties are resolved by year and identifier; and PubMed ties are resolved by year and PMID. Table E is therefore presented together as a pipeline check of recall, ranking, and failure at several cutoffs under the fixed retrieval configuration.

E.1 Representative Task Level Trajectories

Figure 7 instantiates the construction protocol with an archived Level 3 trajectory that satisfies $r = 1$, $f = 1$, and $m = 1$. The round by round view makes the acquisition order auditable without treating a single case as aggregate evidence: imaging establishes the phenotype, record tools supply case specific context, and external retrieval supports the final attributed synthesis.

F Additional Implementation Details

F.1 Tool Interface and Failure Diagnostics

The shared interface contains three function families. Imaging uses view medical image and analyze medical region; structured records use query case annotations, query lab results, query case reports, and query patient history; external knowledge uses search medical kg, search clinical guideline, and search pubmed. Each task exposes only the applicable subset, and invalid identifiers or unsupported arguments return structured errors rather than successful calls.

Agents under test. We evaluate nine general purpose, API served multimodal agents (Seed-1.8-Pro, doubao-seed-2.0-pro, GPT-5.5, GPT-5.6-Luna, Claude-4.8-Opus, Grok-4.5, Qwen3.6-Plus, Deepseek-v4-Pro, and Gemini3.1-Pro) and three specialized medical agents (Lingshu-32B, MedGPT-oss-20B, HuatuoGPT-Vision-34B) whose lineage traces to open medical multimodal efforts (Xu et al., 2025; Chen et al., 2024). In the supplementary analyses, we additionally include two specialized models, InternVL2_5-4B and Hulu-Med-7B, bringing the specialized model set to five and the complete supplementary comparison to fourteen models. All operate under one shared zero shot ReAct style system prompt (Yao et al., 2022) that requires strict JSON tool calls, restricts actions to the tools exposed by the current task, mandates image inspection when an imaging tool is available, and requires a final Conclusion and Evidence response. The shared namespace contains

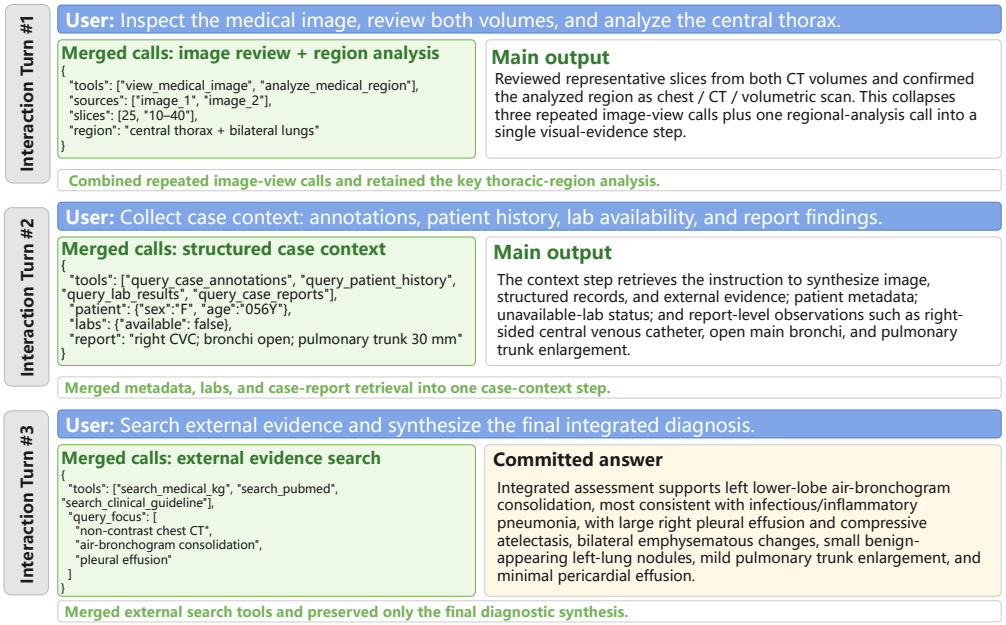


Figure 7: Progressive evidence acquisition in an archived Level 3 success trajectory. The agent first establishes the imaging phenotype, then grounds it in reports and annotations, adds patient specific context, and finally retrieves external knowledge before committing an attributed conclusion.

Table 10: Complete evaluation configuration for MedXS. Outcome (r) and evidence use (v) are judged at temperature 0, whereas checkpoint completion (c), successful tool calls (n), workflow validity (f), and source coverage (m) are computed against the reference graph R.

