MEDROUTER: DEMYSTIFYING KNOWLEDGE DIFFERENCES ACROSS MEDICAL LLMS FOR ROUTING-BASED REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical question answering spans diverse specialties and modalities, and individual medical large language models (LLMs) exhibit distinct strengths across tasks and domains. This heterogeneity suggests that combining specialists may enable broader coverage of medical questions than relying on any single model. However, existing LLM routing methods primarily seek to balance answer quality and inference cost, leaving open how to exploit differences in specialist competence to improve medical reasoning. In this paper, we introduce *MedRouter*, an agentic system that uses an embedding-based multi-label router to select and query specialist LLMs, then passes their responses to a generator to produce the final answer. We further propose *SCALE* (*Specialist Competence-Aware Learning*), a two-stage training framework that first trains the Router with specialist correctness supervision and then optimizes its selections through reinforcement learning. The second stage uses a Performance Gain Reward (PGR) that measures how specialist information affects the generator's answer correctness relative to answering without that information. Experiments on eight text-based and multimodal medical QA benchmarks show that MedRouter outperforms the strongest routing baseline by 8% in average accuracy. Our analysis of specialist outputs further reveals distinct strengths and complementary question-level coverage, motivating learned routing to combine these capabilities for more comprehensive medical reasoning.

1 INTRODUCTION

Medical tasks span clinical specialties and involve information presented as text, images, or both. To support these tasks, a growing range of medical large language models (LLMs) have been developed (Labrak et al., 2024; Sellergren et al., 2025; Wang et al., 2023). These models differ in the sources, domains, coverage, and modalities of their training data, the data available to their developers, and the tasks they are designed to perform. Such differences may shape the knowledge that models acquire and retain (Chang et al., 2024). Therefore, a model may excel in one specialty while lacking knowledge that another model possesses (Xie et al., 2025). On questions that require this missing information, the other model may answer correctly even if its overall accuracy is lower. Choosing a single model based on average performance could then overlook useful capabilities available in other models. We therefore investigate whether existing medical LLMs have complementary question-answering capabilities and how their responses can be selected and combined to improve medical question answering.

To examine this possibility, we analyze specialist responses at the question level, comparing which questions six medical specialist LLMs answer correctly across four text-based QA benchmarks (MedQA, MedMCQA, MedCaseR, and MX-T) and four multimodal QA benchmarks (SLAKE, PathVQA, VQA-RAD, and MX-M). The multimodal evaluation includes only the three vision-capable specialists (see §4). Although MediX-R1 (Mullappilly et al., 2026) achieves the highest accuracy on all four text benchmarks (Figure 1(a)), each specialist correctly answers some questions missed by all other models in the applicable pool (Figure 1(b-g)). To quantify how much additional coverage these differences offer, we compare specialist selection at the dataset and question levels. Selecting the best specialist for each dataset yields 43.6% micro-averaged accuracy across the eight benchmarks, only slightly above the 42.5% achieved with one best global specialist. By contrast, a

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

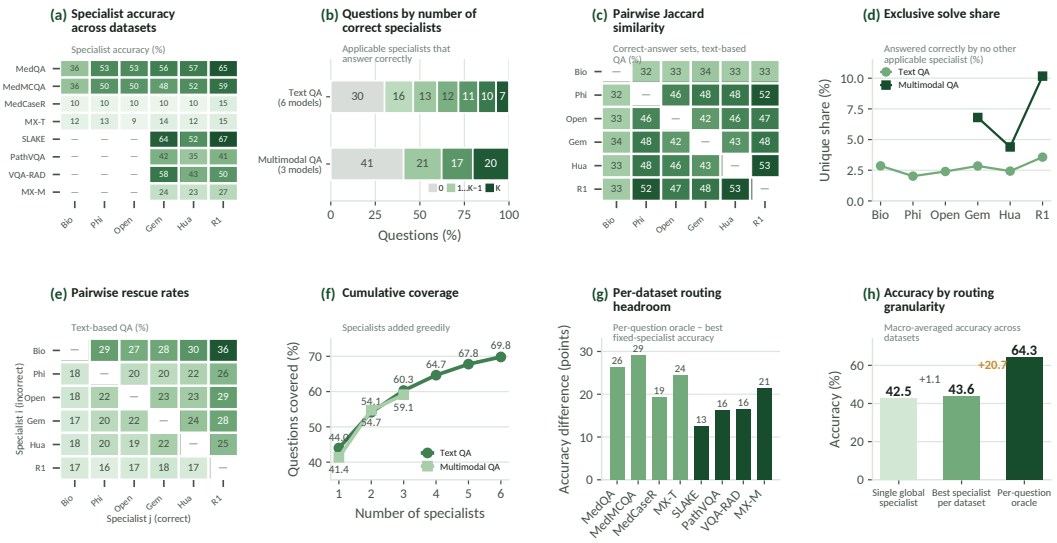

Figure 1: **Question-level complementarity among medical specialist LLMs.** (a) Specialist accuracy across datasets. (b) Distribution of questions by the number of applicable specialists that answer correctly. (c) Pairwise Jaccard similarity between specialists' correct-answer sets on text-based QA. (d) Exclusive solve share: questions answered correctly by one specialist but by no other applicable specialist. (e) Pairwise rescue rates on text-based QA: the fraction of questions answered incorrectly by specialist i that specialist j answers correctly. (f) Cumulative coverage as specialists are greedily added, where a question is covered if at least one selected specialist answers correctly. (g) Per-dataset routing headroom, defined as per-question oracle accuracy minus the best fixed-specialist accuracy. (h) Macro-averaged accuracy of a single global specialist, the best specialist per dataset, and a per-question oracle. The oracle counts a question as correct whenever at least one applicable specialist answers correctly.

per-question oracle that selects a correct specialist response whenever one is available reaches 64.3% (Figure 1(h)). This 20.7-percentage-point gap reflects questions missed by the best fixed specialist for each dataset but answered correctly by another specialist. These findings motivate question-level routing to leverage complementary specialist capabilities for more comprehensive medical reasoning, rather than relying on dataset-level rankings.

Existing routing systems learn to select models at the question level, often to balance answer quality and inference cost (Ding et al., 2024; Chen et al., 2024b; Ong et al., 2025). Some approaches also incorporate responses from selected models into answer generation (Zhang et al., 2025; Qian et al., 2025; Xia et al., 2026). Building on this line of work, we place greater emphasis on leveraging complementary specialist knowledge for medical reasoning. Moreover, many agentic routing systems use a single LLM for both model selection and answer generation, making it difficult to optimize routing to fully exploit the information provided by specialists.

In this paper, we introduce ***MedRouter***, an agentic routing system which separates specialist selection from answer generation. An embedding-based multi-label router selects one or more medical specialist LLMs for each question, and a generator uses their responses together with the original question to produce the final answer. We train the router with ***SCALE*** (***S**pecialist **C**ompetence-**A**ware **LE**arning*), a two-stage procedure. Stage I uses specialist correctness labels to learn which specialists can answer each question independently. Stage II then evaluates the selected responses through the generator and updates the routing policy with reinforcement learning. Its Performance Gain Reward (PGR) is computed from the generator's answer correctness with and without the selected responses. Throughout both stages, only the router is updated; the embedding encoder, specialists, and answer generator remain frozen.

