DECOUPLING MEDICAL LLM REASONING WITH AGENTTOOL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have achieved strong performance on medical question answering, yet single-number accuracy obscures *where* they fail. We introduce a diagnostic evaluation framework that decomposes medical reasoning into three complementary abilities: **Knowledge**, the possession of question-relevant medical knowledge; **Recall**, the ability to retrieve such knowledge when answering a question; and **Inference**, the ability to solve the question when the required knowledge is provided in context. We instantiate this framework on MedQA, MedMCQA, and MMLU-Med with dedicated probes for each ability. Evaluating six LLMs reveals that recall tracks downstream accuracy most consistently among the three measured abilities, whereas recognized knowledge alone shows weak and sometimes negative association with correctness. We further show that standard reinforcement learning (e.g., GRPO) improves primarily medium- and high-recall questions, leaving the low-recall regime largely untouched in our setting. Motivated by these findings, we propose **AgentTool RL**, a reinforcement learning framework that addresses the low-recall bottleneck by making parametric recall an explicit, on-demand tool. The tool runs in an isolated context to generate relevant knowledge without disrupting final-answer reasoning. We further optimize reasoning and recall with decoupled group-wise objectives, using answer correctness for reasoning trajectories and a length-regularized recall-coverage reward for tool-generated recall. Across in-domain and out-of-distribution medical QA benchmarks, AgentTool RL outperforms standard GRPO and strong medical baselines on every dataset.

1 INTRODUCTION

Large language models (LLMs) have made substantial progress on medical question answering (Singhal et al., 2023; Chen et al., 2024b; Singhal et al., 2025), with leading systems approaching expert-level performance on standardized exams such as the USMLE. Yet high benchmark accuracy provides limited insight into *how* these models reason about medicine or *where* they fail. Existing medical QA benchmarks (Jin et al., 2021; Hendrycks et al., 2021; Pal et al., 2022) typically report a single end-to-end accuracy number, which conflates missing knowledge, failed parametric recall, and flawed reasoning.

This limitation matters for training. Recent progress in LLM reasoning has largely relied on reinforcement learning (RL) with verifiable rewards (Shao et al., 2024; Zheng et al., 2025; Lightman et al., 2024), which optimizes answer correctness through a scalar reward signal. Such methods implicitly assume that different failure modes can be improved through the same end-to-end objective. Whether this assumption holds in medical reasoning, where factual knowledge, recall, and inference are tightly intertwined, remains unclear.

We argue that improving medical LLMs requires decomposing medical reasoning into separately measurable abilities. We introduce a diagnostic framework with three meta-abilities: **Knowledge** (K), whether the model possesses the medical knowledge relevant to a question; **Recall** (R), whether it can spontaneously retrieve that knowledge from its parameters given only the question; and **Inference** (I), whether it can derive the correct answer when the relevant knowledge is provided. We measure these abilities using factual probes, open-ended recall, and knowledge-conditioned answering, respectively, and instantiate the framework on MedQA (Jin et al., 2021), MedMCQA (Pal et al.,

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

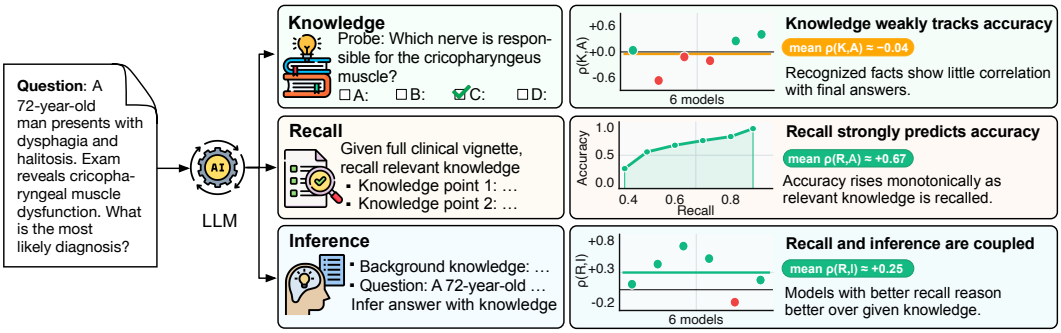


Figure 1: Overview of our diagnostic framework and empirical analysis. The left side illustrates the decomposition of medical reasoning into Knowledge, Recall, and Inference probes, while the right panel shows their correlations with downstream accuracy on MedQA. Here ρ denotes Spearman's rank correlation coefficient computed over binned evaluation results.

2022), and MMLU-Med (Hendrycks et al., 2021). Figure 1 summarizes the framework and main findings.

What the decomposition reveals. Evaluating six representative LLMs, including general-purpose and medical-specialized models, suggests that LLMs often fail not because the relevant knowledge is entirely absent under recognition probes, but because it is not reliably activated in the question context. This is reflected in the weak and sometimes negative correlation between recognized knowledge and accuracy, contrasted with the consistently positive and usually strong recall–accuracy association. Moreover, recall and inference are positively coupled, indicating that poor recall often co-occurs with weaker knowledge-conditioned reasoning. Further experiments also show that standard GRPO-based RL does not fully address this bottleneck in our setting, as its gains concentrate in the medium- and high-recall regimes while remaining limited for low-recall questions.

AgentTool RL for decoupled medical reasoning. Motivated by this diagnosis, we propose AgentTool RL, which exposes parametric knowledge recall as an explicit AgentTool that the model can invoke on demand. The tool runs in an isolated context and returns recalled knowledge as tool result for subsequent reasoning, thereby separating recall generation from answer inference. This separation enables process-specific optimization under reinforcement learning: reasoning trajectories receive an answer-correctness reward, while recall trajectories receive a length-regularized recall-quality reward. We train the resulting policy with GRPO, jointly learning to improve recall quality and to reason over recalled knowledge. Across in-domain and out-of-distribution medical QA benchmarks, AgentTool RL improves over standard GRPO by up to $+3.8$ accuracy points per dataset and yields more consistent gains across recall regimes.

Contributions. We make four contributions:

- We propose a **diagnostic evaluation framework** that decomposes medical LLM reasoning into Knowledge, Recall, and Inference, enabling fine-grained failure analysis.
- We provide an **empirical analysis across six LLMs** showing that, under this diagnostic protocol, recall is the measured ability most consistently associated with downstream accuracy, while recognized knowledge alone weakly predicts correctness; we further show that standard RL yields limited gains in the low-recall regime.
- We introduce **AgentTool RL**, a framework that equips the policy with an on-demand isolated recall tool and optimizes recall generation and final-reward reasoning with separate rewards.
- We **evaluate AgentTool RL** on in-domain and out-of-distribution medical QA benchmarks, where it outperforms standard GRPO on every dataset (average accuracy $66.2 \rightarrow 69.2$ on Qwen3-8B and $71.4 \rightarrow 73.8$ on Qwen3.5-9B), with ablations isolating the contribution of each component.

2 Diagnostic Benchmark for Medical Reasoning

Medical reasoning in LLMs involves multiple interacting abilities. A model may fail a clinical question not because it lacks the relevant knowledge, but because it fails to retrieve it when needed,

or cannot synthesize it into a correct answer. Existing medical QA benchmarks conflate these into a single accuracy metric, making it hard to diagnose *where* and *why* a model fails. To address this, we propose a diagnostic framework that decomposes medical reasoning into three meta-abilities—**Knowledge**, **Recall**, and **Inference**—with evaluation protocols built on three widely-adopted datasets: MedQA, MedMCQA, and MMLU-Med. We also report standard downstream accuracy (A). We do not claim these abilities are strictly orthogonal—in practice they share representations and partially co-vary—but operationalizing them as separately measurable quantities lets us localize failures and design targeted interventions.

2.1 KNOWLEDGE EVALUATION

The knowledge metric measures whether a model has *internalized* the factual knowledge required to answer a given question, independent of whether it can recall or apply the knowledge. For each question q, we construct a structured reference knowledge set using GPT-5-mini, asking for core pathological knowledge, competing hypotheses, discriminative cues, and common pitfalls or confounders. Since this reference set is model-generated, we apply quality control. Specifically, we sample two independent GPT-5-mini responses per question and regard a question as mastered only if both responses answer it correctly, yielding mastery rates of 95.0% on MedQA, 80.9% on MedMCQA, and 89.6% on MMLU-Med. For non-mastered questions, we provide the correct option and a brief explanation as correction guidance, then regenerate the knowledge points.

To measure whether a model recognizes these knowledge points, we generate three multiple-choice probes for each knowledge point. Each probe targets a single atomic fact and is designed for direct recognition rather than multi-step reasoning. For a question q with knowledge points $\{k_1, \ldots, k_n\}$, the knowledge score $K(q)$ is the average probe accuracy:

$$K(q) = \frac{1}{3n} \sum_{i=1}^{n} \sum_{j=1}^{3} \mathbb{1}[\text{model answers MCQ}_{i,j} \text{ correctly}]. \tag{1}$$

Thus, $K(q)$ measures whether the model can recognize the reference knowledge under controlled MCQ probes, as a practical proxy for parametric knowledge possession.

Validating the gold knowledge set. To reduce circularity from a single model family, we judge gold knowledge with Claude-4-5-sonnet rather than GPT-5-mini. The judge labels every gold point for medical correctness and relevance, covering the full pool, and finds 94.8% both correct and relevant. We then calibrate the judge against a medically-trained annotator on 150 stratified questions (50 per dataset), with Cohen's $\kappa = 0.87$.

2.2 RECALL EVALUATION

The recall metric measures whether, given a question, the model can spontaneously *recall* the reference knowledge from its parameters, regardless of whether it can subsequently use that knowledge to derive the correct answer. For each question q, we use a recall-specific prompt that asks the model to enumerate the knowledge points relevant to the question. We do not provide answer options, so the model must perform open-ended parametric recall rather than selecting among visible alternatives.

Let $\mathcal{K}_{\text{gold}} = \{k_1^g, \ldots, k_n^g\}$ denote the gold knowledge points associated with question q, and let $\mathcal{K}_{\text{pred}} = \{k_1^p, \ldots, k_m^p\}$ denote the knowledge points produced by the model. We embed all knowledge points using BAAI/BGE-M3 (Chen et al., 2024a) and compute the recall score as

$$R(q) = \frac{1}{n} \sum_{i=1}^{n} \max_{j \in \{1,\ldots,m\}} \cos\left(\mathbf{e}(k_i^g), \mathbf{e}(k_j^p)\right), \tag{2}$$

where $\mathbf{e}(\cdot)$ is the embedding function and $\cos(\cdot,\cdot)$ is cosine similarity. Intuitively, each gold knowledge point is matched to the most similar model-generated point, and $R(q)$ averages these best-match similarities. This makes $R(q)$ a coverage-oriented measure: it captures whether the model can surface the reference knowledge from its parameters, but does not directly penalize additional irrelevant or hallucinated points. We adopt this design because missing relevant knowledge is the failure mode that recall is intended to diagnose, whereas selecting and using the recalled information is part of the inference process. To confirm that this embedding-based score reflects semantic coverage rather than an artifact of the embedding model, we cross-validate it against a non-embedding

LLM coverage judge that rates coverage by reading each model output against the gold points; the two scores correlate strongly (Spearman $\rho = 0.79$, Appendix E).

2.3 Inference Evaluation

The inference metric measures whether a model can answer the original question *given* the reference knowledge, thereby isolating knowledge use from parametric recall. We prepend the full knowledge set from Section 2.1, including correct-path knowledge, competing hypotheses, discriminative cues, and confounders. Category labels are removed and the points are shuffled to avoid revealing the answer directly. The model is thus given a heterogeneous mix of correct-path and deliberately distracting knowledge, and must itself *select, weigh, and discriminate* which facts are decisive.