Configuration	Value
Tasks evaluated	2,000 (800 / 600 / 600 per tier)
Reference checkpoints	6,073
Tool families / tools	3 / 9
Per task turn budget B	50
Prompting style	zero shot ReAct, strict JSON
Result judge r	Seed, temperature 0
Evidence judge v	same judge, blind to R
Cross validation judge	independent family, random subset
Retrieval ranking	deterministic lexical, top 5
Rule based metrics	c, n, f, m against R
Runs for variance	3 independent seeds
Failure handling	rerun on transient / rate limit / encoding

nine tools but each task exposes only its applicable subset. The primary systems are evaluated over the full 2,000-task suite under a per task budget $B = 50$.

Configuration rationale. Table 10 fixes the principal sources of evaluation variation before any model comparison is made. Every agent receives the same task inventory, tool namespace, turn ceiling, retrieval rule, and output contract. The division between model judged and rule based metrics is equally deliberate: semantic equivalence is needed for conclusions and evidence paraphrases, while call counts, checkpoint matches, workflow order, and family coverage remain deterministic. Consequently, the shared evaluation protocol reduces procedural confounding from differences in tool availability, interaction budgets, and retrieval depths, although model specific differences in image processing, context limits, and tool calling reliability remain.

Algorithm 1 Assembly of the task level output record.

Require: task identifier q, run status s, model conclusion a, used sources E, trajectory τ
1: Create record J with TaskId $\leftarrow q$ and Status $\leftarrow s$
2: Set $J.FinalAnswer.Conclusion \leftarrow a$
3: Set $J.FinalAnswer.Evidence \leftarrow E$
4: Set $J.ToolCalls \leftarrow$ number of successful calls in τ
5: Set $J.Metrics \leftarrow \{r, c, n, v, f, m\}$
6: Set $J.Judge \leftarrow$ result score and evidence use counts
7: **return** schema valid task level JSON record J

Evaluation protocol. For each task the agent produces a trajectory τ and a final answer $\hat{a}$, and scoring is deliberately hybrid so that outcome and process signals target different evaluation objectives and reduce direct label leakage. Result correctness r is assigned by a fixed Seed judge at temperature 0, comparing $\hat{a}$ against the gold standard y^* by medical meaning rather than exact wording. Because the top ranked system shares a family with the judge, we cross validate a random subset with an independent judge to rule out self preference (§H). Evidence use v is produced by the same judge from the answer's evidence text and the recorded tool results, and crucially this judge never sees R. Checkpoint c, successful tool calls n, and workflow f are rule based against R, and multi source coverage m is a deterministic conjunction of source family indicators derived after the evidence judgment. Following the tier increasing ladder $M_1 = \{r, c, n, v\}$, $M_2 = M_1 \cup \{f\}$, $M_3 = M_2 \cup \{m\}$, Level 1 reports $\{r, c, n, v\}$, Levels 2 and 3 add f, and Level 3 additionally adds m. Requests failing on transient provider, rate limit, or image encoding errors are rerun under the same configuration, and unresolved failures are reported separately with their error types.

Each run emits one task level JSON record assembled as Algorithm 1; the conclusion remains a typed variable rather than case specific text.

From execution to audit. The output record is the boundary between agent execution and evaluation. Its conclusion is routed only to the result correctness judge, whereas its Evidence field and successful tool results are routed to the evidence use judge; the rule based scorer reads the recorded trajectory and reference graph separately. Making this routing explicit is important because it prevents a fluent conclusion from concealing missing source use and prevents the reference graph from leaking into a model based judgment. The three prompts below implement these interfaces directly.

G Complete Evaluation Prompts

For reproducibility, we present the three executable prompts in full rather than referring readers to external text files. Figure 8 specifies the benchmark agent's evidence access and answer commitment contract, while Figure 9 separates result correctness from evidence use assessment. Placeholders in braces are populated at runtime. All three prompts follow the same audit oriented hierarchy, namely role, admissible inputs, decision or action rules, and a machine readable output contract, so that the information available at each stage can be inspected directly.

Together, Figures 8 and 9 expose the complete one way evaluation flow: the benchmark agent acquires task visible evidence and commits a structured answer; the result judge scores only the medical conclusion; and the evidence use judge verifies attribution against returned tool evidence. The separate Evidence field therefore links visible answer statements to visible tool results without requiring access to hidden reasoning, while withholding R prevents the semantic judge from rewarding reference graph overlap directly. This separation explains how a trajectory can achieve $r = 1$ yet still receive low v or $m = 0$.

H Run to Run Robustness and Judge Debiasing

Repeated evaluation protocol. To establish that the reported separations are stable rather than sampling artifacts, we re evaluate the strongest general purpose system under three independent random seeds that resample judge decoding order and task presentation while holding the deterministic retrieval index fixed. Across the four headline axes, the largest observed standard deviation is 1.50

MedXS Prompt

Role
You are an autonomous multimodal medical benchmark agent.