We evaluate *MedRouter* using *Qwen3.5-2B* and *Gemma-4-E2B* as answer generators across eight benchmarks. *MedRouter* achieves the highest mean accuracy across benchmarks among the com-

pared methods with both generators, reaching 45.9% and 46.1%, respectively. Both exceed the 43.6% benchmark average of Best-in-Pool, which selects the strongest standalone specialist separately for each dataset. Ablations with *Qwen3.5-2B* show that SCALE outperforms training with either stage alone, supporting specialist correctness supervision followed by generator-level feedback. Further analyses strengthen the motivation for *MedRouter* and demonstrate its advantages. To sum up, our contributions are:

- We characterize question-level complementarity among medical specialist LLMs and quantify the additional correct-answer coverage available beyond selecting the best specialist for each dataset.
- We introduce *MedRouter* for multi-label specialist selection and response integration, and *SCALE*, which trains the router through specialist correctness supervision followed by reinforcement learning from generator-level feedback.
- We evaluate *MedRouter* across eight text-based and multimodal medical QA benchmarks, achieving the highest mean accuracy among the compared methods with both answer generators. Out-of-distribution tests and ablations assess generalization and the contribution of each training stage.

2 Related Work

Reinforcement Learning for Agentic System. Reinforcement learning (RL) (Kaelbling et al., 1996) has recently been widely applied to LLMs to improve reasoning capabilities and agent adaptation (Ouyang et al., 2022; Guo et al., 2025; Liu et al., 2025b;a). Representative RL algorithms include Proximal Policy Optimization (PPO) (Schulman et al., 2017), Group Relative Policy Optimization (GRPO) (Shao et al., 2024a), and Generalized Diffusion Policy Optimization (GDPO) (Liu et al., 2026). In terms of agentic systems, DeepRetrieval (Jiang et al., 2025a), Search-R1 (Jin et al., 2025), and S3 (Jiang et al., 2025b) apply RL to train models to reason about interactions with search engines for information retrieval and question answering. Meanwhile, Formula-R1 (Cao et al., 2025) uses RL to enable LLMs to perform symbolic table reasoning by deriving executable spreadsheet formulas, while DeepCoder (Luo et al., 2025) uses RL for code reasoning and generation tasks.

Router for LLM Selection. A key motivation for LLM selection is to balance response quality and inference efficiency, since more capable models can be costly to run. A substantial body of work has explored this trade-off. HybridLLM (Ding et al., 2024) and C2MAB-V (Dai et al., 2024) use budget-aware routing and cost-sensitive bandits, respectively. TO-Router (Stripelis et al., 2024) and FORC (Šakota et al., 2024) predict query-specific model suitability, while FrugalGPT (Chen et al., 2023), GraphRouter (Feng et al., 2024), RouteLLM (Ong et al., 2025), and RouterDC (Chen et al., 2024b) explore cascades, graph-based selection, preference-based routing, and contrastive learning. More recently, Router-R1 (Zhang et al., 2025) and xRouter (Qian et al., 2025) optimize routing policies with RL. However, our motivation differs: we aim to understand knowledge differences across medical LLMs and develop a system that combines their complementary capabilities to provide more comprehensive medical reasoning.

Medical Reasoning with LLMs. LLMs are increasingly being applied to a broad range of medical reasoning tasks, including diagnosis (Zhou et al., 2025; Yang et al., 2024; Cao et al., 2026b), clinical prediction from electronic health records (Cao et al., 2026a), and clinical evidence synthesis (Wang et al., 2024; 2025). Representative benchmarks include MedQA (Jin et al., 2020) and PubMedQA (Jin et al., 2019) for medical question answering, as well as PathVQA (He et al., 2020) and PMC-VQA (Zhang et al., 2023) for multimodal medical reasoning. To improve medical reasoning capabilities, some studies train LLMs on large-scale medical corpora from diverse sources (Sellergren et al., 2025; Labrak et al., 2024; Wang et al., 2023; Shen et al., 2026). Other studies seek to combine the strengths of multiple models through ensemble learning or multi-agent collaboration. For example, Yang et al. (Yang et al., 2025b) propose an ensemble framework that aggregates predictions from multiple LLMs through voting. CURE (Elshaer & Rashed, 2025) introduces a confidence-driven ensemble that detects uncertainty in a primary LLM and dynamically routes queries to helper models for collaborative reasoning. Another line of work explores multi-agent collaboration, in which LLMs iteratively refine their predictions through self-consistency and collaboration (Shang et al., 2025). MMedAgent-RL (Xia et al., 2026) uses reinforcement learning to coordinate triage and specialist agents for multimodal medical reasoning. Despite this progress, existing approaches

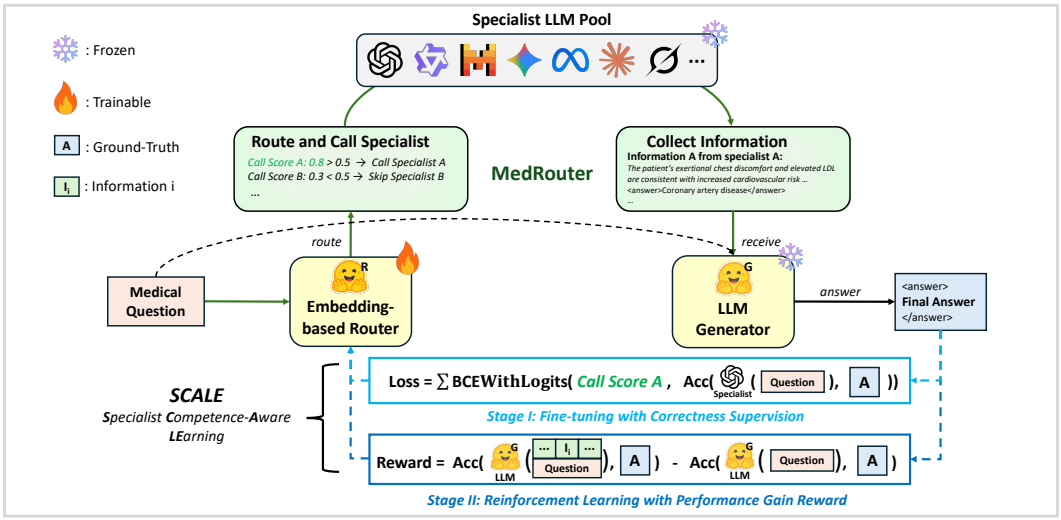

Figure 2: Overview of the *MedRouter* framework and *SCALE* training. Given a medical question, an embedding-based Router predicts call scores and selects specialists from a frozen LLM pool. The frozen LLM generator receives the selected specialists' information and produces the final answer. During training, only the Router is updated: it is first trained with specialist correctness supervision using a binary cross-entropy loss, and then optimized with reinforcement learning using a Performance Gain Reward that compares answer correctness with and without specialist information.

typically rely on a single LLM, combine models through fixed or heuristic strategies, or require costly multi-agent interactions. They do not explicitly learn to exploit the complementary medical knowledge of existing LLMs or to integrate this knowledge effectively for medical reasoning.

