

Extended Abstract

Motivation Medical settings are often high-stakes, with heavy reliance on factuality. As a result, many recent works in medical AI use retrieval and curriculum based approaches (Gargari and Habibi, 2025; Jiménez-Sánchez et al., 2019). Additionally, medical settings makes it easy to translate outcomes into reward signals for reinforcement learning. Evidence-based reasoning in making medical decisions translates into clear reward signals Xiong et al. (2024). For our project, we use the USMLE (United States Medical Licensing Examination) subset of MedQA (Jin et al., 2021) to test whether a memory-use policy can distinguish relevant retrieved content from similar but unhelpful entries. This is particularly valuable in clinical contexts because hallucinated or misretrieved information carries risk of patient harm.

Method We frame retrieval as a learned action rather than a fixed policy: for each question we retrieve candidate passages from a chunked medical textbook corpus by similarity, and the student model learns to decide which of them to incorporate. We instantiate this with a curriculum-distillation-then-RL pipeline. A larger expert model (Qwen2.5-32B-Instruct) demonstrates selective retrieval by annotating training questions with the entries it found useful, its chain-of-thought, and a final answer. We use its correct responses to fine-tune a small student (Qwen2.5-7B-Instruct). We then refine the student’s selection policy with Group Relative Policy Optimization (GRPO), rewarding correct answers while penalizing indiscriminate use of retrieved context so the model is incentivized to only incorporate evidence that helps it reach the correct answer.

Implementation We ran a five-condition incremental ablation (zero-shot, +JSON format, +retrieved memory, +SFT, +GRPO) to attribute performance changes cleanly. For GRPO, we swept $\beta \in \{0, 0.02, 0.05, 0.1\}$, rollout counts $G \in \{8, 16\}$, and 1 vs. 3 epochs. We also evaluated inference-time strategies (i.e., self-consistency voting, cloze scoring, and reasoning-first ordering) and two rounds of DAGger iterative distillation. All conditions were evaluated on the full 1,273-question MedQA-USMLE test set with 95% Wilson confidence intervals.

Results Across SFT, four GRPO configurations, and two rounds of DAGger iterative distillation, we show through ablation testing and qualitative chain-of-thought analysis that that no variant beat the initial zero-shot baseline of 58.9%. The SFT model achieved a 51.5% accuracy under chain-of-thought evaluation and 58.4% under letter-only evaluation, which suggests that the structured output format imposed constraints on our model that resulted in decreased performance. The highest accuracy achieved by a GRPO variant was 52.2%. DAGger performed below baseline with a highest accuracy of 53.2%. We noticed that GRPO-trained models selected exactly one memory entry in 96.6% of cases, with 75% of those cases selecting the first memory entry regardless of the β value. The Spearman correlation between entry selection and retrieval relevance was near zero across all variants, which indicates a rewards-hacking problem where the model prioritizes the position of the memory entry over the content-relevance.

Discussion There are several limitations to our approach, primarily as a result of our compute and time constraints. Our memory bank was constructed using all-MiniLM-L6-v2, which is a general embedding model that is not optimized for medical text. Future research could look into using a domain-adapted embedding model. Additionally, our binary accuracy rewards model is a poor fit for a task like MedQA; we often see all rollouts answering challenging questions incorrectly, leading to a reward of 0 across all rollouts, and thus resulting in zero gradient contribution. This binary accuracy could be replaced with a shaped reward that includes retrieval quality. Our results suggest the 7B student model has a very limited ability to perform multi-step reasoning and answer domain-specific questions. Future research could test out a similar architecture with a more advanced model.

Conclusion We framed retrieval as a learned action and tested whether RL-trained selective memory improves small-model medical reasoning. Our systematic ablation across five training conditions showed that no trained configuration improved over the zero-shot no-memory baseline. We identified three primary causes of the model’s poor performance: a reward sparsity issue, an overly constrained output JSON-structure, and reward hacking in which GRPO trained the model to almost always select exactly one retrieved entry (typically the first entry) rather than learning to select all of the most relevant entries.

Memory as an Action Space: Adaptive Retrieval in Small Language Models for Medical Reasoning

Summer Royal
Department of Computer Science
Stanford University
sroyal@stanford.edu

Renn Su
Department of Computer Science
Stanford University
rrsu@stanford.edu

Abstract

Our project explores a memory-as-action framework for retrieval-augmented medical reasoning on small language models. Rather than applying static retrieval policies, we use a 32B "exper" model (Qwen2.5-32B-Instruct) to train a 7B "student" model (Qwen2.5-7B-Instruct) to selectively decide which retrieved medical textbook entries to incorporate when answering USMLE multiple-choice questions. We designed a five-stage pipeline: memory bank construction, retrieval, expert annotation via a 32B teacher, SFT warm-start, and Group Relative Policy Optimization (GRPO). We tested whether a learned memory policy can improve over both a zero-shot baseline and unconditional RAG. We ran a five-condition ablation and additional inference-time modifications including self-consistency voting, cloze scoring, and DAGger iterative distillation and found that no trained configuration improves over the zero-shot no-memory baseline of 58.9%. We identified three structural sources of the poor performance: reward sparsity from binary accuracy signals on challenging MCQ questions, training collapse under aggressive reinforcement learning scaling, and single-entry positional collapse in which GRPO-trained models cite the top-ranked retrieved entry in 75% of cases regardless of its relevance. We characterize and discuss the implications of the following failure modes: decorative citation, reasoning-answer mismatch, and case-similarity substitution.

1 Introduction

LLM-powered systems have become more commonly used in medical settings, assisting in contexts ranging from research to clinical care Haupt and Marks (2023). High-parameter large language models (32B+) excel at medical board-style questions and are able to effectively retrieve relevant content from a curriculum, but they can be expensive to deploy at scale, which disproportionately affects the medically under-resourced who are more reliant on LLM-powered clinical systems in the first place Du et al. (2024). Small models (7B) are less computationally expensive and easier to finetune, but they lack the specialized medical knowledge needed for expert-level performance Magister et al. (2023) and struggle to identify relevant content in large bodies of corpora.

Many AI-powered tools in medicine use *retrieval-augmented generation* (RAG) to ground its outputs in verified, accurate external knowledge. During the retrieval process, the model receives relevant passages from a reference corpus alongside the question to find relevant information to base its answers on Lewis et al. (2020). RAG and its successors, however, apply fixed retrieval policies, meaning they always retrieve, always prepend, and never decide. This creates two compounding failure modes. First, retrieved passages may be superficially similar to the question but factually irrelevant, especially on USMLE-style questions where clinical terminology overlaps broadly with

the textbook corpus. Second, a model that unconditionally conditions on all retrieved context cannot adapt its reliance on memory to the difficulty of the question or the quality of what was retrieved.

Critically, we study *generalizability*. Where prior medical RL succeeds on narrow, domain-specific corpora in which clinical patterns recur, MedQA-USMLE (Jin et al., 2020) spans the breadth of clinical medicine with little topic overlap across questions, so a selection policy cannot lean on a narrow band of reusable evidence. Prior medical work is uniformly performance-oriented—it treats retrieval as beneficial, reports aggregate gains, and is evaluated largely on large models—and to our knowledge none examines whether the learned selection mechanism tracks relevance rather than a surface artifact. We use the high-stakes and evidence-reliant field of medicine to explore where a confidently cited but misselected passage risks patient harm, and where cost pressures push deployment toward the smaller models in which we find the mechanism breaks down. We show that in this broad-span, setting the grounding provided by a learned selection policy is cosmetic, an illusion of robustness rather than the relevance-aware skepticism it is assumed to confer.

Specifically, we explore a memory-as-action framework: a small student model learns a memory policy to selectively retrieve entries from a medical textbook corpus, and we frame retrieval as a learned decision rather than a fixed policy. We use a curriculum distillation pipeline seeded by a 32B expert model and refined with Group Relative Policy Optimization (GRPO) to train our small "student" model to selectively choose which retrieved medical knowledge entries to inform its answer to a given question (Shao et al. (2024a)). Our task is the USMLE subset of MedQA, a benchmark of medical licensing exam questions that demands both specialized domain knowledge and multi-step clinical reasoning (Jin et al., 2021). We use Qwen2.5-7B-Instruct as the student and Qwen2.5-32B-Instruct as the expert, with retrieval from a chunked MedQA textbook corpus.

Our paper provides two central contributions. First, we conduct systematic ablation across five training conditions (zero-shot, SFT, GRPO $\beta=0$, $\beta=0.02$, $\beta=0.05$). Second we diagnose the failure modes that prevent retrieval-augmented distillation from improving over the zero-shot baseline on MedQA USMLE, including reward sparsity under binary accuracy signals and training collapse under aggressive RL scaling. We also outline the implications for future memory-policy work on small models.

2 Related Work

2.1 Adaptive Retrieval and Distillation for Small Models

Prior work has explored two main paradigms for augmenting language models with external knowledge at inference time. Lewis et al. (2020) introduced retrieval-augmented generation (RAG) as a method for grounding language model outputs in external documents. Ho et al. (2023) and Magister et al. (2023) demonstrated that chain-of-thought reasoning can be distilled from large to small models via SFT. Bengio et al. (2009) introduced curriculum learning as a strategy for ordering training examples from simple to complex. RAG and its successors apply fixed retrieval policies with no mechanism for deciding when retrieval is helpful or harmful. More recently, Suzgun et al. (2025) proposed the Dynamic Cheatsheet (DC), a test-time memory mechanism that accumulates relevant prior solved examples across queries using embedding-similarity rules and retrieves a predetermined top-k most similar at inference time. Our previous project extended this approach on small models and found that DC-RS significantly outperforms static or no-memory baselines—but only for tasks where reference corpora exist and transferable patterns can be reused (Prashar et al., 2026). Both paradigms share a key limitation: their retrieval policies are static, consulting memory regardless of relevance. Our work differs by training the model itself to make this decision via an RL reward signal. We explore the open question we take up is whether optimizing retrieval processes directly with a reward yields a more adaptive policy.

