

RouteVerify: A Diagnostic Study of Tool-Augmented Chest X-ray Agents

Haochen Liu^{1,2}, Yukuo Li³, Xia Liu^{1,2}, Qingwei Wang^{1,2},
Shourong Xin^{1,2}, Zhuohui Liu^{1,2}, Long Zhao^{1,2,*}

¹Key Laboratory of Computing Power Network and Information Security,
Ministry of Education, Shandong Computer Science Center
(National Supercomputer Center in Jinan),

Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

²Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing,
Shandong Fundamental Research Center for Computer Science, Jinan, China

³Shandong University of Science and Technology, Qingdao, China
10431250133@stu.qlu.edu.cn, zhaolong@qlu.edu.cn

Abstract

Tool-augmented medical agents promise clinical control and interpretability, yet their specialized outputs are often invoked broadly and compressed into text. It remains unclear whether failures arise from this textual interface or low-fidelity upstream tools. We study this question through RouteVerify, a transparent diagnostic sandbox combining a lightweight linear router, compact evidence cards, and verifier-gated multimodal reinspection. On matched and held-out ChestAgentBench splits, RouteVerify improves answer accuracy by about seven points over naive always-call-tools execution while roughly halving token use. However, it remains about two points below a direct VLM-only baseline and uses nearly twice the tokens. Oracle probes clarify the bottleneck: replacing predicted evidence with benchmark-side gold explanations raises the same downstream text reasoner above 95 percent accuracy, or above 91 percent after removing explicit answer cues. Thus, text cards can express sufficient evidence on this benchmark; the dominant practical limitation is upstream tool quality. An external JSRT/SCR audit supports this diagnosis, revealing limited nodule-classification discrimination and structure-dependent segmentation quality. The remaining raw-versus-sanitized oracle gap also indicates sensitivity to human-authored cues. We present RouteVerify as an empirical boundary study of where current tool-augmented medical agents fail.

1 Introduction

Chest X-ray analysis remains one of the most pervasive imaging examinations in daily clinical practice. More broadly, modern visual question answering and vision-language modeling have evolved from early generic VQA benchmarks (Antol et al., 2015; Hudson and Manning, 2019) to large-scale contrastive and instruction-following multimodal

models (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023b). Within medical imaging, this trajectory has rendered direct chest X-ray question answering increasingly feasible, with recent systems showing strong zero-shot or few-shot capabilities (Li et al., 2023a; Moor et al., 2023; Lee et al., 2025; Bannur et al., 2024; Chen et al., 2024). Parallel to this end-to-end paradigm, a prominent trend in the medical AI community has emerged: constructing tool-augmented reasoning agents (Fallahpour et al., 2025). By integrating general vision-language backbones with specialized medical pipelines—such as dedicated classifiers and segmenters—these modular agents promise enhanced observability, precise evidence extraction, and more controllable clinical reasoning paths.

However, this modular architecture also introduces an underexamined operational paradox. Once a case is equipped with heterogeneous local tools, the system must decide which tools to invoke, how to format their outputs, and when to bypass them for raw multimodal re-inspection. Existing multi-tool agents predominantly adopt a crude scheduling strategy, executing all available tools for every query and forwarding verbose outputs downstream (Fallahpour et al., 2025). More generally, recent agent work has shown that LLMs can plan over tools, APIs, and environments (Yao et al., 2023; Schick et al., 2023; Li et al., 2023c; Qin et al., 2024; Patil et al., 2024; Liu et al., 2024), while adaptive-computation and dynamic-routing research has argued that expensive modules should be invoked selectively rather than uniformly (Jeong et al., 2024; Wang et al., 2018; Xin et al., 2020; Elbayad et al., 2020; Yin et al., 2022). Yet these ideas have not been carefully examined for clinically complex, case-level medical reasoning. This gap raises a basic question for the current paradigm: when specialized tools are mediated through intermediate text, do they actually deliver a net advantage over direct end-to-end visual inference?

*Corresponding author.

Rather than engineering another complex agent to claim a marginal gain, we treat tool scheduling itself as the object of diagnosis. We establish a transparent, controlled empirical sandbox named ROUTEVERIFY to examine the operational boundary and representational bottlenecks of modular medical reasoning. To isolate the effect of representation alignment from confounding model capacity, we intentionally avoid heavy new neural components. Instead, ROUTEVERIFY uses a minimalist linear router (TF-IDF features with Logistic Regression), compact text-based evidence cards, and a verifier-gated multimodal escalation path. This design functions as a controlled testbed for tracing how medical context is preserved, compressed, or lost as it moves across modalities.

Our empirical evaluations on ChestAgentBench reveal a clear pattern. Under a runnable specialized tool subset (classification and segmentation), ROUTEVERIFY substantially reduces the damage from naive always-call-tools execution, moving from 0.4600 to 0.5500 accuracy on a matched 300-question subset while reducing average token usage from 1845.64 to 888.75. The same pattern persists on a 749-question held-out split, where accuracy moves from 0.4860 to 0.5527 and tokens drop from 1820.28 to 872.48. Crucially, however, a direct VLM-only baseline from the same API family still performs better, reaching 0.5728 accuracy with only 447.33 tokens.

