MedSR: Aligning Language Models to Reason and Search for Medical Question Answering with Group Relative Policy Optimization

Zaifu Zhan[1,2], Jun Wang[1], Yiran Song[1], Min Zeng[1], Kai Yu[1], Mengyuan Cui[1], Mingquan Lin[1], Feng Xie[1], and Rui Zhang[1,*]

[1]*Division of Computational Health Sciences, Department of Surgery, University of Minnesota, Minneapolis, MN, USA*

[2]*Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA*

[*]*Corresponding author. E-mail: ruizhang@umn.edu*

Large language models can answer medical questions from parametric memory, but this behavior gives limited control over whether the model seeks evidence or uses retrieved information appropriately. We present MedSR, a medical search-and-reasoning framework that fine-tunes language models to produce auditable reason–search–answer trajectories with constrained final answers. MedSR uses group relative policy optimization with LoRA adaptation, an external search tool, and rule-based answer rewards, without a separately trained reward model. We evaluate three compact models on MedQA-USMLE, HeadQA, and PubMedQA, comparing direct prompting, chain-of-thought prompting, self-consistency, few-shot prompting, untrained search prompting, and GRPO-trained search behavior. Four retrieval corpora are studied: PubMed, medical textbooks, StatPearls, and Wikipedia. The results were task dependent: trained search gave the clearest gains for Qwen2.5-3B and Llama-3.2-3B on MedQA and HeadQA, while the PubMed corpus was most closely aligned with PubMedQA. In this benchmark setting, updating the model to generate and use search queries was more effective than simply appending retrieved passages to the prompt, and performance depended on the retrieval corpus, reward design, retrieval depth, and adapter capacity.

Keywords: Medical Question Answering; Retrieval-Augmented Generation; Reinforcement Learning; Group Relative Policy Optimization; Large Language Models.

1. Introduction

Since GPT-3[1] and its successors demonstrated that scaling Transformer-based models[2,3] could produce surprisingly capable text generators, the NLP community has pushed these models into increasingly specialized domains. Medicine stands out among them, not because question answering is new, but because the cost of a wrong answer is unusually high. General-purpose and domain-adapted models now perform well on clinical knowledge benchmarks and medical examination questions.[4–9] Recent medical foundation-model studies have also emphasized efficient adaptation, quantization, dataset composition, retrieval-based conditioning, and reliability in medical reasoning.[10–14] Yet answers still often come from the model's parametric

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

memory alone, a limitation in a field where evidence shifts with new studies, guidelines, and drug approvals.

Retrieval-augmented models such as REALM,[15] DPR,[16] and RAG[17] were developed to reduce this dependence on memorized knowledge by pairing generation with external documents. Later systems showed that retrieval can support both open-domain question answering and few-shot language modeling when the retrieved passages are well matched to the input.[18–20] In medical QA, the documents may come from PubMed abstracts, textbooks, clinical summaries, or broad reference sources. Corpus choice and retrieval depth matter more than early RAG applications often assumed, and recent biomedical in-context learning work has also shown that retrieval strategy affects downstream performance.[13,21,22] The retrieval behavior, however, is usually fixed before the model sees a question. A developer chooses the query template, corpus, top-k, and truncation length. That recipe may be adequate for some questions and wasteful for others. It can also retrieve text that matches the surface wording of a long question while missing the fact needed to answer it.

We turned to reinforcement learning for a practical reason: medical QA already has a usable reward signal. Benchmarks such as MedQA provide a ground-truth answer, so a full trajectory–reasoning, search, evidence reading, and final answer–can be scored without training a separate reward model. Policy-gradient methods such as PPO established a practical route for optimizing generated behavior beyond next-token likelihood,[23] and GRPO removes the need for a learned critic in settings with group-level rewards.[24] Recent work has also shown that models can interleave reasoning with browsing, tool use, and search.[25–29] At the same time, medical LLM reviews and self-correction studies suggest that explicit reasoning traces do not by themselves guarantee correct medical answers.[14,30,31] Current medical RAG pipelines rarely make that connection. Retrieval is still usually a preprocessing step, and the learning objective seldom treats the query, evidence, and final medical answer as one trajectory.

MedSR studies this formulation in medical QA. It fine-tunes compact instruction-tuned language models to learn what to search and how to use retrieved evidence before committing to a constrained answer. MedSR uses GRPO with LoRA adaptation, a medical search tool, and rule-based answer rewards to optimize the reason–search–answer trajectory. We evaluate MedSR on MedQA-USMLE,[32] HeadQA,[33] and PubMedQA,[34] across three base models and four retrieval corpora. These benchmarks support controlled comparison, but they are narrower than real clinical language tasks; recent holistic evaluations such as MedHELM emphasize that medical LLM assessment should span more diverse task types and output formats.[35] We compare trained retrieval with direct prompting, chain-of-thought, self-consistency, few-shot prompting, and untrained search prompting.

2. Materials and Methods

2.1. *Framework overview*

Fig. 1 shows how a question passes through MedSR. The policy is the language model whose search-and-answer behavior we optimize. Its pretrained weights remain frozen while LoRA adapters are updated. Each experiment connects the model to one of the four retrieval corpora shown in the search-tool block. During a rollout, the model generates a query, receives top-k

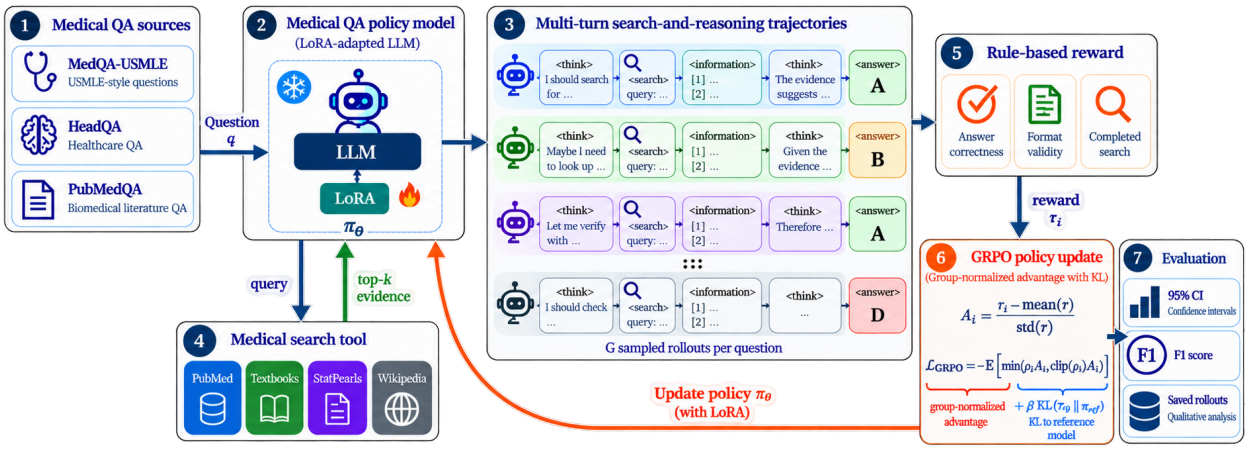


Fig. 1. **MedSR architecture for learning medical search and reasoning with GRPO.** Questions from three datasets (1) enter a LoRA-adapted model (2), which samples multiple trajectories per question (3) using top-k evidence from one corpus per experiment (4). Answer correctness and trajectory-format checks determine rewards (5); completed search is part of the format check. GRPO updates the adapters under KL regularization while base weights remain frozen (6). Evaluation reports answer metrics and retains rollout traces (7).

snippets from that corpus, and continues reasoning before producing an answer in the required format.