2.2 Reinforcement Learning for Retrieval and Context Selection

Prior work has explored reinforcement learning as a mechanism for training language models to make better sequential decisions. Notably, a growing body of work casts model retrieval and context into the reinforcement learning problem space, typically with Group Relative Policy Optimization (GRPO; Shao et al. (2024b)) following DeepSeek-R1 Guo et al. (2025) and building on architectures such as Reinforcement Learning with Human Feedback (RLHF) Ouyang et al. (2022) and verbal-

reinforcement methods such as Reflexion Shinn et al. (2023) that do not target retrieval. There are two broader (non-exhaustive, but relevant to this project) categories of work in this space that fall into optimizing *query generation* and *answer selection*. For query generation, Search-R1 Jin et al. (2025) treats the LM as an agent whose actions include interleaved reasoning and live search calls, and R1-Searcher Song et al. (2025) teaches when to invoke search mid-reasoning. For answer selection, DynamicRAG Sun et al. (2025) reranks via an LLM-judged answer-quality reward and Context-Picker (Zhu et al., 2025) selects a minimally viable subset and adds offline evidence distillation to overcome reward sparsity. Our work more thoroughly examines question answering by exploring whether a learned selection policy tracks relevance or exploits surface artifacts.

2.3 Medical Question Answering and Inference-Time Reasoning

There is a growing body of work grounded in applying deep learning and reinforcement learning topics to specific issues in the medical setting. Medicine is a high-stakes domain that relies on factuality, so there have been several approaches in using retrieval or corpus-based learning. Prior work applies RAG to clinical tasks such as diagnostic imaging and clinical trial eligibility screening (Gargari and Habibi, 2025; Xiong et al., 2024). Curriculum learning has also been explored for visual tasks such as bone fracture detection (Jiménez-Sánchez et al., 2019). By extension, medical settings allow for the existence of objectively correct answers makes outcome-based rewards readily available. Evidence-based reasoning in making medical decisions translates into clear reward signals Xiong et al. (2024). Due to these reward signals, GRPO has been explored for use in clinical question answering as it estimates its advantage baseline from a group of sampled rollouts rather than a learned value function, sidestepping the unstable value estimation such a sparse binary reward would induce under conditions such as PPO Dai et al. (2026); Shao et al. (2024b). GRPO has been adopted for medical reasoning, including domain-specific (e.g., cardiovascular) question answering and clinical foundation models (Ahmadi et al., 2026; Dai et al., 2026). For this project, we adopt GRPO but apply it to the memory-selection decision rather than to answer generation to investigate its effects on GRPO.

3 Method

3.1 Memory as an Action

A natural approach to retrieval-augmented medical reasoning is static RAG where we always retrieve the top- k passages and prepend them to the prompt. This is simple but has two failure modes:

1. Retrieved passages may be irrelevant or misleading despite high cosine similarity. This problem is especially relevant for our task where USMLE questions can have a lot of keywords that match parts of the corpus text that are irrelevant for answering the question.
2. A model that always receives the same five retrieved entries cannot adapt its reliance on memory to the difficulty of the question or the quality of the retrieval since it treats all retrieved context as equally trustworthy.

We instead explore and characterize selective memory use as a *decision problem*. At inference time, the model must determine which retrieved entries to incorporate. We are modeling this as a reinforcement learning problem by turning retrieval into an action space and turning answer accuracy into a reward signal, and we trained this policy using GRPO Shao et al. (2024a).

Problem formulation. More specifically,

3.2 Pipeline Overview

Our pipeline consists of five stages executed sequentially; each depends on outputs from the prior (see Figure 1):

1. **Memory bank construction:** We chunked the MedQA textbook corpus and built a cosine similarity index for retrieval.
2. **Retrieval:** For each question, we retrieved the top- k corpus entries and ranked them.

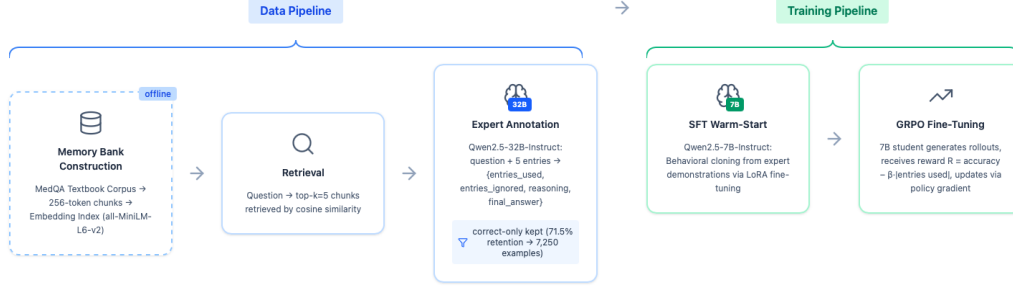


Figure 1: Memory-Augmented MedQA Pipeline

- Expert annotation:** Used a 32B "expert" to demonstrate selective retrieval. The expert is given the question and retrieved entries and then outputs a structured response indicating which entries it used, step-by-step reasoning, and a final answer. We discarded any expert outputs that weren't parsable.
- SFT warm-start:** Finetuned the 7B student on the expert-annotated examples to initialize the policy in the correct structured-output format and transfer the expert's selective behavior before reinforcement learning training started. The student is trained by minimizing the token-level negative log-likelihood of the expert demonstrations,

$$\mathcal{L}_{\text{SFT}}(\theta) = - \sum_{(x, o^*) \in \mathcal{D}} \sum_{t=1}^{|o^*|} \log \pi_{\theta}(o_t^* | x, o_{<t}^*), \quad (1)$$

where o^* is a parseable, expert-correct annotation. This is behavior cloning of the expert's selective-retrieval policy: it initializes π_{θ} in the structured output format and the expert's selection behavior so that GRPO begins from a policy that produces parseable rollouts with enough correct answers to yield a non-zero gradient.

- GRPO fine-tuning:** Applied group relative policy optimization to refine the student's memory-use policy using an accuracy-and-efficiency reward.

The SFT stage bootstraps the student into a regime where it reliably produces well-structured JSON outputs and uses only selectively uses retrieved memory as needed so that GRPO can begin from a policy that generates diverse rollouts with a sufficient frequency of correct answers to produce a meaningful gradient signal. Attempting GRPO without this initialization would have risked training collapse from the start; a base model rarely produces parseable JSON files (especially with the QWEN model, which seemingly had difficulty adhering to a consistent structured JSON format), so that nearly all rollouts would receive the parse-failure penalty and no accuracy reward.

3.3 Expert Annotation via Curriculum Distillation

The expert (Qwen2.5-32B-Instruct) receives each training question, its four answer choices, and the top-5 retrieved entries ranked by cosine similarity, and must produce a single JSON object with four fields: entries_used (a list of entry ranks the expert found helpful), entries_ignored (the remaining ranks), reasoning (chain-of-thought), and final_answer (a single letter A, B, C, or D).

We use curriculum distillation Bengio et al. (2009) where we filter out any responses where the expert's selective retrieval led to a correct answer choice so that the SFT dataset does not have any noise from the expert's failed attempts. to expert-correct responses before including examples in the SFT dataset. The 32B expert achieves 71.5% accuracy on the training set; the remaining $\sim 29\%$ of questions where the expert errs are dropped, yielding 7,250 clean SFT examples ($\sim 71\%$ retention rate from 10,178 training questions).

3.4 GRPO Memory Policy Training

After SFT, we apply GRPO Shao et al. (2024a) to optimize the student’s memory-use policy with respect to accuracy. For each training prompt, GRPO samples rollouts from the current policy, computes per-rollout rewards, and updates the model using the group-relative advantage. Advantage is calculated as: the reward minus the mean reward across the number of rollouts for that prompt. Since the advantage is computed relative to the group mean rather than an estimated value function, GRPO requires no learned critic network, which is beneficial here since our sparse binary accuracy reward would make stable value estimation difficult. While prior work often reports aggregate accuracy after process-level signals, our sparse binary reward setup surfaces errors and isolates diagnostic signals.

The reward function combines the following terms:

$$R = \alpha \cdot \mathbf{1}[\text{correct}] - \beta \cdot |\text{entries_used}| \quad (2)$$

where $\alpha \cdot \mathbf{1}[\text{correct}]$ is a binary reward signal for answering the question correctly and $-\beta \cdot |\text{entries_used}|$ penalizes overuse of retrieved entries to encourage the model to cite only genuinely helpful context rather than using all five entries by default. β is the key hyperparameter that controls the accuracy-vs-efficiency tradeoff. Note that $\beta=0$ removes this term entirely.