The resulting diagnosis is more specific than our initial hypothesis. Oracle explanation probes show that when the same downstream text reasoner receives high-fidelity dataset-side evidence cards, it reaches 0.9533 accuracy in a raw setting and 0.9167 after explicit removal of answer-revealing cues. The text interface therefore has ample semantic bandwidth for this benchmark. The dominant failure source lies upstream: low-accuracy local tools produce noisy, weakly aligned evidence cards, and the downstream model cannot reliably recover from that noise. The remaining 3.66-point gap between raw and sanitized oracle cards further suggests some shortcut sensitivity to human-authored phrasing, consistent with a broader literature on shortcut learning and spurious linguistic priors in both language and multimodal models (Geirhos et al., 2020; McCoy et al., 2019; Agrawal et al., 2018; Dancette et al., 2021). Our results therefore shift the narrative away from blind tool accumulation and away from blaming textual mediation itself; the harder problem is how to generate evi-

dence cards whose factual quality is competitive with direct multimodal perception.

In summary, the core conceptual and empirical contributions of this work are threefold:

- We present ROUTEVERIFY, a transparent diagnostic sandbox for studying tool-mediated chest X-ray reasoning under explicit accuracy, token, latency, and tool-usage constraints.
- Through matched-subset and held-out evaluation on ChestAgentBench, we document a persistent gap in which direct end-to-end VLM inference remains stronger and cheaper than the tested tool-mediated pipelines.
- We provide route-level, category-level, and rescue/harm diagnostics that make the current boundary of text-mediated tool use explicit and motivate future work on richer multimodal interfaces and uncertainty-aware arbitration.

2 Related Work

Recent chest X-ray VLMs increasingly support both radiology report generation and visual question answering directly from images (Li et al., 2023a; Moor et al., 2023; Lee et al., 2025; Banur et al., 2024; Chen et al., 2024; Zhang et al., 2025; Zhao and Liu, 2026), following a broader arc from generic VQA and visual reasoning benchmarks (Antol et al., 2015; Hudson and Manning, 2019) to large-scale multimodal pretraining (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023b). Recent report-generation methods further explore multi-scale cross-modal semantic alignment to connect visual evidence with specialized medical language (Zhang et al., 2025; Zhao and Liu, 2026). In parallel, tool-augmented agents, from ReAct and Toolformer (Yao et al., 2023; Schick et al., 2023; Madaan et al., 2023; Shinn et al., 2023) to more explicit tool-use benchmarks and API-centric agent systems (Li et al., 2023c; Qin et al., 2024; Patil et al., 2024; Liu et al., 2024), combine large models with external tools for more explicit reasoning traces. Medical systems such as MedRAX (Fallahpour et al., 2025) instantiate this paradigm in radiology, while JointCoder shows how a medical task can be decomposed into stage-wise multi-agent clinical workflows (Liu et al., 2026). Complementarily, the Knowledge-Evolvable Assistant augments an LLM with retrieved experience, reflection, and external medical

knowledge for Chinese patient guidance, illustrating knowledge-mediated clinical reasoning beyond imaging (Lu et al., 2025). More directly related to our routing focus, work on adaptive computation and dynamic routing argues that expensive modules should be invoked selectively rather than uniformly (Jeong et al., 2024; Wang et al., 2018; Xin et al., 2020; Elbayad et al., 2020; Yin et al., 2022). Finally, our diagnostic framing is also informed by work on shortcut learning and modality-specific biases, which has shown that strong benchmark performance can be driven by spurious heuristics rather than faithful evidence use (Geirhos et al., 2020; McCoy et al., 2019; Agrawal et al., 2018; Dancette et al., 2021). Our study sits at the intersection of these strands and asks a narrower question: once medical tool outputs are converted into text and re-consumed downstream, does the resulting pipeline preserve enough visual information to justify its additional cost?

3 Method

3.1 Problem Definition

Let a chest X-ray case be represented by an image set $\mathcal{I} = \{I_1, \dots, I_m\}$ and a multiple-choice question q with candidate answer set $\mathcal{A} = \{A, B, C, D, E, F\}$. For a ground-truth answer $y \in \mathcal{A}$, the agent predicts a final answer $\hat{y} \in \mathcal{A}$, possibly using a pathology classifier and a segmentation model in addition to direct multimodal inference. Routing selects a computation route $r = \pi(q, \mathcal{I}) \in \mathcal{R}$ under budgets on token cost, latency, and tool calls:

$$\begin{aligned} \max_{\pi} \quad & \mathbb{E}[\mathbf{1}(\hat{y} = y)] \\ \text{s.t.} \quad & \mathbb{E}[\tau] \leq B_{\tau}, \\ & \mathbb{E}[\ell] \leq B_{\ell}, \\ & \mathbb{E}[n] \leq B_n, \end{aligned} \tag{1}$$

where B_{τ} , B_{ℓ} , and B_n denote the budgets for token cost, latency, and external tool calls. The goal is therefore not a new perception backbone, but a routing study over an existing tool set.

3.2 Framework Overview

Figure 1 shows the overall pipeline. The router first predicts whether segmentation and multimodal verification are worth paying for, after which the selected tools are run and compressed into compact evidence cards. A text reasoner answers from the cards first, and a multimodal verifier re-inspects

the original image only when confidence is insufficient. Tool outputs are cached at the case level for repeated questions. The design is intentionally simple so the experiments isolate where tool mediation helps and where it only adds overhead.

3.3 Structured Evidence Cards

Raw tool outputs are verbose and heterogeneous, so we convert them into structured evidence cards. The classification card keeps the top pathology, classifier prediction score, top findings, and a top-1/top-2 margin; the segmentation card keeps processed organs, coarse area ratios, and simple center summaries. Throughout the method section, we distinguish classifier *prediction scores*, router *route probabilities*, and model *answer confidence*. The key purpose of this design is diagnostic: it creates a controlled textual bottleneck through which we can test how much visual evidence survives tool mediation.