The inference score $I(q)$ is the accuracy on the original multiple-choice question under this shuffled knowledge context:

$$I(q) = \mathbb{1}[\text{model answers } q \text{ correctly} \mid \text{shuffled knowledge context}]. \tag{3}$$

Thus, $I(q)$ serves as a controlled measure of knowledge-conditioned reasoning, with the recall step factored out by construction.

3 Empirical Analysis

Using the diagnostic benchmark described in Section 2, we evaluate six models spanning different scales, architectures, and training paradigms: Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Qwen3-8B, Qwen3-14B, HuatuoGPT-o1-8B (Chen et al., 2024b), and MedGemma1.5-4B (Sellergren et al., 2026). We examine the interplay among the three meta-abilities—Knowledge (K), Recall (R), and Inference (I)—and their relationships with downstream accuracy (A).

3.1 Correlation Analysis

Table 1 reports the Spearman correlations among K, R, I, and A. For each correlation $\rho_{X,Y}$, we sort questions by X, partition them into 40 equal-frequency bins, average X and Y within each bin, and compute Spearman's ρ over the resulting bin-level pairs. We use this procedure throughout our diagnostic correlation analyses. The main table includes only the correlations central to our discussion; additional results include the knowledge-provision counterfactual (Appendix A), full correlation results (Appendix B), and sensitivity to the number of bins (Appendix D).

Table 1: Spearman correlation coefficients among Knowledge (K), Recall (R), Inference (I), and downstream Accuracy (A), computed over 40 equal-frequency bins on each benchmark. Bold marks $|\rho| \geq 0.50$; green/red additionally encode the sign for color readers.

	MedQA				MedMCQA				MMLU-Med			
Model	$\rho_{K,A}$	$\rho_{K,R}$	$\rho_{R,A}$	$\rho_{R,I}$	$\rho_{K,A}$	$\rho_{K,R}$	$\rho_{R,A}$	$\rho_{R,I}$	$\rho_{K,A}$	$\rho_{K,R}$	$\rho_{R,A}$	$\rho_{R,I}$
Qwen2.5-7B	0.03	0.34	**0.89**	0.08	0.44	−0.46	**0.64**	0.36	0.07	−0.29	**0.79**	**0.52**
Qwen2.5-14B	**−0.52**	0.02	**0.67**	0.37	0.38	**−0.60**	**0.73**	0.40	**−0.80**	**−0.60**	**0.90**	**0.83**
Qwen3-8B	−0.09	0.13	**0.67**	**0.63**	**0.51**	−0.45	**0.86**	**0.63**	−0.01	−0.40	**0.61**	**0.54**
Qwen3-14B	−0.16	0.03	**0.63**	0.45	**0.73**	−0.37	**0.78**	**0.71**	−0.49	−0.45	**0.82**	**0.60**
HuatuoGPT-o1-8B	0.20	−0.20	**0.56**	−0.18	**0.50**	**−0.76**	**0.77**	**0.85**	0.25	**−0.89**	**0.57**	0.47
MedGemma1.5-4B	0.32	−0.25	**0.62**	0.14	**0.69**	**−0.71**	**0.75**	**0.54**	0.24	**−0.53**	0.32	0.49

Knowledge possession alone weakly predicts accuracy. Although $\rho_{R,A}$ is consistently positive and usually strong, $\rho_{K,A}$ is much weaker and frequently near zero or negative, and it remains below 0.50 in absolute value on every MedQA cell. This indicates that, under our operational measure, possessing the reference knowledge under controlled MCQ probes does not reliably translate into answering the original question. The gap is precisely the phenomenon our recall axis is designed to isolate.

Knowledge and recall are often negatively correlated. $\rho_{K,R}$ is frequently negative, particularly on the more knowledge-centric MedMCQA and MMLU-Med. One possible explanation is a *local confidence trap*: in these knowledge-centric benchmark, models with stronger knowledge recognition

may rely on a small set of high-confidence knowledge associations, leading to narrower recalled sets and lower overall coverage. In contrast, the weaker $K-R$ tension on MedQA may reflect its more diagnosis-oriented clinical-vignette format, where answering a question requires integrating patient cues and activating broader reasoning chains rather than recognizing individual knowledge items in isolation. Although this interpretation remains speculative, the empirical pattern is clear: strong knowledge recognition does not necessarily translate into strong open-ended recall, suggesting that the two abilities should be evaluated separately.

Recall is the measured ability most consistently associated with downstream accuracy. The $\rho_{R,A}$ column is positive in every model–dataset cell and usually large: it ranges from $+0.32$ to $+0.90$, with 17 of 18 cells at or above $+0.50$. Thus, among our measured abilities, recall tracks downstream accuracy most reliably: models that surface the relevant knowledge are substantially more likely to answer correctly, regardless of scale or training paradigm. This consistency across six heterogeneous models suggests that recall, as measured by our diagnostic protocol, is a strong and actionable correlate of medical QA accuracy, beyond knowledge recognition alone.

Recall and inference are positively correlated. $\rho_{R,I}$ is generally positive and often large, indicating that models with weaker recall also tend to perform worse on inference, even when the relevant knowledge is explicitly provided. This suggests that recall and inference are partially coupled: models that fail to surface relevant knowledge may also be less effective at identifying, selecting, and using that knowledge once it is available. This finding highlights a limitation of retrieval-only approaches. Simply supplying missing knowledge may not fully close the performance gap, because low-recall models also tend to be weaker at reasoning over the provided knowledge. Improving the model's own recall may therefore be necessary, rather than relying on external retrieval alone.

3.2 Reinforcement Learning Yields Limited Gains in the Low-Recall Regime

Given the strong association between recall and downstream accuracy, we next ask whether current reinforcement learning methods can address the recall bottleneck. To investigate, we train Qwen3-8B and Qwen2.5-7B with GRPO (Shao et al., 2024) and examine how RL changes answer accuracy across different levels of initial recall ability. Specifically, we group questions by the *base model*'s recall score and measure post-RL accuracy within each group. We use the pre-RL recall score for stratification because RL can change instruction-following behavior and thus affect responses to the recall-elicitation prompt, making direct pre- and post-RL recall comparisons difficult to interpret. As shown in Figure 2, the gains from RL are concentrated in the *medium-to-high recall regime*. When the base model already recalls a substantial portion of the relevant knowledge, RL can improve how that knowledge is used to derive the final answer. In contrast, for questions in the *low-recall regime*, RL provides little improvement, suggesting that an end-to-end correctness signal alone is insufficient to recover knowledge that the model fails to recall, thereby limiting further gains in answer accuracy. This pattern motivates a training approach that makes recall an explicit optimization target, rather than leaving it implicitly entangled within a single end-to-end answer-correctness objective.

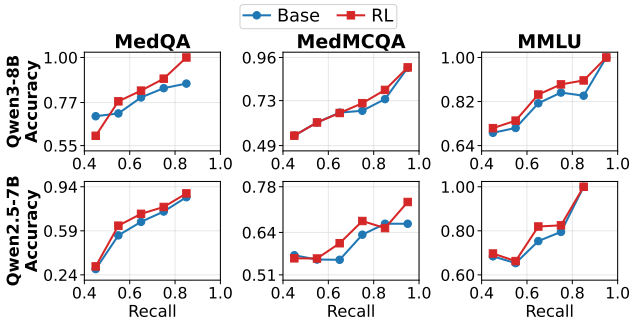


Figure 2: Effect of GRPO-based reinforcement learning on accuracy across recall score intervals.

4 Method

Building on the above findings, we propose a two-stage AgentTool training framework. **Stage 1** uses supervised fine-tuning (SFT) to familiarize the model with two response modes: answering directly without tool use, and invoking a recall tool before producing the final answer. **Stage 2** applies reinforcement learning to the resulting agentic policy, improving recall generation and final-answer reasoning while leaving the model free to decide on demand whether to invoke the AgentTool. Figure 3 overviews the Stage 2 pipeline.

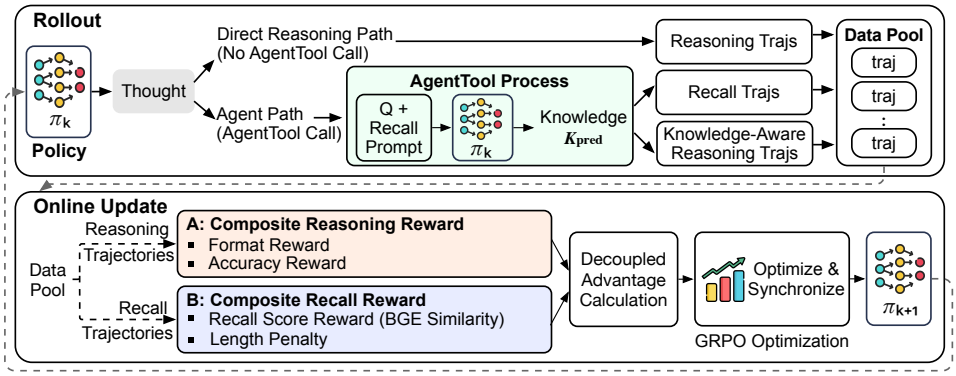


Figure 3: Overview of AgentTool RL. The policy either answers directly or invokes AgentTool, which performs recall in an isolated context and returns $\mathcal{K}_{\text{pred}}$ for subsequent reasoning. Reasoning and recall trajectories are optimized with separate rewards and decoupled advantages under GRPO.

4.1 Stage 1: Behavior Cloning

The goal of Stage 1 is to initialize the policy with the basic interaction patterns required for AgentTool use. We construct a cold-start SFT dataset with two types of trajectories. **Direct reasoning trajectories**: given the original question q and answer options, the model directly produces the reasoning process and final answer without invoking the AgentTool. **AgentTool trajectories**: the model first invokes the AgentTool, receives a tool response containing recalled knowledge, and then produces the final answer conditioned on the returned knowledge. AgentTool trajectories also contain supervision for the internal recall process. Under the recall prompt p_{recall}, the model takes the question q as input and generates the recalled knowledge as the tool response. We pair direct and AgentTool trajectories over the same training questions, so behavior cloning teaches response modes rather than changing question coverage. Thus, Stage 1 initializes a unified policy for direct answering, tool execution, and tool-conditioned answering. This stage is intended to initialize the prescribed response and tool-use formats rather than to inject additional medical capability.

4.2 Stage 2: AgentTool RL

After behavior cloning, the policy is exposed to both direct reasoning and AgentTool-based reasoning modes. Stage 2 further optimizes this agentic policy with RL. For each question q, a rollout either answers directly or invokes the AgentTool during reasoning. If invoked, the tool runs in an isolated recall context (p_{recall}, q) to generate $\mathcal{K}_{\text{pred}}$, which is returned to the main context. We organize the resulting data into two trace types: **reasoning trajectories**, covering the complete answer-generation process in both direct and AgentTool paths, and **recall trajectories**, covering only the recall process.

Reward Design. We assign rewards according to trajectory type. Each reasoning trajectory receives

$$r_{\text{main}} = r_{\text{fmt}} + r_{\text{ans}}, \tag{4}$$

where r_{fmt} rewards the required output format and r_{ans} rewards answer correctness. This is applied uniformly to both direct and AgentTool reasoning trajectories, so the policy is free to invoke the tool whenever it improves the final answer. In contrast, each recall-process trajectory receives a separate recall reward:

$$r_{\text{recall}} = R(q) - \frac{|\mathcal{K}_{\text{pred}}|}{C}, \tag{5}$$

where $R(q)$ is the recall score defined in Eq. 2, $|\mathcal{K}_{\text{pred}}|$ denotes the number of recalled knowledge points, and C is a normalization constant. Together, this length-regularized objective rewards coverage of the reference knowledge while discouraging excessive or loosely related enumeration.