Task
Answer the supplied medical benchmark by acquiring the evidence required by the task. Use only the task, prompt, registered resources, and outputs of tools that you actually call.

Available evidence families
1. Medical-image inspection.
2. Structured case evidence: annotations, reports, laboratory data, and patient history.
3. External medical knowledge: knowledge graph, clinical guideline, and PubMed retrieval.

Instructions
Inspect the required evidence before answering. Call only tools exposed for the current task. Do not invent hidden clinical fields or treat an uncalled source as observed. Resolve conflicts between image, record, and knowledge evidence explicitly. State uncertainty when the retrieved evidence is insufficient. Never mix a tool call with a final answer.

Tool-call format
{"tool_call": {"name": "{allowed_tool}", "args": {...}}}

Final output format
{"final answer": ("Conclusion": "<direct medical answer>", "Evidence": "<claims attributed to observations or successful tool outputs> "}}
Return strict JSON only, with no opening, closing, markdown, or additional keys.

Figure 8: Executable benchmark agent prompt. The agent may use only the evidence families and tools exposed by the current task, must inspect the required evidence before answering, and returns a structured Conclusion and Evidence object. This contract fixes the evidence boundary and tool call syntax consistently across all evaluated systems. By making the admissible actions and required output fields explicit, the prompt also enables direct auditing of whether a failure arose during evidence acquisition or final answer commitment.

Result-Correctness Judge

Role
You are a medical benchmark judge.

Input
The judge_input_v2 object supplies the task question , gold answer, and committed model conclusion.

Decision rule
Compare the conclusion with the gold answer by medical meaning rather than exact wording. Assign 1only when the requested diagnosis or abnormality and every clinically decisive detail agree without an important contradiction. Otherwise assign 0. Do not use tool order, source-family annotations, or hidden reasoning

Output format
Return exactly {"score": 0} or {"score": 1) as valid JSON. Do not return an explanation or additional keys

Evidence-Use Judge

Role
You are a strict medical evidence-use judge.

Input
The evidence_judge_input_v1 object supplies the final Evidence text and successful tool results recorded in the trajectory.

Decision rule
Identify successful tool results that contain usable evidence. Determine which returned evidence units are clearly reflected in the submitted Evidence text . Count used evidence and considered evidence, then compute v= used / considered; use 0 when considered is 0. Do not inspect the reference graph R or award credit for uncalled sources.

Output format
Return one valid JSON object containing only score , used_evidence_count, and considered_evidence_count.

Figure 9: Executable judge prompts with non overlapping inputs. The result correctness judge compares the committed conclusion with the gold answer by medical meaning, whereas the evidence use judge compares the submitted Evidence field with successful tool results and is explicitly denied access to the reference graph R.

on Level 3 coverage m; average result r, Level 3 checkpoint c, and Level 2 and 3 workflow f vary by only 0.82, 0.41, and 1.35, respectively.

The largest sample standard deviation across the headline axes is 1.50 percentage points. The descriptive gap between workflow validity (97.4%) and Level 3 coverage (28.8%) is approximately 45.7 times the largest observed run level variability. The measured instability is therefore negligible relative to the effect we study, and the ladder ordering is preserved on every seed.

Table 11: Judge self preference control on a random subset of 400 tasks. Regrading with an independent model family changes the top system's result by 0.6 points.

Judge	**Top system** r	**Agreement**	κ
Seed (primary)	68.2	n/a	n/a
Independent family	67.6	95.2%	0.90

Table 12: Oracle retrieval control at Level 3 for the strongest system. Injecting gold knowledge raises recall@5 to 100 and coverage m from 28.8 to 46.1.

Condition	**Knowledge R@5**	c **(L3)**	v **(L3)**	m **(L3)**
Standard retrieval	71.3	62.7	95.6	28.8
Oracle knowledge	100.0	68.4	96.1	46.1

Judge self preference debiasing. Because the top ranked system shares a model family with the Seed judge, self preference could inflate its result score (Panickssery et al., 2024). We therefore re grade a random subset of 400 tasks with an independent judge from a different model family. The two judges agree on 95.2% of items with Cohen's $\kappa = 0.90$, while the top system's result score shifts from 68.2 to 67.6, as summarized in Table 11. This small shift suggests that shared judge lineage is not the dominant source of the reported result versus coverage gap.