3.4 Lightweight Routing

A naive multi-tool agent invokes both tools for every question even when spatial analysis is unnecessary. We instead map the question into TF-IDF features $\mathbf{x}$ and use two logistic regressions to predict whether segmentation is useful (p_{seg}) and whether multimodal verification should be available (p_{mm}). The main route family contains four actions: `verify_cls_lo`, `verify_cls_hi`, `verify_cls_seg_lo`, and `verify_cls_seg_hi`, where low/high indicate different verification priors.

3.5 Verifier-Gated Multimodal Escalation

If every routed example is always escalated, the system loses efficiency; if escalation is never allowed, text-side reasoning fails on harder spatial or relational questions. We therefore let the text reasoner first output a provisional answer, an answer-confidence estimate, and an escalation flag. When confidence is insufficient, a multimodal verifier directly re-inspects the original image and replaces the draft answer. In this paper, the verifier is both a fallback and a diagnostic probe: it reveals how much information the text-mediated route fails to preserve and must recover through direct visual access.

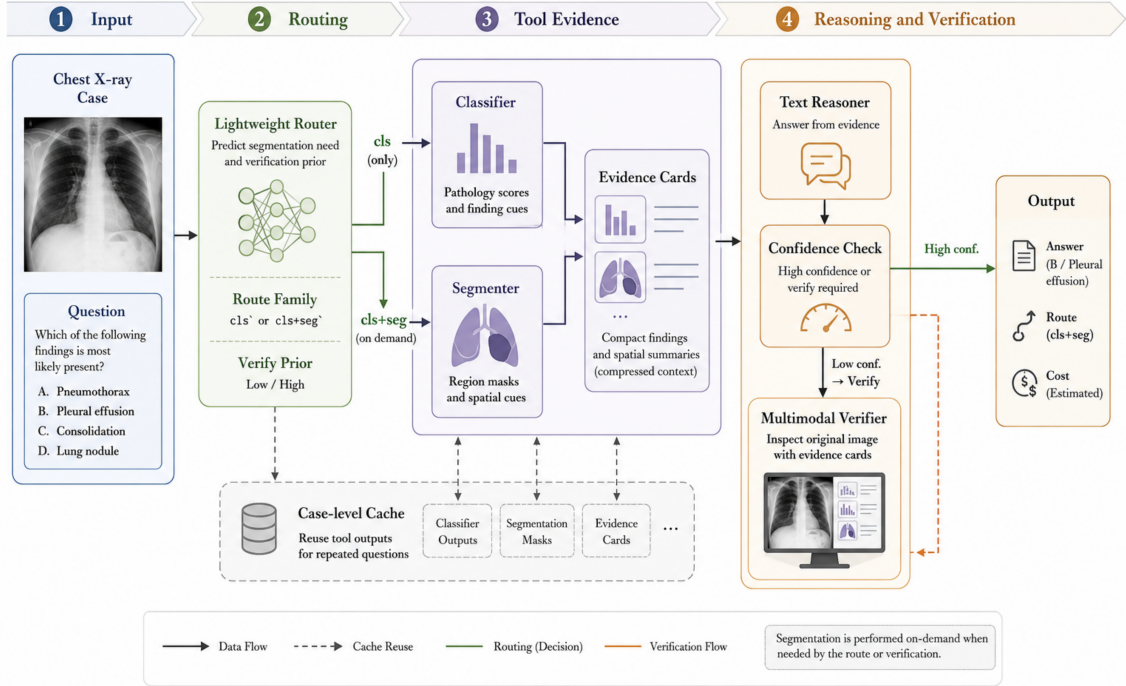


Figure 1: Overview of ROUTEVERIFY. The router selects whether segmentation is used and whether multimodal verification is likely to be needed; tool outputs are compressed into evidence cards; and multimodal reinspection is triggered only when the text-side answer is insufficiently confident.

4 Experiments

4.1 Experimental Setup

We study three questions: how much selective scheduling recovers over naive tool chaining (**RQ1**), which pipeline components matter most (**RQ2**), and what larger-scale evaluation plus negative results reveal about the current boundary of tool-augmented chest X-ray reasoning (**RQ3**). We evaluate on ChestAgentBench with a 70/30 case_id-level split, report a matched 300-question subset and a held-out 749-question split, and reserve 10- and 25-question subsets for smoke tests. Metrics are answer accuracy, tokens, latency, tool count, segmentation rate, and escalation rate.

Main comparisons use VLM-only, AllTools-Raw, and ROUTEVERIFY. Targeted ablations add AllTools-Struct, AllTools-Verify-Struct, and RouteStruct, while exploratory variants include EvidenceVerify, CalibVerify, and several VLM-first hybrids. The text backend is qwen-max; both the direct multimodal baseline and the verifier use qwen-vl-plus. We evaluate a verified local tool subset consisting of chest X-ray classification and

segmentation; the grounding tool is excluded because of dependency and offline model-resolution issues. Unless otherwise noted, the random seed is 42.

4.2 Main Results

Matched 300-question subset (RQ1). Table 1 reports the main controlled comparison on both the matched 300-question subset and the held-out 749-question split.

Relative to AllTools-Raw, ROUTEVERIFY recovers 9.0 absolute accuracy points on matched 300 while cutting average tokens by 51.8%. Yet it still remains 1.0 point below VLM-only (0.5500 vs. 0.5600) while using roughly twice the tokens. Selective scheduling therefore repairs much of the damage from naive tool chaining, but does not erase the gap to direct end-to-end multimodal inference.

Figure 2 shows the same pattern visually: AllTools-Raw is strictly dominated, while ROUTEVERIFY narrows the gap to VLM-only without eliminating it.