Inspired by Search-R1,[26] MedSR treats search as an action inside the RL trajectory and redesigns the setting for medical QA. We replace general web search with medical evidence corpora, constrain the final answer to the benchmark label space, and score the completed trajectory with medical QA rewards. MedQA-USMLE and HeadQA require a single option letter, while PubMedQA requires *yes*, *no*, or *maybe*. We also store the generated query, retrieved snippets, reasoning text, and final answer for error analysis. The reward is assigned only after the answer is generated, so the model must learn which query and evidence-use pattern lead to a correct medical answer.

During training, the same model samples multiple trajectories for each question. Completed trajectories receive rewards based on answer correctness and format validity. GRPO compares rewards within the group and updates the LoRA adapters using the resulting advantages, with KL regularization against a frozen reference model. Supervision comes from benchmark answers rather than manually annotated search traces. The evaluation block summarizes answer-level metrics and saved trajectories for qualitative analysis.

Our contribution is the medical QA implementation and evaluation of this search-as-action approach: dataset-specific answer constraints, parameter-efficient adaptation, and an analysis of evidence sources and failure modes across three models, three datasets, and four corpora.

2.2. *Datasets*

We use MedQA-USMLE,[32] HeadQA,[33] and PubMedQA[34] labeled splits. HeadQA covers multiple healthcare disciplines, including medicine, nursing, and pharmacology. We split PubMedQA labeled into 500 training examples and 500 test examples. We treat MedQA and HeadQA as multiple-choice tasks requiring a single option letter, whereas PubMedQA requires *yes*, *no*, or *maybe*. We convert each example into a multiturn tool-use record containing the prompt, ground-truth answer list, original question, default search query, and search-tool metadata; for PubMedQA, we append abstract context to the question text before prompting.

2.3. *Prompt and tool-use trajectory*

We define the learning unit as a structured tool-use trajectory. Following the search-as-action design used in Search-R1,[26] the policy uses `<think>` for reasoning, `<search>` for the query, `<information>` for retrieved evidence, and `<answer>` for the final response. MedQA and HeadQA require a single option letter; PubMedQA requires *yes*, *no*, or *maybe*. With this format, search and evidence remain auditable, and the same reward function can score answer correctness and minimal trajectory validity across datasets.

2.4. *Retrieval corpora*

We evaluated four MedRAG-style corpora: PubMed, medical textbooks, StatPearls, and Wikipedia.[21] They contained 23,898,701, 125,847, 365,650, and 1,179,078 snippets, respectively, spanning article-level evidence, educational exposition, clinical summaries, and broad background knowledge. Dense retrieval used E5-base-v2 embeddings[36] and FAISS Flat indexes.[37] Queries and passages used E5 prefixes, embeddings were L2-normalized, and the retrieval service returned ranked snippets with title/identifier, score when available, and truncated text. Main GRPO experiments used top-$k = 2$ with a 900-character cap per document to limit rollout context. Search-prompt baselines retained the baseline script's top-$k = 3$ setting with the same per-document cap. These comparisons therefore reflect both fine-tuning and retrieval-depth settings; the top-k ablation measures sensitivity within the trained system.

2.5. *GRPO fine-tuning*

We use GRPO as a critic-free reinforcement learning objective.[24] For each prompt x, the rollout engine samples a group of $G = 4$ complete tool-use trajectories $\{y_i\}_{i=1}^{G}$, where each trajectory includes generated reasoning tokens, search actions, retrieved observations, and the final answer. We compute the scalar reward r_i only after the full trajectory is completed. Instead of training a value function, GRPO estimates the advantage by normalizing rewards within the group:

$$A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^{G})}{\text{std}(\{r_j\}_{j=1}^{G}) + \epsilon}.$$

The normalization compares trajectories sampled for the same question rather than absolute reward scales across datasets.

We update the actor with a clipped policy-gradient objective over generated tokens, with a KL regularizer to keep the LoRA-adapted policy close to the frozen reference model:

$$\mathcal{L}_{\text{GRPO}} = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\left(\rho_{i,t}A_i, \text{clip}(\rho_{i,t}, 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}})A_i\right) + \beta D_{\text{KL}}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right),$$

where $\rho_{i,t} = \pi_\theta(y_{i,t} \mid x, y_{i,<t})/\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})$, $\beta = 0.001$, and π_{ref} is the frozen base model. We apply the loss only to generated assistant tokens and treat retrieved text as an external observation. Main experiments used training splits for optimization and test splits for evaluation; ablations varied retrieval depth, reward design, corpus/retriever condition, LoRA rank, and training duration in the Llama-3.2-3B-Instruct PubMedQA setting.

Implementation used the verl GRPO trainer. We adapted the actor with LoRA on all linear modules, using rank 64 and alpha 32 in the main experiments.[38] We used the frozen reference model for KL regularization. Training used bfloat16 precision, gradient checkpointing, remove-padding optimization, FSDP parameter and optimizer offload, and vLLM rollout generation.[39] The learning rate was 3×10^{-5}; train, mini-batch, and validation batch sizes were 4; each prompt produced 4 rollouts; and maximum prompt, response, and tool-response lengths were 2048, 512, and 1024 tokens, respectively.

2.6. *Reward function*

We extract the final answer from the last `<answer>...</answer>` span and evaluate only that answer, not the reasoning text. Responses receive zero answer reward if the answer is absent or outside the dataset-specific answer space. Valid answers are lowercased and stripped of punctuation and whitespace before exact matching.

The main experiments use a shaped reward that combines exact answer accuracy with a small trajectory-format term:

$$r = \begin{cases} 1.0, & \text{if the answer is correct and the format is valid,} \\ 0.9, & \text{if the answer is correct but the format is invalid,} \\ 0.1, & \text{if the answer is incorrect but the format is valid,} \\ 0, & \text{otherwise.} \end{cases}$$

The format criterion required a valid answer, at least one search span, at least one information span, and balanced trajectory tags. Thus, "Completed search" in Fig. 1 refers to the search/information tag check within the format term, not a separate main-experiment reward or a judgment of evidence relevance. We keep this term small so that reward primarily reflects answer correctness. Table 1 summarizes ablated reward variants.