For each prompt x , GRPO samples a group of G responses $\{o_1, \dots, o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$, scores each with Eq. 2, and computes a group-relative advantage

$$\hat{A}_i = R(o_i) - \frac{1}{G} \sum_{j=1}^G R(o_j), \quad (3)$$

which is assigned to every token of o_i .¹ The policy is updated by maximizing the clipped surrogate objective

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x, \{o_i\}} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min \left(\rho_{i,t} \hat{A}_i, \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon) \hat{A}_i \right) - \lambda_{\text{KL}} D_{\text{KL}}(\pi_{\theta} \parallel \pi_{\text{ref}}) \right], \quad (4)$$

where $\rho_{i,t} = \pi_{\theta}(o_{i,t} \mid x, o_{i,<t}) / \pi_{\theta_{\text{old}}}(o_{i,t} \mid x, o_{i,<t})$ is the per-token importance ratio and π_{ref} is the frozen SFT checkpoint.

A key structural challenge in applying GRPO here is *reward sparsity*; on many prompts, all G rollouts receive the same outcome (all correct or all incorrect), making the group-relative advantage zero and thus producing zero gradient update. For the USMLE questions that are very challenging and require a lot of domain-specific knowledge, it’s often the case that all rollouts get the question wrong. We partially address this by using sampling temperature 1.0 during rollout generation (to increase diversity within a rollout group) and increasing the number of rollouts as part of our ablation study (to increase the probability of mixed-outcome groups), but reward sparsity remains a structural limitation of binary rewards. This is especially an issue for our specific task given that this model is already performing close to its practical limits on this benchmark given its size. There is little variance left to exploit, and GRPO’s signal is not relevant when the model simply does not have the capacity to answer the more challenging questions.

3.5 Alternate Approaches and Post-Training Variants

To ensure comprehensiveness and to characterize the impact of post-training choices on MedQA performance, we explore four additional strategies:

Dagger. As an alternative to one-shot curriculum distillation, we also evaluate DAgger (Ross et al., 2011), an iterative imitation-learning algorithm. Each round, we roll out the current student, identify the questions it answers incorrectly, have the 32B expert relabel those states, and aggregate the new demonstrations into the training set for the next round. This addresses the distribution-shift weakness of pure behavior cloning, in which the student encounters states (here, harder questions) underrepresented in the initial expert set. We run and report results on DAgger as an additional comparison.

¹The original formulation (Shao et al., 2024b) additionally normalizes by the group standard deviation; we use the unnormalized group-mean baseline.

Post-Training Modifications. Our inference explores three post-training approaches, namely Cloze Scoring, Self-Consistency Voting, and Reasoning-First Prompting. Each isolates a candidate cause of the model’s sub-baseline accuracy (i.e., decoding, output format, or latent knowledge) rather than introducing a new training signal.

Cloze scoring. Rather than generating an answer letter, we score each option by the likelihood the model assigns to its text conditioned on the question and select the highest-scoring option Brown et al. (2020); Holtzman et al. (2021). This bypasses lossy free-form generation and surface-form competition, isolating whether the bottleneck is the model’s parametric knowledge rather than its decoding or output formatting.

Self-consistency voting. We sample k independent chain-of-thought trajectories at temperature $T=0.7$ and take the majority answer Wang et al. (2022). The method exploits the tendency of correct reasoning paths to agree more often than incorrect ones; it tests whether greedy decoding leaves a recoverable correct-answer mode unexploited.

Reasoning-first ordering. We reorder the supervised output schema so the chain-of-thought precedes the committed answer letter, rather than emitting the letter first Kojima et al. (2022); Wei et al. (2022). We wanted to make sure that the model’s reasoning was not a post-hoc reverse engineering of the chosen answer. To explore causality between reasoning and multiple choice answer, we note that if reasoning-first improves accuracy, the original format imposed a measurable cost.

3.6 Alternatives Considered

Always-retrieve RAG (no selection). Using a zero-shot model with all five retrieved entries always prepended is simpler and eliminates the SFT and RL stages. However, we recognized that it cannot adapt retrieval to question difficulty, and it is maximally vulnerable to using conflicting or irrelevant memory: all five entries are always attended to, so non-related information can send a unclear signal. Our selective framework is designed to be more robust by training the model to actively reject unhelpful context.

PPO instead of GRPO. PPO requires a learned value network to estimate baselines. GRPO replaces this with group-relative advantages, removing the critic entirely. In our sparse-reward setting (where rollout groups can easily be all-correct or all-incorrect), a learned value function would face significant variance and potentially destabilize training. GRPO’s group baseline is a simpler and more robust choice for binary accuracy rewards.

4 Experimental Setup

4.1 Dataset and Task

We evaluated the student model on the USMLE subset of MedQA Jin et al. (2020) questions, a compilation of United States Medical Licensing Exam multiple-choice questions. These questions require multi-step clinical reasoning and specialized medical domain knowledge. The dataset contains 10,178 training examples and 1,273 evaluation examples, each consisting of a clinical vignette and four answer choices (A, B, C, and D). We use the full 1,273-example evaluation set as our authoritative test set for all reported accuracy figures. We used a smaller 200-question pilot evaluation in early development while we tested out different strategies, but we solely report our results on the full dataset.

4.2 Models

We use **Qwen2.5-7B-Instruct** as the student model and **Qwen2.5-32B-Instruct** as the expert. For one of our ablation conditions, we ran our full pipeline on a Qwen2.5-72B-Instruct for comparison. All the models used in our project are Instruct-tuned variants. We did not use base (non-instruct) checkpoints anywhere in the pipeline so that all comparisons are Instruct-vs-Instruct and differences cannot be attributed to instruction-following capability.

4.3 Hyperparameters

Memory bank: We chunked the MedQA textbook corpus into 256-token passages with a 32-token sliding word overlap. Each chunk was embedded using all-MiniLM-L6-v2 with L2 normalization, and we retrieved the top $k = 5$ entries per question by cosine similarity at inference time.

SFT: We fine-tuned using LoRA with rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05, training for 2 epochs with a learning rate of 2×10^{-4} and weight decay 0.01. The effective batch size was 32 (per-device batch size 4 with gradient accumulation over 4 steps), and training ran across 2 GPUs with gradient checkpointing enabled and a maximum sequence length of 2,048 tokens.

GRPO: We fixed $\alpha=1.0$ and $\gamma=0.5$ throughout, and swept $\beta \in \{0.0, 0.02, 0.1\}$ and rollout count $G \in \{8, 16\}$ as part of our ablation. We evaluated both 1 and 3 training epochs. The KL coefficient was set to 0.04 and sampling temperature to 1.0, with a learning rate of 5×10^{-7} , per-device batch size 2, and gradient accumulation over 2 steps. LoRA configuration matched SFT ($r=16$, $\alpha=32$). Rollouts were generated via vLLM in colocated mode across 8 GPUs.

4.4 Ablation Conditions

We use a five-condition incremental ablation in which each condition adds exactly one pipeline component over the prior, enabling clean attribution of performance changes to individual components:

1. **Zero-shot:** Letter-only prompt, no memory, base Qwen2.5-7B-Instruct weights.
2. **+JSON format:** Structured JSON output prompt matching the training schema; no memory; base weights. Isolates the cost of the output format constraint.
3. **+Retrieved memory:** JSON prompt with all 5 retrieved entries provided; base weights. Isolates the effect of retrieval without any learned selection.
4. **+SFT:** Student fine-tuned on expert-generated examples (7,250 examples, $\sim 71\%$ retention from expert-correct filtering).
5. **+GRPO:** RL-trained checkpoint with reward R from Equation 2.

Additional Post-Training Modifications

In addition to our five-condition ablation, we conducted the following standard experiments at inference time to evaluate the impact of post-training decisions on the model’s MedQA performance:

1. **Cloze-scoring:** Cloze-scoring bypasses lossy free-form generation by scoring option-text likelihood directly Holtzman et al. (2021).
2. **Self-consistency voting:** Sample k independent CoT trajectories ($T = 0.7$) and take the majority answer Wang et al. (2022).
3. **Reasoning-first on SFT:** Move CoT reasoning ahead of the committed letter, as opposed to a post-hoc rationalization Kojima et al. (2022); Wei et al. (2022).
4. **Dagger:** Use expert-guided demonstrations throughout training to compare against curriculum distillation Ross et al. (2011).

All conditions are evaluated on the same 1,273-question set. We report the accuracy with 95% Wilson confidence intervals. All training and evaluation runs on Modal Labs cloud infrastructure.

5 Results

5.1 Quantitative Evaluation

Table 1 reports held-out accuracy on the full MedQA USMLE evaluation set ($n = 1,273$) for all training conditions and post-training modifications. The 32B expert achieves 71.5% accuracy on the training set and serves as a reference for the knowledge gap the distillation pipeline attempts to bridge.

No trained configuration improves over the zero-shot no-memory baseline under the consistent chain-of-thought (CoT) evaluation protocol. The base model achieves 58.9% without memory. Every

Table 1: MedQA-USMLE held-out results ($n=1,273$). The primary evaluation protocol is chain-of-thought JSON (CoT); letter-only rows indicate conditions where the model is asked to only output the a multiple choice letter. *Used* is the mean number of selected memory entries out of five retrieved. CI=95% Wilson interval. Dashes indicate no confidence interval was computed. Other than for the teacher, all baselines refer to the 7B model.