Method	Matched 300			Held-out 749		
	Acc.	Tokens	Time (s)	Acc.	Tokens	Time (s)
VLM-only	0.5600	444.86	0.4212	0.5728	447.33	0.6617
AllTools-Raw	0.4600	1845.64	8.5371	0.4860	1820.28	7.0657
Ours	0.5500	888.75	7.9829	0.5527	872.48	6.7465

Table 1: Main comparison on matched and held-out evaluation.

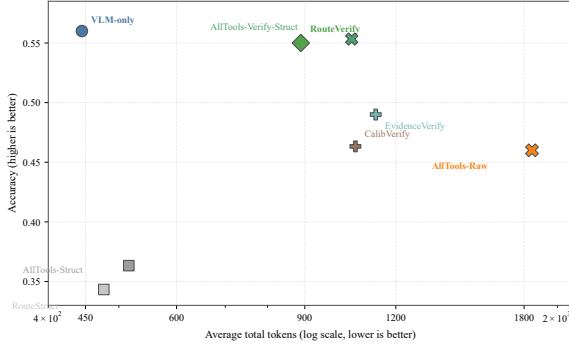


Figure 2: Accuracy-token tradeoff on matched 300.

Held-out 749-question evaluation (RQ3). The held-out results tell the same story more clearly: ROUTEVERIFY remains substantially stronger than AllTools-Raw, but it still trails VLM-only by 2.01 accuracy points (0.5527 vs. 0.5728) while using about $1.95\times$ as many tokens.

Seed robustness (RQ3). The same directional pattern also holds on two re-sampled 300-question subsets: on seed 7, ROUTEVERIFY reaches 0.5500 versus 0.5200 for AllTools-Raw; on seed 123, it reaches 0.5367 versus 0.4900.

Alternative scheduler controls (RQ1). To test whether the result is specific to the linear router, we hold the downstream *routed-verify* pipeline fixed and replace only its scheduler. A pseudo-random router matches the coarse segmentation and escalation rates of ROUTEVERIFY, a rule router uses hand-written spatial and diagnostic keywords, and an LLM router uses qwen-max on question text alone. We also construct a stricter random control that exactly matches the learned router’s four route-family counts (16/85/60/139 for *verify_cls_lo*, *verify_cls_hi*, *verify_cls_seg_lo*, and *verify_cls_seg_hi*) and average it over five seeds. Table 2 shows that the learned router remains both more accurate and more token-efficient. The rate-matched result is particularly informative: matching route frequencies does not recover question-conditioned route selection.

4.3 Ablation Studies

Structured evidence and routing (RQ2). Table 3 disentangles the effects of structured evidence compression, verifier-enabled escalation, and routing on the 300-question subset.

Table 3 isolates three effects. Compression alone is not enough: replacing raw tool outputs with structured evidence cuts tokens sharply, but accuracy drops from 0.4600 to 0.3633, and routing without verification falls further to 0.3433. Multimodal verification is the main source of recovery: adding it to the structured always-tools pipeline raises accuracy to 0.5533, essentially matching ROUTEVERIFY at 0.5500. Once verification is available, routing mainly reduces unnecessary segmentation and prompt budget.

Decoupling Modality and Noise via Oracle Explanations. The oracle rows in Table 3 provide a direct control for separating modality from evidence quality. We replace tool-predicted cards with dataset-side gold explanations while keeping the downstream interface textual and keeping the same answer format. Under this intervention, accuracy jumps from 0.3433 for RouteStruct and 0.3633 for AllTools-Struct to 0.9533 with a raw gold card and 0.9167 with a sanitized gold card. This gap is too large to attribute to routing alone. It shows that the text card has sufficient semantic bandwidth to support high-quality clinical reasoning when the evidence it carries is faithful. The limiting factor in the low-performing structured pipelines is therefore not that “text cannot express medical image evidence,” but that the local classifier and segmenter populate the card with noisy, incomplete, and sometimes misleading signals. In that sense, the downstream model is not starved by textual compression; it is starved, and occasionally poisoned, by weak upstream predictions. The raw-versus-sanitized gap is also informative. Removing overt answer cues lowers accuracy by 3.66 points, from 0.9533 to 0.9167, while still leaving performance far above any deployable tool pipeline. We interpret this residual drop as evidence of shortcut

Scheduler	Accuracy	Avg. Tokens	Avg. Tools	Seg. Rate	Esc. Rate
Learned ROUTEVERIFY router	0.5500	888.75	1.66	0.6633	0.8833
Random	0.4633	1584.43	1.41	0.5500	0.9700
Rule	0.4700	1685.36	1.63	0.7733	0.9700
LLM	0.4733	1696.58	1.68	0.8200	0.9667
Rate-matched random (5 seeds)	0.4560 $\pm$ 0.0152	1593.56	1.43	0.5713	0.9700

Table 2: Alternative scheduler controls on the matched 300-question subset. The strict random control matches the sampled four-way route distribution exactly; its realized segmentation rate can still be reduced by tool-applicability filtering.