Decoupled Advantage Computation. For each question q, we compute advantages separately for reasoning and recall-process trajectories. All reasoning trajectories form one group and use r_{main}, while AgentTool recall trajectories form another group and use r_{recall}. Within each group, the advantage is computed by normalizing rewards:

$$\hat{A} = \frac{r - \mu_r}{\sigma_r}, \tag{6}$$

where μ_r and σ_r are the mean and standard deviation of rewards within the corresponding group. The training batch is the union of the two groups.

GRPO Optimization. We optimize the policy with GRPO (Shao et al., 2024) over this union of reasoning and recall trajectories. The objective improves both final-answer reasoning and recall quality while keeping their credit assignment separated. Compared with a single end-to-end correctness signal, this design offers three advantages. First, AgentTool performs recall in an isolated context (p_{recall}, q) and returns an explicit $\mathcal{K}_{\text{pred}}$, preventing knowledge enumeration from disrupting the main reasoning chain. Second, recall trajectories form a separate advantage group, allowing recall quality to be optimized directly via r_{recall} rather than indirectly through final correctness. Third, the separate recall reward targets low-recall questions that end-to-end RL often fails to improve.

5 Experiments

5.1 Experimental Setup

Training Data. We use 19,178 QA pairs from the MedQA and MedMCQA training splits provided by Liu et al. (2025a). Among them, 8,000 questions are annotated with generated reference knowledge points. For Stage 1 (behavior cloning), we sample 4,000 of these questions and generate two paired cold-start trajectories for each question using GPT-5-mini: one direct reasoning trajectory and one AgentTool trajectory. For Stage 2 (AgentTool RL), we rollout on all 19,178 training questions and apply the recall reward only to questions with generated reference knowledge. More details and data construction prompts are reproduced in Appendix G and Appendix F.

Benchmarks. We evaluate on five medical QA datasets spanning in-domain and out-of-distribution settings. **In-domain:** MedQA-USMLE (Jin et al., 2021) and MedMCQA (Pal et al., 2022) test sets. **Out-of-distribution:** MMLU-Med (Hendrycks et al., 2021), PubMedQA (Jin et al., 2019), and MedXpert (Zuo et al., 2025).

Baselines. We compare AgentTool RL with three groups of baselines. First, we evaluate off-the-shelf general and medical LLMs, including Qwen3-8B, HuatuoGPT-o1-8B (Chen et al., 2024b), MedGemma1.5-4B, and AlphaMed-8B. Second, we construct a Self-Ask-style RL baseline on Qwen3-8B by combining Self-Ask prompting (Press et al., 2023) with GRPO training. This baseline serves as a single-context factual-priming comparison: it allows the model to generate intermediate questions and reasoning tokens that may elicit parametric knowledge, but it keeps recall-like content and final reasoning in the same context and optimizes only final-answer correctness, without an isolated recall tool or a recall-specific reward. Third, we compare with standard training stages on the same Qwen3-8B backbone, including SFT (Stage 1) and GRPO. To assess backbone robustness, we also repeat the full pipeline on the stronger Qwen3.5-9B model. Implementation details for AgentTool RL and baselines are provided in Appendix H.

5.2 Main Results

Table 2 reports accuracy on five medical QA benchmarks. Within the Qwen3-8B backbone, **AgentTool RL achieves the best result on every benchmark**, improving average accuracy from 64.1% to 69.2%. It outperforms standard GRPO by $+3.0$ points and Self-Ask by $+2.7$ points. These comparisons distinguish two alternatives: GRPO applies end-to-end RL without decomposition, while Self-Ask introduces additional intermediate reasoning within the same context to elicit potentially missing parametric knowledge. AgentTool RL improves over both, suggesting that its gains do not come merely from RL training or generating additional recall-oriented reasoning. Instead, the results support the benefit of explicitly separating recall from final-answer reasoning and optimizing the recall process with a dedicated reward. Compared to GRPO, AgentTool RL brings greater gains in MedQA ($+3.8$), MedMCQA ($+3.5$) and MedXpert ($+3.8$), where answering often depends on retrieving clinical knowledge before reasoning. The gains are smaller on PubMedQA ($+1.8$), where the input already provides the relevant biomedical abstract, and on MMLU-Med ($+1.8$), where the Qwen3-8B base model already reaches 82.6%, leaving less room for improvement from addressing recall failures. The same trend holds on the stronger Qwen3.5-9B backbone: AgentTool RL again ranks first among Qwen3.5-9B variants on every benchmark, with the largest gains on MedXpert ($+4.8$) and MedMCQA ($+3.2$). Together, these results show that AgentTool RL provides consistent benefits across backbones, especially when parametric recall remains an important bottleneck.

Table 2: Main results (accuracy, %) on five medical QA benchmarks, grouped by category. **Bold** / underline mark the best / second best per column across all models.

	In-domain		Out-of-distribution			
Model	MedQA	MedMCQA	MMLU-Med	PubMedQA	MedXpert	**Avg.**
Off-the-shelf models						
Qwen3-8B	78.1	64.6	82.6	77.3	17.8	64.1
HuatuoGPT-o1-8B	72.6	60.4	75.5	79.2	16.8	60.9
MedGemma1.5-4B	69.1	59.8	69.7	67.6	16.4	56.5
AlphaMed-8B	76.2	64.7	75.9	80.4	22.1	63.9
Non-tool agentic baseline						
Qwen3-8B + Self-Ask	82.8	66.2	83.4	79.6	20.6	66.5
Qwen3-8B training variants						
Qwen3-8B + SFT	78.4	64.7	82.3	77.6	17.8	64.2
Qwen3-8B + GRPO	82.5	65.7	83.5	79.1	20.4	66.2
AgentTool RL (Qwen3-8B)	86.3	69.2	85.3	**80.9**	24.2	69.2
Qwen3.5-9B training variants						
Qwen3.5-9B + SFT	87.9	71.4	85.4	79.0	26.7	70.1
Qwen3.5-9B + GRPO	90.5	72.9	85.9	79.5	28.0	71.4
AgentTool RL (Qwen3.5-9B)	**92.5**	**76.1**	**87.0**	80.6	**32.8**	**73.8**

5.3 Analysis

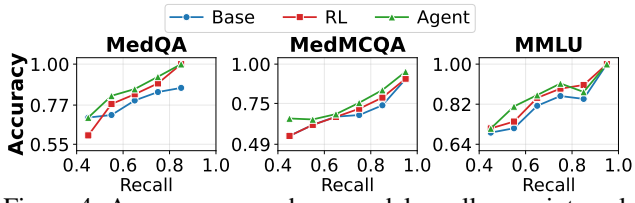


Figure 4: Accuracy across base-model recall score intervals for standard GRPO versus AgentTool RL.

AgentTool RL improves the low-recall regime. Section 3.2 showed that standard GRPO mainly improves medium- and high-recall questions, while leaving low-recall cases largely unchanged. We repeat the same stratified analysis for AgentTool RL by binning test questions according to the *base model*'s recall score and comparing its accuracy with GRPO. As shown in Figure 4, AgentTool RL yields consistent gains across recall intervals, with the largest gains concentrated in the low- and middle-recall bins where GRPO shows limited progress. This pattern suggests that AgentTool provides more direct supervision for recall failures. When relevant knowledge is not recalled, end-to-end answer supervision offers only an indirect signal, whereas AgentTool isolates recall and optimizes its coverage with a dedicated reward, improving performance where standard GRPO is least effective.

5.4 Ablation Study

Each component contributes. Table 3 shows consistent gains from each proposed component. Adding AgentTool to standard RL improves all five benchmarks, increasing average accuracy from 66.2 to 67.9, indicating that separating recall from the main reasoning context is beneficial. Adding r_{recall} yields further gains, reaching 86.3 on MedQA, 69.2 on MedMCQA, 85.3 on MMLU-Med, 80.9 on PubMedQA, and 24.2 on MedXpert. Together, these results suggest that AgentTool provides the structural separation for recall, while r_{recall} directly improves knowledge coverage.

On-demand invocation is necessary. The *Forced Recall* variant is an always-call AgentTool baseline: during both training and evaluation, every rollout is required to invoke the recall tool. It uses the same recall tool and recall reward, but removes the policy's ability to decide when the tool should be invoked. It performs substantially worse than the full method on all benchmarks and also underperforms the GRPO baseline on MMLU-Med, showing that recall is not universally helpful. When the model can directly answer the question, an unconditional recall step may introduce noisy or redundant context and interfere with direct reasoning. The learned policy instead invokes the tool selectively, with benchmark-specific invocation rates reported in Appendix I. This result also argues against a pure "more test-time tokens" explanation: Forced Recall generates an additional recall trajectory for every question, yet it is consistently worse than the learned on-demand policy.

Table 3: Ablation study (accuracy, %). Starting from standard end-to-end RL, we cumulatively add the AgentTool and the recall reward; the third row is our full method.

	In-domain		Out-of-distribution		
Method	MedQA	MedMCQA	MMLU-Med	PubMedQA	MedXpert
Qwen3-8B + RL	82.5	65.7	83.5	79.1	20.4
+ AgentTool	84.9	67.5	84.9	79.3	22.9
+ AgentTool + r_{recall} (full)	**86.3**	**69.2**	**85.3**	**80.9**	**24.2**
Forced Recall + r_{recall}	83.3	66.1	82.1	79.3	21.3

Therefore, the gains of AgentTool come not from always recalling more knowledge or using more reasoning tokens, but from learning when recall should be invoked.

6 Related Work

Medical QA and diagnostic evaluation of LLMs. LLMs have been adapted to medical question answering either by prompting general-purpose models (Singhal et al., 2023; Nori et al., 2023; Singhal et al., 2025) or by domain-specific training (Chen et al., 2024b; Wu et al., 2024; 2025). Progress, however, is still measured primarily by accuracy on exam-style benchmarks (Jin et al., 2021; Pal et al., 2022; Hendrycks et al., 2021; Jin et al., 2019; Zuo et al., 2025), which conflates heterogeneous failure modes. Complementary diagnostic studies probe individual capabilities, including factual recall (Kandpal et al., 2023; Petroni et al., 2019), chain-of-thought faithfulness (Turpin et al., 2023), calibration (Kadavath et al., 2022), test-time scaling (Huang et al., 2025), and shortcut learning (Zi et al., 2026). Closest to our work, Thapa et al. (2025) disentangle knowledge from reasoning and show that the two can fail independently. We extend this perspective with a three-way decomposition into *knowledge*, *recall*, and *inference*, isolating recall as a distinct axis and linking this diagnosis to a targeted training intervention.

RL and agentic reasoning in medicine. Reinforcement learning with verifiable rewards has become a common recipe for improving medical LLM reasoning, as seen in HuatuoGPT-o1 (Chen et al., 2024b), Fleming-R1 (Liu et al., 2025b), AlphaMed (Liu et al., 2025a), and Med-U1 (Zhang et al., 2025). However, these methods largely follow an end-to-end training paradigm: the model generates a complete reasoning trajectory and is optimized mainly by final-answer correctness. As shown in Section 3.2, such a signal provides weak credit assignment for low-recall questions, where the model fails before reasoning begins by not surfacing the relevant knowledge. A related agentic line augments medical reasoning with retrieval or search (Lu et al., 2025; Xu et al., 2025; Wang et al., 2025; Shi et al., 2025; Li et al., 2025), where deciding when and how to query external sources becomes a key component (Jeong et al., 2024). This line is complementary to our work but targets a different bottleneck. Search and RAG mainly expand the *knowledge* available to the model through non-parametric evidence from external corpora, whereas our goal is to improve *parametric recall*: the model's ability to surface knowledge it already possesses. In this sense, query formulation in retrieval systems can itself be viewed as a recall-dependent step, since the model must first identify what information is needed before it can retrieve it. Rather than relying on external retrieval, we isolate recall in a separate context and optimize it directly with an explicit recall reward, making recall a separately trainable sub-task rather than an implicit byproduct of end-to-end answer optimization.