3. Experimental Setup

3.1. *Baselines*

We run baselines with vLLM separately from the GRPO trainer.[39] All baselines are inference-only prompting methods; none are trained with GRPO. Direct answering asks the model to return the constrained answer without retrieved evidence. Chain-of-thought prompting

Table 1. **Reward formulations for GRPO fine-tuning of medical question-answering models.** Answer + format is the main-experiment reward; the other formulations vary the contributions of answer correctness, trajectory format, search use, or response length. All apply the same dataset-specific answer-validity check before scoring.

Variant	Reward definition
Answer only	$r = 1$ if correct; 0 otherwise
Answer + format	$r = 0.9$ if correct, plus 0.1 if format is valid
Format-heavy	$r = 0.8$ if correct, plus 0.2 if format is valid
Search-light	$r = 0.85$ if correct, plus 0.1 if format is valid, plus 0.05 if search is used
Length penalty	$r = 0.9$ if correct, plus 0.1 if format is valid, minus 0.02 if over length

allows intermediate reasoning before the final answer,[40] and self-consistency samples 5 chain-of-thought completions at temperature 0.7 and uses majority vote.[41] Few-shot prompting provides one, three, five, or ten labeled examples in the prompt. Untrained search prompting appends retrieved snippets to the prompt but does not update model weights. Direct and few-shot baselines use a 64-token cap; chain-of-thought and search baselines use 512 tokens. We rescore all outputs with the same answer extractor and normalization rules.

3.2. *Evaluation*

We use exact-match accuracy after answer extraction as the primary metric, with Wilson 95% confidence intervals.[42] Micro-F1 pools decisions across labels and equals accuracy in this single-label evaluation; macro-F1 averages F1 equally over observed reference and predicted labels, while weighted-F1 weights each label by its reference frequency. Empty extracted answers are assigned an invalid label for F1 calculation. The format score checks search-trajectory structure as defined above. Baselines are scored through an answer-only representation, so their zero format scores do not imply invalid answer labels or failure to follow the baseline prompt. Shaped reward combines accuracy and format; it is not a clinical-quality score. We compute metrics from saved per-example outputs. Code for the MedSR experiments is available at `https://github.com/Learner4everrr/MedSR.git`.

4. Results

4.1. *Main comparison across models, datasets, and corpora*

Table 2 presents the central comparison by benchmark, method, corpus, and model, reporting accuracy, Wilson 95% confidence intervals, F1 scores, and format score. The relative ranking of methods varied across datasets and models, so we focus on trends that were consistent across multiple settings. Direct answering, chain-of-thought, self-consistency, and few-shot prompting remained competitive for several model–dataset pairs, particularly when the answer format was simple or the base model already appeared to know the relevant fact. Longer reasoning alone did not consistently improve accuracy. Chain-of-thought and self-consistency sometimes helped, but they also produced verbose generations that still ended with the wrong option or PubMedQA label.

Untrained search prompting did not consistently exceed the best prompt-only baseline, even for PubMedQA. Review of the saved generations suggested that retrieved text sometimes introduced a related but answer-irrelevant finding or increased format errors. MedSR fine-tuning performed better when the model learned to issue the query, read the returned snippet, and keep the final answer inside the required answer space.

The strongest gains appeared for Qwen2.5-3B and Llama-3.2-3B-Instruct on MedQA and HeadQA. DeepSeek-R1-Distill-Qwen-1.5B changed less, which is consistent with the qualitative rollouts: this smaller reasoning-distilled model was less stable in following the tool-use format. Corpus choice changed the ranking as well. PubMed aligned best with PubMedQA, while textbooks, StatPearls, and Wikipedia were more useful for some examination-style questions. Detailed model-specific values are reported in Table 2, and cross-condition trends are shown in Fig. 2.

4.2. *Qualitative success and failure cases*

We reviewed saved test-set generations from trained-search experiments. Table 3 shows cases stratified by retrieval quality and final-answer correctness. In successful cases, the generated query often compressed a long medical question into a disease, intervention, mechanism, or study finding that matched the retrieved evidence. The model then used the evidence to select a constrained answer rather than copying a sentence from the retrieved passage.

The failure cases were heterogeneous. Some rollouts retrieved relevant passages but selected an incorrect option, indicating that evidence access did not guarantee correct evidence interpretation. Other rollouts retrieved broad or off-topic snippets, leaving the final answer primarily dependent on the model's parametric memory. A third class involved trajectory or answer-format errors, such as placing explanatory text inside the answer tag or returning evidence-like phrases instead of a single option letter or PubMedQA label. Looking at the saved trajectory made these errors separable: a reviewer could see whether the failure began with the query, the retrieved text, the interpretation of relevant evidence, or the final answer format.

4.3. *Error analysis*

The saved rollouts made the error analysis more specific than aggregate accuracy. Several failures started at the query step. The model sometimes searched with an overly broad disease name, copied too much of the question, or omitted the key intervention, population, or outcome; the retrieved evidence then looked plausible but did not resolve the question. A different failure came from corpus mismatch. Literature-style snippets were not always optimal for examination questions that required canonical teaching facts, while educational summaries could miss PubMedQA-specific study details.

Other errors happened after apparently useful retrieval. The model could find topically relevant text but map it to the wrong option or yes/no/maybe label, especially when the evidence contained negation, comparative wording, or a subtle distinction between association and causation. In a smaller set of cases, the final answer appeared to come from parametric memory even when retrieved evidence suggested another label. We also saw plain format

Table 2. **Performance of MedSR on MedQA-USMLE, HeadQA, and PubMedQA.** Search uses retrieved context without weight updates; Ours denotes MedSR. Acc: accuracy; CI: Wilson 95% confidence interval; Micro, Macro, Wt: micro-, macro-, and weighted-F1; Fmt: search-trajectory format score. Metrics are percentages. Bold identifies the highest values within each dataset–model block. Text, StatP, and Wiki denote textbooks, StatPearls, and Wikipedia.