Method	Accuracy	Avg Total Tokens	Avg Duration (s)
AllTools-Raw	0.4600	1845.64	8.5371
AllTools-Struct	0.3633	516.01	4.9120
AllTools-Verify-Struct	0.5533	1043.79	7.2078
RouteStruct	0.3433	476.75	5.6080
Ours	0.5500	888.75	7.9829
Raw Gold	0.9533	527.62	4.1997
Sanitized Gold	0.9167	527.10	4.1644

Table 3: Core ablation on matched 300, augmented with oracle explanation controls. The first five rows are runnable pipelines built from the available local tools and show that verification recovers much of the loss introduced by naive tool chaining, but that purely tool-derived structured cards remain fragile. The two gray rows replace predicted tool outputs with dataset-side gold textual explanations while keeping the downstream text reasoning interface fixed. Their large gains show that the evidence-card modality itself is not the main bottleneck; the decisive issue is the factual quality of the upstream evidence that populates the card.

sensitivity to human-authored linguistic traces. The sanitized result is therefore the more meaningful ceiling for a text-mediated medical-agent pipeline that must rely on medical content rather than annotation style.

Verifier design (RQ2). Changing the verifier does not alter the conclusion. Relative to Ours at 0.5500 accuracy and 888.75 tokens, EvidenceVerify drops to 0.4900 with 1126.09 tokens, and CalibVerify drops further to 0.4633 with 1056.10 tokens. Offline replay with stricter escalation thresholds also produces only marginal changes. Prompt-side verifier refinements therefore do not address the dominant failure mode, which the oracle explanation probe identifies as upstream evidence quality rather than insufficient textual bandwidth.

Evidence-budget-matched raw control (RQ2). We additionally test whether AllTools-Raw is weak merely because its JSON evidence is long. ALLTOOLS-RAW-BUDGETMATCHED keeps the same tools and text reasoner but truncates raw evidence to approximately the structured-card budget. Under the same current runtime, the full raw re-run reaches 0.4500 accuracy with 1956.73 tokens, whereas the budget-matched variant falls to 0.3567 with 462.87 tokens; AllTools-Struct is similarly

short (516.01 tokens) and reaches 0.3633. Thus verbosity is not the dominant explanation: naive truncation removes useful signal without repairing evidence quality, while multimodal verification is what restores performance to 0.5533 for AllTools-Verify-Struct.

4.4 Diagnostic Analysis

Question-type behavior. Category-level results show that the behavior of ROUTEVERIFY is not uniform across reasoning patterns.

Figure 3 shows that ROUTEVERIFY improves over AllTools-Raw on six of the seven MedRAX categories, with the largest shifts on localization, diagnosis, and characterization. Relative to VLM-only, it remains close across categories, which supports the paper’s main claim: routed tool use reduces much of the inefficiency of naive always-call-tools execution, but only rarely exceeds a strong direct VLM answer.

Method disagreement analysis. Pairwise overlap further clarifies the discussion. On the matched 300-question subset, ROUTEVERIFY answers 52 questions correctly that AllTools-Raw misses, whereas AllTools-Raw answers only 25 that ROUTEVERIFY misses; on the held-out 749-question split, this margin widens to 114 versus 64.

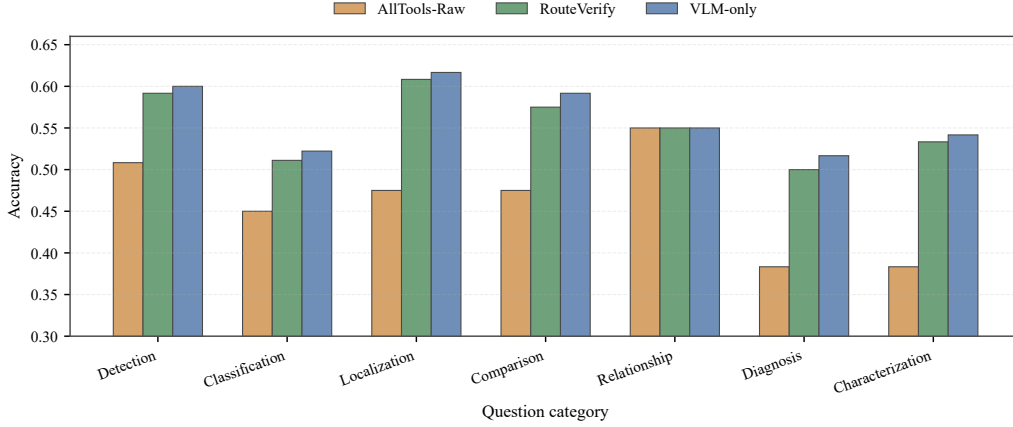


Figure 3: Category-wise accuracy on matched 300.

Split	matched 300	held-out 749
Pre-verify text accuracy	0.3733	0.4126
Final routed accuracy	0.5500	0.5527
Escalation rate	0.8833	0.8812
Rescued cases among escalated	69 (26.0%)	148 (22.4%)
Harmed cases among escalated	16 (6.0%)	43 (6.5%)
Net corrected cases	+53	+105

Table 4: Internal rescue/harm accounting for ROUTEVERIFY.

Relative to VLM-only, however, the gap reverses to 8 versus 5 on the matched subset and 22 versus 7 on the held-out split. The text-side answer before multimodal escalation reaches only 0.3733 on the matched subset and 0.4126 on the held-out split, so verification is doing meaningful work inside the routed pipeline.

Table 4 makes this dependence explicit. The text-only stage is weak on its own, and verification is responsible for a large share of the routed system’s final accuracy. On both splits, about 88% of cases are escalated and verification yields a positive net gain, but it also reverses 6.0%–6.5% of escalated cases. The routed agent therefore depends on verification to recover information lost in text-mediated tool summaries, while the verifier itself remains imperfect.