Model	Training setting	Eval method	Accuracy	95% CI	Used
<i>Baselines</i>					
Base	No memory	Letter	59.9%	—	0.0
Base	No memory	CoT	58.9%	56.2–61.6	0.0
Base	All five retrieved entries	CoT	57.9%	55.2–60.6	5.0
Teacher	All five retrieved entries	CoT	71.4%	—	5.0
<i>Supervised fine-tuning</i>					
SFT	2 epochs	Letter	58.4%	—	1.2
SFT	2 epochs	CoT	51.5%	48.8–54.3	1.0
SFT	2 epochs	Reason-first vote ($k=16$)	44.6%	41.9–47.4	1.5
<i>Group Relative Policy Optimization</i>					
GRPO	$\beta=0.0$, e1/g8	Letter	58.3%	—	1.2
GRPO	$\beta=0.05$, e1/g8	CoT	50.0%	47.3–52.8	1.1
GRPO	$\beta=0.02$, e3/g16	CoT	51.8%	49.1–54.6	1.0
GRPO	$\beta=0.05$, e3/g16	CoT	52.1%	49.3–54.8	1.0
GRPO	$\beta=0.10$, e3/g16	CoT	52.2%	49.4–54.9	1.0
GRPO	$\beta=0.10$, e3/g16	Self-consistency vote ($k=16$)	50.5%	47.8–53.3	1.1
<i>DAGger iterative distillation</i>					
DAGger	Round 1, 7,856 records	CoT	52.2%	49.5–55.0	3.1
DAGger	Round 2, 8,171 records	CoT	53.2%	50.4–55.9	3.7

trained variant including SFT (51.5%), all GRPO configurations (50.0–52.2%), and two rounds of DAGger iterative distillation (52.2–53.2%), falls 5 to 9 percentage points below our baseline. GRPO with $\beta=0$ holds approximately even with the zero-shot model under the letter-only evaluation protocol (58.3% vs. 59.9%), but no configuration achieves a statistically meaningful improvement over the baseline under any consistent protocol.

Under CoT evaluation, the base model without memory achieves 58.9% (95% CI: 56.2–61.6). Providing all five retrieved corpus entries unconditionally yields 57.9% (55.2–60.6), a small but consistent decrease, indicating that unfiltered retrieval provides no benefit on MedQA USMLE.

The SFT model trained on 7,250 expert-correct examples achieves 51.5% accuracy under CoT evaluation, which is 7.4 points below the no-memory baseline. Under letter-only evaluation, which matches the SFT training format (reasoning traces were discarded and only the final answer letter was supervised), accuracy recovers to 58.4%, near the letter-only baseline of 59.9%. Applying a reason-first self-consistency vote over $k=16$ samples further degrades accuracy to 44.6%.

β Selectivity Parameter Sweeping: We swept $\beta \in \{0.0, 0.02, 0.05, 0.1\}$ and training intensity at two scales: 1 epoch / 8 rollouts and 3 epochs / 16 rollouts. Under letter evaluation, $\beta=0$ holds near the zero-shot letter baseline at 58.3%, while any $\beta > 0$ collapses letter accuracy between 48% and 51% and average memory use approaches zero. Under the consistent CoT evaluation, all GRPO configurations fall between 50.0% to 52.2%. Scaling from 1 epoch with 8 rollouts to 3 epochs with 16 rollouts yields a consistent but small lift of approximately +2 percentage points at every β , with no configuration approaching the zero-shot baseline:

β	1 epoch & 8 rollouts (letter)	3 epoch & 16 rollouts (letter)	Δ
0.02	48.1%	49.9%	+1.8
0.05	48.1%	50.6%	+2.5
0.10	48.0%	50.9%	+2.9

GRPO’s group-relative advantage signal requires mixed-outcome rollout groups; if all rollouts for a given prompt are either correct or all rollouts are all incorrect, the advantage is zero and no gradient update occurs. On MedQA USMLE, the 7B student is already near its maximum performance

capacity, and the hardest questions that require a useful gradient signal are exactly the ones where all rollouts fail (since all rollouts get those questions incorrect). Increasing rollout count from 8 to 16 and training from 1 to 3 epochs yielded a consistent but small +2 point lift across all β values, but it still did not outperform the baseline. Thus, our results suggest that binary accuracy rewards are a poor fit for tasks where the model lacks the underlying capacity to answer correctly regardless of retrieval behavior.

5.1.1 Analyzing post-training modifications

Three inference-time strategies were evaluated in an attempt to better understand the architectural causes of our model’s poor accuracy.

Self-consistency voting: Majority vote over $k=16$ chain-of-thought samples ($T=0.7$) produced no improvement. For the best GRPO variant ($\beta=0.1$, e3/g16), vote-only accuracy of 50.3% was within noise of greedy decoding at 52.2%. For SFT, voting was clearly worse than greedy (44.6% vs. 51.5%). Cloze-only scoring was substantially weaker for both models (37–40%), and agreement between the vote and cloze signals was low (53% for GRPO, 39% for SFT).

Dagger iterative distillation: Two rounds of iterative distillation—in which the 32B expert relabels training examples the student currently fails—produced consistent round-over-round improvement: Round 1 (7,856 examples, 2,965 relabeled by the expert) reached 52.2% (49.5–55.0); Round 2 (8,171 examples, 2,730 relabeled) reached 53.2% (50.4–55.9). The expert recovered approximately 62% of the student’s failures in each round. Despite the progressive gains, both DAgger rounds remain approximately 6 percentage points below the no-memory baseline.

Reward sparsity: A structural property of the GRPO training signal is that binary accuracy rewards on 4-option MCQs frequently produce degenerate rollout groups in which all G rollouts share the same outcome, yielding zero group-relative advantage and no gradient update. This is especially prevalent on harder MedQA USMLE questions where the 7B student is near its performance ceiling.

5.2 Qualitative Analysis

Our results suggest that selective retrieval was learned but does not track relevance. All trained models under stable training conditions consistently produce well-formed JSON with parse rates of 97.9–100% and cite an average of 1.0–1.5 out of 5 retrieved entries per question. However, Table 2 shows that the Spearman correlation between the number of entries selected and their retrieval relevance is near zero across every variant ($\rho \approx -0.01$ to -0.06 , all $p > .03$).

Table 2: Memory selectivity for trained variants. Selectivity is measured as the Spearman correlation ρ between the number of memory entries selected and their retrieval relevance. Values near zero indicate the model did not learn to use memory more selectively. p -values in parentheses.

Model	Training setting	Eval	ρ (p)
SFT	2 epochs	CoT	−0.020 (.47)
SFT	2 epochs	Reason-first vote	−0.014 (.61)
GRPO	$\beta=0.05$, e1/g8	CoT	−0.045 (.11)
GRPO	$\beta=0.02$, e3/g16	CoT	−0.049 (.08)
GRPO	$\beta=0.05$, e3/g16	CoT	−0.033 (.23)
GRPO	$\beta=0.10$, e3/g16	CoT	−0.045 (.11)
GRPO	$\beta=0.10$, e3/g16	Self-consistency vote	−0.045 (.11)
DAgger	Round 1	CoT	−0.057 (.04)
DAgger	Round 2	CoT	−0.045 (.11)

Among GRPO-trained models, 96.6% of responses cite exactly one memory entry (aggregated across all β values; see Appendix Table 4). Of those single-entry selections, the overwhelming preference of the first-ranked (position-1) entry is consistent across all β configurations (Appendix Table 5). Rather than learning to select entries by content relevance, GRPO-trained models cited exactly one entry in 96.6% of cases, with 75.0% of selections being the first entry (position-1) entry regardless of β . The Spearman correlation between entry selection and retrieval relevance is near zero across all trained variants ($\rho \approx -0.01$ to -0.06 , all $p > .03$). This indicates the model learned a reward-minimizing heuristic in which it cites one entry (usually the first) rather than a content-driven memory policy. We

introduced the memory-use penalty β to discourage indiscriminate full-context use, but it instead reinforced minimal citation without improving selectivity.

Table 3 summarizes failure modes identified through manual inspection of model outputs across SFT and GRPO configurations. The dominant patterns are single-entry collapse toward the highest-ranked retrieval result, decorative citation of memory without extracting actionable facts, and reasoning–answer mismatches in which the chain of thought supports one option but the model commits to another.

Table 3: Qualitative failure taxonomy for the MedQA SFT $\rightarrow$ GRPO selective-retrieval pipeline. Manual inspection identifies five recurring failure patterns. The dominant pathology is not the absence of reasoning, but failure to make retrieval causally useful for multiple-choice medical QA. For exact questions and verbatim memory usage, see the appendix.

Failure mode	Representative output	Interpretation
Reward hacking: selecting a single entry	entries_used: [1], entries_ignored: [2,3,4,5]	The model selects exactly one retrieved entry in 96.6% of cases (see Table 4), reflecting reward-shaped retrieval minimization rather than adaptive evidence use.
Default to using the first entry	entries_used: [1], entries_ignored: [2,3,4,5] <i>all five entries have near identical cosines (0.687, 0.685, 0.684, 0.683, 0.682)</i>	The model disproportionately defaults to the first retrieved entry in 75.0% of cases (see Table 5), suggesting a positional heuristic rather than selection by relevance.
Matching on irrelevant commonalities	<i>"Entry 1 provides a context of a patient with a Swan-Ganz catheter, but it does not provide specific normal values for right ventricular pressures. Therefore, the correct answer is A"</i>	The question asked what blood pressure a Swan-Ganz catheter would read in the ventricle of a healthy man with no known heart conditions. The retrieved entry was about intubating a burn victim with a Swan-Ganz catheter to evaluate his hemodynamic volume status. The irrelevant Swan-Ganz catheter is the only similarity between these two entries.
Reasoning–answer mismatch	<i>"... a muscarinic antagonist, which is option C... the most appropriate answer is C."</i> with final_answer: "A"	The model reaches an intermediate conclusion supporting one option but outputs a different final answer, indicating weak option grounding.
Case-similarity substitution	<i>"The case described in entry 1 matches the symptoms of ADHD, including impaired attention, which is a key feature of the disorder."</i>	The model justifies answers by surface case resemblance rather than identifying the specific principle that distinguishes the correct option. (An entry about a youth ADHD case is misapplied to a youth autism case.)