3 METHODOLOGY

We introduce *MedRouter*, an agentic system that routes medical questions to a pool of specialist LLMs and aggregates their responses to support medical reasoning, as illustrated in Figure 2. MedRouter separates specialist selection from answer generation into two components: an embedding-based Router and an LLM generator. We formulate specialist selection as a multi-label prediction problem and propose *SCALE* (*Specialist Competence-Aware Learning*), a two-stage framework for training the Router.

3.1 MEDROUTER: SPECIALIST SELECTION AND ANSWER GENERATION

Specialist selection. Let q denote a medical question, $\mathcal{S} = \{s_1, \ldots, s_K\}$ a pool of K frozen specialist LLMs, and A the ground-truth answer. The embedding encoder E_ψ maps q to a representation, which is fed into a multi-label MLP Router g_ϕ to produce one logit per specialist:

$$\mathbf{z}(q) = g_\phi(E_\psi(q)) \in \mathbb{R}^K, \qquad p_i(q) = \sigma(z_i(q)) \,. \tag{1}$$

Here, $p_i(q)$ is the predicted call score for specialist s_i. Separate sigmoid outputs allow more than one specialist to be selected for the same question. At inference, the Router selects all specialists whose scores meet or exceed the threshold $\tau = 0.5$:$\mathcal{S}_q = \{s_i \in \mathcal{S} : p_i(q) \geq \tau\}$. If no specialist meets the threshold, it selects the highest-scoring specialist, i.e., $\mathcal{S}_q = \{\arg\max_{s_i \in \mathcal{S}} p_i(q)\}$.

Answer generation. Each selected specialist produces a response $s_i(q)$. We collect these responses as $I_q = \{s_i(q) : s_i \in S_q\}$ and provide them, together with the original question, to the generator: $\hat{A}_q = G(q, I_q)$. The generator produces the final answer from this input rather than returning a selected specialist's answer directly.

3.2 SPECIALIST COMPETENCE-AWARE LEARNING

Specialist correctness provides a starting point for learning which models to consult, but it does not fully describe how their responses affect the generator. A specialist may reach an incorrect final answer yet provide information that helps the generator answer correctly. *SCALE* therefore first trains the router using individual specialist correctness labels, then refines its selections using generator-level feedback. Throughout training, the embedding encoder, specialist LLMs, and answer generator remain frozen; only the Router parameters ϕ are updated.

3.2.1 STAGE I: LEARNING SPECIALIST COMPETENCE

Let $\mathcal{D} = \{(q_j, A_j)\}_{j=1}^{N}$ denote the training set, where A_j is the reference answer to question q_j. For a model response r, let $C(r, A) \in \{0, 1\}$ indicate whether its answer is correct with respect to reference A. Each specialist receives a binary label $y_{ji} = C(s_i(q_j), A_j)$ for each training question. We train the router with binary cross-entropy on these labels:

$$\mathcal{L}_{\mathrm{FT}} = \frac{1}{NK} \sum_{j=1}^{N} \sum_{i=1}^{K} \mathrm{BCEWithLogits}(z_{ji}, y_{ji}), \tag{2}$$

where N is the number of training questions and K is the number of specialists. A question may have several positive labels, allowing the router to learn the competence of multiple specialists rather than choose a single training target. This stage evaluates specialists independently; it does not use the generator's answer to score their responses.

3.2.2 STAGE II: LEARNING FROM GENERATOR OUTCOMES

Starting from the Stage-I router, we use reinforcement learning to evaluate specialist selections through the frozen generator.

Routing actions. During training, we sample each specialist independently according to its call probability, $a_{qi} \sim \mathrm{Bernoulli}(p_i(q))$. The resulting action $\mathbf{a}_q = (a_{q1}, \ldots, a_{qK})$ specifies the selected set $S_q(\mathbf{a}_q) = \{s_i \in S : a_{qi} = 1\}$. Let $I_q(\mathbf{a}_q)$ denote the responses returned by these specialists. Sampling allows training to evaluate different selections, whereas inference uses the threshold-based rule described above.

Performance Gain Reward. For a training question with reference answer A, the generator produces answers with and without the selected responses:

$$\hat{A}_q^{+} = G(q, I_q(\mathbf{a}_q)), \qquad \hat{A}_q^{-} = G(q, \varnothing). \tag{3}$$

We define the Performance Gain Reward (PGR) as

$$R_{\mathrm{PGR}}(q, \mathbf{a}_q) = C(\hat{A}_q^{+}, A) - \frac{1}{2} C(\hat{A}_q^{-}, A). \tag{4}$$

The reward is 1 when only the answer with specialist responses is correct, 0.5 when both answers are correct, -0.5 when only the answer without specialist responses is correct, and 0 when both are incorrect. The selected set is therefore evaluated through the generator's answer, rather than through the correctness of its individual specialists.

Policy optimization. The independent Bernoulli decisions give the following log-probability for a routing action:

$$\log \pi_\phi(\mathbf{a}_q \mid q) = \sum_{i=1}^{K} \left[a_{qi} \log p_i(q) + (1 - a_{qi}) \log(1 - p_i(q)) \right]. \tag{5}$$

We update the router using REINFORCE with a learned value baseline (Williams, 1992; Sutton et al., 1999):

$$\mathcal{L}_{\mathrm{policy}} = -\mathbb{E}_{q, \mathbf{a}_q} \left[(R_{\mathrm{PGR}}(q, \mathbf{a}_q) - V_\omega(q)) \log \pi_\phi(\mathbf{a}_q \mid q) \right], \tag{6}$$

where the expectation is over training questions and sampled routing actions. The value head $V_\omega(q)$ estimates the expected reward for question q and serves as a baseline to reduce gradient variance. Subtracting this baseline gives the feedback used to increase or decrease the probability of the sampled selection.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUPS

Router and Generator LLMs. MedRouter consists of an embedding-based router and an LLM answer generator. We use the frozen *Qwen3-VL-Embedding-2B* encoder (Li et al., 2026) to obtain a 2,048-dimensional last-token representation of each question. The trainable multi-label routing head applies layer normalization, a linear projection to 256 dimensions, a GELU activation, and a final linear layer that outputs one logit for each of the six specialists. For the generator, we evaluate *Qwen3.5-2B* (Yang et al., 2025a) and *Gemma-4-E2B* (Team et al., 2026).

Medical Specialist LLMs. Our routing system operates over a pool of medical specialists comprising both text-only LLMs and vision-language models. The text-only specialists are BioMistral-7B (Labrak et al., 2024), MediPhi-3.8B (Corbeil et al., 2025), and OpenBioLLM-8B (Aaditya, 2024); the multimodal specialists are MedGemma-4B (Sellergren et al., 2025), HuatuoGPT-Vision-7B (Chen et al., 2024a), and MediX-R1-8B (Mullappilly et al., 2026). Although broadly comparable in parameter scale, these specialists differ in their training corpora, architectures, and training procedures, which may contribute to their distinct performance profiles.

Datasets. We evaluate MedRouter and the baselines on four text-based medical QA benchmarks: MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), MedCaseReasoning (Wu et al., 2025), and the text subset of MedXpertQA (Zuo et al., 2025). For multimodal medical QA, we use SLAKE (Liu et al., 2021), PathVQA (He et al., 2021), VQA-RAD (Lau et al., 2018), and the multimodal subset of MedXpertQA (Zuo et al., 2025). Since MedXpertQA does not provide a training set, we use both its text and multimodal subsets exclusively for out-of-distribution (OOD) evaluation.