Intervention rescue/harm relative to direct VLM. The preceding accounting asks whether the verifier improves the internal text draft. A stricter diagnostic instead treats VLM-only as the default and asks whether replacing its answer with the final tool-mediated answer would help. On matched 300, only 5 interventions are beneficial, 8 are harmful, and 287 preserve the same correctness outcome; on held-out 749, the correspond-

ing counts are 7, 22, and 720. A gate trained on matched 300 and selected for held-out accuracy chooses no interventions at its best threshold, thereby reverting to VLM-only (0.5728 accuracy and 447.33 tokens). Tool use therefore rescues many weak *text-card drafts*, but offers only a sparse and difficult-to-identify rescue region over a strong direct multimodal default. This distinction explains why verification is valuable inside ROUTEVERIFY without making the complete tool pipeline superior to VLM-only.

External audit of upstream tools. To directly test the evidence-quality diagnosis, we evaluate the exact classifier and anatomical-structure segmenter used by ROUTEVERIFY on all 247 clean JSRT DICOM radiographs. On the only directly overlapping image-level target, Nodule, the classifier reaches 0.6232 AUROC and 0.5789 accuracy. Segmentation Dice is structure-dependent, ranging from 0.6291–0.6308 for the clavicles to 0.7486/0.8671 for the left/right lungs and 0.8809 for the heart. These independent ground-truth measurements support the conclusion that imperfect tool outputs can contaminate downstream evidence cards. Appendix A reports the full audit scope and cautions that this external evaluation does not by itself establish case-level downstream causality on ChestAgentBench.

Targeted robustness repair as a boundary test. We also tested a modality-aware applicability gate that suppresses chest-only tools on visually incompatible inputs. Its effect is negligible or negative: gating both classification and segmentation hurts matched accuracy (0.5400→0.5200), while segmentation-only gating changes matched accu-

racy only marginally (0.5400→0.5433) and slightly lowers held-out accuracy (0.5527→0.5501). Input-tool mismatch is therefore real, but it is not the dominant source of the remaining gap to direct VLM inference.

Exploratory reliability probes. Small-sample auxiliary probes also do not show a reliability advantage: on 50 jointly evaluable examples, the unsupported visual claim rate is 0.52 for ROUTEVERIFY versus 0.50 for VLM-only, and on a 100-example calibration probe the expected calibration error is 0.4205 versus 0.3965.

5 Conclusion

This paper studies tool scheduling as a diagnostic lens on the current multi-tool medical-agent paradigm. Across both matched and held-out evaluation, ROUTEVERIFY consistently reduces the deficit of naive always-call-tools execution, but it still does not surpass a strong direct multimodal baseline. Oracle explanation probes sharpen the diagnosis: text evidence cards are not themselves the core bottleneck. When the downstream text reasoner is given high-fidelity dataset-side explanations, accuracy rises to 0.9533 in a raw setting and 0.9167 after removal of overt answer cues, showing that textual mediation is expressive enough for this benchmark. The main practical bottleneck is therefore the quality of the upstream tools that populate the card, not the existence of the card itself. The residual raw-versus-sanitized gap further suggests some shortcut sensitivity to human-authored cues. Future work should therefore focus less on abandoning text interfaces and more on improving evidence extraction quality, grounding, and shortcut-resistant supervision for medical-agent pipelines.

Limitations

This study has several limitations. First, the experiments are restricted to the currently runnable tool subset, namely classification and segmentation; a stronger localization or grounding tool was not available in the present environment, so our conclusions should not be read as a final verdict on all possible medical-agent tool stacks. Second, under real tool outputs, the routed system still remains slightly below the native VLM baseline while incurring higher token and latency cost. We interpret this result as evidence about the present fidelity of the upstream tools rather than as a blanket rejection of text-mediated evidence cards, especially in

light of the oracle explanation controls. Third, the main answer-level comparisons are conducted on a single benchmark. Appendix A reports a narrower external JSRT/SCR audit of the classifier and segmenter, but it covers only the Nodule output and five annotated structures and does not establish case-level downstream associations or substitute for multi-center validation across institutions and reporting styles.

Ethics Statement

This work is a research study on medical question answering and should not be used for autonomous clinical decision making. The evaluated systems may produce incorrect, incomplete, or overconfident answers. Their outputs should therefore be treated only as experimental benchmark results rather than clinical recommendations. Any real-world deployment would require substantially stronger validation, calibration, safety constraints, governance, and human oversight.

Acknowledgements

This study was supported by Shandong Provincial Natural Science Foundation (ZR2024MF088), the National Natural Science Foundation of China (No.61806105), Key Research and Development Program of Shandong Province (Major Science and Technology Innovation Project) (No. 2026CXGC010604), Program of New Twenty Policies for Universities of Jinan (No.202333008), the Pilot Project for Integrated Innovation of Science, Education, and Industry of Qilu University of Technology (Shandong Academy of Sciences) (2024ZDZX08) and the Pilot Project for Integrated Innovation of Science, Education, and Industry of Qilu University of Technology (Shandong Academy of Sciences) (No.2025ZDZX01).