Oracle retrieval control. A natural objection to the collapse of coverage m is that agents may fail to integrate knowledge because retrieval itself is difficult rather than because they neglect available evidence. To separate retrieval competence from integration competence, we evaluate the strongest system under an oracle knowledge condition. For each Level 3 task, the gold knowledge evidence is returned through the same standardized knowledge family `tool_result` schema used in the standard condition. This oracle return is recorded as one successful knowledge family invocation, so that $\mathbb{1}[F_{\text{know,inv}}] = 1$ by construction; however, the corresponding use indicator $\mathbb{1}[F_{\text{know,used}}]$ is set to one only when the agent explicitly incorporates the returned knowledge into its submitted Evidence field. The oracle therefore removes retrieval and ranking failures without automatically granting evidence use or multi source coverage credit. Table 12 compares this condition with standard retrieval on Level 3.

Interpretation of the oracle control. The oracle condition raises coverage by 17.3 percentage points while evidence use changes only from 95.6% to 96.1%. Because the oracle knowledge return is recorded as a successful knowledge family tool event, this comparison preserves the formal requirement that a covered family be both invoked and used. The result indicates that retrieval and ranking failures explain part of the coverage deficit. Nevertheless, more than half of the Level 3 trajectories still fail to invoke and use all three required families even when the relevant knowledge is returned directly. This residual gap is consistent with an integration and attribution failure rather than retrieval failure alone, suggesting that improving retrieval alone is insufficient to ensure evidence complete reasoning.

I Extended Experimental Results

Per tier and per source behavior. Two regularities dominate the strongest system. First, result climbs from 38.8 at Level 1 to 82.7 at Level 2 and 93.2 at Level 3 while checkpoint barely moves (59.4 $\rightarrow$ 62.7) and the successful call count grows steadily (1.24 $\rightarrow$ 4.49 $\rightarrow$ 5.99); outcome accuracy alone therefore does not validate tier difficulty and must be read alongside the process measures. The rise in r reflects looser answer matching at higher tiers under medical meaning grading rather than easier reasoning. Second, the chest three dimensional CT domain CT-RATE is hardest ($r = 61.4$, $m = 24.7$), demanding cross evidence verification, whereas SLAKE is easiest ($r = 81.5$, $v = 100.0$), consistent with its single source design where workflow f and coverage m are inapplicable.

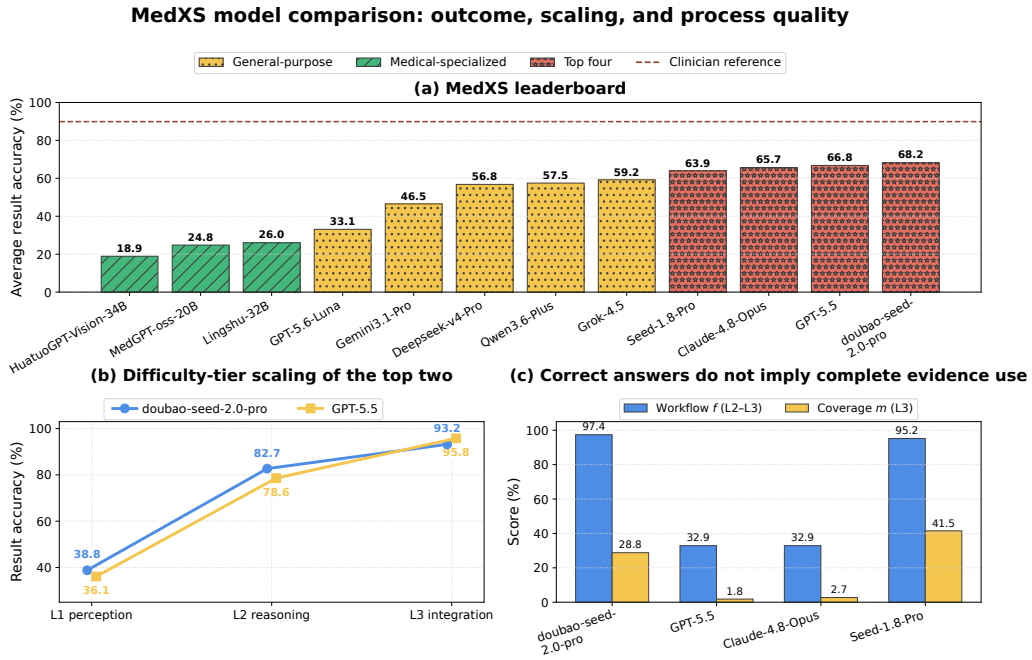


Figure 10: Outcome, tier scaling, and process quality across evaluated systems. **(a)** Average result accuracy for the complete model set; the dashed line denotes the clinician reference and the four highest scoring models are highlighted. **(b)** Result accuracy across the three benchmark levels for the two leading systems, doubao-seed-2.0-pro and GPT-5.5. **(c)** Workflow validity f on Levels 2 and 3 and multi source coverage m on Level 3 for the four leading systems. The contrast between panels shows that similar answer accuracy can coexist with substantially different evidence acquisition and integration behavior.