		Qwen2.5-3B						Llama-3.2-3B-Instruct						DeepSeek-R1-Distill-Qwen-1.5B					
Experiment	Corpus	Acc	95% CI	Micro	Macro	Wt	Fmt	Acc	95% CI	Micro	Macro	Wt	Fmt	Acc	95% CI	Micro	Macro	Wt	Fmt
MedQA																			
Direct	-	32.3	29.8–34.9	32.3	30.6	38.6	0.0	51.2	48.5–54.0	51.2	40.4	50.9	0.0	23.6	21.4–26.1	23.6	16.9	25.7	0.0
CoT	-	31.0	28.5–33.6	31.0	24.9	37.5	0.0	48.9	46.1–51.6	48.9	34.2	51.5	0.0	14.2	12.4–16.2	14.2	12.8	19.2	0.0
SC-CoT	-	44.2	41.5–47.0	44.2	29.4	44.6	0.0	47.7	44.9–50.4	47.7	40.1	50.5	0.0	24.5	22.2–26.9	24.5	17.0	25.6	0.0
1-shot	-	41.6	39.0–44.4	41.6	33.9	43.0	0.0	50.5	47.8–53.3	50.5	40.6	51.2	0.0	16.0	14.1–18.1	16.0	10.5	21.2	0.0
3-shot	-	43.2	40.5–45.9	43.2	35.2	44.7	0.0	51.0	48.2–53.7	51.0	40.5	51.0	0.0	13.4	11.7–15.4	13.4	14.1	17.6	0.0
5-shot	-	39.2	36.6–41.9	39.2	33.2	42.0	0.0	48.1	45.3–50.8	48.1	38.5	48.5	0.0	12.1	10.4–14.0	12.1	9.1	16.0	0.0
10-shot	-	31.2	28.7–33.8	31.2	24.8	37.7	0.0	41.5	38.8–44.2	41.5	29.7	37.6	0.0	14.5	12.6–16.5	14.5	10.7	18.8	0.0
Search	PubMed	29.7	27.2–32.3	29.7	28.0	35.5	0.0	47.0	44.2–49.7	47.0	31.1	46.9	0.0	11.3	9.7–13.2	11.3	8.8	15.6	0.0
Search	Text	30.3	27.9–32.9	30.3	28.6	36.2	0.0	48.6	45.9–51.4	48.6	32.3	48.6	0.0	12.6	10.9–14.6	12.6	11.7	17.6	0.0
Search	StatP	32.5	30.0–35.1	32.5	29.8	37.8	0.0	48.5	45.8–51.3	48.5	32.2	48.3	0.0	13.0	11.2–14.9	13.0	10.4	18.2	0.0
Search	Wiki	32.8	30.3–35.5	32.8	30.2	37.9	0.0	45.0	42.3–47.8	45.0	30.6	45.9	0.0	12.5	10.8–14.4	12.5	13.6	16.9	0.0
Ours	PubMed	49.9	47.1–52.6	49.9	8.2	49.6	0.1	54.2	51.5–56.9	54.2	44.3	55.7	95.1	24.7	22.5–27.2	24.7	14.0	24.9	0.0
Ours	Text	49.5	46.7–52.2	49.5	38.6	49.0	98.7	**56.2**	**53.5–58.9**	**56.2**	**56.1**	**56.3**	**100**	26.5	24.1–29.0	26.5	7.4	26.6	0.0
Ours	StatP	48.4	45.7–51.1	48.4	36.7	46.8	99.5	54.7	51.9–57.4	54.7	54.6	54.8	**100**	**26.8**	**24.4–29.3**	**26.8**	9.4	12.9	**95.8**
Ours	Wiki	**51.3**	**48.6–54.0**	**51.3**	**40.8**	**51.2**	**99.6**	53.8	51.1–56.5	53.8	53.3	53.7	99.3	25.8	23.5–28.3	25.8	**21.0**	**26.8**	0.0
HeadQA																			
Direct	-	47.4	45.6–49.3	47.4	40.7	51.3	0.0	56.9	55.0–58.7	56.9	45.1	56.6	0.0	36.4	34.6–38.2	36.4	21.4	37.4	**0.0**
CoT	-	50.9	49.0–52.8	50.9	36.1	54.4	0.0	55.8	54.0–57.7	55.8	38.4	57.7	0.0	32.3	30.6–34.1	32.3	20.0	35.0	**0.0**
SC-CoT	-	57.0	55.2–58.9	57.0	37.8	57.0	0.0	56.3	54.5–58.2	56.3	45.9	57.4	0.0	37.4	35.6–39.2	37.4	**30.0**	37.5	**0.0**
1-shot	-	45.2	43.3–47.1	45.2	33.7	42.9	0.0	52.8	50.9–54.6	52.8	35.8	53.7	0.0	31.1	29.4–32.8	31.1	13.4	33.6	**0.0**
3-shot	-	49.5	47.7–51.4	49.5	40.7	51.5	0.0	55.0	53.1–56.8	55.0	54.6	54.6	0.0	31.4	29.7–33.1	31.4	22.8	34.2	**0.0**
5-shot	-	51.3	49.4–53.1	51.3	41.9	53.0	0.0	54.1	52.2–55.9	54.1	53.8	53.9	0.0	28.7	27.0–30.4	28.7	18.3	32.1	**0.0**
10-shot	-	51.5	49.7–53.4	51.5	43.1	54.4	0.0	52.6	50.7–54.4	52.6	42.1	52.7	0.0	28.3	26.6–30.0	28.3	16.0	32.1	**0.0**