7 Conclusion

We introduced a diagnostic framework that decomposes medical reasoning into Knowledge, Recall, and Inference. Across three benchmarks, Recall is most consistently associated with accuracy, while GRPO shows limited gains on low-recall questions. Motivated by this, AgentTool RL isolates parametric recall as an explicit tool and optimizes recall and reasoning with decoupled rewards. Across five medical QA benchmarks, it consistently outperforms GRPO and strong medical baselines, especially in low-recall settings. A limitation of our study is its reliance on model-generated reference knowledge, which may introduce annotation noise into the diagnostic analysis. Future work can improve factuality verification, incorporate external retrieval, and test generalization beyond medicine.

AI USE STATEMENT

In this work, generative AI tools were used to construct synthetic experimental artifacts. Specifically, GPT-5-mini was used to generate reference knowledge points, knowledge-probe questions, and cold-start SFT trajectories, and to provide an auxiliary coverage judgment for validating the embedding-based recall metric. Claude-4.5-Sonnet was used to assess the medical correctness and relevance of the generated reference knowledge. These uses are described in Sections 2 and 5 and Appendices E–G. Generative AI tools were also used to assist literature discovery and improve the readability of the manuscript. They were not used for research ideation, hypothesis formulation, experimental design, data analysis, result interpretation, method implementation, or the development or proof of mathematical claims. All AI-generated experimental artifacts and AI-assisted text were reviewed by the authors. In addition, the knowledge-quality evaluation was calibrated against a medically trained annotator on 150 stratified questions. The authors take full responsibility for the content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work studies medical question answering using existing public benchmarks and does not involve prospective clinical studies or direct interaction with patients. The evaluated models and the proposed method are intended for research and educational benchmarking, not for clinical deployment or medical decision-making. Model-generated reasoning and recalled knowledge may contain factual errors, hallucinations, or biases, and benchmark accuracy should not be interpreted as evidence of clinical competence. We follow the applicable licenses and usage conditions of all datasets and models used in this study.

REPRODUCIBILITY STATEMENT

We provide the complete diagnostic and training methodology in Sections 2–5. Appendix F contains the prompt templates used for knowledge construction, recall evaluation, and AgentTool interaction; Appendix G describes the cold-start data construction procedure; Appendix H reports the training and evaluation hyperparameters; and Appendix I reports tool-invocation statistics. Dataset splits, backbones, reward definitions, and decoding settings are specified in Section 5 and Appendix H.

REFERENCES

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *arXiv preprint arXiv:2402.03216*, 2024a.

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*, 2024b.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. m1: Unleash the Potential of Test-Time Scaling for Medical Reasoning with Large Language Models. *arXiv preprint arXiv:2504.00869*, 2025.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Proceedings of NAACL*, 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.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of EMNLP*, 2019.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. *International Conference on Machine Learning*, 2023.

Weitao Li, Boran Xiang, Xiaolong Wang, Zhinan Gou, Weizhi Ma, and Yang Liu. UR^2: Unify RAG and Reasoning through Reinforcement Learning. *arXiv preprint arXiv:2508.06165*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *ICLR*, 2024.

Che Liu, Haozhe Wang, Jiazhen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel Rueckert, and Rossella Arcucci. Beyond distillation: Pushing the limits of medical LLM reasoning with minimalist rule-based RL. *arXiv preprint arXiv:2505.17952*, 2025a.

Chi Liu, Derek Li, Yan Shu, Robin Chen, Derek Duan, Teng Fang, and Bryan Dai. Fleming-R1: Toward Expert-Level Medical Reasoning via Reinforcement Learning. *arXiv preprint arXiv:2509.15279*, 2025b.

Keer Lu, Zheng Liang, Youquan Li, Jiejun Tan, Xili Wang, Da Pan, Shusen Zhang, Guosheng Dong, Bin Cui, Yunhuai Liu, and Wentao Zhang. Med-R^3: Enhancing Medical Retrieval-Augmented Reasoning of LLMs via Progressive Reinforcement Learning. *arXiv preprint arXiv:2507.23541*, 2025.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 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*, 2022.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? In *EMNLP*, 2019.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. MedGemma 1.5 Technical Report. *arXiv preprint arXiv:2604.05081*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

Yucheng Shi, Tianze Yang, Canyu Chen, Quanzheng Li, Tianming Liu, Xiang Li, and Ninghao Liu. SearchRAG: Can Search Engines Be Helpful for LLM-based Medical Question Answering? *arXiv preprint arXiv:2502.13233*, 2025.

Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

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.

Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, Shriya Reddy, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. Disentangling reasoning and knowledge in medical large language models. *arXiv preprint arXiv:2505.11462*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *NeurIPS*, 2023.

Lehan Wang, Yi Qin, Honglong Yang, and Xiaomeng Li. Proactive reasoning-with-retrieval framework for medical multimodal large language models. *arXiv preprint arXiv:2510.18303*, 2025.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. PMC-LLaMA: Towards Building Open-source Language Models for Medicine. *Journal of the American Medical Informatics Association*, 2024.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. MedReason: Eliciting Factual Medical Reasoning Steps in LLMs via Knowledge Graphs. *arXiv preprint arXiv:2504.00993*, 2025.

Kaishuai Xu, Wenjun Hou, Yi Cheng, and Wenjie Li. RAR^2: Retrieval-Augmented Medical Reasoning via Thought-Driven Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Xiaotian Zhang, Yuan Wang, Zhaopeng Feng, Ruizhe Chen, Zhijie Zhou, Yan Zhang, Hongxia Xu, Jian Wu, and Zuozhu Liu. Med-U1: Incentivizing Unified Medical Reasoning in LLMs via Large-scale Reinforcement Learning. *arXiv preprint arXiv:2506.12307*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Xing Zi, Xinying Zhou, Jinghao Xiao, Catarina Moreira, and Mukesh Prasad. Shattering the shortcut: A topology-regularized benchmark for multi-hop medical reasoning in LLMs. *arXiv preprint arXiv:2603.12458*, 2026.

Yuxin Zuo et al. MedXpert: An Expert-Level Medical Question Answering Benchmark. *arXiv preprint*, 2025.

A Counterfactual Evidence for the Recall Bottleneck

The strong R–A correlation reported in Section 3.1 can reflect either an addressable recall bottleneck or a correlation induced by question difficulty. A common reviewer concern is that hard questions may simultaneously yield low recall and low accuracy, even if recall itself is not on the causal pathway. We therefore use the inference score $I(q)$ as a behavioral counterfactual: because $I(q)$ measures accuracy *when gold knowledge is provided*, it estimates how much performance could improve if the missing knowledge were made available.

Knowledge-provision gain. We define

$$\Delta(q) = I(q) - A(q), \tag{7}$$

the gain from adding gold knowledge to the context. If a question is hard mainly because of reasoning complexity, providing gold knowledge should not systematically help low-R questions more than high-R questions. If missing recall is a binding constraint, however, Δ should be largest in the low-recall regime, producing a negative $\rho_{R,\Delta}$. This is a behavioral counterfactual, not a randomized intervention: Δ can still be affected by how salient or complete the gold knowledge is on different items. We therefore treat the result as suggestive rather than causally identifying.

Shared dependence on $\mathcal{K}_{\text{gold}}$. A second caveat is that $R(q)$, $I(q)$, and our SFT targets are all defined against the same gold knowledge set. This creates a mechanical link: when R is low, the gold knowledge supplied for I contains more items the model failed to retrieve, so Δ has more room to increase. We therefore do not claim that this analysis identifies a recall→accuracy causal effect. The pattern is still informative for two reasons. First, Δ would not be consistently positive, nor reach the observed magnitudes (up to ~ 0.4 in individual model–dataset cells), if the supplied gold points were not answer-bearing. Second, the pattern appears across model families and benchmarks, rather than only for one model's idiosyncratic failures. We therefore read the result as evidence that

gold-knowledge injection helps most where native recall was most incomplete, which is the pattern predicted by the recall-bottleneck hypothesis.

Empirical pattern. Table 4 reports $\rho_{R,\Delta}$ across all six models and three benchmarks, computed over the same 40 equal-frequency bins by R used in Section 3. Recomputing the analysis with the alternative bin counts in Appendix D, $\{20, 60, 80, 100\}$, preserves the sign and qualitative magnitude of every cell. The dominant pattern is negative: **15 of 18 model–dataset cells** have $\rho_{R,\Delta} < 0$. The signal is especially consistent on the Qwen family, where all 12 cells are negative across two model generations (Qwen2.5, Qwen3) and two scales (7B/8B, 14B), with a median $\rho_{R,\Delta} = -0.40$ and a minimum of -0.92 on Qwen2.5-7B/MedQA. Subject to the shared-gold-set caveat above, this provides *counterfactual support*—not formal identification—that recall is an addressable factor rather than only a correlate of difficulty.

Table 4: Spearman correlation $\rho_{R,\Delta}$ between recall score R and knowledge-provision gain $\Delta = I - A$, computed over 40 equal-frequency bins by R. Negative values indicate that injecting gold knowledge yields larger gains precisely where recall was weakest. Bold marks $|\rho| \geq 0.30$; † marks model–dataset combinations where a secondary inference deficit dampens the gain (see Appendix A).

Model	MedQA	MedMCQA	MMLU-Med
Qwen2.5-7B	$\mathbf{-0.92}$	-0.21	$\mathbf{-0.35}$
Qwen2.5-14B	-0.26	$\mathbf{-0.46}$	$\mathbf{-0.33}$
Qwen3-8B	$\mathbf{-0.38}$	$\mathbf{-0.51}$	$\mathbf{-0.38}$
Qwen3-14B	$\mathbf{-0.43}$	$\mathbf{-0.51}$	$\mathbf{-0.46}$
HuatuoGPT-o1-8B	$\mathbf{-0.68}$	$+0.07^{\dagger}$	-0.06
MedGemma1.5-4B	$\mathbf{-0.58}$	$+0.21^{\dagger}$	$+0.25^{\dagger}$

Interpreting the three exceptions. The three cells where $\rho_{R,\Delta} \geq 0$ (HuatuoGPT-o1-8B on MedMCQA and MedGemma1.5-4B on MedMCQA/MMLU-Med) are not inconsistent with this reading, although the explanation is post hoc. They occur for smaller medical-specialized models on short factoid datasets, where Section 3.1 already suggests a secondary inference deficit: these models do not fully exploit provided knowledge, so Δ can remain small even when R is low. The same models show strongly negative $\rho_{R,\Delta}$ on MedQA, where clinical-vignette reasoning makes missing knowledge more costly.

Overall, the counterfactual analysis suggests that recall is most addressable when inference itself is not the limiting factor. When inference is also weak, improving recall alone is insufficient. This distinction motivates our method (Section 4): Stage 1 teaches both recall-tool use and tool-conditioned answering (Section 4.1), while Stage 2 optimizes recall generation and final-answer reasoning with separate rewards.

B Full Correlation Results

Table 5 reports the complete Spearman correlations among K, R, I, and A. It extends the main-text table (Table 1) with $\rho_{K,I}$ and $\rho_{I,A}$. All coefficients use the same 40 equal-frequency bins as the main diagnostic analysis.