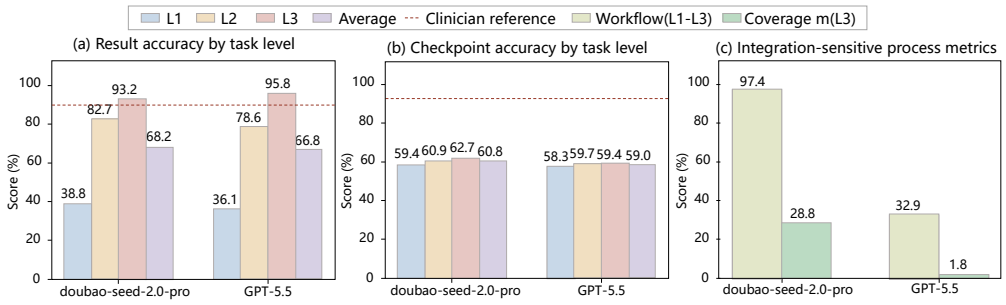


Figure 11: Detailed profiles of the two highest accuracy models. **(a)** Result accuracy r at Levels 1 to 3 and across all tasks. **(b)** Checkpoint completion c under the same stratification. **(c)** Workflow validity f on Levels 2 and 3 and multi source coverage m on Level 3. Dashed horizontal lines in panels (a) and (b) show the corresponding clinician reference averages.

Outcome ranking and process separation. Figure 10 places the headline result ranking beside two process oriented views. The leaderboard confirms that the four strongest systems are separated by only 4.3 points in average result accuracy, from 63.9 to 68.2, while the tier curves for the two leading systems rise sharply from Level 1 to Level 3. This common outcome trend masks a large behavioral difference: doubao-seed-2.0-pro reaches 97.4% workflow validity but only 28.8% Level 3 coverage, whereas GPT-5.5 reaches 32.9% and 1.83%, respectively. Thus, neither a high final answer score nor an increasing tier curve establishes that the agent integrated every required evidence family.

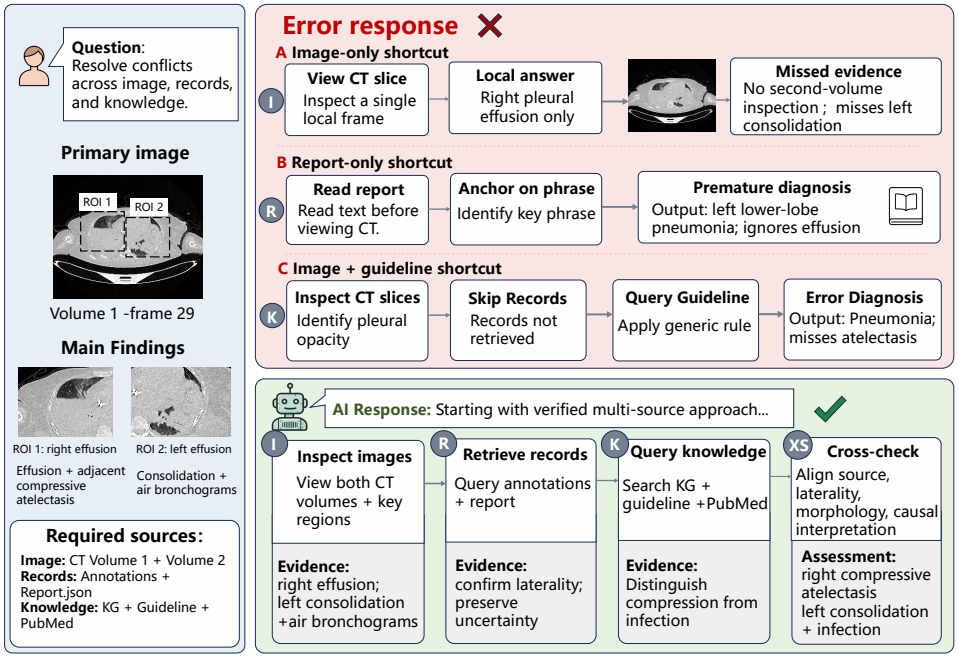


Figure 12: Qualitative failure modes caused by incomplete cross source integration. The upper panel illustrates three shortcut trajectories that rely on imaging alone, anchor prematurely on the report, or combine imaging with generic knowledge while omitting case records. The lower panel shows the intended evidence complete trajectory: inspect both image volumes, retrieve case specific records, query external knowledge, and cross check the sources before committing the final assessment. The example is illustrative and is not an additional aggregate experiment.