Search	PubMed	33.0	31.3–34.8	33.0	33.1	41.5	0.0	56.8	55.0–58.7	56.8	38.0	57.0	0.0	33.7	31.9–35.5	33.7	25.4	38.1	**0.0**
Search	Text	35.4	33.6–37.2	35.4	34.6	43.3	0.0	55.7	53.9–57.6	55.7	37.2	55.8	0.0	32.6	30.8–34.3	32.6	24.6	37.0	**0.0**
Search	StatP	35.4	33.7–37.3	35.4	34.4	43.0	0.0	53.5	51.6–55.4	53.5	35.6	53.5	0.0	33.0	31.2–34.8	33.0	29.6	37.0	**0.0**
Search	Wiki	35.4	33.7–37.3	35.4	34.1	42.7	0.0	53.4	51.5–55.3	53.4	35.9	53.9	0.0	30.8	29.1–32.5	30.8	19.8	34.7	**0.0**
Ours	PubMed	63.3	61.5–65.1	63.3	50.9	63.7	**100**	62.4	60.5–64.2	62.4	62.3	62.4	99.8	37.2	35.4–39.0	37.2	25.2	37.8	**0.0**
Ours	Text	63.4	61.6–65.2	63.4	41.9	63.0	99.7	62.5	60.6–64.3	62.5	31.2	62.5	99.6	35.9	34.1–37.7	35.9	13.3	36.8	**0.0**
Ours	StatP	65.4	63.6–67.2	65.4	**64.9**	65.1	**100**	**62.9**	**61.1–64.7**	**62.9**	**62.8**	**62.9**	**99.9**	36.6	34.8–38.4	36.6	8.7	37.2	**0.0**
Ours	Wiki	**65.5**	**63.7–67.2**	**65.5**	52.0	**65.2**	100	61.8	59.9–63.6	61.8	41.5	62.3	98.1	**37.7**	**36.0–39.6**	**37.7**	17.1	**38.5**	**0.0**
PubMedQA																			
Direct	-	60.4	56.0–64.6	60.4	31.1	57.8	0.0	69.6	65.4–73.5	69.6	36.8	67.2	0.0	60.0	55.6–64.2	60.0	35.8	61.6	**0.0**
CoT	-	59.4	55.0–63.6	59.4	32.0	59.0	0.0	70.4	66.3–74.2	70.4	36.9	67.2	0.0	31.2	27.3–35.4	31.2	23.9	40.6	**0.0**
SC-CoT	-	54.4	50.0–58.7	54.4	**41.0**	53.8	0.0	69.4	65.2–73.3	69.4	48.3	66.3	0.0	65.8	61.5–69.8	65.8	38.1	64.9	**0.0**
1-shot	-	65.2	60.9–69.2	65.2	34.7	62.4	0.0	68.0	63.8–71.9	68.0	47.1	64.5	0.0	53.6	49.2–57.9	53.6	31.8	56.4	**0.0**
3-shot	-	57.8	53.4–62.1	57.8	30.8	55.6	0.0	52.0	47.6–56.3	52.0	25.8	50.3	0.0	49.6	45.2–54.0	49.6	30.3	51.4	**0.0**
5-shot	-	67.2	63.0–71.2	67.2	35.1	64.2	0.0	64.8	60.5–68.9	64.8	42.1	59.6	0.0	51.0	46.6–55.4	51.0	30.2	53.8	**0.0**
10-shot	-	67.0	62.8–71.0	67.0	34.1	63.2	0.0	56.4	52.0–60.7	56.4	36.6	52.3	0.0	48.0	43.7–52.4	48.0	29.4	51.8	**0.0**
Search	PubMed	62.6	58.3–66.7	62.6	34.2	60.9	0.0	71.6	67.5–75.4	71.6	49.7	68.0	0.0	41.0	36.8–45.4	41.0	24.3	46.3	**0.0**
Search	Text	41.4	37.2–45.8	41.4	23.2	38.3	0.0	64.0	59.7–68.1	64.0	45.4	61.7	0.0	26.0	22.3–30.0	26.0	19.4	35.9	**0.0**
Search	StatP	45.0	40.7–49.4	45.0	28.0	44.5	0.0	64.8	60.5–68.9	64.8	45.5	62.3	0.0	26.6	22.9–30.6	26.6	20.0	36.6	**0.0**
Search	Wiki	35.4	31.3–39.7	35.4	20.2	30.4	0.0	63.6	59.3–67.7	63.6	45.8	61.3	0.0	33.8	29.8–38.1	33.8	24.4	44.3	**0.0**
Ours	PubMed	**73.4**	**69.4–77.1**	**73.4**	38.3	**69.8**	**99.8**	**74.6**	**70.6–78.2**	**74.6**	**51.6**	**70.5**	**100**	**70.6**	**66.5–74.4**	**70.6**	**48.2**	66.4	**0.0**
Ours	Text	70.8	66.7–74.6	70.8	24.7	67.6	99.4	64.2	59.9–68.3	64.2	45.1	61.1	99.8	67.0	62.8–71.0	67.0	34.1	63.0	**0.0**
Ours	StatP	**73.4**	**69.4–77.1**	**73.4**	38.2	69.7	**99.8**	70.8	66.7–74.6	70.8	36.4	66.7	99.8	**70.6**	**66.5–74.4**	**70.6**	36.2	**66.7**	**0.0**
Ours	Wiki	68.0	63.8–71.9	68.0	28.5	64.9	99.6	73.0	68.9–76.7	73.0	38.2	69.3	99.8	68.6	64.4–72.5	68.6	46.5	64.4	**0.0**