Providing knowledge weakens, but does not erase, dependence on native recall. Comparing the two recall columns, $\rho_{R,I}$ is smaller than $\rho_{R,A}$ in 16 of the 18 model–dataset cells, with the largest gaps on MedQA (e.g., Qwen2.5-7B $0.89 \to 0.08$, HuatuoGPT-o1-8B $0.56 \to -0.18$, MedGemma1.5-4B $0.62 \to 0.14$). Because I is measured after inserting gold knowledge into the context, this drop means that many low-recall questions become easier once the missing facts are supplied. This is the lever AgentTool RL targets: it turns recall into an explicit, separately rewarded step for the cases where the base model fails to surface the needed knowledge (Section 3.2). The decoupling is only partial, however. $\rho_{R,I}$ remains positive in most cells and even exceeds $\rho_{R,A}$ for HuatuoGPT-o1-8B on MedMCQA ($0.77 \to 0.85$), indicating that low recall often co-occurs with weaker use of provided knowledge. Thus knowledge provision relieves recall failure but does

Table 5: Complete Spearman correlation coefficients among Knowledge (K), Recall (R), Inference (I), and downstream Accuracy (A), computed over 40 equal-frequency bins on each benchmark. Inference accuracy is measured by augmenting questions with shuffled gold knowledge points. Bold marks $|\rho| \geq 0.50$; green/red additionally encode the sign for color readers.

Model	$\rho_{K,A}$	$\rho_{K,R}$	$\rho_{R,A}$	$\rho_{R,I}$	$\rho_{K,I}$	$\rho_{I,A}$
MedQA						
Qwen2.5-7B	0.03	0.34	**0.89**	0.08	0.11	0.27
Qwen2.5-14B	**−0.52**	0.02	**0.67**	0.37	−0.21	**0.58**
Qwen3-8B	−0.09	0.13	**0.67**	**0.63**	−0.28	**0.57**
Qwen3-14B	−0.16	0.03	**0.63**	0.45	−0.21	**0.63**
HuatuoGPT-o1-8B	0.20	−0.20	**0.56**	−0.18	−0.11	0.20
MedGemma1.5-4B	0.32	−0.25	**0.62**	0.14	0.02	0.23
MedMCQA						
Qwen2.5-7B	0.44	−0.46	**0.64**	0.36	0.10	**0.54**
Qwen2.5-14B	0.38	**−0.60**	**0.73**	0.40	0.16	**0.61**
Qwen3-8B	**0.51**	−0.45	**0.86**	**0.63**	0.19	**0.80**
Qwen3-14B	**0.73**	−0.37	**0.78**	**0.71**	0.18	**0.83**
HuatuoGPT-o1-8B	**0.50**	**−0.76**	**0.77**	**0.85**	0.25	**0.68**
MedGemma1.5-4B	**0.69**	**−0.71**	**0.75**	**0.54**	0.21	**0.56**
MMLU-Med						
Qwen2.5-7B	0.07	−0.29	**0.79**	**0.52**	0.17	**0.54**
Qwen2.5-14B	**−0.80**	**−0.60**	**0.90**	**0.83**	**−0.66**	**0.71**
Qwen3-8B	−0.01	−0.40	**0.61**	**0.54**	−0.24	**0.79**
Qwen3-14B	−0.49	−0.45	**0.82**	**0.60**	−0.19	**0.80**
HuatuoGPT-o1-8B	0.25	**−0.89**	**0.57**	0.47	−0.23	**0.58**
MedGemma1.5-4B	0.24	**−0.53**	0.32	0.49	−0.02	0.33

not make recall irrelevant; improving parametric recall remains complementary to retrieval-based approaches.

Inference ability predicts accuracy, especially on simpler reasoning formats. $\rho_{I,A}$ is positive in every cell and is strongest on MedMCQA ($+0.54$ to $+0.83$) and MMLU-Med ($+0.33$ to $+0.80$). Once a model can answer correctly with the relevant knowledge in context, that ability transfers most directly to benchmarks whose questions often hinge on a compact knowledge association. On MedQA, $\rho_{I,A}$ is weaker and more variable ($+0.20$ to $+0.63$), consistent with its clinical vignettes requiring integration of multiple cues beyond the supplied knowledge points.

Knowledge is most positively associated with inference on knowledge-dense questions. MedMCQA is the only dataset where $\rho_{K,I}$ is consistently positive ($+0.10$ to $+0.25$), matching its consistently positive $\rho_{K,A}$. This is expected for a knowledge-dense, near-factual-lookup format, where recognizing relevant facts is closer to solving the question. On MedQA and MMLU-Med, $\rho_{K,I}$ is near zero or negative (down to -0.66 for Qwen2.5-14B on MMLU-Med), echoing the weak $\rho_{K,A}$ there. Recognizing isolated facts is therefore not enough to use them within a reasoning chain, even when they are placed in context. This dissociation motivates treating recall as a separate axis.

Figure 5 visualizes the recall–accuracy relationship behind the positive $\rho_{R,A}$ column. After grouping samples into 0.1-width recall bins, mean accuracy generally increases with recall across model–benchmark combinations. This provides a direct view of the aggregate trend: models are more likely to answer correctly when they surface the relevant knowledge from their parameters.

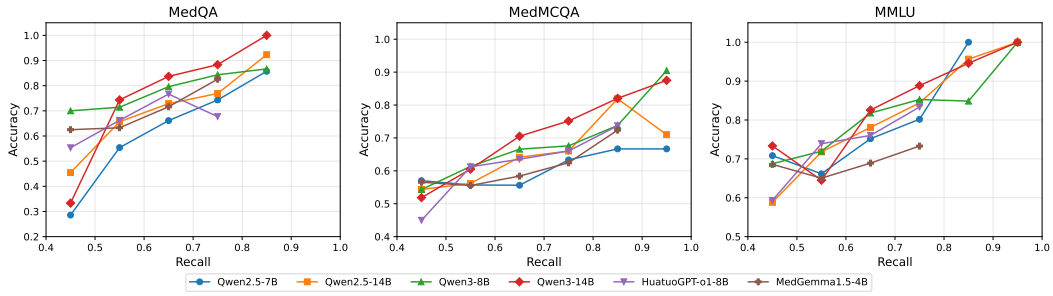


Figure 5: Recall score versus downstream accuracy across six models on three medical QA benchmarks. Samples are grouped into bins of width 0.1 by recall score and the mean accuracy within each bin is plotted; bins with fewer than 30 samples are omitted to limit small-bin noise. A consistent positive association between recall and accuracy is observed for every model–benchmark combination.

C Descriptive Statistics of Diagnostic Scores

Table 6 reports the item-level mean and variance of $K(q)$, $R(q)$, and $I(q)$ before any binning. These statistics provide context for interpreting the correlation results in Section 3.1. The Knowledge score is high across all models and datasets, reflecting that the recognition-style probe is often solvable by current medical and general-domain LLMs. However, $K(q)$ is not identical across models or benchmarks, and its weak or negative association with downstream accuracy therefore should not be read as evidence that medical knowledge is irrelevant. Rather, it shows that recognizing the reference facts under controlled probes is not sufficient to ensure that the same facts are retrieved and used in the original question context. Recall scores occupy a lower range than Knowledge scores, consistent with recall being the bottleneck isolated by our framework, while Inference scores retain substantial variance because $I(q)$ is a binary outcome measured after providing the gold knowledge context.

Table 6: Item-level descriptive statistics of Knowledge (K), Recall (R), and Inference (I) before binning. Means and variances are computed over individual questions within each model–dataset pair.

Model	**Dataset**	$K(q)$		$R(q)$		$I(q)$	
		Mean	Var.	Mean	Var.	Mean	Var.
Qwen2.5-7B	MedQA	0.891	0.002	0.628	0.004	0.775	0.174
Qwen2.5-14B	MedQA	0.923	0.001	0.652	0.004	0.838	0.136
Qwen3-14B	MedQA	0.940	0.001	0.665	0.004	0.903	0.088
Qwen3-8B	MedQA	0.918	0.001	0.653	0.004	0.863	0.118
HuatuoGPT-o1-8B	MedQA	0.887	0.002	0.595	0.003	0.834	0.139
MedGemma1.5-4B	MedQA	0.865	0.003	0.628	0.003	0.782	0.170
Qwen2.5-7B	MedMCQA	0.879	0.003	0.642	0.006	0.661	0.224
Qwen2.5-14B	MedMCQA	0.915	0.002	0.657	0.006	0.721	0.201
Qwen3-14B	MedMCQA	0.930	0.002	0.672	0.005	0.771	0.177
Qwen3-8B	MedMCQA	0.908	0.003	0.658	0.006	0.736	0.194
HuatuoGPT-o1-8B	MedMCQA	0.877	0.003	0.614	0.004	0.684	0.216
MedGemma1.5-4B	MedMCQA	0.834	0.005	0.634	0.004	0.610	0.238
Qwen2.5-7B	MMLU-Med	0.907	0.003	0.648	0.006	0.757	0.184
Qwen2.5-14B	MMLU-Med	0.941	0.002	0.669	0.006	0.831	0.140
Qwen3-14B	MMLU-Med	0.956	0.001	0.680	0.005	0.857	0.123
Qwen3-8B	MMLU-Med	0.938	0.002	0.673	0.006	0.840	0.134
HuatuoGPT-o1-8B	MMLU-Med	0.906	0.003	0.609	0.004	0.790	0.166
MedGemma1.5-4B	MMLU-Med	0.853	0.005	0.635	0.004	0.701	0.214

D SENSITIVITY TO THE NUMBER OF BINS

The diagnostic correlations in Table 1, Table 5, and Table 4 are computed by sorting questions on the conditioning score X, partitioning them into equal-frequency bins, averaging both scores within each bin, and taking the Spearman rank correlation between the paired bin means (Section 3.1). This convention is separate from the recall-metric validation in Appendix E, which correlates two per-question scalar recall scores directly. We use the bin-level statistic by design. The goal of the analysis is not to build an item-level classifier that predicts the binary outcome $A(q)$ from a single scalar score, but to ask whether groups of questions with higher measured Knowledge, Recall, or Inference exhibit higher expected accuracy. This regime-level question matches our motivating diagnosis: whether low-recall regions behave differently from medium- and high-recall regions, and whether RL improves those regions differently. We therefore interpret the reported ρ values as correlations between conditional means, not as item-level predictive effect sizes. At the item level, $A(q)$ is binary and highly noisy: two questions with similar recall scores can differ in reasoning complexity, distractor structure, option sensitivity, guessing, and residual noise in the automatically generated knowledge set. A raw item-level rank correlation therefore answers a different and substantially noisier question, while binning reduces item-level variance and highlights aggregate monotonic trends.

The binning step trades off two competing concerns: too few bins discards resolution and can oversmooth local structure, while too many bins makes each per-bin mean noisy and moves the estimate closer to an item-level statistic. We adopt 40 bins throughout the main text as a compromise.

Table 7: Sensitivity of the three bin-level correlations to the number of equal-frequency bins. For every model and benchmark we recompute $\rho_{R,A}$, $\rho_{K,A}$, and $\rho_{K,R}$ at $\{20, 40, 60, 80, 100\}$ bins; the **40**-bin column (boldface header) is the setting used throughout the main text (Table 1). The tested bin counts show the same qualitative pattern: $\rho_{R,A}$ is positive in all 90 cells, $\rho_{K,A}$ stays weak and dataset-dependent, and $\rho_{K,R}$ is negative throughout MedMCQA and nearly always negative on MMLU-Med. Magnitudes vary with bin count because the estimator trades off aggregation against per-bin variance.