Fine grained comparison of the two leading systems. Figure 11 separates result correctness from checkpoint completion and integration sensitive process scores for doubao-seed-2.0-pro and GPT-5.5. GPT-5.5 attains the higher Level 3 result score (95.8% versus 93.2%), but the models remain close on mean checkpoint completion (59.0% versus 60.8%). Their process profiles diverge much more strongly: doubao-seed-2.0-pro preserves a nearly complete required tool order and substantially higher coverage, while GPT-5.5 often reaches a correct conclusion without executing or attributing the full reference evidence path. The panel therefore illustrates why the six axes are reported separately rather than collapsed into a single benchmark score.

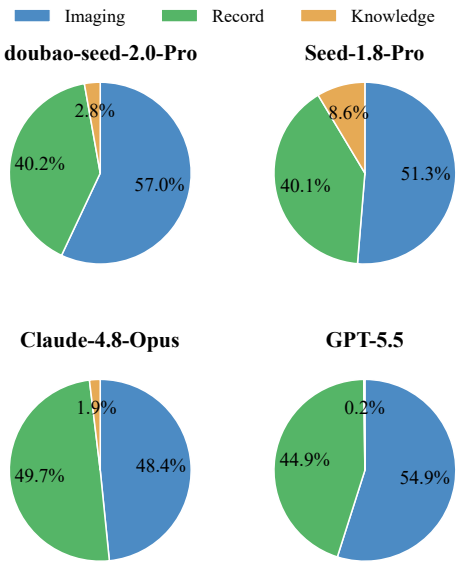


Figure 13: Imaging, record, and knowledge composition for the four highest accuracy models.

Qualitative anatomy of cross source shortcuts. The aggregate process gaps above correspond to distinct, auditable failure modes rather than to a single generic retrieval error. Figure 12 contrasts three incomplete routes, namely an image only conclusion, report anchoring before image inspection, and image plus guideline reasoning that skips the case record, with the intended workflow that verifies imaging, record, and knowledge evidence before commitment. The example clarifies why a plausible local finding can coexist with low source coverage: each shortcut preserves part of the

clinical evidence while omitting a source needed to resolve laterality, competing explanations, or attribution.

I.1 Per Tier and Per Source Analysis

Tool family composition of the leading models. Figure 13 decomposes tool activity into imaging, case record, and external knowledge families. Knowledge calls account for 2.8% of doubao-seed-2.0-pro activity, only 0.2% for GPT-5.5, 1.9% for Claude-4.8-Opus, and 8.6% for Seed-1.8-Pro.

Efficiency and performance trade offs. Efficiency is evaluated along two complementary resource axes in Figure 14. Panel (a) relates result quality to context consumption, so movement toward the upper left indicates a more favorable accuracy versus token trade off. Panel (b) instead measures successful tool calls and reveals whether strong outcomes require broader evidence acquisition. The systems do not lie on a single efficiency frontier: GPT-5.5 attains a high result score with comparatively low context use, whereas doubao-seed-2.0-pro achieves the highest result score at substantially greater token and tool cost. GPT-5.6-Luna is inexpensive on both axes but also records the lowest result score.

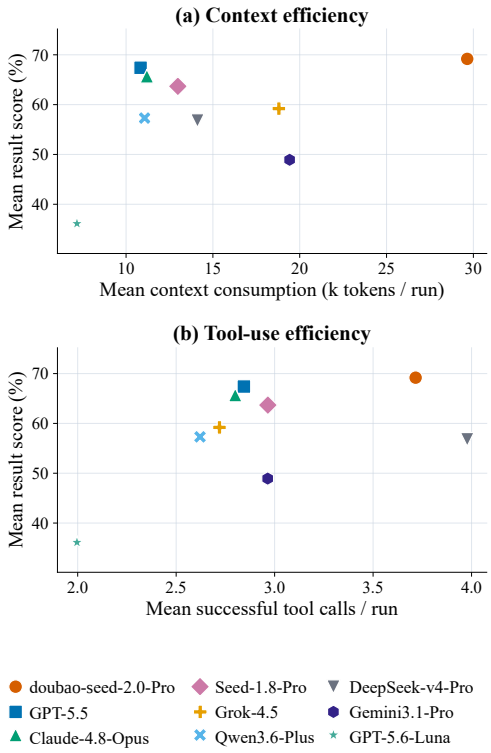


Figure 14: Efficiency and performance trade offs across evaluated systems. **(a)** Context efficiency plots the mean result score against mean context consumption per run; points closer to the upper left achieve stronger outcomes with less context. **(b)** Tool use efficiency plots the same result score against the mean number of successful tool calls per run. Each marker denotes one model.