Baselines. We compare MedRouter against three groups of baselines: non-routing methods, ML-based routing methods, and LLM-based routing.

- **Non-routing methods.** These methods use a single generator to answer each question without consulting the specialist pool. **Direct** prompts the generator to answer without retrieval or fine-tuning. **RAG** provides the three most similar training questions and their gold answers as in-context examples. We encode questions with *Qwen3-VL-Embedding-2B*, L2-normalize the resulting embeddings, and retrieve training examples based on cosine similarity. **SFT** fine-tunes the generator on the training split, computing loss only on the completion. For multimodal models, images are included during fine-tuning. **RL** trains the generator with GRPO (Shao et al., 2024b) on the training split using a correctness-only reward: $+1$ for an answer judged correct in the `<answer>` block, 0 for an answer judged incorrect, and -1 if the block is missing or contains the prompt placeholder.
- **ML-based routing methods.** These baselines select specialists or aggregate their responses using machine-learning-based and voting-based strategies. The ML-based Routers use *Qwen3-VL-Embedding-2B* as the text encoder. **Ensemble** queries every specialist, normalizes their answers by lowercasing them and removing framing text, and selects the answer with the largest group of matching responses. **KNN Router** embeds each question, retrieves its $k = 50$ nearest training examples, and selects the specialist that answered the largest number of retrieved examples correctly. **MLP Router** uses representations from the frozen encoder as input to a feed-forward network, which predicts the probability that each specialist will answer correctly. It selects all specialists whose scores meet the threshold $\tau = 0.5$, defaulting to the highest-scoring specialist if none meets the threshold. Its output head is multi-label, allowing it to select multiple specialists for a question. **Encoder Router** fine-tunes the encoder end-to-end using a multiclass classification head and cross-entropy loss, allowing us to assess the effect of encoder fine-tuning relative to the MLP Router.
- **LLM-based routing methods.** These methods use an LLM to select specialists, aggregate their responses, or generate the final answer. **LLM Ensemble** provides the question and all specialist responses to an LLM, which generates the final answer. Unlike string voting, it can reconcile paraphrases, discount implausible responses, or select a minority answer, but it queries the entire specialist pool for every question. **Prompt LLM** receives the question and a description of each specialist, selects one specialist, and returns its answer verbatim. It serves as a direct prompting-based LLM routing baseline. **Router-R1** (Zhang et al., 2025) is an RL-trained LLM router optimized using only a final-answer reward: -1 for a missing or placeholder answer, 0 for

Table 1: Accuracy (%) on text-based and multimodal medical QA benchmarks. Within each LLM generator block, the best result for each dataset and the average is shown in **bold**. The relative improvement of *MedRouter* over the second-highest average in the same block is shown in **green**.

	Text-Based Medical QA				Multimodal Medical QA				
	MedQA	**MedMCQA**	**MedCaseRea.**	**MedX.-Text**	**SLAKE**	**PathVQA**	**VQA-RAD**	**MedX.-MM**	**Average**
Best-in-Pool	65.2	59.3	15.1	15.4	67.0	41.7	57.9	27.4	43.6
ML Routing Methods									
Ensemble	61.6	59.0	15.7	14.5	66.2	39.5	50.8	27.0	41.8
KNN Router	61.0	56.5	12.3	16.0	70.0	41.3	56.7	26.1	42.5
MLP Router	57.7	52.2	10.7	13.7	70.7	40.9	56.1	24.9	40.9
Encoder Router	62.5	55.9	14.7	14.9	71.6	40.0	54.3	27.3	42.6
Qwen3.5-2B									
Non-Routing Methods									
Direct	52.0	48.8	7.4	14.6	59.2	37.3	51.6	26.1	37.1
RAG	51.5	43.7	11.4	15.5	70.7	37.8	44.1	25.9	37.5
SFT	51.1	47.8	9.1	16.9	**74.0**	**44.9**	45.4	25.6	39.4
RL	51.7	48.9	10.1	14.4	60.2	42.3	53.9	26.0	38.4
Routing Baselines									
Ensemble	54.7	50.4	14.2	14.7	59.2	37.5	53.2	25.9	38.7
Prompt LLM	53.2	49.0	10.4	13.1	65.6	39.1	**56.0**	24.5	38.9
Router-R1	50.8	46.1	12.4	16.0	52.8	34.9	46.0	20.3	34.9
MedRouter	**71.8**	**59.9**	**19.6**	**17.3**	72.0	43.4	55.1	**27.9**	**45.9** (+7.7%)
Gemma-4-E2B									
Non-Routing Methods									
Direct	63.0	49.3	10.7	14.7	50.6	34.7	48.1	24.1	36.9
RAG	58.1	47.7	12.8	15.4	53.6	39.7	55.0	25.1	38.4
SFT	50.4	45.4	8.8	14.2	53.3	40.1	51.9	23.0	35.9
RL	53.3	44.2	9.4	13.3	48.1	37.3	52.7	20.3	34.8
Routing Baselines									
LLM Ensemble	65.1	57.4	16.7	15.2	51.5	36.4	**55.4**	24.9	40.3
Prompt LLM	54.7	50.1	11.2	12.9	48.3	**45.3**	52.6	26.6	37.7
Router-R1	52.6	46.7	12.3	14.5	46.5	42.9	40.3	25.8	35.2
MedRouter	**74.3**	**63.0**	**20.3**	**16.5**	**67.9**	44.1	53.9	**29.1**	**46.1** (+8.2%)

an answer judged incorrect, and $+1$ for an answer judged correct. The LLM actively selects and calls specialists iteratively, incorporates the returned information, and generates the final answer, keeping routing and generation within a single model.

Training and Evaluation. We use the full test set of each benchmark for evaluation. For training, we use 10,000 examples each from MedQA, MedMCQA, and MedCaseReasoning, and 4,918, 19,654, and 1,793 examples from SLAKE, PathVQA, and VQA-RAD, respectively. This yields 30,000 text-based and 26,365 multimodal training examples, which we combine during training. We randomly sample 200 examples from each dataset's training split to form a 1,200-example development set, corresponding to approximately 2.1% of the selected training examples. The combined test sets contain 8,803 text-only and 10,231 image-based questions, for a total of 19,034 questions. We use *GPT-5.6-luna* as an LLM judge to assess whether each generated answer is semantically consistent with the ground-truth answer in its medical context. Further details on training, evaluation, and the *MedRouter* and baseline configurations are provided in Appendix D.