		$\rho_{R,A}$					$\rho_{K,A}$					$\rho_{K,R}$				
Data	**Model**	20	**40**	60	80	100	20	**40**	60	80	100	20	**40**	60	80	100
MedQA	Qwen2.5-7B	0.97	0.89	0.88	0.83	0.77	−0.09	0.03	0.18	0.13	0.23	0.17	0.34	0.37	0.29	0.25
	Qwen2.5-14B	0.93	0.67	0.58	0.56	0.43	−0.39	−0.52	−0.40	−0.27	−0.28	0.19	0.02	0.04	0.10	0.05
	Qwen3-8B	0.81	0.67	0.67	0.52	0.48	−0.10	−0.09	−0.13	−0.02	−0.02	−0.07	0.13	0.17	0.16	0.18
	Qwen3-14B	0.92	0.63	0.46	0.60	0.55	−0.41	−0.16	−0.12	−0.14	−0.23	−0.10	0.03	−0.26	−0.05	−0.08
	HuatuoGPT-o1-8B	0.53	0.56	0.51	0.27	0.42	0.21	0.20	0.25	0.18	0.13	−0.20	−0.20	−0.28	−0.17	−0.24
	MedGemma1.5-4B	0.75	0.62	0.67	0.56	0.57	0.68	0.32	0.28	0.04	0.11	−0.37	−0.25	−0.10	−0.02	0.03
MedMCQA	Qwen2.5-7B	0.77	0.64	0.52	0.41	0.38	0.68	0.44	0.41	0.42	0.41	−0.63	−0.46	−0.30	−0.40	−0.35
	Qwen2.5-14B	0.75	0.73	0.75	0.62	0.63	0.44	0.38	0.41	0.20	0.22	−0.64	−0.60	−0.52	−0.52	−0.47
	Qwen3-8B	0.92	0.86	0.77	0.58	0.51	0.78	0.51	0.37	0.45	0.48	−0.15	−0.45	−0.26	−0.19	−0.20
	Qwen3-14B	0.87	0.78	0.80	0.73	0.75	0.91	0.73	0.50	0.38	0.49	−0.18	−0.37	−0.37	−0.35	−0.47
	HuatuoGPT-o1-8B	0.93	0.77	0.65	0.53	0.40	0.77	0.50	0.50	0.59	0.42	−0.88	−0.76	−0.70	−0.61	−0.56
	MedGemma1.5-4B	0.87	0.75	0.46	0.57	0.45	0.78	0.69	0.45	0.51	0.42	−0.88	−0.71	−0.58	−0.48	−0.55
MMLU-Med	Qwen2.5-7B	0.98	0.79	0.76	0.65	0.53	0.34	0.07	0.09	0.01	0.12	−0.18	−0.29	−0.18	−0.13	0.07
	Qwen2.5-14B	0.94	0.90	0.68	0.70	0.60	−0.38	−0.80	−0.66	−0.57	−0.61	−0.53	−0.60	−0.67	−0.60	−0.43
	Qwen3-8B	0.29	0.61	0.54	0.43	0.50	−0.11	−0.01	−0.09	−0.15	−0.23	−0.55	−0.40	−0.36	−0.39	−0.25
	Qwen3-14B	0.92	0.82	0.70	0.57	0.68	−0.01	−0.49	−0.20	−0.21	−0.19	−0.34	−0.45	−0.28	−0.25	−0.24
	HuatuoGPT-o1-8B	0.68	0.57	0.42	0.34	0.18	0.29	0.25	0.16	0.04	0.03	−0.98	−0.89	−0.84	−0.77	−0.66
	MedGemma1.5-4B	0.51	0.32	0.26	0.08	0.18	0.29	0.24	0.13	0.09	0.16	−0.82	−0.53	−0.36	−0.19	−0.27

Table 7 recomputes all three bin-level correlations—$\rho_{R,A}$, $\rho_{K,A}$, and $\rho_{K,R}$—at $\{20, 40, 60, 80, 100\}$ bins for every model and benchmark, with the 40-bin column corresponding to the values reported in the main text (Table 1). Across this sweep, the same qualitative patterns as Section 3.1 remain visible:

- $\rho_{R,A}$ **is positive everywhere.** Recall tracks accuracy in all 90 cells (6 models $\times$ 3 datasets $\times$ 5 bin counts); the smallest value is $+0.08$ (MedGemma1.5-4B / MMLU-Med / 80 bins) and the sign never flips. The headline "recall is the meta-ability most consistently predictive of accuracy" is therefore robust across the bin counts we tested.

- $\rho_{K,A}$ **stays weak and dataset-dependent.** At every bin count it is small and mixed-sign on MedQA and MMLU-Med (frequently negative for the Qwen models) and only consistently positive on MedMCQA, exactly the pattern that motivates our claim that recognized knowledge alone is a poor predictor of correctness.
- $\rho_{K,R}$ **is negative on MedMCQA and nearly always negative on MMLU-Med.** It is negative in all 30 MedMCQA cells and in 29 of 30 MMLU-Med cells; the only exception is a small positive value for Qwen2.5-7B at 100 bins ($+0.07$). This preserves the qualitative "local confidence trap" pattern across the sweep while avoiding over-interpreting a near-zero edge case.

What *does* change most with the bin count is magnitude and small-bin noise: many correlations attenuate as the bins grow finer (e.g., $\rho_{R,A}$ for Qwen2.5-7B on MedQA moves $0.97 \to 0.77$ from 20 to 100 bins), because the estimator trades off aggregation against per-bin variance. A few weak cells fluctuate near zero, but these fluctuations do not change the main patterns above within the tested sweep. Forty bins are a practical compromise in this range: enough bins to expose non-monotonic structure, few enough that each bin mean is not dominated by item-level noise. We accordingly use 40 bins for the diagnostic correlation analyses while avoiding any interpretation of the absolute magnitudes as unbinned item-level correlations.

E Convergent Validation of the Recall Metric

The recall score $R(q)$ (Eq. 2) maps model-generated knowledge points to the gold set via BGE-M3 embedding similarity. To verify that $R(q)$ reflects semantic knowledge coverage rather than only an artifact of the embedding model, we cross-validate it against a non-embedding LLM coverage judge (GPT-5-mini), which reads the generated recall text and the gold knowledge points directly.

Independence and scope of the validation. This validation checks whether the embedding-based recall score agrees with a second, non-embedding way of measuring coverage. The LLM judge is given two inputs: the fixed gold knowledge points and the recall text generated by an evaluated model. It then decides, for each gold point, whether that point is stated or entailed by the model's recall text.

The validation is therefore independent of the embedding scorer, but not fully independent of the gold-point generator, because GPT-5-mini is used both to generate the initial gold knowledge points and to make this coverage judgment. We treat the LLM-judged score only as a *convergent check* of the BGE-M3 recall metric, not as the primary evaluation signal. This choice does not introduce model-under-test favoritism, since all evaluated models in our diagnostic analysis are non-GPT systems (Qwen, HuatuoGPT-o1, and MedGemma). The remaining limitation is narrower: the coverage judge and the gold-point generator belong to the same model family, so future work could make this validation stricter by using a different-family judge for the coverage decision as well.

Per-point coverage judgment. The key design choice is that the judge never emits a numeric score directly—LLMs are poorly calibrated at producing an absolute value such as "0.73." Instead, it makes a sequence of *binary* decisions, one per gold knowledge point, which we then aggregate into a float outside the model. Concretely, for a question q with gold set $\mathcal{K}_{\text{gold}} = \{k_1^g, \ldots, k_n^g\}$ and the model's recalled text containing $\mathcal{K}_{\text{pred}}$, we present the judge with the full recalled text and one gold point k_i^g at a time, and ask whether that point is semantically covered—i.e., stated, paraphrased, or clearly entailed—anywhere in the recalled text. This yields a binary indicator

$$c_i = \mathbb{I}\big[\, k_i^g \text{ is covered by } \mathcal{K}_{\text{pred}} \,\big] \in \{0, 1\}, \qquad i = 1, \ldots, n, \tag{8}$$

where the judgment is made by the LLM rather than by embedding similarity. The per-question LLM-judged recall is then the *coverage fraction*, computed by us, not by the judge:

$$R_{\text{LLM}}(q) = \frac{1}{n} \sum_{i=1}^{n} c_i \;\in\; [0, 1]. \tag{9}$$

This mirrors $R(q)$ exactly in structure—both average a per-gold-point match over the same set $\mathcal{K}_{\text{gold}}$—and differs only in the matching mechanism: continuous max-cosine for $R(q)$ versus a discrete LLM coverage call for $R_{\text{LLM}}(q)$. Because the average runs over n binary terms, $R_{\text{LLM}}(q)$ is itself fine-grained ($n+1$ distinct levels for an n-point question), so no discretization or binning of

either score is required. Asking the judge for many easy binary calls, rather than one hard numeric estimate, is both more reliable and yields a smoother aggregate.

Agreement. Because both scores are scalar recall-coverage scores on $[0, 1]$—with $R_{\text{LLM}}(q)$ taking $n+1$ possible values for an n-point question—we report their rank correlation rather than a binary agreement rate. Across 1000 questions spanning all three benchmarks and six models, the BGE-based $R(q)$ and the LLM-judged $R_{\text{LLM}}(q)$ correlate at Spearman $\rho = 0.79$. The two estimates rely on different matching mechanisms—embedding geometry versus direct semantic coverage judgment—so their agreement supports the use of $R(q)$ as a recall-coverage proxy over the gold knowledge points. Because the coverage judge is not independent of the gold-generation model family, we interpret this result as evidence of measurement convergence rather than as a separate ground-truth annotation. Separate factuality audits of generated knowledge remain complementary to this validation.

F Prompt Templates

This appendix documents the prompts used throughout the paper. The knowledge-point and MCQ-probe generation prompts build the diagnostic benchmark (Section 2); the recall prompt p_{recall} is used both for the recall metric (Section 2.2) and as the isolated-context recall generator at training and inference time (Section 4); the answering prompt and tool definition drive the agent policy. The listings preserve the operative instructions, with long lines rewrapped and redundant formatting boilerplate omitted to fit the page.

F.1 Gold Knowledge-Point Generation

Used with GPT-5-mini to construct the gold knowledge set $\mathcal{K}_{\text{gold}}$ for each question. For non-mastered questions, the optional {hints} field supplies the correct option and a brief explanation as correction guidance, and the knowledge points are regenerated.

```
You are a **Golden Knowledge Space Generator** for medical question answering.

Your task is NOT to solve the question and NOT to select a final diagnosis.
Your task is to construct a **gold-standard knowledge space** that represents
what a careful, unbiased medical reasoning process SHOULD recall and consider.
This knowledge space is used for **coverage-based evaluation and RL**.

You must explicitly include:
1) knowledge required for the correct reasoning path
2) knowledge belonging to reasonable but ultimately incorrect paths
3) knowledge that discriminates between competing paths
4) common pitfalls and confounders

## Output Structure (Strict)
Your output must contain exactly four sections, using these fixed tags and no
additional text:

<core_path_knowledge> ... </core_path_knowledge>
<competing_hypotheses_knowledge> ... </competing_hypotheses_knowledge>
<discriminative_knowledge> ... </discriminative_knowledge>
<pitfalls_and_confounders> ... </pitfalls_and_confounders>

## Global Output Rules
- Each item must be a concrete, standalone, and concise medical knowledge point
  (disease,
  syndrome, symptom, sign, test, imaging pattern, drug class, toxin, mechanism).
- Do NOT write explanations or reasoning; do NOT include knowledge details.
- Do NOT repeat the same item across sections.
- Do NOT state which diagnosis is "most likely". Points should be suitable as
  knowledge graph nodes.

## Section-Specific Instructions
1. <core_path_knowledge>: points strictly required to justify the correct path
   (minimal but sufficient); no alternatives or exclusions here.
2. <competing_hypotheses_knowledge>: reasonable alternative diagnoses that fit
   part of the evidence but are ultimately incorrect; span multiple disease
   families if applicable (neurodegenerative, psychiatric, metabolic, toxic,
   vascular); no trivial/irrelevant options.
3. <discriminative_knowledge>: points used to distinguish the core path from
   competing hypotheses (specific signs, imaging/lab findings, treatment
   response, temporal patterns); each should help rule in/out a hypothesis.
```

```
4. <pitfalls_and_confounders>: common sources of diagnostic error (drug effects,
   overlapping syndromes, mimicking secondary conditions, cognitive confounders).

## Bias Control (Hard Constraints)
- Do NOT frame content around a single narrative or assume the answer is obvious.
- Do NOT collapse competing hypotheses into one category.
- Treat exclusion knowledge as equally important as confirmatory knowledge.

The question is: {question}
[Here are some hints which may be helpful: {hints}]
```