failures: explanatory text, retrieved phrases, or multiple candidate labels were placed inside `<answer>...</answer>` where only one constrained answer was allowed.

The error analysis separates the behaviors improved by MedSR from the behaviors that remain weakly supervised. GRPO fine-tuning can reward trajectories that search and end with valid answers, but the reward only observes final-answer correctness and coarse format validity. It does not directly supervise whether the query is medically optimal, whether each retrieved passage is clinically relevant, or whether the reasoning faithfully follows the evidence. Future reward designs should consider retrieval quality, evidence attribution, contradiction handling,

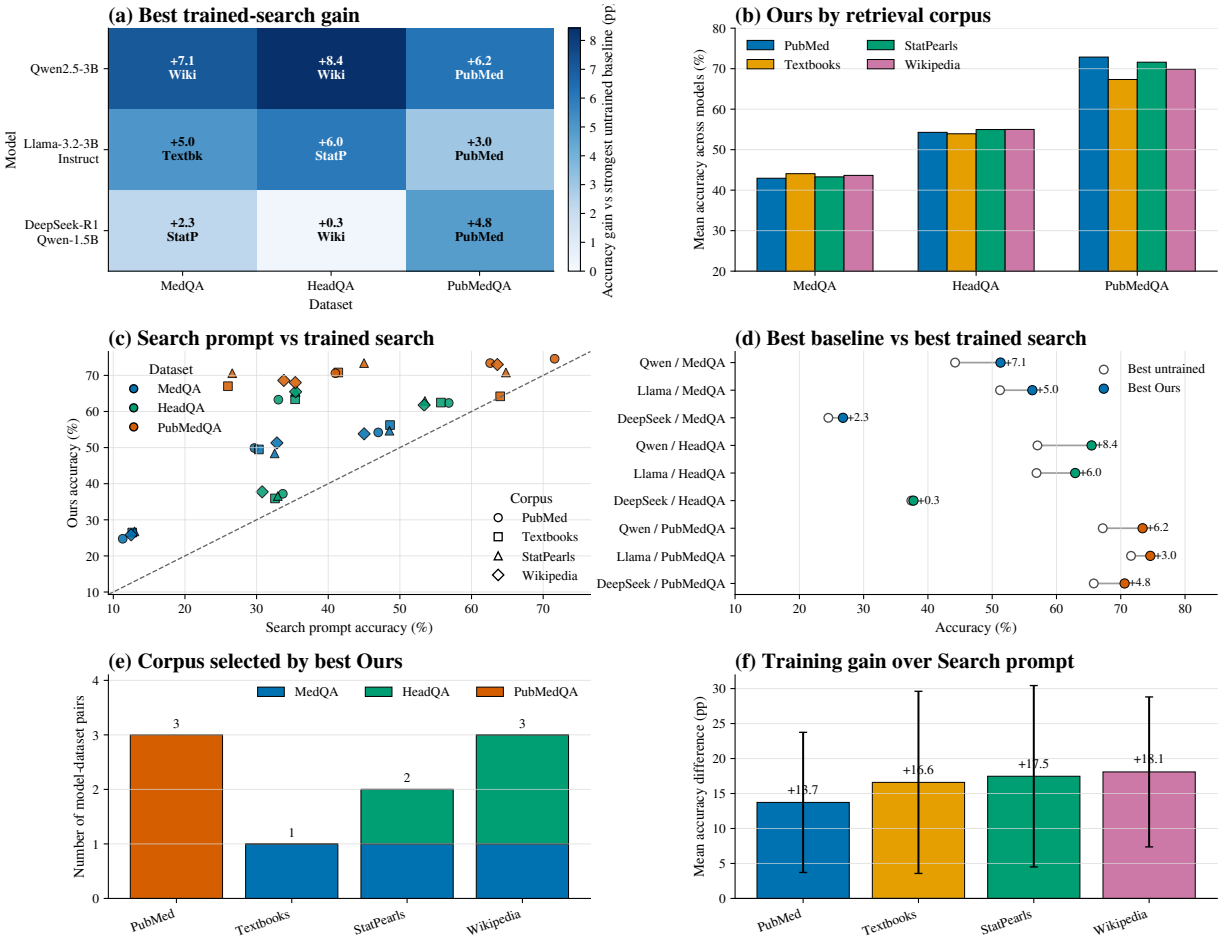


Fig. 2. **Comparison of MedSR and inference-only baselines across models, datasets, and retrieval corpora.** (a) Best MedSR minus best baseline accuracy per model–dataset pair. (b) MedSR accuracy averaged across models for each dataset and corpus. (c) Search versus MedSR accuracy, paired by model, dataset, and corpus. (d) Best baseline and MedSR accuracies. (e) Number of model–dataset pairs for which each corpus yields the highest MedSR accuracy; stacked colors denote datasets. (f) MedSR minus Search accuracy by corpus, averaged over nine model–dataset pairs; error bars show one standard deviation across pairs. Accuracy differences are in percentage points. Best values denote observed test-set maxima.

and clinician-rated utility.

4.4. *Ablation studies*

The ablations use the Llama-3.2-3B-Instruct PubMedQA setting and vary retrieval depth, reward formulation, corpus/retriever condition, LoRA rank, and fine-tuning duration. These experiments were designed to separate three contributors to performance: the quality of evidence exposed to the policy, the reward signal used to select higher-quality trajectories, and the adapter capacity available for changing search-and-answer behavior.

Retrieval depth was non-monotonic in Fig. 3(a). Top-$k = 2$ was strongest, while larger k values often reduced accuracy. This pattern is relevant for medical QA because additional

Table 3. **Retrieved evidence and answer outcomes in selected MedSR test-set trajectories.** Qwen2.5-3B examples illustrate correct answers with aligned evidence, answer or format errors despite aligned retrieval, and retrieval failures. Questions and retrieved passages are excerpted; reference and generated answers are shown for each case.

Dataset / corpus	Retrieval quality	Example question	Top retrieved document/snippet shown to the model	Target / model answer	Outcome
PubMedQA / PubMed	Aligned	Bone-anchor fixation for percutaneous bladder neck suspension in female stress urinary incontinence.	PubMed hit `0322_22353`: "Does bone anchor fixation improve the outcome of percutaneous bladder neck suspension in female stress urinary incontinence?"	yes / yes	Correct answer after aligned retrieval.
HeadQA / StatPearls	Aligned	"Form extracellular fibers with high tensile strength."	StatPearls article `19708_0`: "Biochemistry, Collagen Synthesis – Introduction. Collagen ... provides structural support ... resistance to stretching."	B / B	Correct option after aligned retrieval.
HeadQA / Textbooks	Aligned	"Fibrocartilage."	Histology Ross: "Fibrocartilage is a combination of dense connective tissue and cartilage ... large bundles of type I collagen."	C / C	Correct option after aligned retrieval.
PubMedQA / PubMed	Aligned	Optimal acid suppression time for healing severe reflux oesophagitis.	PubMed hit `0661_16389`: "A model of healing of Los Angeles grades C and D reflux oesophagitis: is there an optimal time of acid suppression..."	yes / span	Retrieval aligned, but final answer violated the yes/no/maybe format.
PubMedQA / PubMed	Aligned	Whether a practicing surgeon can detect early lymphedema reliably.	PubMed hit `0477_21998`: "Can a practicing surgeon detect early lymphedema reliably? ... circumference changes ... limb volume."	maybe / no	Retrieval aligned, but the model made a semantic yes/no/maybe error.
HeadQA / PubMed	Failed/empty	"The linear molecule of DNA associated with proteins is."	Query: "..."; PubMed hits `0975_2954` and `0976_10335`: contents ". .".	C / D	Retrieval failed; final answer was incorrect despite valid format.
MedQA-USMLE / Wikipedia	Off-topic/generic	Renal biopsy with intravascular spindle-shaped vacuoles after unstable angina treatment.	Wikipedia hit: "Urinary tract infection. Children"; second hit: "Kidney stone disease. Ureteroscopic surgery."	B / D	Retrieval was off topic and final answer was incorrect.
MedQA-USMLE / Wikipedia	Off-topic/generic	Cocaine-associated chest pain after acute coronary syndrome treatment.	Wikipedia hit: "Heart. Imaging"; second hit: "Heart. Heart failure."	D / A	Retrieval was off topic and final answer was incorrect.
HeadQA / Textbooks	Failed/incom	"The linear molecule of DNA associated with proteins is."	Biochemistry Lippincott: "Linear and circular DNA molecules"; Cell Biology Alberts: "Polycomb group ... proteins."	C / B	Retrieval was incomplete and final answer was incorrect.

snippets can add corroborating evidence, but they can also introduce partially related findings, population differences, or terminology that distracts the model from the required label. The corpus and retriever comparison in Fig. 3(c) showed a similar dependence on evidence alignment: PubMed was strongest for PubMedQA, whereas other corpora were competitive in exam-style settings. MedSR therefore depends on evidence source and retrieval depth, beyond the binary choice of using retrieval.

In the reward comparison in Fig. 3(b), answer correctness remained the most reliable training signal. Small format and search-shaping terms helped stabilize valid trajectories, but heavier format emphasis did not reliably improve accuracy. We use tags to make trajectories

auditable, but keep the reward focused on the final medical answer. In Fig. 3(d), rank 64 was the largest rank tested and performed best. This does not establish adapter-capacity saturation; higher ranks may further improve performance.

The checkpoint analysis in Fig. 3(e) showed that most improvement appeared within the first 500–1250 steps, followed by noisier later checkpoints. One possible explanation is early reinforcement of valid search-and-answer behavior, although accuracy alone cannot establish this mechanism. These curves are retrospective test-set sensitivity analyses. Repeated examination of test performance introduces a risk of selection bias; checkpoint and hyperparameter selection should use a separate validation split before final testing.