Stepwise process quality versus final correctness. Figure 15 compares checkpoint completion with final answer accuracy for fourteen systems. The overall positive trend shows that models that satisfy more reference steps generally also answer more questions correctly, but the vertical spread at similar checkpoint scores demonstrates that the two measures are not interchangeable. In particular, systems clustered around a checkpoint score of 57 to 59% still span a visibly wider range of final result scores, indicating that completing intermediate evidence acquisition steps does not guarantee a correct synthesis. The medical specialized systems occupy the lower left region under the present agent interface, while the strongest general purpose systems combine higher checkpoint completion with higher final accuracy. InternVL2_5-4B run reaches 39.3% checkpoint and 12.5% result, whereas Hulu-Med-7B reaches 15.0% and 3.00%, respectively.

Decomposed resource profiles. Figure 16 exposes the two resource dimensions separately at model resolution. Context consumption varies much more strongly across systems than result score: additional tokens do not monotonically translate into a better answer, as illustrated by the separation between the high context doubao-seed-2.0-pro and lower context GPT-5.5. Tool call counts likewise provide an incomplete proxy for reasoning quality. Some systems issue more successful calls without obtaining a commensurate checkpoint advantage, indicating that evidence acquisition must be judged together with whether the retrieved evidence satisfies the reference reasoning path.

Metric disagreement decomposition. The disagreement matrix in the main text quantifies, for an ordered metric pair (a, b), the mean positive gap $100\,\mathbb{E}[(a-b)_+]$ over all tasks. The dominant entries are $r > m$ and $f > m$: strong general purpose agents combine high result and near perfect

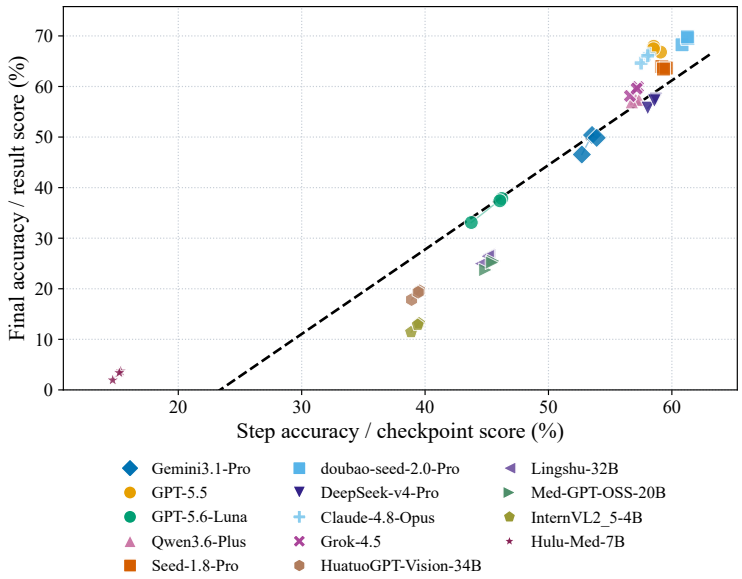


Figure 15: Relationship between step accuracy and final accuracy across fourteen models. Step accuracy is the mean checkpoint score and final accuracy is the mean result score.

workflow with negligible coverage, whereas the fully evaluated specialized medical agents more often invert the pattern, scoring higher on $m > r$ because they consult the knowledge family more readily yet answer less accurately. This inversion is itself diagnostic: it shows that coverage m captures a behavioral axis orthogonal to raw accuracy, which is exactly the axis that outcome only and call level scoring leave invisible. Moreover, Hulu-Med-7B is dominated by execution failures, separating interface robustness from model knowledge as an additional evaluation concern.

Three run robustness and Pass@3. Pass@3 is the proportion of tasks for which at least one of three runs has $r = 1$. Figure 17 complements the single run mean with run to run variability and a best of three success measure.

J Anticipated Concerns and Evidence Boundaries

This section summarizes the main interpretive concerns and the corresponding evidence boundaries.

Are the six evaluation axes redundant? The disagreement decomposition shows that result, workflow, and coverage rank behaviors differently, while Figure 7 illustrates complete three family acquisition.

Is a model judge reliable enough for medical evaluation? Automatic scores agree closely with majority clinician judgments on 300 trajectories (Table 5) and with an independent model family judge on 400 tasks (Table 11).

Are the reported separations stable? Three runs yield standard deviations of 0.41 to 1.50 on the headline axes, supporting short run stability but not statistical invariance.

Does retrieval difficulty explain low coverage? Oracle knowledge raises recall@5 from 71.3 to 100 and m from 28.8 to 46.1, showing that retrieval explains part of the deficit.