F.2 KNOWLEDGE MCQ-PROBE GENERATION

Given a single atomic knowledge point, generates three recognition probes used to compute the knowledge score $K(q)$ (Eq. 1).

```
You are a medical education expert. Generate THREE independent,
knowledge-focused multiple-choice questions (MCQs) based on ONE given medical
knowledge point.

Goal: the three questions test the SAME knowledge point from different angles to
ensure deep understanding rather than superficial pattern recognition. They test
ONLY factual or conceptual knowledge, NOT clinical reasoning, case
interpretation, or diagnostic inference.

Strict requirements:
1. Each question tests the SAME single, atomic knowledge point.
2. Generate EXACTLY THREE MCQs.
3. They must differ meaningfully in structure or cognitive perspective.
4. Do NOT include clinical vignettes, patient cases, or symptom combinations.
5. Do NOT require multi-step reasoning or inference.
6. Do NOT restate the knowledge point verbatim as the question stem.
7. All questions answerable using general medical knowledge.
8. Each question has exactly four options (A-D); only one is correct.
9. Avoid superficial paraphrases -- differ in wording and testing angle.

Guidelines: neutral, educational language; plausible but clearly incorrect
distractors; include at least one medium- and one higher-difficulty question;
at least one tests deeper conceptual understanding (not just definition recall).

Output format:
<knowledge_mcqs>
Question 1:
A. ...   B. ...   C. ...   D. ...
Correct answer: X
(Questions 2 and 3 follow the same format)
</knowledge_mcqs>

The knowledge point to test: {knowledge_point}
```

F.3 RECALL ENUMERATION (p_{RECALL})

Elicits open-ended parametric recall without answer options, yielding $\mathcal{K}_{\text{pred}}$ for the recall score $R(q)$ (Eq. 2). The structured <reason>/<knowledge_points> format mirrors iterative human recall.

```
You are a medical education expert. Generate detailed background knowledge
points related to the following question stem from a multiple-choice medical QA
item. The goal is to support conceptual understanding, NOT to solve the question.

Strict requirements:
1. Do NOT state or imply which option is correct.
2. Do NOT infer, reconstruct, compare, or evaluate hidden answer choices.
3. Do NOT provide diagnostic conclusions, final judgments, or recommendations.
4. Avoid strong directive language such as: "the most likely", "pathognomonic",
   "diagnostic of", "therefore", "thus the answer is", "should choose".
5. Use neutral, descriptive, educational language only.
6. Provide all background knowledge points relevant to the question.
7. Do NOT miss any important knowledge domains related to the scenario.
8. Focus on knowledge domains, mechanisms, and conceptual frameworks -- not
   answer selection.

### Output Structure
Simulate a human-like recall process using multiple reasoning segments:

<reason>
```

```
A reasoning segment explaining what conceptual domain the question brings to mind.
</reason>
<knowledge_points>
1. Knowledge Point 1
2. Knowledge Point 2
</knowledge_points>
<reason>
Next reasoning step expanding to related knowledge points based on the previous
step.
</reason>
<knowledge_points>
3. Knowledge Point 3
4. Knowledge Point 4
</knowledge_points>
(these reason + knowledge_points blocks may repeat as needed)

### Structural Rules
* Each <reason> block corresponds to a specific conceptual direction.
* Each <knowledge_points> block includes only knowledge triggered by the
  immediately preceding reasoning block.
* Do NOT emit one large reasoning section followed by all knowledge; the
  structure should reflect iterative recall.

### Content Constraints
* <reason>: describe cognitive recall direction, not problem-solving; do not
  mention options, eliminate choices, indicate likelihood, or conclude.
* <knowledge_points>: each item is a concise point about a medical concept; do
  not explain in detail -- state each point briefly.
```

F.4 AGENTTOOL DEFINITION

The recall tool exposed to the policy at Stage 2. The empty parameter schema reflects that the tool takes no arguments—it recalls knowledge for the current question in an isolated context.

```
{
  "name": "knowledgeRecall",
  "description": "An optional self-recall function that the model may invoke
    before answering when recalling background medical knowledge is likely to
    improve accuracy, completeness, or grounding. The model should decide
    dynamically whether to call this function. Do not call it when the question
    can be answered directly or when recall is likely to add noise, irrelevant
    information, or unnecessary latency. When used, this function makes the
    recalled knowledge visible as an explicit intermediate step before final
    reasoning.",
  "parameters": {"type": "object", "properties": {}}
}
```

F.5 ANSWERING PROMPT

The system prompt for direct answering and for tool-conditioned answering after a recall call. The fixed <think>/<answer> format supports the format reward r_{fmt} and exact-match answer extraction.

```
You are a highly trained medical expert with strong knowledge in basic sciences,
clinical medicine, and evidence-based practice. You are capable of applying
medical knowledge to clinical scenarios, including diagnosis, pathophysiology,
and treatment decisions.

For the following multiple-choice question, select one correct answer. Your
answer should only include the selected serial number (one option letter, such
as A, B, C, or D). The reasoning process and answer are enclosed within
<think> </think> and <answer> </answer> tags, respectively, i.e.,
<think> reasoning process here </think><answer> Your choice here </answer>.
```

For the inference metric $I(q)$ (Section 2.3), the same answering prompt is used, but the gold knowledge points are inserted into the context with their category labels removed and their order randomized, reducing the chance that the model can infer the answer from the structure of the provided evidence rather than selecting and applying the relevant facts.

Table 8: Composition of the Stage 1 cold-start SFT dataset. The two trajectory types are paired over the same 4,000 questions. Each AgentTool trajectory simultaneously supplies supervision for tool execution (the recall segment) and tool-conditioned answering (the final-answer segment).

Trajectory type	# Trajectories	Behaviors supervised
Direct reasoning	4,000	Direct answering
AgentTool	4,000	Tool execution; tool-conditioned answering
Total	**8,000**	

G Cold-Start Data Construction

The Stage 1 behavior-cloning dataset (Section 4.1) initializes the policy with the three behaviors required by the AgentTool interface: direct answering, recall-tool execution, and tool-conditioned answering. We construct it from the training pool of Liu et al. (2025a) (19,178 MedQA/MedMCQA training-split QA pairs), of which 8,000 questions have generated reference knowledge points. From this knowledge-annotated subset, we sample 4,000 unique questions and generate two paired trajectories for each question using GPT-5-mini: one direct reasoning trajectory and one AgentTool trajectory. This yields 4,000 direct reasoning trajectories and 4,000 AgentTool trajectories over the same 4,000 question stems. The purpose of the GPT-generated SFT data is to teach the model the prescribed response format and tool-interaction protocol, rather than to inject new medical knowledge or reasoning capabilities. We therefore use a relatively small cold-start dataset with limited question coverage. As shown in Table 2, direct-reasoning SFT yields virtually no improvement over the base model (64.2% vs. 64.1% average accuracy), suggesting that the GPT-generated supervision alone does not account for the downstream gains. The improvements of AgentTool RL are therefore better attributed to the subsequent reinforcement-learning stage and its explicit optimization of tool invocation and recall quality, rather than conventional knowledge distillation from GPT-generated trajectories.

Direct reasoning trajectories. For each sampled question, GPT-5-mini produces a chain-of-thought reasoning trace followed by the final answer, conditioned on the question stem and answer options. We retain a trajectory only if its final answer matches the gold label, so that the cold-start policy is cloned exclusively from correct reasoning.

AgentTool trajectories. Each AgentTool trajectory interleaves three segments: (i) a tool call to `recall`; (ii) a tool response containing recalled knowledge, generated under the recall prompt p_{recall} in the isolated recall context (p_{recall}, q); and (iii) a final answer produced by reasoning over the returned knowledge. The final-answer segment is again filtered for correctness. This yields supervision for both the tool-execution behavior and the tool-conditioned answering behavior from a single trajectory, while keeping the recall context isolated from the main reasoning context exactly as at RL and inference time.

Balancing. We use an equal 1:1 split of paired direct and AgentTool trajectories (4,000 each) so that Stage 1 does not bias the policy toward always or never invoking the tool. Because the two trajectory types are generated from the same questions, the split balances response modes rather than changing question coverage, leaving the decision of *when* the tool is beneficial to be shaped by Stage 2 RL (Section 4.2). Table 8 summarizes the composition.

H Implementation Details

We provide the hyperparameters and infrastructure used for both training stages. Unless otherwise noted, the details below describe the Qwen3-8B experiments; the Qwen3.5-9B experiments use the same two-stage training pipeline, hyperparameters, and evaluation protocol.

Stage 1 (SFT). For AgentTool RL, we fine-tune for 2 epochs on the 8,000 paired cold-start trajectories with the AdamW optimizer, a peak learning rate of 1×10^{-5} under a cosine schedule with 10% linear warmup, a global batch size of 128, and a maximum sequence length of 8,192 tokens.

Stage 2 (AgentTool RL). We optimize with GRPO (Shao et al., 2024). We use the full 19,178-question training pool for RL, with generated reference knowledge available for 8,000 of those questions. We use a prompt batch size of 128 questions and train for 8 epochs. For each question we sample $N = 8$ rollouts; rollouts that invoke the tool additionally spawn one recall generation in the isolated context. Reasoning trajectories and recall trajectories form two separate advantage groups (Eq. 6), with rewards r_{main} (Eq. 4) and r_{recall} (Eq. 5) respectively; r_{recall} is computed only when the question has generated reference knowledge, and the length-normalization constant is $C = 100$. Specifically, each RL update is formed as the union of two sample types. The reasoning subset contains the main answer-generation rollouts and is scored only by r_{main}; the recall subset contains the isolated recall generations produced by tool calls and is scored only by r_{recall}. We do not add r_{main} and r_{recall} for any single sample. Advantages are normalized within the corresponding sample type: reasoning samples are normalized among the N main rollouts for the same question, while recall samples are normalized among the recall generations spawned by tool-using rollouts for that question. The tool-invocation decision itself belongs to the main reasoning trajectory and is therefore optimized through r_{main}; r_{recall} updates only the isolated recall generation. When a question has no recall sample, it contributes no recall loss. When it has exactly one recall sample, the recall-group advantage is set to 0, avoiding an undefined or zero-variance normalization; the main reasoning group is unaffected. Questions without generated reference knowledge remain in the RL prompt pool and still contribute the main answer reward through their final-answer trajectories. If such a rollout invokes the tool, the isolated recall generation is used only as intermediate context for the final answer. We use a constant learning rate of 5×10^{-6}, a rollout temperature of 1.0, a sampling top-p of 1.0, a KL coefficient of 0.005, an entropy coefficient of 0, and greedy decoding at evaluation. The recall reward uses BAAI/BGE-M3 embeddings, matching the recall metric in Section 2.2.

Reward values. For reasoning trajectories, the format component is $r_{\text{fmt}} = -1$ if the required `<think>`/`<answer>` structure or final-answer format is invalid, and $r_{\text{fmt}} = 0$ otherwise. The answer component is $r_{\text{ans}} = 1$ if the formatted final answer exactly matches the gold option and 0 otherwise. Thus $r_{\text{main}} \in \{-1, 0, 1\}$: malformed responses receive -1, well-formed but incorrect responses receive 0, and well-formed correct responses receive 1.