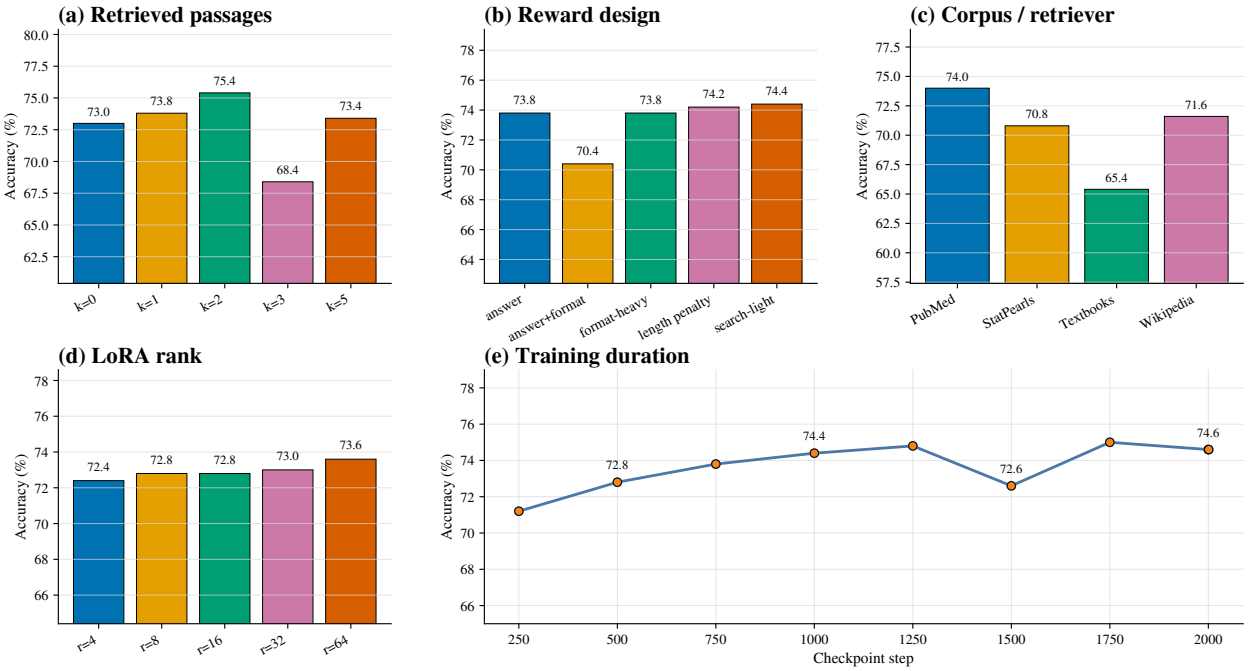


Fig. 3. **Sensitivity of MedSR to retrieval configuration, reward formulation, adapter rank, and fine-tuning duration.** All comparisons use Llama-3.2-3B-Instruct on PubMedQA. (a) Passages retrieved per query; $k = 0$ disables retrieval. (b) Reward formulations defined in Table 1. (c) Corpus choice with the E5-base-v2 retriever fixed. (d) LoRA adapter rank. (e) Test accuracy across fine-tuning checkpoints. All accuracies are percentages.

5. Discussion

Untrained search prompting was a relevant baseline for PubMedQA because the task is tied to abstracts, but adding retrieved text often left the model with the same answer-selection problem and, in some cases, extra distractors. Even snippets that are not copied may influence an answer through a plausible but irrelevant association or a population mismatch. MedSR uses GRPO-style relative rewards[24] to favor higher-scoring trajectories, although correct final answers alone do not establish faithful use of the evidence.

The corpus results also show that no single retrieval source was best across all benchmarks.

PubMed aligned best with PubMedQA, while textbooks, StatPearls, and Wikipedia were useful for different examination-style questions. Increasing top-k did not monotonically help either. Corpus and depth determine the evidence distribution that GRPO can learn from. A unified index spanning all four corpora would test access to mixed evidence sources; explicit source selection would additionally require a source-aware search action.

For medical informatics, the retrieval trace helps locate failures that final-answer accuracy cannot distinguish. The qualitative cases include relevant retrieval followed by an incorrect answer, sometimes involving negation, comparative wording, or association versus causation. These observations motivate evidence-attribution and contradiction checks, but do not establish the frequency of each failure type.

One practical use of MedSR is post-hoc audit. For an incorrect PubMedQA prediction, the saved trajectory helps reviewers examine the query, retrieved text, and evidence interpretation. A format checker followed by constrained regeneration could also reduce malformed answers. Such corrections should be evaluated separately from the original model outputs because they change the inference procedure.

The results also show that a model score should be interpreted as a system-level measurement rather than only as a property of the base LLM. For example, the same Llama-3.2-3B-Instruct model scored 45.0% with untrained Wikipedia search prompting and 56.2% with MedSR using textbooks on MedQA. A medical QA system built for literature questions may need PubMed-like evidence, while an examination-focused system may align better with educational or summary resources. More retrieval is not automatically better if additional passages add distractors or shift the model away from the answer format.

The study has limits that follow directly from this experimental design. We train and evaluate on structured QA benchmarks rather than live clinical decision support tasks, and final answers are constrained to benchmark-specific answer spaces. Because the rewards are based on benchmark labels, MedSR does not evaluate whether the model can communicate uncertainty, adapt to patient-specific context, or avoid harmful open-ended clinical advice. The experiments also use compact open models; larger models may follow the tool format more reliably, rely differently on parametric memory, or require different retrieval depths. Retrieved snippets should not be read as clinical guidelines or as sufficient evidence for care decisions. The retriever also remains fixed: MedSR trains the language model's search-and-answer behavior, not the dense encoder or corpus index. Deployment would require local validation of corpus choice and retrieval depth, a separate validation split for checkpoint selection, safety evaluation, and clinician review.

6. Conclusion

We introduced MedSR, a reinforcement learning framework for medical question answering that trains language models to reason with external medical search. Across three benchmarks, three compact models, and four retrieval corpora, MedSR improved many model–dataset combinations and exposed an auditable reason–search–answer trace. In these experiments, appending retrieved passages to a frozen LLM was less effective than training the model to decide what to search for and how to use the returned evidence. For medical QA, search should

be evaluated as part of the learned answering process rather than only as an inference-time prompt addition.

References

1. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, Language models are few-shot learners, in *Advances in Neural Information Processing Systems*, eds. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin (Curran Associates, Inc., 2020).
2. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Curran Associates, Inc., 2017).
3. J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, eds. J. Burstein, C. Doran and T. Solorio (Association for Computational Linguistics, Minneapolis, Minnesota, June 2019).
4. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, Large language models encode clinical knowledge, *Nature* **620**, 172 (2023).
5. K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis *et al.*, Toward expert-level medical question answering with large language models, *Nature medicine* **31**, 943 (2025).
6. H. Nori, N. King, S. M. McKinney, D. Carignan and E. Horvitz, Capabilities of gpt-4 on medical challenge problems, *arXiv preprint arXiv:2303.13375* **abs/2303.13375** (2023).
7. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* **36**, 1234 (02 2020).
8. R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon and T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, *Briefings in Bioinformatics* **23**, p. bbac409 (11 2022).
9. X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, M. G. Flores, Y. Zhang *et al.*, GatorTron: A large clinical language model to unlock patient information from unstructured electronic health records, *arXiv preprint arXiv:2203.03540* **abs/2203.03540** (2022).
10. Z. Zhan and R. Zhang, Towards better multi-task learning: A framework for optimizing dataset combinations in large language models, in *Findings of the Association for Computational Linguistics: NAACL 2025*, eds. L. Chiruzzo, A. Ritter and L. Wang (Association for Computational Linguistics, Albuquerque, New Mexico, April 2025).
11. Z. Zhan, S. Zhou, H. Zhou, Z. Liu and R. Zhang, EPEE: Towards efficient and effective foundation models in biomedicine, *npj Health Systems* **3**, p. 30 (2026).
12. Z. Zhan, S. Zhou, M. Zeng, K. Yu, M. Song, X. Chen, J. Wang, Y. Hou and R. Zhang, Quantized large language models in biomedical natural language processing: Evaluation and recommendation, *arXiv preprint arXiv:2509.04534* **abs/2509.04534** (2025).
13. Z. Zhan, S. Zhou, X. Zhou, Y. Xiao, J. Wang, J. Deng, H. Zhu, Y. Hou, Y. Song, M. Lin and R. Zhang, Retrieval-augmented in-context learning for multimodal large language models in disease classification, *Journal of Biomedical Informatics* **178**, p. 105017 (2026).
14. Z. Zhan, M. Cui and R. Zhang, Can large language models self-correct in medical question an-