4.2 Main Results

Table 1 reports accuracy across eight medical QA benchmarks. MedRouter achieves an average accuracy of 45.9% with *Qwen3.5-2B* and 46.1% with *Gemma-4-E2B*, representing relative improvements of 7.7% and 8.2%, respectively, over the strongest routing baseline, Encoder Router (42.6%). Both variants also exceed the 43.6% average of Best-in-Pool, which uses the strongest individual specialist for each dataset. The gains are especially pronounced on text-based QA and extend to both out-of-distribution MedXpertQA subsets. With *Qwen3.5-2B*, MedRouter improves accuracy on MedQA, which draws on medical licensing exams to assess broad medical knowledge, from the best routing baseline's 62.5% to 71.8%, a relative gain of 14.9%. On MedCaseReasoning, which evaluates diagnosis from clinical case reports often involving rare or complex conditions and long-tail reasoning challenges, it achieves a relative accuracy gain of 24.8%. These results demonstrate that *MedRouter* can effectively improve medical QA performance by selecting and integrating specialist responses at the question level. Beyond these performance gains, our experiments also reveal several valuable insights.

Table 2: Ablation study of *SCALE* with *Qwen3.5-2B* as the generator. The best result in each column is shown in **bold**. The relative decrease in average accuracy from the full *SCALE* is shown in **red**.

	Text-Based Medical QA				Multimodal Medical QA				
	MedQA	MedMCQA	MedCaseRea.	MedX.-Text	SLAKE	PathVQA	VQA-RAD	MedX.-MM	Average
MedRouter w/ SCALE	**71.8**	59.9	**19.6**	**17.3**	**72.0**	**43.4**	**55.1**	**27.9**	**45.9**
w/ Stage I FT only	69.3	57.2	17.6	16.5	69.3	41.1	51.4	26.0	43.6 (**-5.0%**)
w/ Stage II RL only	70.1	59.1	17.7	16.7	71.2	41.3	54.9	27.2	44.8 (**-2.4%**)
w/ Stage II RL using correctness reward	70.6	**60.0**	18.4	16.6	69.7	42.6	52.5	27.0	44.7 (**-2.6%**)

Leveraging Knowledge Encoded in Specialist LLMs. The limited gains from directly post-training the generator on our available QA training data suggest that this approach alone is insufficient in our setting. Many current LLMs have already been trained on large amounts of publicly available web data. Meanwhile, existing specialist LLMs may encode medical knowledge acquired from sources not represented in our QA training data. This motivates us to draw on knowledge encoded in existing LLMs rather than relying solely on public datasets. For example, *MedGemma* was trained in part on licensed and de-identified medical datasets (Sellergren et al., 2025). Rather than attempting to transfer all such knowledge into a single generator, *MedRouter* accesses it through specialist responses to improve performance on medical reasoning tasks.

Decoupling Routing and Generation. Our results suggest that jointly learning specialist selection and answer generation within a single LLM can be challenging. For example, Router-R1 (Zhang et al., 2025) uses an RL-trained LLM router with a final-answer reward to select specialists and reason over their responses iteratively. In our experiments, this training is unstable: the final-answer reward provides only an indirect and sparse signal for routing, and jointly optimizing selection and reasoning may make it difficult for the model to learn which specialist to consult for each question. The routing policy can consequently collapse toward a narrow set of specialists, degrading performance. Table 1 also shows that the ML-based routing methods generally outperform the LLM-based routing baselines in our evaluation. Therefore, in designing *MedRouter*, we use an embedding-based router that can be trained directly to predict specialist competence and optimize selections for downstream answer accuracy. A separate LLM generator integrates the selected responses to address diverse questions without also having to make routing decisions. By decoupling routing from generation, we allow each component to focus on a distinct role and achieve better performance.

MedRouter does not achieve the best result on every multimodal benchmark. Specialist responses may omit relevant visual evidence or introduce conflicting information, and our pool contains only three vision-capable specialists. Moreover, because the generator remains frozen, training the router does not directly improve the generator's visual understanding. These factors may limit its ability to answer image-based questions correctly.

4.3 Ablation Study of *SCALE*

To assess the effectiveness of *SCALE*, we ablate each training stage and the Performance Gain Reward (PGR), as shown in Table 2. The full two-stage framework achieves the highest average accuracy of 45.9% and the best result on seven of the eight benchmarks.

Using Stage I fine-tuning alone reduces average accuracy to 43.6%, while using Stage II reinforcement learning alone yields 44.8%. Together, these results suggest that correctness supervision provides a useful initialization, while subsequent RL further improves specialist selection for the downstream generator, making the two stages complementary. The gain over Stage I alone is consistent with the motivation for Stage II: whether a specialist answers correctly is an imperfect measure of whether its response contains useful information, as the case study in Section F illustrates. Conversely, the lower performance of RL from scratch suggests that learning from final-answer feedback alone is more difficult. In our training runs, we find this setting is also more prone to concentrating calls on a narrow set of specialists, which may limit routing diversity and performance.

Replacing PGR with a correctness-only reward ($+1$ for a correct final answer and 0 otherwise) reduces the average accuracy from 45.9% to 44.7%. Unlike the correctness-only reward, PGR uses the generator's answer without specialist information as a question-specific reference, distinguishing cases where specialist information corrects an otherwise incorrect answer from those where the generator can already answer correctly on its own. By explicitly rewarding the improvement at-

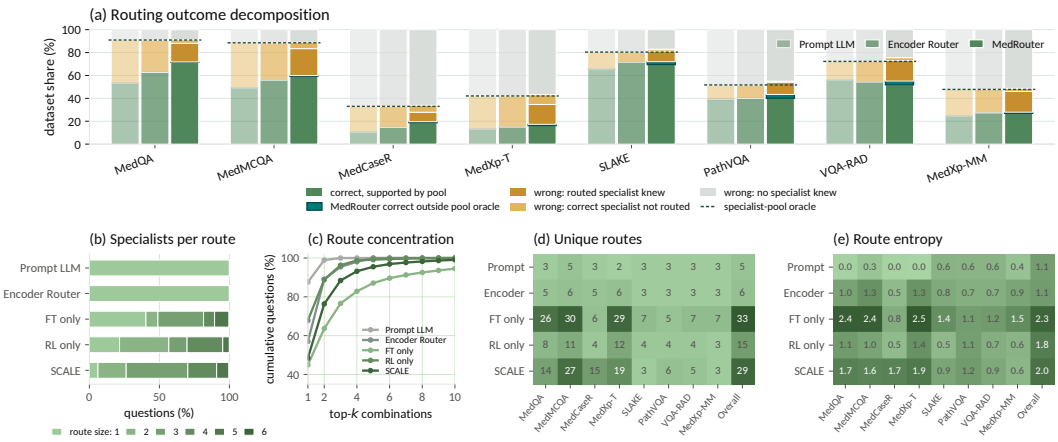

Figure 3: Routing outcomes and specialist selection patterns across eight benchmarks. (a) Decomposition of outcomes for Prompt-LLM, Encoder Router, and *MedRouter*. The dashed line marks the fraction of questions for which at least one specialist knows the answer. (b) Distribution of the number of specialists selected per question. (c) Cumulative fraction of questions assigned to the k most frequent specialist combinations. (d) Number of unique routes on each benchmark and overall. (e) Entropy of the route distribution on each benchmark and overall.

tributable to specialist information, PGR provides a more informative learning signal for routing and may reduce variance in RL updates, leading to more stable training and higher final accuracy.