For each case outlined in the table 3 the appendix contains the verbatim question and answer options, the gold and model-selected options, and the complete set of retrieved memory candidates with their retrieval cosine similarities (see Section C).

Failure Mode Analysis: Ultimately, data suggests that no trained configuration (SFT, RPO across four values of β , or DAgger iterative distillation) can improve over the zero-shot no-memory baseline. We found that qualitative analysis of reasoning traces further explain causes of failure:

1. **Reward hacking causes concerns:** Through our failure modes, we further showed that GRPO-trained models commonly cited exactly one entry (96.6% of cases), with 75% of selections being the first entry (position-1) entry, even when cosines are within 0.01 of one another. This indicates the model learned a reward-minimizing heuristic in which it cites one entry (usually the first) rather than a content-driven memory policy.
2. **Model capacity bottlenecks are visible in reasoning:** We show how our 7B student model has errors in the downstream inference process that do not appear to have a causal relationship with SFT or curriculum distillation. Failure modes such as reasoning–answer mismatches (outputting the wrong answer compared to a correct reasoning trace) and using irrelevant commonalities (notably with reasoning relationships such as "therefore") suggest that failures can partially be attributed to the model’s own capabilities to follow instructions and internalize reasoning.

3. **Surface similarity does not imply structural relevance:** Many observed failures happen as cosine retrieval returns passages that are topically close but structurally unhelpful (e.g., similar patient demographics, a tool like the Swan-Ganz catheter used in two different procedures). The retrieved entries often match the question on surface features, such as shared symptoms, procedures, patient demographics, or disease labels, but do not contain the causal mechanism, diagnostic distinction, or factual value required to answer the question (see: failure modes of case-similarity and matching on irrelevant commonalities). The student then treats this surface overlap as sufficient evidence. This indicates that the retrieval-selection pipeline breaks down at the boundary between superficial similarity and informational relevance: it retrieves cases that look similar, but does not reliably identify whether those cases support the specific reasoning step required by the answer.

6 Discussion

6.1 Limitations

The memory bank was constructed using all-MiniLM-L6-v2 which is a small general-purpose embedding model that is **not** optimized for medical text. We noticed that retrieved entries that are ranked highly by cosine similarity frequently share surface-level clinical terminology with the question without containing the specific factual content needed to answer it (see Appendix Table 4), which is likely a result of the overly limited capacity of our model. A medical-domain embedding model (BioSentBERT, MedCPT, etc.) could substantially improve retrieval precision.

A four-option multiple-choice task with a small 7B model has limited outcome variance. A dense reward that incorporates partial credit, calibration, retrieval quality, or any nuance beyond a binary answer accuracy, would provide a stronger gradient signal for GRPO and reduce our reward sparsity problem Ahmadi et al. (2026). This is especially relevant for our task where we have a diversity of question difficulties including some questions that require multi-step reasoning and complex domain-specific knowledge.

Relatedly, our results suggest that the 7B student is operating close to its maximum performance capacity on the MedQA USMLE task, which suggests that even a well-learned memory policy cannot make up for the underlying lack of domain-specific knowledge. We believe that the poor accuracy is partly due Qwen’s limited capacity, as opposed to a weakness in our retrieval policy, since prior work on the same MedQA-USMLE achieved a much higher baseline performance of 90-95% using OpenAI’s o4 Mini High and Gemini 2.5 Pro models Arkangel AI (2025). We were constrained to using Qwen2.5-32B-Instruct and Qwen2.5-7B-Instruct due to compute limitations. Our initial hypothesis (that selective retrieval can bridge the 7B–32B performance gap) may not hold when the student lacks the parametric knowledge to reason over retrieved context correctly. The fact that DAGger iterative distillation showed small but consistent round-over-round improvement (52.2% → 53.2%) suggests that additional supervision can incrementally make progress, but the rate of improvement implies diminishing returns well before the baseline.

The memory-use penalty β was unstable; any $\beta > 0$ immediately resulted in formatting failures and near-constant use of 1 memory entry rather than gradual, progressive selectivity. This suggests the boundary between stable and unstable training is effectively at $\beta = 0$, meaning it is not possible for the penalty to encourage meaningful selectivity without destabilizing the output format. A smoother penalty schedule, such as a schedule that anneals β from zero over training rather than applying it at full strength from the first update, may help obtain the intended tradeoff between accuracy and retrieval efficiency.

6.2 Future Directions

One possible next step could include replacing the binary accuracy with a shaped reward that includes retrieval quality. For example, future researchers could look into rewarding the model when selected entries contain n -gram overlap with the gold reasoning chain, or penalizing selection of entries whose content is provably irrelevant. This could address the sparsity problem while more precisely targeting the causal link between retrieval and correct reasoning.

The qualitative failure taxonomy in Table 3 shows that the dominant failure mode is not an inability to reason, but an inability to make retrieval causally useful: retrieved entries are cited at a surface

level without extracting specific factual support. Future architectures might train a separate retrieval critic that scores entries on their inferential utility rather than lexical overlap, or use chain-of-thought supervision that explicitly links retrieved facts to intermediate reasoning steps.

Since our memory policy can only be as good as the entries that it retrieves, future research could benefit from using a domain-adapted embedding model and a larger medical corpus. Future research could look into improving the quality of the candidate set that we pass to the policy by running contrastive or hard-negative retrieval finetuning on USMLE questions.

Our DAGger results suggest that iterative expert relabeling of the student’s failures produces some small but consistent improvement, so future research could look at a more aggressive DAGger schedule with more rounds or relabeling a larger percentage of the training dataset.

7 Conclusion

In this paper, we reframe memory retrieval as an action that a small-parameter "student" model (Qwen2.5-7B) can learn from a large-parameter "expert" model (Qwen2.5-32B). We tested to see whether an expert model could train the small model to selectively incorporate these retrieved medical textbook entries to answer USMLE multiple-choice questions. Our pipeline included curriculum distillation from the 32B expert model, SFT warm-start, and GRPO with varies memory-use penalties. We found that no trained configuration improved over the zero-shot no-memory baseline that had an accuracy of 58.9%.

We performed systematic ablation tests to isolate the various causes of the model’s poor performance, and we identified three key sources. First, binary accuracy rewards on a four-option multiple choice task with a very limited small 7B model produce a severe reward sparsity issue. The challenging questions where a gradient signal would be most useful are the exact questions where all rollouts fail. Second, supervised training on JSON-structured outputs imposed a format constraint that limited the model’s accuracy at inference time. Chain-of-thought acted as post-hoc rationalization rather than causally integrated reasoning. Lastly, GRPO trained the model into almost always selecting the first retrieved entry rather than acting as a content-driven selection policy as evidenced by the fact that 96.6% of responses cited exactly one entry and 75% of which were the first-ranked retrieval regardless of relevance.

To explore the role of retrieval in medical contexts, we provide quantitative and qualitative insights that show how curriculum learning and retrieval-based medical systems can fail in a generalized and multi-step task such as MedQA. Our results and analysis suggest that small models can systematically fail at complex medical tasks, especially in an action space where reward structures are sparse. While there are promising medical applications of LLMs built on retrieval and distillation, our findings suggest that these systems should not be treated as inherently more reliable simply because they cite retrieved evidence. As smaller and cheaper medical AI models become more widespread and commercialized, future work should ensure that these systems not only retrieve evidence, that evidence is selected, interpreted, and used in a causally meaningful way.

8 Team Contributions

Both project partners contributed equally to the ideation and project design.

- **Summer Royal:**

For this report: Wrote second half of extended abstract, the regular abstract, introduction, most of methods, experimental setup, parts of results/analysis, discussion, and conclusion sections.

For the code: Fixed accuracy bug in fetch_labbench (answer shuffling), built MedQA data pipeline and memory bank for milestone baseline, baseline testing setup, uploaded and organized curriculum text, built embed_curriculum.py with JSON support, managed source papers and full SFT pipeline, fixed eval bug, increased rollout epochs and GPU count for inference scaling, updated system prompt, fixed timeout problems, and incorporated TA feedback.

- **Renn Su:**

For this report: Wrote the first half of the extended abstract, related works, parts of methods,

parts of results/analysis, and helped edit the other sections. Set up the second half of the implementation pipeline and helped debug the first half of the pipeline, ran SFT evaluations. For the code: Initial memory baseline setup, built GRPO pipeline with Modal and caching support, DC-RS and memory verification variants of SFT, VMG variant, GRPO implementation first pass, fixed GRPO stage 4 (pinned TRL, patched broken optional-dep imports), stage 2 teacher outputs with SFT train/eval and teacher stats, LoRA and Modal optimizations for step 3 SFT, Modal pre-run naming conventions and checkpoints, all model baselines, teacher runs, evaluation harness support, COT modification on JSON output, results for GRPO/SFT and post-training variants

9 Changes from Proposal

We document the following changes from our proposal:

1. **From new architecture to failure mode analysis:** Our initial goal was to propose a new pipeline that utilizes memory as an action space to enhance performance on MedQA. As our results landed consistently below baseline, we pivoted to a thorough analysis of failure modes and the pitfalls of deploying similar architectures in medical settings.
2. **Addition of DAGger and post-training modifications:** We found that using an ablation test was not enough to more thoroughly comprehend the failure modes and definitively explain poor performance. As a result, we broadened our scope of experiments to include DAGger, Cloze-scoring, self-consistency voting, and reasoning-first prompting.
3. **Increased focus on chain of thought:** To better understand our model performance and provide traceability, we enabled and analyzed chain of thought and model traces. We provide details in the qualitative analysis and appendix sections.