swering? an exploratory study, *arXiv preprint arXiv:2604.00261* **abs/2604.00261** (2026).
15. K. Guu, K. Lee, Z. Tung, P. Pasupat and M. Chang, Retrieval augmented language model pre-training, in *Proceedings of the 37th International Conference on Machine Learning*, eds. H. D. III and A. Singh, Proceedings of Machine Learning Research, Vol. 119 (PMLR, 13–18 Jul 2020).
16. V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen and W.-t. Yih, Dense passage retrieval for open-domain question answering, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, eds. B. Webber, T. Cohn, Y. He and Y. Liu (Association for Computational Linguistics, Online, November 2020).
17. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel and D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in *Advances in Neural Information Processing Systems*, eds. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin (Curran Associates, Inc., 2020).
18. G. Izacard and E. Grave, Leveraging passage retrieval with generative models for open domain question answering, in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, eds. P. Merlo, J. Tiedemann and R. Tsarfaty (Association for Computational Linguistics, Online, April 2021).
19. S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. De Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. Rae, E. Elsen and L. Sifre, Improving language models by retrieving from trillions of tokens, in *Proceedings of the 39th International Conference on Machine Learning*, eds. K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu and S. Sabato, Proceedings of Machine Learning Research, Vol. 162 (PMLR, 17–23 Jul 2022).
20. G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel and E. Grave, Atlas: Few-shot learning with retrieval augmented language models, *Journal of Machine Learning Research* **24**, 1 (2023).
21. G. Xiong, Q. Jin, Z. Lu and A. Zhang, Benchmarking retrieval-augmented generation for medicine, in *Findings of the Association for Computational Linguistics: ACL 2024*, eds. L.-W. Ku, A. Martins and V. Srikumar (Association for Computational Linguistics, Bangkok, Thailand, August 2024).
22. Z. Zhan, J. Wang, S. Zhou, J. Deng and R. Zhang, MMRAG: Multi-mode retrieval-augmented generation with large language models for biomedical in-context learning, *Journal of the American Medical Informatics Association* **32**, 1505 (10 2025).
23. J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov, Proximal policy optimization algorithms, *arXiv preprint arXiv:1707.06347* **abs/1707.06347** (2017).
24. Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li *et al.*, Deepseekmath: Pushing the limits of mathematical reasoning in open language models, *arXiv preprint arXiv:2402.03300* **abs/2402.03300** (2024).
25. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike and R. Lowe, Training language models to follow instructions with human feedback, in *Advances in Neural Information Processing Systems*, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (Curran Associates, Inc., 2022).
26. B. Jin, H. Zeng, Z. Yue, J. Yoon, S. O. Arik, D. Wang, H. Zamani and J. Han, Search-r1: Training LLMs to reason and leverage search engines with reinforcement learning, *Second Conference on Language Modeling* **2025** (2025).
27. R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders *et al.*, Webgpt: Browser-assisted question-answering with human feedback, *arXiv preprint arXiv:2112.09332* **abs/2112.09332** (2021).
28. S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan and Y. Cao, React: Synergizing

reasoning and acting in language models, *The Eleventh International Conference on Learning Representations* **2023** (2023).
29. T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda and T. Scialom, Toolformer: Language models can teach themselves to use tools, in *Advances in Neural Information Processing Systems*, eds. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine (Curran Associates, Inc., 2023).
30. S. Zhou, Z. Xu, M. Zhang, C. Xu, Y. Guo, Z. Zhan, Y. Fang, S. Ding, J. Wang, K. Xu *et al.*, Large language models for disease diagnosis: A scoping review, *npj Artificial Intelligence* **1**, p. 9 (2025).
31. Z. Zhan, Y. Hou, K. Yu, M. Zeng, A. Burgun, X. Chen and R. Zhang, An underexplored frontier: Large language models for rare disease patient education and communication–a scoping review, *arXiv preprint arXiv:2604.14179* **abs/2604.14179** (2026).
32. D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang and P. Szolovits, What disease does this patient have? a large-scale open domain question answering dataset from medical exams, *Applied Sciences* **11** (2021).
33. D. Vilares and C. Gómez-Rodríguez, HEAD-QA: A healthcare dataset for complex reasoning, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, eds. A. Korhonen, D. Traum and L. Màrquez (Association for Computational Linguistics, Florence, Italy, July 2019).
34. Q. Jin, B. Dhingra, Z. Liu, W. Cohen and X. Lu, PubMedQA: A dataset for biomedical research question answering, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, eds. K. Inui, J. Jiang, V. Ng and X. Wan (Association for Computational Linguistics, Hong Kong, China, November 2019).
35. S. Bedi, H. Cui, M. Fuentes *et al.*, Holistic evaluation of large language models for medical tasks with MedHELM, *Nature Medicine* **32**, 943 (2026).
36. L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder and F. Wei, Text embeddings by weakly-supervised contrastive pre-training, *arXiv preprint arXiv:2212.03533* **abs/2212.03533** (2022).
37. J. Johnson, M. Douze and H. Jégou, Billion-scale similarity search with gpus, *IEEE Transactions on Big Data* **7**, 535 (2021).
38. E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, LoRA: Low-rank adaptation of large language models, *International Conference on Learning Representations* **2022** (2022).
39. W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang and I. Stoica, Efficient memory management for large language model serving with pagedattention, in *Proceedings of the 29th Symposium on Operating Systems Principles*, , SOSP '23 Vol. 29 (Association for Computing Machinery, New York, NY, USA, 2023).
40. J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le and D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in *Advances in Neural Information Processing Systems*, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (Curran Associates, Inc., 2022).
41. X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery and D. Zhou, Self-consistency improves chain of thought reasoning in language models, *The Eleventh International Conference on Learning Representations* **2023** (2023).
42. E. B. Wilson, Probable inference, the law of succession, and statistical inference, *Journal of the American Statistical Association* **22**, 209 (1927).