References

- Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. [Don’t just assume; look and answer: Overcoming priors for visual question answering](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4971–4980.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian

- Borgeaud, and 8 others. 2022. [Flamingo: A visual language model for few-shot learning](#). In *Advances in Neural Information Processing Systems* 35.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. [VQA: Visual question answering](#). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.
- Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaurya Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, and 2 others. 2024. [MAIRA-2: Grounded radiology report generation](#). *arXiv preprint arXiv:2406.04449*.
- Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis P. Langlotz. 2024. [CheXagent: Towards a foundation model for chest X-ray interpretation](#). In *AAAI 2024 Spring Symposium Series: Clinical Foundation Models*.
- Corentin Dancette, Rémi Cadène, Damien Teney, and Matthieu Cord. 2021. [Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1574–1583.
- Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. 2020. [Depth-adaptive transformer](#). In *International Conference on Learning Representations*.
- Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. 2025. [MedRAX: Medical reasoning agent for chest X-ray](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 15661–15676. PMLR.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. [Shortcut learning in deep neural networks](#). *Nature Machine Intelligence*, 2:665–673.
- Drew A. Hudson and Christopher D. Manning. 2019. [GQA: A new dataset for real-world visual reasoning and compositional question answering](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709.
- Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. [Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.
- Seowoo Lee, Jiwon Youn, Hyungjin Kim, Mansu Kim, and Soon Ho Yoon. 2025. [CXR-LLaVA: A multi-modal large language model for interpreting chest X-ray images](#). *European Radiology*, 35:4374–4386.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. [LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day](#). In *Advances in Neural Information Processing Systems* 36.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. [BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023c. [API-Bank: A comprehensive benchmark for tool-augmented LLMs](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3102–3116, Singapore. Association for Computational Linguistics.
- Kangjun Liu, Zhenyu Li, Jianlei Wang, Hongjiao Guan, Ying Lian, Guoqiang Chen, Tao Xin, and Wenpeng Lu. 2026. [JointCoder: Exploring automated ICD coding on real-world Chinese EHRs with a multi-agent framework](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 263–274, San Diego, California, United States. Association for Computational Linguistics.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. [AgentBench: Evaluating LLMs as agents](#). In *International Conference on Learning Representations*.
- Wenpeng Lu, Kangjun Liu, Jianlei Wang, Xueping Peng, Tao Shen, Fa Zhu, Weiyu Zhang, Jiabing Zhu, Tao Xin, and Athanasios V. Vasilakos. 2025. [Advancing Chinese conversation-based patient guidance with a benchmark and knowledge-evolvable assistant](#). *IEEE Journal of Biomedical and Health Informatics*. Online ahead of print.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon,

- Nouha Dziri, Shrimai Prabhunoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-Refine: Iterative refinement with self-feedback](#). In *Advances in Neural Information Processing Systems* 36.
- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. [Med-Flamingo: A multimodal medical few-shot learner](#). In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 353–367. PMLR.
- Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2024. [Gorilla: Large language model connected with massive APIs](#). In *Advances in Neural Information Processing Systems* 37.
- Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. [ToolLLM: Facilitating large language models to master 16000+ real-world APIs](#). In *International Conference on Learning Representations*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language models can teach themselves to use tools](#). In *Advances in Neural Information Processing Systems* 36.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflexion: Language agents with verbal reinforcement learning](#). In *Advances in Neural Information Processing Systems* 36.
- Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Kenji Komatsu, Mitsuhiro Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. 2000. [Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules](#). *American Journal of Roentgenology*, 174(1):71–74.
- Bram van Ginneken, Mikkel B. Stegmann, and Marco Loog. 2006. [Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database](#). *Medical Image Analysis*, 10(1):19–40.
- Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E. Gonzalez. 2018. [SkipNet: Learning dynamic routing in convolutional networks](#). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 409–424.
- Ji Xin, Raphael Tang, Jaesun Lee, Yaoliang Yu, and Jimmy Lin. 2020. [DeeBERT: Dynamic early exiting for accelerating BERT inference](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2246–2251, Online. Association for Computational Linguistics.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing reasoning and acting in language models](#). In *International Conference on Learning Representations*.
- Hongxu Yin, Arash Vahdat, Jose M. Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. 2022. [A-ViT: Adaptive tokens for efficient vision transformer](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10809–10818.
- Zhihao Zhang, Long Zhao, Dun Lan, Yuyao Wang, Linfeng Jiang, and Xiangjun Dong. 2025. [Multi-scale based cross-modal semantic alignment network for radiology report generation](#). In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1538–1543, Vienna, Austria. IEEE.
- Long Zhao and Xia Liu. 2026. [Radiology report generation via multi-path modeling with medical language priors and semantic guidance](#). In *ICASSP 2026–2026 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain. IEEE.

A External Tool-Quality Audit

The main experiments infer an upstream evidence-quality bottleneck from end-to-end behavior and oracle text controls. To assess this bottleneck directly, we audit the exact classifier and anatomical-structure segmenter used by ROUTEVERIFY on all 247 clean JSRT DICOM radiographs, comprising 154 nodule-positive and 93 nodule-negative cases (Shiraishi et al., 2000; van Ginneken et al., 2006). We retain the same model objects, pre-processing procedures, output indices, and fixed thresholds as in the main pipeline. We use the clean original DICOM images and exclude JPEG versions containing burned-in contours.

JSRT/SCR is suitable for this audit because the same radiographs provide both case-level pulmonary-nodule labels and independent anatomical contours. Nodule is the only JSRT image-level label that directly matches a classifier output. SCR provides ground-truth contours for five overlapping segmentation structures: the left lung, right lung, heart, left clavicle, and right clavicle. The lungs and heart are exposed in the ROUTEVERIFY evidence card, while the clavicles are included as smaller, thinner, and more challenging structures. The card-exposed mediastinum cannot be independently evaluated because SCR provides no corresponding annotation.

Classifier audit. Table 5 reports performance for the Nodule output at the same fixed 0.5 threshold used by the audit pipeline. Discrimination is limited: AUROC is 0.6232, accuracy is 0.5789, and both false-positive and false-negative errors are substantial.

Target	AUROC	AUPRC	Accuracy	Sens.	Spec.	F1
Nodule	0.6232	0.7631	0.5789	0.5519	0.6237	0.6204

Table 5: JSRT classifier audit on 247 radiographs. The confusion matrix is TP=85, FP=35, TN=58, and FN=69.