10 Disclaimer of AI Assistance

- We did all of the project ideation ourselves without using AI. We came up with the project idea and design decisions without consulting AI.
- We designed the architecture for our pipeline by ourselves without using AI.
- We forked the initial repo from a previous Dynamic Cheatsheet project that we (Renn and Summer) worked on.
- We wrote this report ourselves without using AI.
- We used Claude Code to implement the majority of our extensions. Again, we came up with the designs for the extensions ourselves, but we used Claude Code to write and debug the code.

References

Arash Ahmadi, Parisa Masnadi, Sarah Sharif, Charles Nicholson, David Ebert, and Mike Banad. 2026. Improving Heart-Focused Medical Question Answering in LLMs via Variance-Aware Rubric Rewards with GRPO. *arXiv preprint arXiv:2606.05174* (2026).

Arkangel AI. 2025. Arkangel AI real-time multi-LLM agent answers clinicians' medical questions with 90% accuracy. <https://arkangel.ai/en/research/arkangel-ai-real-time-multi-llm-agent-answers-clinicians-medical-questions-with-90-accuracy>. Accessed 2026-06-08.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

- David Dai, Peilin Chen, Chanakya Ekbote, and Paul Liang. 2026. Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training. *Advances in Neural Information Processing Systems* 38 (2026), 37406–37453.
- Delong Du, Richard Paluch, Gunnar Stevens, and Claudia Müller. 2024. Exploring patient trust in clinical advice from AI-driven LLMs like ChatGPT for self-diagnosis. *arXiv preprint arXiv:2402.07920* (2024).
- Omid Kohandel Gargari and Gholamreza Habibi. 2025. Enhancing medical AI with retrieval-augmented generation: a mini narrative review. *Digital health* 11 (2025), 20552076251337177.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).
- Claudia E Haupt and Mason Marks. 2023. AI-generated medical advice—GPT and beyond. *Jama* 329, 16 (2023), 1349–1350.
- Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*. 14852–14882.
- Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn’t always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 7038–7051.
- Amelia Jiménez-Sánchez, Diana Mateus, Sonja Kirchhoff, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Miguel A González Ballester, and Gemma Piella. 2019. Medical-based deep curriculum learning for improved fracture classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 694–702.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Serkan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516* (2025).
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *arXiv preprint arXiv:2009.13081* (2020). <https://arxiv.org/abs/2009.13081>
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* 11, 14 (2021), 6421.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
- Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 1773–1781.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

- Jaanak Prashar, Renn Su, and Summer Royal. 2026. Small Models Think Big: Toward Effective Memory Distillation for Small Co-Scientists. <https://web.stanford.edu/class/cs224n/project.html> Stanford CS224N Custom Project.
- Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 627–635.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, et al. 2024a. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300* (2024).
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems* 36 (2023), 8634–8652.
- Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592* (2025).
- Qingyun Sun, Jiaqi Yuan, Shan He, Xiao Guan, Haonan Yuan, Xingcheng Fu, Jianxin Li, and Philip S Yu. 2025. DyG-RAG: Dynamic Graph Retrieval-Augmented Generation with Event-Centric Reasoning. *arXiv preprint arXiv:2507.13396* (2025).
- Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. 2025. Dynamic Cheatsheet: Test-Time Learning with Adaptive Memory. *arXiv preprint arXiv:2504.07952* (2025).
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*. 6233–6251.
- Siyan Zhu, Chengdong Xu, Kaiqiang Ke, and Chao Yu. 2025. Context-Picker: Dynamic context selection using multi-stage reinforcement learning. *arXiv preprint arXiv:2512.14465* (2025).

A Appendix

B Additional Experiments

C Qualitative Failure Case Studies

We provide the full input–output records for the representative failure cases summarized in the main text. The entry actually cited by the model is marked **Used**. Each case is drawn from the SFT→GRPO selective-retrieval model at the indicated KL coefficient β .

Table 4: Distribution of the number of memory entries selected by GRPO variants on the MedQA-USMLE held-out set. Each row reports the count and percentage of examples using exactly k memory entries out of five retrieved entries.

Condition	N	0 used	1 used	2 used	3 used	4 used	5 used
$\beta = 0.02$	1,273	17 (1.3)	1,212 (95.2)	40 (3.1)	0 (0.0)	1 (0.1)	3 (0.2)
$\beta = 0.05$	1,273	14 (1.1)	1,230 (96.6)	27 (2.1)	0 (0.0)	0 (0.0)	2 (0.2)
$\beta = 0.10$	1,273	8 (0.6)	1,247 (98.0)	17 (1.3)	0 (0.0)	0 (0.0)	1 (0.1)
Aggregate	3,819	39 (1.0)	3,689 (96.6)	84 (2.2)	0 (0.0)	1 (0.0)	6 (0.2)

Percentages are shown in parentheses.

Table 5: Retrieval-position usage by GRPO variants on the MedQA-USMLE held-out set. Each column reports the percentage of examples in which the model selected the retrieved memory entry at that ranked position.

Condition	N	Position 1	Position 2	Position 3	Position 4	Position 5
$\beta = 0.02$	1,273	75.2	12.0	12.3	3.1	0.4
$\beta = 0.05$	1,273	74.9	10.8	12.1	3.4	0.5
$\beta = 0.10$	1,273	75.1	10.0	12.0	3.5	0.5
Aggregate	3,819	75.0	10.9	12.1	3.3	0.4

Values are percentages. Percentages can sum to more than 100 across positions because an example may use more than one memory entry.

C.1 Decorative retrieval

Example ID: medqa_usmle_eval_00503 **Source run:** $\beta = 0.05$

Failure mode: The cited entry is acknowledged by the model to contain no answer-bearing fact, yet is still reported as the basis for the prediction.

Question. While in the ICU, a 62-year-old male undergoes placement of a Swan-Ganz catheter to evaluate his right heart pressures. All pressures are found to be within normal limits, and the cardiology fellow records a pulmonary wedge pressure of 10 mmHg. Which of the following are normal values for the pressures that will be obtained from this patient’s right ventricle?

Answer options.

- A. 25/10 mmHg [selected]
- B. 25/5 mmHg [gold]
- C. 10/0 mmHg
- D. 100/70 mmHg

Gold answer: B **Model selection:** A **Outcome:** incorrect

Retrieved memory candidates.

Rank	Cosine	Used	Memory entry (textbook case)
1	0.653	✓	A 57-year-old man is admitted to the burn unit after he was brought to the emergency room following an accidental fire in his house. His past medical history is unknown due to his current clinical condition. Currently, his blood pressure is 75/40 mmHg, pulse rate is 140/min, and respiratory rate is 17/min. The patient is subsequently intubated and started on aggressive fluid resuscitation. A Swan-Ganz catheter is inserted to clarify his volume status. Which of the following hemodynamic parameters would you expect to see in this patient?
2	0.581		A 37-year-old male presents to your clinic with shortness of breath and lower extremity edema. He was born in Southeast Asia and emigrated to America ten years prior. Examination demonstrates 2+ pitting edema to the level of his knees, ascites, and bibasilar crackles, as well as an opening snap followed by a mid-to-late diastolic murmur. The patient undergoes a right heart catheterization that demonstrates a pulmonary capillary wedge pressure (PCWP) of 24 mmHg. The patient is most likely to have which of the following?
3	0.565		A 69-year-old woman is admitted to the hospital with substernal, crushing chest pain. She is emergently moved to the cardiac catheterization lab where she undergoes cardiac angiography. Angiography reveals that the diameter of her left anterior descending artery (LAD) is 50% of normal. If her blood pressure, LAD length, and blood viscosity have not changed, which of the following represents the most likely change in LAD flow from baseline?
4	0.552		A 71-year-old man is admitted to the hospital one hour after he was found unconscious. His pulse is 80/min and systolic blood pressure is 98 mm Hg; diastolic blood pressure cannot be measured. He is intubated and mechanically ventilated with supplemental oxygen at a tidal volume of 450 mL and a respiratory rate of 10/min. Arterial blood gas analysis shows: PCO ₂ 43 mm Hg O ₂ saturation 94% O ₂ content 169 mL/L Pulmonary artery catheterization shows a pulmonary artery pressure of 15 mm Hg and a pulmonary capillary wedge pressure of 7 mm Hg. Bedside indirect calorimetry shows a rate of O ₂ tissue consumption of 325 mL/min. Given this information, which of the following additional values is sufficient to calculate the cardiac output in this patient?
5	0.545		A 75-year-old man comes to the physician because of a 4-month history of progressive shortness of breath and chest pressure with exertion. Cardiac examination shows a crescendo-decrescendo systolic murmur that is heard best in the second right intercostal space. Radial pulses are decreased and delayed bilaterally. Transesophageal echocardiography shows hypertrophy of the left ventricle and a thick, calcified aortic valve. The area of the left ventricular outflow tract is 30.6 mm ² . Using continuous-wave Doppler measurements, the left ventricular outflow tract velocity is 1.0 m/s, and the peak aortic valve velocity is 3.0 m/s. Which of the following values most closely represents the area of the stenotic aortic valve?