4.4 Analysis of Routing

Figure 3(a) decomposes the outcomes of different routing strategies according to whether the specialist pool contains a knowledgeable specialist and whether the routed information leads to a correct answer. *MedRouter* achieves a larger proportion of correct answers supported by the specialist pool. Notably, *MedRouter* also answers 332 questions correctly even when no specialist in the pool produces the correct answer. This suggests that the generator can benefit from complementary information in imperfect specialist responses and synthesize a correct answer beyond the capability of any individual specialist; we provide an illustrative example in the case study in Appendix F. Nevertheless, effectively integrating specialist information remains a bottleneck: some questions are still answered incorrectly despite routing to a knowledgeable specialist, as the LLM generator is frozen and is not optimized within our framework.

Figures 3(b)–(e) further examine specialist selection and route diversity. *SCALE* often selects multiple specialists and uses 29 distinct routes across the benchmarks. Although the fine-tuning-only variant uses more distinct routes and exhibits higher route entropy than *SCALE*, *SCALE* still maintains substantially greater route diversity and higher overall entropy than the RL-only variant. These results indicate that the combined training procedure preserves diverse specialist combinations while avoiding the substantial reduction in routing diversity observed with RL alone.

5 Conclusion

In this paper, we examined how medical specialist LLMs differ in question-answering performance and found that their strengths are complementary at the question level. Motivated by this finding, we introduced *MedRouter*, which separates specialist selection from answer generation, and *SCALE*, a two-stage framework that trains the router using specialist correctness supervision followed by reinforcement learning based on the downstream value of specialist responses. Across eight text-based and multimodal medical QA benchmarks, MedRouter improves average accuracy over the strongest routing baselines while keeping the specialists and generator frozen. These results highlight the potential of question-level routing to make better use of existing medical LLMs and motivate further work on combining their capabilities for more complex medical reasoning tasks.

REFERENCES

Aaditya. Openbiollm-llama3-8b. https://huggingface.co/aaditya/Llama3-OpenBioLLM-8B, 2024.

Lang Cao, Jingxian Xu, Hanbing Liu, Jinyu Wang, Mengyu Zhou, Haoyu Dong, Shi Han, and Dongmei Zhang. Fortune: Formula-driven reinforcement learning for symbolic table reasoning in language models. *arXiv preprint arXiv:2505.23667*, 2025.

Lang Cao, Qingyu Chen, and Yue Guo. Ehr-rag: Bridging long-horizon structured electronic health records and large language models via enhanced retrieval-augmented generation. *arXiv preprint arXiv:2601.21340*, 2026a.

Lang Cao, Yuhao Shen, Tianyang Luo, Simo Du, Hao Peng, and Yue Guo. Guideskill: Evolving executable llm agent skills for guideline-grounded clinical reasoning. *arXiv preprint arXiv:2607.26160*, 2026b.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How do large language models acquire factual knowledge during pretraining? In *Advances in Neural Information Processing Systems*, volume 37, pp. 60626–60668, 2024. doi: 10.52202/079017-1939. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6fdf57c71bc1f1ee29014b8dc52e723f-Abstract-Conference.html.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision: Towards injecting medical visual knowledge into multimodal llms at scale, 2024a.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Shuhao Chen, Weisen Jiang, Baijiong Lin, James T. Kwok, and Yu Zhang. RouterDC: Query-based router by dual contrastive learning for assembling large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 66305–66328, 2024b. doi: 10.52202/079017-2120. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7a641b8ec86162fc875fb9f6456a542f-Abstract-Conference.html.

Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, François Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. A modular approach for clinical slms driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment, 2025.

Xiangxiang Dai, Jin Li, Xutong Liu, Anqi Yu, and John Lui. Cost-effective online multi-llm selection with versatile reward models. *arXiv preprint arXiv:2405.16587*, 2024.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid LLM: Cost-efficient and quality-aware query routing. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b47d93c99fa22ac0b377578af0a1f63a-Abstract-Conference.html.

Ziad Elshaer and Essam A Rashed. Cure: Confidence-driven unified reasoning ensemble framework for medical question answering. *Big Data and Cognitive Computing*, 9(12):299, 2025.

Tao Feng, Yanzhen Shen, and Jiaxuan You. Graphrouter: A graph-based router for llm selections. *arXiv preprint arXiv:2410.03834*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

Pengcheng Jiang, Jiacheng Lin, Lang Cao, R. Tian, S. Kang, Z. Wang, Jimeng Sun, and Jiawei Han. Deepretrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. *arXiv preprint arXiv: 2503.00223*, 2025a.

Pengcheng Jiang, Xueqiang Xu, Jiacheng Lin, Jinfeng Xiao, Zifeng Wang, Jimeng Sun, and Jiawei Han. s3: You don't need that much data to train a search agent via rl. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21610–21628, 2025b.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 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. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua 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)*, pp. 2567–2577, 2019.

Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement learning: A survey. *J. Artif. Intell. Res.*, 4:237–285, 1996. URL https://api.semanticscholar.org/CorpusID:1708582.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5848–5864. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.348. URL https://aclanthology.org/2024.findings-acl.348/.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):1–10, 2018.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging*, pp. 1650–1654, 2021.

Hanbing Liu, Lang Cao, Yuanyi Ren, Mengyu Zhou, Haoyu Dong, Xiaojun Ma, Shi Han, and Dongmei Zhang. Bingo: Boosting efficient reasoning of llms via dynamic and significance-based reinforcement learning. *arXiv preprint arXiv:2506.08125*, 2025a.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026.

Yihao Liu, Shuocheng Li, Lang Cao, Yuhang Xie, Mengyu Zhou, Haoyu Dong, Xiaojun Ma, Shi Han, and Dongmei Zhang. Superrl: Reinforcement learning with supervision to boost language model reasoning. *arXiv preprint arXiv:2506.01096*, 2025b.

Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, et al. Deepcoder: A fully open-source 14b coder at o3-mini level. *Notion Blog*, 1, 2025.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Jaseel Muhammad Kaithakkodan, Omair Mohamed, Mohamed Zidan, Fahad Khan, Salman Khan, Rao Anwer, and Hisham Cholakkal. MediX-R1: Open ended medical reinforcement learning. In *Findings of the Association for Computational Linguistics: EMNLP 2026*. Association for Computational Linguistics, 2026. URL https://medix.cvmbzuai.com/. Accepted for publication.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5503a7c69d48a2f86fc00b3dc09de686-Abstract-Conference.html.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022. URL https://api.semanticscholar.org/CorpusID:246426909.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 248–260, 2022.

Cheng Qian, Zuxin Liu, Shirley Kokane, Akshara Prabhakar, Jielin Qiu, Haolin Chen, Zhiwei Liu, Heng Ji, Weiran Yao, Shelby Heinecke, Silvio Savarese, Caiming Xiong, and Huan Wang. xrouter: Training cost-aware llms orchestration system via reinforcement learning. *arXiv preprint arXiv:2510.08439*, 2025.

Marija Šakota, Maxime Peyrard, and Robert West. Fly-swat or cannon? cost-effective language model choice via meta-modeling. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 606–615, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017. URL https://api.semanticscholar.org/CorpusID:28695052.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. doi: 10.48550/arXiv.2507.05201. URL https://arxiv.org/abs/2507.05201.

Kexin Shang, Chia-Hsuan Chang, and Christopher C Yang. Collaboration among multiple large language models for medical question answering. In *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)*, pp. 388–396. IEEE, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024a. URL https://arxiv.org/abs/2402.03300.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b.