Segmentation audit. Table 6 shows that segmentation fidelity is strongly structure-dependent. Dice decreases to 0.7486 for the left lung and approximately 0.63 for the smaller clavicles. Even among large card-exposed structures, the asymmetry between left- and right-lung performance indicates that geometric evidence should not be treated as uniformly reliable.

These findings provide direct external evidence that upstream tool fidelity is a major bottleneck

Structure	Dice
Left lung	0.7486
Right lung	0.8671
Heart	0.8809
Left clavicle	0.6308
Right clavicle	0.6291

Table 6: SCR segmentation audit on the five independently annotated structures that overlap the deployed segmenter.

in the evaluated configuration. Classification errors and structure-dependent segmentation errors are converted into textual findings and geometric summaries and can therefore propagate through the evidence card. This interpretation is consistent with the oracle evidence control, where replacing predicted evidence with high-fidelity benchmark-side explanations raises downstream accuracy to 0.9533, or 0.9167 after explicit answer cues are removed.

We restrict this conclusion to the audited Nodule output and five independently annotated anatomical structures. The experiment does not validate all classifier labels or all 14 segmentation channels. Because JSRT/SCR is external to ChestAgentBench, this audit measures independent tool fidelity but does not by itself establish case-level downstream associations.

B Prompt and Evidence Templates

Routing note. The main ROUTEVERIFY router is a learned logistic-regression model over TF-IDF question features rather than an LLM prompt. The prompts below therefore cover only the modules that invoke text or multimodal models.

B.1 Structured Evidence Card

The structured evidence card is a newline-delimited key-value summary produced from classifier and segmenter outputs. The main fields used in the routed system are:

- Classification card: classification_top_pathology, classification_top_score, classification_top_findings, classification_margin_top1_top2.
- Per-image classification card: classification_image_i_top_pathology, classification_image_i_top_findings, classification_image_i_margin_top1_top2.

- Segmentation card: segmentation_processed_organ, segmentation_lung_area_ratio_left_to_right, and organ summaries of the form Organ:center=(...), conf=...

A typical evidence card passed to the text reasoner looks as follows:

```
classification_top_pathology: Lung
Opacity
classification_top_score: 0.68
classification_top_findings: Lung
Opacity=0.68, Mass=0.60, Hernia=0.57
classification_margin_top1_top2: 0.08
segmentation_processed_organ: Left Lung,
Right Lung, Heart, Mediastinum
segmentation_lung_area_ratio_left_to_
right: 1.0389
LeftLung: center=(0.50,0.75), conf=0.31
RightLung: center=(0.47,0.28), conf=0.30
Heart: center=(0.68,0.58), conf=0.18
Mediastinum: center=(0.21,0.53), conf=0.07
```

The AllTools-Raw baseline uses the same downstream text reasoner but replaces this card with a raw JSON serialization of tool outputs.

B.2 Text Reasoner Prompt

System prompt.

You answer Chest X-ray multiple-choice questions from tool evidence. You are allowed to use only the question and the evidence card below. If the evidence is incomplete or ambiguous, set need_visual_check to true. Return compact JSON with keys answer, confidence, need_visual_check, and rationale. Do not return markdown.

User prompt template.

```
Question:
[QUESTION]
Evidence card:
[EVIDENCE CARD]
```

Expected JSON schema.

- answer: one of A-F
- confidence: a float in [0, 1]
- need_visual_check: boolean
- rationale: one short sentence

B.3 Direct Multimodal Baseline Prompt

System prompt.

You are a medical imaging expert. Answer the multiple choice question with only one letter (A/B/C/D/E/F).

User prompt template. The user message is the original multiple-choice question. The chest X-ray image or image set is provided as visual input to the multimodal backend.

B.4 Evidence-Aware Multimodal Verifier Prompt

System prompt.

You are a medical imaging expert verifying a chest X-ray multiple-choice answer. Use the image as primary evidence. Use the draft answer and tool evidence card as auxiliary context. If the evidence card conflicts with the image, trust the image. Return only the single best answer letter (A/B/C/D/E/F).

User prompt template.

```
Question:
[QUESTION]
Draft text-only answer: [DRAFT ANSWER]
Draft confidence: [DRAFT CONFIDENCE]
Draft reasoner requested visual check:
[true/false]
Draft rationale: [DRAFT RATIONALE]
Tool evidence card:
[EVIDENCE CARD]
Use the image as primary evidence. Use the draft answer and tool evidence only as auxiliary context. If they conflict with the image, trust the image. Return only the best answer letter.
```

When explanation collection is enabled, the verifier uses the same instruction family but returns compact JSON with fields answer, confidence, rationale, and visual_evidence.

B.5 Exploratory Prompt Variants

Several exploratory variants reported in the paper reuse the same overall prompt structure with small schema changes:

- **Calibrated verifier.** The text-side reasoner is asked to additionally output option_scores, evidence_sufficiency, and contradiction_flag.

- **VLM-first hybrid.** The initial multi-modal model outputs answer, confidence, need_tool_check, rationale, and visual_evidence; tool calls are triggered only when need_tool_check is true.
- **Candidate-restricted arbitration.** The multi-modal adjudicator is instructed to choose only from the candidate letters proposed by the direct VLM branch and the tool branch.
- **Option-pruned rejudge.** The text-side module first produces a small plausible option set from tool evidence, after which the multimodal model is restricted to that subset.

These variants are implemented in the same code path and differ mainly in output schema and whether the multimodal model is asked to answer freely or only within a restricted candidate set.