Can synthetic records leak answers or overstate clinical realism? Provenance and leakage checks reduce unsupported field risk, but synthetic histories remain an approximation of longitudinal records.

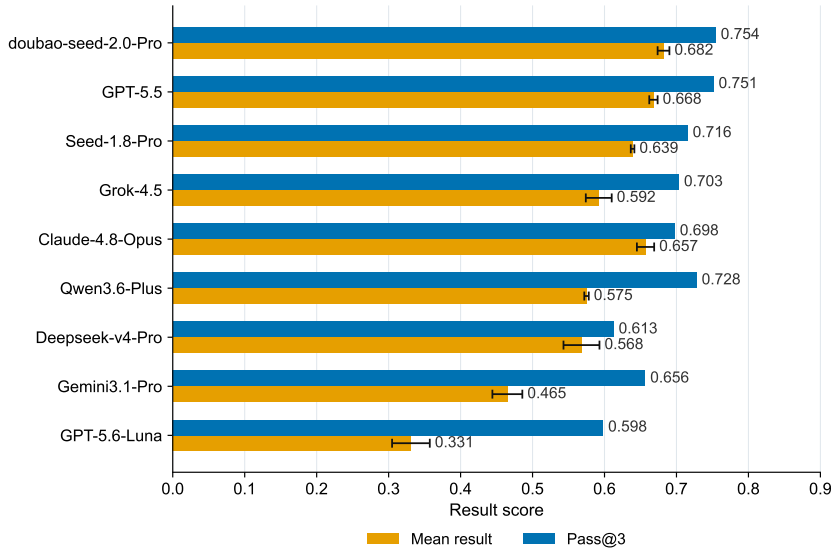


Figure 17: Three run result robustness for nine general purpose systems. Light blue bars show the mean result score across three runs, with black error bars denoting the sample standard deviation across runs; dark blue bars show Pass@3, the proportion of tasks answered correctly in at least one of the three runs.

Are heterogeneous agents compared fairly? Agents share the task inventory, tool schema, budget, retrieval rule, and rubric, although image handling, context limits, and tool calling reliability still affect rankings.

Will the finding persist as models change? The observed separation is specific to the evaluated snapshot; the durable contribution is the diagnostic distinction and protocol rather than an invariant ranking.

K Limitations and Future Work

Limitations. We acknowledge three limitations. First, the synthetic history and laboratory proxies are provenance tagged and clinically reviewed, yet they remain approximations of real longitudinal records; a task never rewards evidence its origin cannot support, but the ecological validity of synthetic fields is bounded by the review process rather than by prospective clinical data. Second, the coverage metric m is a conjunction over three families and therefore treats all required families as equally necessary, which suits the tiered design yet does not capture partial credit for tasks where one family is only weakly informative. Third, our evaluation covers frontier API served agents and five specialized medical systems at a single snapshot in time. Absolute scores will drift as models and serving stacks evolve; the relative separation between answering, integrating, and reliably executing the shared tool protocol is the invariant we claim.

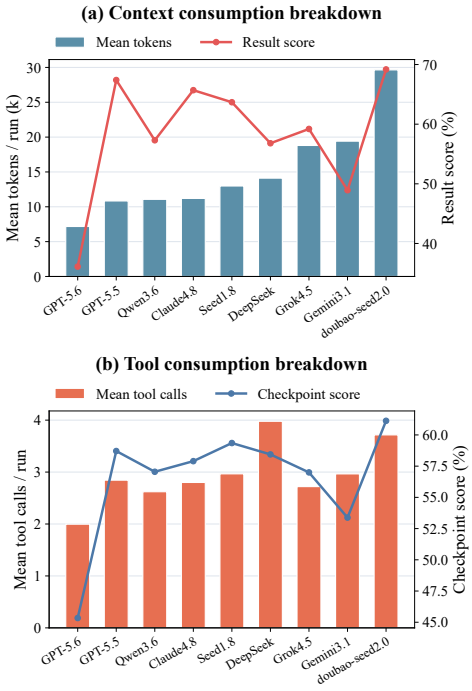


Figure 16: Model level efficiency breakdown. **(a)** Mean context consumption and result score for nine systems. **(b)** Mean successful tool calls and checkpoint score for the same systems. The associations are descriptive rather than causal.

Future work. These limitations suggest three directions. First, we plan to substitute synthetic proxies with de identified longitudinal records under appropriate governance. Second, we intend to generalize m into graded coverage that weights families by their annotated informativeness on each task, while preserving the non inflation property of Proposition 1. Third, we will study interventions that directly target *tool tunnel vision*, including retrieval forcing curricula and reward shaping on the coverage axis, to test whether the gap between valid tool ordering and genuine cross source integration can be closed rather than merely measured.