Yuhao Shen, Lang Cao, Simo Du, Yuqing Wang, Juexiao Zhou, Hao Peng, and Yue Guo. Medguidex: Internalizing decision logic from executable guidelines into large language models for clinical reasoning. *arXiv preprint arXiv:2605.26567*, 2026.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Dimitris Stripelis, Zhaozhuo Xu, Zijian Hu, Alay Dilipbhai Shah, Han Jin, Yuhang Yao, Jipeng Zhang, Tong Zhang, Salman Avestimehr, and Chaoyang He. Tensoropera router: A multi-model router for efficient llm inference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 452–462, 2024.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. Accelerating clinical evidence synthesis with large language models. *arXiv preprint arXiv: 2406.17755*, 2024.

Zifeng Wang, Lang Cao, Qiao Jin, Joey Chan, Nicholas Wan, Behdad Afzali, Hyun-Jin Cho, Chang-In Choi, Mehdi Emamverdi, Manjot K. Gill, Sun-Hyung Kim, Yijia Li, Yi Liu, Hanley Ong, Justin Rousseau, Irfan Sheikh, Jenny J. Wei, Ziyang Xu, Christopher M. Zallek, Kyungsang Kim, Yifan Peng, Zhiyong Lu, and Jimeng Sun. A foundation model for human-ai collaboration in medical literature mining. *arXiv preprint arXiv: 2501.16255*, 2025.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J. Tao, Min Woo Sun, Alejandro Lozano, and James Zou. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports, 2025.

Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Zihan Dong, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Linjun Zhang, Shujie Liu, Yan Lu, and Huaxiu Yao. MMedAgent-RL: Optimizing multi-agent collaboration for multimodal medical reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2awntLXwR6.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. Medical foundation large language models for comprehensive text analysis and beyond. *NPJ digital medicine*, 8(1):141, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–29, 2024.

Han Yang, Mingchen Li, Huixue Zhou, Yongkang Xiao, Qian Fang, Shuang Zhou, and Rui Zhang. Large language model synergy for ensemble learning in medical question answering: design and evaluation study. *Journal of Medical Internet Research*, 27:e70080, 2025b.

Haozhen Zhang, Tao Feng, and Jiaxuan You. Router-R1: Teaching LLMs multi-round routing and aggregation via reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4719. URL https://papers.nips.cc/paper_files/paper/2025/hash/ceaa137fce916aba5c65fceb1309088b-Abstract-Conference.html.

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

Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Yi Fang, Sirui Ding, Jiashuo Wang, Kaishuai Xu, et al. Large language models for disease diagnosis: A scoping review. *npj Artificial Intelligence*, 1(1):9, 2025.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80961–80990. PMLR, 2025. URL https://proceedings.mlr.press/v267/zuo25a.html.

Contents of Appendix

A AI Use Statement **16**

B Ethics Statement **16**

C Limitations **16**

D Experimental Setup Details **16**

E Performance of Medical Specialist LLMs **17**

F Case Study: Integrating Complementary Specialist Information **18**

G Prompts Used in This Work **21**

A AI USE STATEMENT

In this work, we used generative AI tools for interpreting experimental results and assisting with the writing of the manuscript. We have not used generative AI tools for generating datasets, conducting experiments without author verification, or making final research decisions, and other required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for brainstorming, literature search, improving readability, polishing text, organizing paper structure, formatting tables, and formatting references. We have reviewed all AI-assisted work. We manually checked technical claims, citations, experimental results, and manuscript content. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

B ETHICS STATEMENT

This work studies how to route medical questions to specialist LLMs and combine their responses to support medical reasoning. Our experiments use public medical question-answering benchmarks and model-generated responses for research evaluation. These systems are not intended to provide clinical diagnoses, treatment recommendations, or other professional medical advice, and their outputs may be inaccurate, incomplete, or inconsistent. MedRouter should therefore be viewed as a research framework rather than a clinical decision-making tool. Any real-world clinical use would require validation by qualified medical professionals, appropriate privacy protections, and rigorous assessment of safety and potential harm.

C LIMITATIONS

This work focuses on routing medical questions to specialist LLMs and leveraging their responses to improve answer correctness, but several limitations remain. MedRouter depends on the availability and reliability of its specialist models, and its performance may vary across datasets and deployment settings. Moreover, our evaluation is limited to medical question answering, whereas many real-world medical tasks involve more complex forms of reasoning and information synthesis. Such tasks may require gathering, organizing, and integrating specialist input over multiple steps. Future work could extend MedRouter to these settings and investigate more effective ways to exploit specialist information for complex medical reasoning.

D EXPERIMENTAL SETUP DETAILS

Specialist Response Generation. Each specialist answers every question once offline, and we cache its response for use during training and evaluation. This ensures that all routing methods receive the same specialist information. We generate responses with temperature 0.0, a maximum of 1536 tokens, and thinking mode disabled. The prompt requires the format `<reasoning>...</reasoning><answer>...</answer>`. When a specialist is selected, the generator receives only its reasoning, not its final answer.

***MedRouter* Training.** In Stage I of fine-tuning, we train the policy head with binary cross-entropy to predict whether each specialist answers a question correctly. We use AdamW with a learning rate of 2×10^{-5} and a batch size of 16. Training runs for up to 3 epochs, with evaluation every 100 steps and early stopping after 3 evaluations without improvement. The resulting checkpoint initializes Stage II and also serves as a standalone baseline. In Stage II of RL, we randomly initialize the value head. We use AdamW with learning rates of 4×10^{-5} for the policy and 1×10^{-4} for the value head, a batch size of 128 questions, an entropy coefficient of $\beta = 0.01$, and gradient clipping at a norm of 1.0. Batches reflect the proportions of the training sources. Training runs for up to 10 epochs, with development-set evaluation and checkpointing every 100 steps.

Baselines. We compare *MedRouter* with several routing baselines. The KNN router retrieves the $k = 50$ nearest training questions using frozen *Qwen3-VL-Embedding-2B* embeddings, while the MLP router uses the same architecture as the policy head in *MedRouter* over the same embedding features. RAG use original question to retrieve the top $k = 3$ training questions using the same

Table 3: Accuracy (%) of individual models on text-based and multimodal medical QA benchmarks. Best-in-Pool selects the highest-scoring medical specialist separately for each dataset. Bold values indicate the best result among individual medical specialists in each column.