Model rationale (verbatim). *“The question asks for normal right ventricular pressures. Entry 1 provides a context of a patient with a Swan-Ganz catheter, but it does not provide the specific pressures for the right ventricle. However, based on the provided answer choices, the closest match to normal right ventricular pressures is option A.”*

C.2 Reasoning-answer mismatch

Example ID: medqa_usmle_eval_00765 **Source run:** $\beta = 0.1$

Failure mode: The free-text rationale explicitly concludes option C, but the structured final_answer field emits A.

Question. A 7-year-old boy is brought to your office with complaints of wheezing and dyspnea. Laboratory work reveals eosinophilia and positive skin tests for allergens. Which of the following types of drugs would be an effective treatment in this patient?

Answer options.

- A. Beta-2 antagonist [selected]
- B. Histone deacetylase inhibitor
- C. Muscarinic antagonist [gold]
- D. Anticoagulant

Gold answer: C Model selection: A Outcome: incorrect

Retrieved memory candidates.

Rank	Cosine	Used	Memory entry (textbook case)
1	0.733		A 17-year-old boy presents to the office with allergic rhinitis. He reports symptoms of sneezing, nasal congestion, itching, and postnasal drainage every September at the start of the school year. He has a family history of childhood asthma and eczema. He has not tried any medications for his allergies. Which of the following medications is the most appropriate next step to manage the patient's symptoms?
2	0.678		A 14-year-old girl with a history of severe persistent asthma presents to her pediatrician after a recent hospital discharge for asthma exacerbation. Her mother is concerned that her daughter continues to wheeze and cough multiple nights per week. She is also concerned that her daughter frequently uses the bathroom to urinate despite no recent change in her diet. She has allergies to pollen and shellfish, but her mother denies any recent exposure. The patient's medications include albuterol, salmeterol, and both inhaled and oral prednisone. What alternative drug can the pediatrician recommend for this patient?
3	0.660	✓	A 20-year-old man presents to your office with dyspnea, reporting nocturnal cough. You note expiratory wheezing on auscultation. Chest x-ray reveals increased anteroposterior diameter. Past medical history is significant for multiple episodes of "bronchitis" as a child. Which of the following drugs would be most effective for long-term treatment of this patient?
4	0.659		A 19-year-old boy presents to the emergency department with difficulty breathing, which began 1 hour ago. He has had persistent bronchial asthma since 3 years of age and has been prescribed inhaled fluticasone (400 μ g/day) by his pediatrician. He has not taken the preventer inhaler for the last 2 weeks and visited an old house today that had a lot of dust accumulated on the floor. On physical examination, his temperature is 36.8°C (98.4°F), the pulse is 110/min, and the respiratory rate is 24/min. There are no signs of respiratory distress, and chest auscultation reveals bilateral wheezing. Which of the following medications is most likely to provide quick relief?
5	0.644		An 8-year-old girl is brought to the physician by her mother because of a 6-month history of an episodic dry cough, shortness of breath, and chest tightness. She has seasonal allergic rhinitis. Physical examination shows high-pitched expiratory wheezes throughout both lung fields. Pulmonary function testing shows an FEV1 of 70% (N $\geq$ 80%). Which of the following drugs would be most effective at reducing bronchial inflammation in this patient?

Model rationale (verbatim). *"The patient described in entry 3 has a history of severe persistent asthma, which aligns with the patient in the question. The most effective long-term treatment for severe persistent asthma is a muscarinic antagonist, which is option C. However, since the question asks for the most effective long-term treatment, and considering the context of severe persistent asthma, the most appropriate answer is C."*

C.3 Case-similarity substitution

Example ID: medqa_usmle_eval_00309 **Source run:** $\beta = 0.02$

Failure mode: The model justifies its prediction purely by surface resemblance to a retrieved case, importing that case's management (and a procedure absent from the options) rather than identifying the distinguishing principle.

Question. A 56-year-old man is brought to the emergency room after a motor vehicle accident. The patient's vitals are as follows: blood pressure 80/40 mm Hg, heart rate 111/min, respiratory rate 39/min, and temperature 37.1°C (98.8°F). On physical examination, the patient is unconscious with a GCS of 9/15 and is cyanotic. There are open fractures of the left femur and left tibia, a likely shoulder dislocation, multiple contusions on the limbs and thorax, and a puncture wound on the left side of his chest. There are no breath sounds on the left side and there is hyperresonance to percussion on the left. Preparations are made for an emergency needle thoracostomy to be performed to treat this patient's likely tension pneumothorax. Which one of the following is the best choice to provide informed consent for this procedure?

Answer options.

A. The patient's closest relatives (spouse, child or parent), who must be brought to the hospital as fast as possible *[selected]*

B. One of the witnesses at the scene of the car accident, who should be brought to the hospital as fast as possible

C. The hospital ethics committee

D. Informed consent is not needed in this case *[gold]*

Gold answer: D Model selection: A Outcome: incorrect

Retrieved memory candidates.

Rank	Cosine	Used	Memory entry (textbook case)
1	0.725	✓	A 24-year-old male is rushed to the emergency department after sustaining several gunshot wounds to the chest. He was found nonresponsive in the field and was intubated en route to the hospital. His vital signs are as follows: temperature is 98.8 deg F (37.1 deg C), blood pressure is 87/52 mmHg, pulse is 120/min, and respirations are 16/min. Physical examination is significant for decreased breath sounds and dullness to percussion over the right lung. A chest radiograph in the emergency department shows a large fluid collection in the right thoracic cavity. After aggressive fluid resuscitation is initiated, an emergent chest-tube was placed in the emergency department. The chest tube puts out 700 cc of frank blood and 300 cc/hr over the next 5 hours. A follow up post-chest tube insertion chest radiograph demonstrates significant residual right hemothorax. Which of the following is the next best step in management of this patient?
2	0.721		A 24-year-old man is brought to the emergency department after being involved in a motor vehicle accident as an unrestrained driver. He was initially found unconscious at the scene but, after a few minutes, he regained consciousness. He says he is having difficulty breathing and has right-sided pleuritic chest pain. A primary trauma survey reveals multiple bruises and lacerations on the anterior chest wall. His temperature is 36.8°C (98.2°F), blood pressure is 100/60 mm Hg, pulse is 110/min, and respiratory rate is 28/min. Physical examination reveals a penetrating injury just below the right nipple. Cardiac examination is significant for jugular venous distention. There is also an absence of breath sounds on the right with hyperresonance to percussion. A bedside chest radiograph reveals evidence of a collapsed right lung with depression of the right hemidiaphragm and tracheal deviation to the left. Which of the following is the most appropriate next step in the management of this patient?
3	0.698		A 45-year-old man is brought to the emergency department after a car accident with pain in the middle of his chest and some shortness of breath. He has sustained injuries to his right arm and leg. He did not lose consciousness. His temperature is 37°C (98.6°F), pulse is 110/min, respirations are 18/min, and blood pressure is 90/60 mm Hg. He is alert and oriented to person, place, and time. Examination shows several injuries to the upper extremities and chest. There are jugular venous pulsations 10 cm above the sternal angle. Heart sounds are faint on cardiac examination. The lungs are clear to auscultation. An ECG is shown. Which of the following is the most appropriate next step in management?
4	0.690		A 25-year-old man is admitted to the emergency department because of an episode of acute psychosis with suicidal ideation. He has no history of serious illness and currently takes no medications. Despite appropriate safety precautions, he manages to leave the examination room unattended. Shortly afterward, he is found lying outside the emergency department. A visitor reports that she saw the patient climbing up the facade of the hospital building. He does not respond to questions but points to his head when asked about pain. His pulse is 131/min, respirations are 22/min, and blood pressure is 95/61 mm Hg. Physical examination shows a 1-cm head laceration and an open fracture of the right tibia. He opens his eyes spontaneously. Pupils are equal, round, and reactive to light. Breath sounds are decreased over the right lung field, and the upper right hemithorax is hyperresonant to percussion. Which of the following is the most appropriate next step in management?
5	0.686		A 21-year-old man presents to the emergency department after sustaining a stab wound to the neck at a local farmer's market. The patient is otherwise healthy and is complaining of pain. The patient is able to offer the history himself. His temperature is 97.6°F (36.4°C), blood pressure is 120/84 mmHg, pulse is 90/min, respirations are 15/min, and oxygen saturation is 98% on room air. Physical exam demonstrates a 3 cm laceration 1 cm inferior to the mastoid process on the right side. The patient's breath sounds are clear and he is protecting his airway. No stridor or difficulty breathing is noted. Which of the following is the most appropriate next step in the management of this patient?

Model rationale (verbatim). *“The scenario described in the question matches the scenario in entry 1. Therefore, the most appropriate next step in management is to obtain informed consent for the emergency needle thoracostomy.”*

C.4 Default positional bias

Example ID: medqa_usmle_eval_00734 **Source run:** $\beta = 0.05$

Failure mode: With retrieval scores effectively tied across all candidates, the model deterministically selects the first-ranked entry and answers incorrectly.

Question. A 10-year-old girl is brought to a clinic by her parents after 1 week of difficulty breathing, dry cough, and low-grade fevers. She has also developed shallow erythematous ulcers inside of her mouth. There is no history of chills, rigors, or headaches. Her appetite has decreased. The girl has been HIV-positive since birth and is being treated with HAART. Her baseline viral load is $<50/\text{mL}$. On general examination, the girl appears ill. Diffuse inspiratory crackles are heard on chest auscultation with normal heart sounds. No lymphadenopathy or hepatosplenomegaly is noted on physical examination. A blood sample is drawn, which shows a CD4 count of $100/\text{mm}^3$ and a viral load of $25,050/\text{mL}$. What is the most likely mechanism responsible for the lack of responsiveness to therapy?