GRPO baseline. The standard-GRPO baseline starts from the direct-only SFT checkpoint trained on the 4,000 direct reasoning trajectories and is trained end-to-end with only the format and answer-correctness rewards ($r_{\text{fmt}} + r_{\text{ans}}$), without the agent tool or the recall reward. It uses the same 19,178-question RL prompt pool, optimizer, learning rate, rollout count $N = 8$, and decoding settings as AgentTool RL. This avoids initializing the non-tool baseline with tool-use behavior while keeping the RL question pool matched. The third-party models (HuatuoGPT-o1-8B, MedGemma1.5-4B, AlphaMed-8B) are evaluated from their released checkpoints without further training.

Evaluation. Unless otherwise noted, accuracy is exact-match on the final selected option. For the diagnostic benchmark, knowledge probes, recall enumeration, and inference-context answering all use greedy decoding to remove sampling variance from the reported scores.

I TOOL INVOCATION DYNAMICS

Table 9 reports the final test-time invocation rates on each evaluation benchmark. The learned policy calls the tool frequently on MedQA, MedMCQA, MMLU-Med, and MedXpert, but much less often on PubMedQA, indicating that it does not collapse to an always-call strategy. Figure 6 further shows that the AgentTool invocation ratio increases over the course of RL training, indicating that the policy learns to make more frequent use of the isolated recall process when it is useful. The curve is not monotonic, however: several training windows show clear drops in invocation rate. This suggests that the learned behavior is not a blind shift toward always calling the tool; instead, the policy continues to suppress tool use on examples where additional recall is likely unnecessary or potentially noisy.

Table 9: Test-time AgentTool invocation rates (%) of the full AgentTool RL policy. Rates are measured as the fraction of evaluation examples on which the policy invokes the recall tool under the same decoding setting used for accuracy evaluation.

	In-domain		**Out-of-distribution**		
	MedQA	MedMCQA	MMLU-Med	PubMedQA	MedXpert
Invocation rate	86.88	76.90	77.93	17.00	78.00

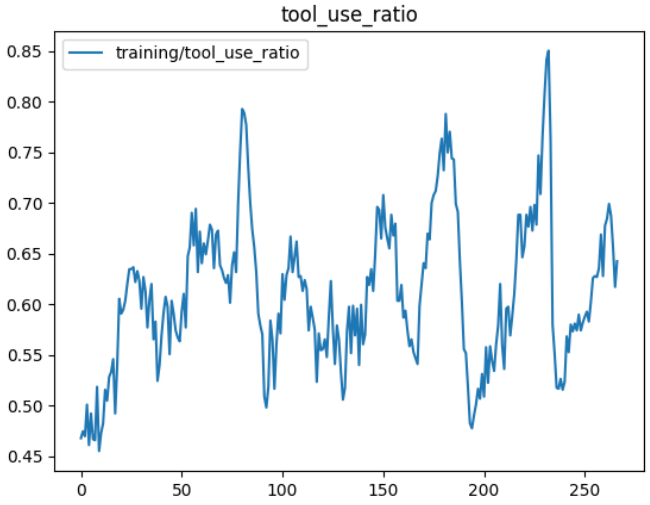


Figure 6: Training trajectory of the AgentTool invocation ratio. The curve reports the fraction of sampled rollouts in each logged training window that invoke the recall tool, i.e., the recall-call rate.

J Case Study: How the Recall Tool Changes Reasoning

Because tool invocation is decided dynamically by the policy, the $N = 8$ rollouts sampled for a single question during RL naturally contain both tool-using and tool-free trajectories on an *identical* question stem. This yields a clean within-question contrast that isolates the effect of the recall tool from question difficulty. We present two contrasting cases: Case 1, where the tool corrects an error the tool-free trajectory makes, and Case 2, where the tool instead misleads a trajectory that answers correctly without it. For each case we trim the input to the question and options only (the system prompt and tool schema are omitted) and reproduce both complete model responses unaltered, with the pivotal reasoning steps highlighted.

J.1 Case 1: Recall Helps (ground truth D)

Input.

```
Question: A 57-year-old woman is admitted to the intensive care unit for
management of shock. Her pulse is feeble and blood pressure is 86/45 mm Hg.
The patient undergoes pulmonary artery catheterization which shows an elevated
pulmonary capillary wedge pressure and increased systemic vascular resistance.
Which of the following additional findings is most likely in this patient?

Options:
A. Cold skin due to loss of intravascular fluid volume
B. Bradycardia due to neurologic dysfunction
C. Mottled skin due to release of endotoxins
```

```
D. Confusion due to decreased stroke volume
```

With recall tool (answers `D`, correct). The policy first identifies the diagnostic findings and decides to recall the hemodynamic profiles of the shock subtypes:

<think> *This question involves a patient with shock who has specific hemodynamic findings from pulmonary artery catheterization. The key findings are: shock (hypotension, feeble pulse), elevated pulmonary capillary wedge pressure (PCWP), and increased systemic vascular resistance (SVR). I should recall the different types of shock and their hemodynamic profiles to properly answer this question.* </think>

It then issues a no-argument call to the recall tool, which returns recalled knowledge in the isolated context:

```
<tool_call>{"name": "knowledgeRecall", "arguments": {}}</tool_call>
<tool_response>
The recalled knowledge:
1. Classification of shock (hypovolemic, cardiogenic, distributive, obstructive)
2. Pulmonary artery catheterization (Swan-Ganz catheter) and its measurements
3. Pulmonary capillary wedge pressure (PCWP) as a surrogate for left atrial pressure
4. Systemic vascular resistance (SVR) and its determinants
5. Cardiogenic shock pathophysiology and hemodynamic profile
6. Left ventricular failure and backward failure concept
7. Compensatory vasoconstriction in hypoperfusion states
8. Frank-Starling mechanism and ventricular filling pressures
... (25 items total; abbreviated for space)
</tool_response>
```

Conditioned on the recalled shock classification, the policy reasons through the options as an explicit *exclusion* over hemodynamic profiles and lands on `D`:

<think> *This hemodynamic profile (elevated PCWP, increased SVR, low cardiac output) is classic for cardiogenic shock. A. Cold skin due to loss of intravascular fluid volume—this describes hypovolemic shock; in hypovolemic shock, PCWP would be LOW (decreased preload), not elevated. This doesn't match. B. Bradycardia—neurogenic shock has LOW SVR, not increased. C. Mottled skin from endotoxins—distributive/septic shock has LOW SVR. D. Confusion due to decreased stroke volume—in cardiogenic shock, decreased stroke volume lowers cerebral perfusion, causing confusion; this matches the hemodynamic profile perfectly. The answer is D.* </think> <answer>D</answer>

Without recall tool (answers `A`, incorrect). On the identical question the tool-free trajectory reaches the *same* diagnosis (cardiogenic shock) and initially favors `D`, but then talks itself into the distractor `A`:

<think> *The patient presents with shock, elevated PCWP, and increased SVR. This clinical picture suggests cardiogenic shock. Evaluating the options: A. Cold skin due to loss of intravascular fluid volume—the cold skin is due to peripheral vasoconstriction from increased SVR, but "loss of intravascular fluid volume" is more typical of hypovolemic shock; this is not the best answer. D. Confusion due to decreased stroke volume—decreased cardiac output and cerebral perfusion; this fits.* Wait, let me reconsider option A. The cold, clammy skin results from sympathetic activation and vasoconstriction. A correctly identifies that the cold skin stems from the increased SVR and decreased cardiac output, even though the mechanism is not exactly loss of intravascular fluid volume. *A stands out as the most accurate answer since it reflects the compensatory vasoconstriction causing cold extremities. D is possible but less specific.* </think> <answer>A</answer>

J.2 Case 2: Recall Hurts (ground truth `A`)

Input.

```
Question: Which of the following condition masks low serum haptoglobin in hemolysis?
```

```
Options:
A. Bile duct obstruction
B. Liver disease
C. Malnutrition
D. Pregnancy
```

With recall tool (answers `D`, incorrect). The policy calls the recall tool, which returns a list emphasizing hepatic *synthesis* and acute-phase behavior of haptoglobin but omitting the impaired-*clearance* mechanism that the question hinges on:

```
<tool_call>{"name": "knowledgeRecall", "arguments": {}}</tool_call>
<tool_response>
The recalled knowledge:
1. Haptoglobin structure and synthesis (hepatic production)
2. Haptoglobin function as hemoglobin scavenger
...
11. Haptoglobin synthesis in inflammation
12. Acute phase reactants and haptoglobin
...
21. Liver disease and haptoglobin synthesis
... (27 items total; no item on cholestasis / impaired hepatic clearance)
</tool_response>
```

Steered by the recalled "acute-phase reactant" frame, the policy dismisses the correct option and settles on `D`:

`<think>` *The question asks which condition would "mask" low haptoglobin, i.e., keep it normal or elevated. A. Bile duct obstruction—this affects bilirubin metabolism but* does NOT directly affect haptoglobin synthesis. *B. Liver disease—impaired synthesis. D. Pregnancy—haptoglobin can be elevated due to its nature as an acute phase reactant.* The classic answer in medical education is that PREGNANCY elevates haptoglobin, which can mask the expected decrease in hemolysis. The answer should be D. `</think> <answer>`D`</answer>`

Without recall tool (answers `A`, correct). On the identical question the tool-free trajectory reasons directly from the clearance mechanism and answers correctly:

`<think>` *Haptoglobin binds free hemoglobin and the complex is cleared by the liver, so haptoglobin is LOW in hemolysis. A. Bile duct obstruction:* haptoglobin has a short half-life and is normally cleared through the liver; in cholestasis/bile duct obstruction, impaired hepatic clearance leads to INCREASED haptoglobin, which masks the low level expected in hemolysis. *B. Liver disease decreases synthesis, exacerbating rather than masking the low level. C. Malnutrition decreases production. D. Pregnancy does not typically elevate haptoglobin in a way that masks hemolysis. The answer is A.* `</think> <answer>`A`</answer>`

J.3 Analysis

The two cases expose the same lever acting in opposite directions: what the recall step contributes is not new facts but a *reweighting* of which considerations dominate the option comparison. In Case 1, both trajectories reach the identical diagnosis (cardiogenic shock) and diverge only at option discrimination. The tool-free trajectory is drawn to a superficially correct detail—cardiogenic shock does cause cold, vasoconstricted skin—and in reconsidering overlooks that option `A` *attributes* that cold skin to "loss of intravascular fluid volume," the mechanism of hypovolemic rather than cardiogenic shock, reversing a correct initial judgment. The recalled per-subtype PCWP/SVR profiles give a discriminative frame against which `A` is ruled out cleanly, suppressing the over-thinking self-reversal.

Case 2 is the failure mode of the same mechanism. Here the tool-free trajectory answers correctly by reasoning directly from the clearance pathway (bile-duct obstruction impairs hepatic clearance,

so haptoglobin accumulates and masks hemolysis). The recalled list, however, is complete-looking (27 items) yet omits precisely this clearance mechanism while foregrounding hepatic *synthesis* and acute-phase behavior; anchored on that frame, the tool-using trajectory explains the correct option away and drifts to the "pregnancy is an acute-phase reactant" distractor. Recall thus helps when it supplies a discriminative structure the question demands and hurts when its salient items steer attention toward a plausible-but-wrong frame. This bidirectional effect is consistent with the modest aggregate margin—tool-using rollouts are correct 72.5% of the time versus 71.1% for tool-free rollouts on the same question pool—and motivates the length-regularized recall-quality reward (Section 4.2), which rewards recall for covering the knowledge that actually resolves the question rather than for enumerating loosely related facts.