Answer options.

A. Ineffective CD8 cytotoxicity

B. env gene mutation [selected]

C. pol gene mutation [gold]

D. Decreased viral replication

Gold answer: C **Model selection:** B **Outcome:** incorrect

Retrieved memory candidates.

Rank	Cosine	Used	Memory entry (textbook case)
1	0.687	✓	A 39-year-old man comes to the emergency department because of a 2-day history of fever, chills, dyspnea, and a non-bloody productive cough. He was diagnosed with HIV infection 4 years ago and has been on highly active antiretroviral therapy since then. His temperature is 38.8°C (101.8°F). Examination shows crackles over the left lower lung base. His CD4+ T-lymphocyte count is 520/mm ³ ($N \geq 500$). An x-ray of the chest shows an infiltrate in the left lower lobe. Sputum cultures grow colonies with a narrow zone of green hemolysis without clearing on blood agar. The most likely causal pathogen of this patient's condition produces which of the following virulence factors?
2	0.685		A 43-year-old man with HIV comes to the physician because of fever and night sweats over the past 15 days. During this period, he has also had headaches and generalized weakness. He has no cough or shortness of breath. He has hypertension controlled with lisinopril and is currently receiving triple antiretroviral therapy. He has smoked one pack of cigarettes daily for the past 15 years and drinks one to two beers on weekends. He is a known user of intravenous illicit drugs. His temperature is 39°C (102°F), pulse is 115/min, respirations are 15/min, and blood pressure is 130/80 mm Hg. Examination shows several track marks on the forearms. The lungs are clear to auscultation. A holosystolic murmur that increases on inspiration is heard along the left sternal border. The remainder of the physical examination shows no abnormalities. Laboratory studies show a leukocyte count of 12,800/mm ³ and an erythrocyte sedimentation rate of 52 mm/h. His CD4+ T-lymphocyte count is 450/mm ³ (normal ≥ 500). Which of the following is the most likely sequela of the condition?
3	0.684		A 28-year-old woman presents to her physician for follow-up. She was found to be HIV-positive 9 months ago. Currently she is on ART with lamivudine, tenofovir, and efavirenz. She has no complaints and only reports a history of mild respiratory infection since the last follow-up. She is also allergic to egg whites. Her vital signs are as follows: the blood pressure is 120/75 mm Hg, the heart rate is 73/min, the respiratory rate is 13/min, and the temperature is 36.7°C (98.0°F). She weighs 68 kg (150 lb), and there is no change in her weight since the last visit. On physical examination, she appears to be pale, her lymph nodes are not enlarged, her heart sounds are normal, and her lungs are clear to auscultation. Her total blood count shows the following findings: Erythrocytes 3.2 x 10 ⁶ /mm ³ Hematocrit 36% Hgb 10 g/dL Total leukocyte count 3,900/mm ³ Neutrophils 66% Lymphocytes 24% Eosinophils 3% Basophils 1% Monocytes 7% Basophils 0 Platelet count 280,000/mm ³ Her CD4+ cell count is 430 cells/ μ L. The patient tells you she would like to get an influenza vaccination as flu season is coming. Which of the following statements is true regarding influenza vaccination in this patient?
4	0.683		A 16-year-old girl is brought to the physician because of a 1-month history of fever, headaches, and profound fatigue. Her temperature is 38.2°C (100.8°F). Examination shows splenomegaly. Laboratory studies show: Leukocyte count 13,000/mm ³ (15% atypical lymphocytes) Serum Alanine aminotransferase (ALT) 60 U/L Aspartate aminotransferase (AST) 40 U/L Heterophile antibody assay negative EBV viral capsid antigen (VCA) antibodies negative HIV antibody negative In an immunocompromised host, the causal organism of this patient's symptoms would most likely cause which of the following conditions?
5	0.682		A 7-year-old girl is brought to the physician by her mother because of a 2-week history of generalized fatigue, intermittent fever, and progressively worsening shortness of breath. Physical examination shows pallor, jugular venous distention, and nontender cervical and axillary lymphadenopathy. Inspiratory stridor is heard on auscultation of the chest. The liver is palpated 3 cm below the right costal margin. Her hemoglobin concentration is 9.5 g/dL, leukocyte count is 66,000/mm ³ , and platelet count is 102,000/mm ³ . An x-ray of the chest shows a mediastinal mass. A bone marrow aspirate predominantly shows leukocytes and presence of 35% lymphoblasts. Which of the following additional findings is most likely in this patient?

Model rationale (verbatim). *“The patient’s symptoms and laboratory findings suggest an opportunistic infection in an HIV-positive patient. The most likely causal pathogen of the patient’s condition is a virus that can cause pneumonia in immunocompromised patients, such as *Pneumocystis jirovecii*.”*

C.5 Case-similarity substitution (autism vs. ADHD)

Example ID: medqa_usmle_eval_00637 **Source run:** $\beta = 0.05$

Failure mode: The question describes an autism-spectrum presentation, but the model anchors on a surface-similar inattentive-child case and imports its ADHD-keyed attribute instead of identifying the distinguishing principle.

Question. A 12-year-old girl is brought to your psychiatry office by her parents, who are concerned that she has not made any friends since starting middle school last year. The girl tells you that she gets nervous around other children, as she cannot tell what they are feeling, when they are joking, or what she has in common with them. Her teachers describe her as “easily distracted” and “easily upset by change.” When asked about her hobbies and interests, she states that “marine biology” is her only interest and gives you a 15-minute unsolicited explanation of how to identify different penguin species. Mental status exam is notable for intense eye contact, flat affect, and concrete thought process. Which of the following is true regarding this child’s disorder?

Answer options.

A. Boys are more commonly affected than girls [gold]

B. Impaired attention is a key feature of the disorder [selected]

C. Intellectual disability is a key feature of the disorder

D. Typical age of onset is 3-5 years of age

Gold answer: A **Model selection:** B **Outcome:** incorrect

Retrieved memory candidates.

Rank	Cosine	Used	Memory entry (textbook case)
1	0.626	✓	A 5-year-old girl is brought to the physician because her mother has found her to be inattentive at home and has received multiple complaints from her teachers at school. She does not complete her assignments and does not listen to her teachers' instructions. She refuses to talk to her parents or peers. Her mother says, "She ignores everything I say to her!" She prefers playing alone, and her mother reports that she likes playing with 5 red toy cars, repeatedly arranging them in a straight line. She avoids eye contact with her mother and the physician throughout the visit. Physical and neurological examination shows no abnormalities. Which of the following is the most likely diagnosis?
2	0.621		A 20-year-old woman is brought to the physician by her mother because she has been worried about her daughter's strange behavior for the past 2 years. She does not have any friends and spends most of her time alone in her room. She usually wears a long, black, hooded cloak. She is anxious at college because she is uncomfortable around other people, and her academic performance is poor. She collects rare crystals and says that they support her "sixth sense." Mental status examination shows slow, hesitant speech, and she avoids eye contact. Which of the following is the most likely diagnosis?
3	0.617		An 8-year-old girl is brought to the physician by her parents because they are concerned with her behavior. She has temper outbursts six or seven times per week, which last anywhere between 5 minutes to half an hour or until she becomes tired. According to her father, she screams at others and throws things in anger "when things don't go her way." He says these outbursts started when she was 6 and a half years old and even between the outbursts, she is constantly irritable. She had been suspended from school three times in the past year for physical aggression, but her grades have remained unaffected. She appears agitated and restless. Physical examination shows no abnormalities. During the mental status examination, she is uncooperative and refuses to answer questions. What is the most likely diagnosis in this child?
4	0.608		A 17-year-old boy is brought to the physician by his mother because of increasingly withdrawn behavior for the last two years. His mother reports that in the last 2–3 years of high school, her son has spent most of his time in his room playing video games. He does not have any friends and has never had a girlfriend. He usually refuses to attend family dinner and avoids contact with his siblings. The patient states that he prefers being on his own. When asked how much playing video games means to him, he replies that "it's okay." When his mother starts crying during the visit, he appears indifferent. Physical and neurologic examinations show no other abnormalities. On mental status examination, his thought process is organized and logical. His affect is flattened. Which of the following is the most likely diagnosis?
5	0.602		A 10-year-old boy is brought to the physician by his parents because they are concerned about his "strange behavior". The parents state that he has always been a lonely kid without many friends, but recently he has been having behavioral problems that seem to be unprovoked and are occurring more frequently. The child throws tantrums for no reason and does not respond to punishment or reward. He also has a "strange obsession" with collecting rocks that he finds on his way to and from school to the point where his room is filled with rocks. He plays alone in his room, lining the rocks up, organizing them by size, shape, or color, and he will randomly bark or make high-pitched noises without provocation. His teachers say he daydreams a lot and is very good at art, being able to recreate his favorite cartoon characters in great detail. On physical assessment, the patient does not make eye contact with the physician but talks incessantly about his rock collection. The child's grammar and vocabulary seem normal but his speech is slightly labored, and he can't seem to tell that the physician is not really interested in hearing about his rock collection. Which of the following is the most likely diagnosis?

Model rationale (verbatim). *"The case described in entry 1 matches the symptoms of ADHD, including impaired attention, which is a key feature of the disorder."*