	Text-Based Medical QA					Multimodal Medical QA				
	MedQA	**MedMCQA**	**MedCaseRea.**	**MedX.-Text**	**Average**	**SLAKE**	**PathVQA**	**VQA-RAD**	**MedX.-MM**	**Average**
Best-in-Pool	65.2	59.3	15.1	15.4	38.7	67.0	41.7	57.9	27.4	48.5
General-Purpose LLMs										
Qwen3.5-2B	52.0	48.8	7.4	14.6	30.7	59.2	37.3	51.6	26.1	43.6
Gemma-4-E2B	63.0	49.3	10.7	14.7	34.4	50.6	34.7	48.1	24.1	39.4
Medical Specialist LLMs										
BioMistral-7B	35.9	36.0	10.0	12.2	23.5	–	–	–	–	–
MediPhi-3.8B	53.1	50.2	10.3	12.7	31.6	–	–	–	–	–
OpenBioLLM-8B	52.6	49.9	9.6	9.3	30.3	–	–	–	–	–
MedGemma-4B	56.0	47.7	9.6	14.5	31.9	63.6	**41.7**	**57.9**	23.9	**46.8**
HuatuoGPT-V-7B	56.6	52.2	9.5	12.4	32.7	51.8	34.9	42.6	22.6	38.0
MediX-R1-8B	**65.2**	**59.3**	**15.1**	**15.4**	**38.7**	**67.0**	40.9	49.9	**27.4**	46.3

embeddings and provides them as context for generation. For Router-R1, we remove the cost reward and optimize the router using only the answer-correctness reward. The SFT and RL baselines use the default training configurations provided by VERL (Sheng et al., 2024).

Infrastructure. All training and evaluation runs use a single NVIDIA GH200 GPU. Specialists and the generator are served with vLLM (Kwon et al., 2023).

E Performance of Medical Specialist LLMs

Table 3 reports the performance of each model across the eight medical QA benchmarks. Among the medical specialists, MediX-R1-8B performs best on all four text-based datasets, achieving an average accuracy of 38.7%. On multimodal QA, the strongest specialist varies by dataset: MediX-R1-8B leads on SLAKE and MedXpertQA-MM, while MedGemma-4B leads on PathVQA and VQA-RAD and achieves the highest multimodal average of 46.8%. Selecting the best specialist separately for each multimodal dataset raises the average to 48.5%, compared with 46.8% for the best single specialist. For text-based QA, the dataset-level Best-in-Pool average equals MediX-R1-8B's average; the potential benefit of routing within these datasets instead arises from the question-level differences analyzed in Figure 1. Text-only specialists are not evaluated on multimodal datasets because they do not accept image inputs

F Case Study: Integrating Complementary Specialist Information

We examine a MedXpertQA-Text question for which the baseline `Qwen3.5-2B` generator answers incorrectly in isolation but reaches the correct answer when provided with routed specialist reasoning traces. Notably, all routed specialists select incorrect final options, yet their intermediate reasoning contains complementary diagnostic cues that the generator can integrate to recover the correct solution. The quoted outputs retain their original wording; `[...]` denotes omitted passages.

Question: MedXpertQA-Text #1464

A 56-year-old female recent immigrant from rural China is evaluated after her husband brings her to the physician with a two-day history of confusion, fever, headaches, and malaise. Her vaccination history is unknown. She presents with a temperature of 39.1 °C (102.4 °F), nuchal rigidity, and photophobia. CSF analysis reveals a neutrophil count of 1,500/mm^3. While initial cultures on **Thayer–Martin and sheep blood agar show no growth**, subsequent heating of the sheep blood agar yields multiple non-hemolytic, opaque, cream-colored colonies. What is the most likely morphological and staining characteristic of the causative organism?

(A) Gram-negative coccobacillus (B) Gram-positive, lancet-shaped diplococcus (C) Gram-variable, filamentous organism (D) Gram-negative, facultative intracellular bacillus (E) Gram-negative diplococcus (F) Gram-positive, pleomorphic rod (G) Gram-positive, spore-forming rod (H) Gram-negative, encapsulated bacillus (I) Gram-positive, branching bacillus (J) Gram-negative, aerobic coccus

Ground truth: (A) Gram-negative coccobacillus.

The culture phenotype points to *Haemophilus influenzae*, a fastidious Gram-negative coccobacillus that requires X (hemin) and V (NAD) factors for growth. Heating blood during the preparation of chocolate agar lyses red blood cells, making these growth factors available, while also inactivating heat-labile NADases that degrade factor V. The absence of growth on selective Thayer–Martin agar further weighs against *Neisseria meningitidis*.

Without specialist information, the baseline `Qwen3.5-2B` generator incorrectly selects **(B) Gram-positive, lancet-shaped diplococcus**, consistent with *Streptococcus pneumoniae* rather than *H. influenzae*.

All three routed specialists below produce incorrect final option selections. Their intermediate reasoning traces, however, contain distinct diagnostic signals.

Specialist Outputs

OpenBioLLM-8B.
Raw Output:

```
The most likely organism in this scenario is Haemophilus influenzae
    , which is known to grow better on enriched media such as Fildes
    ' agar or chocolate agar, and can be identified by its
    characteristic 'drumstick' appearance on Gram stain.
[...]
Therefore, the most likely morphological and staining
    characteristic of the causative organism is a Gram-negative,
    aerobic coccus.
```

Analysis. OpenBioLLM-8B correctly identifies the causative organism as *H. influenzae*, but subsequently mischaracterizes its morphology. *H. influenzae* is a Gram-negative coccobacillus rather than a coccus, and the "drumstick" appearance is classically associated with the terminal spores of *Clostridium tetani*. Thus, the useful signal in this trace is the organism-level identification, whereas its morphological characterization is misleading.

MedGemma-4B.

Raw Output:

```
This is highly suggestive of *Legionella* species.
[...]
(E) Gram-negative diplococcus: This is characteristic of *Neisseria
     gonorrhoeae* or *Neisseria meningitidis*, which would typically
     show growth on Thayer-Martin agar.
[...]
(H) Gram-negative, encapsulated bacillus: This is characteristic of
     *Haemophilus influenzae* or *Klebsiella pneumoniae*, which
    would typically show growth on blood agar.
[...]
Therefore, the most likely answer is (D).
```

Analysis. MedGemma-4B incorrectly concludes (D), but its reasoning provides useful elimination evidence: it notes that *Neisseria* species would typically grow on Thayer–Martin agar, whereas no such growth is observed in the vignette. It also considers *H. influenzae* in its differential diagnosis. Although its claim that *H. influenzae* would typically grow on unheated blood agar is incorrect, the trace nevertheless contributes a useful cue for distinguishing *H. influenzae* from *Neisseria*.

MediX-R1-8B.
Raw Output:

```
[...] it allows growth of organisms like Neisseria that are
    inhibited at room temperature or require specific conditions.
[...]
Therefore, the most likely morphological and staining
    characteristic is Gram-negative diplococcus.
```

Analysis. MediX-R1-8B incorrectly interprets the growth after heating as evidence for *Neisseria* and consequently selects option (E). Unlike the preceding traces, this reasoning provides a misleading diagnostic signal that the generator must reject.

The specialist traces therefore contain a mixture of useful and misleading evidence. OpenBioLLM-8B supplies the correct organism-level hypothesis (*H. influenzae*) but an incorrect morphology, whereas MedGemma-4B supplies useful negative evidence against *Neisseria* despite reaching an incorrect final answer. MediX-R1-8B instead reinforces an incorrect competing diagnosis. Recovering the correct answer requires selectively retaining the useful intermediate cues, rejecting the specialists' erroneous conclusions, and mapping *H. influenzae* to its correct coccobacillary morphology.

Generator With Routed Specialist Information

Raw Generator Output:

```
Heating denatures inhibitors (like hemolysin) that prevent growth
    at room temperature, allowing the organism to thrive.
[...]
While *Neisseria meningitidis* is a strong candidate for meningitis
     in this demographic, the specific culture behavior (growth only
     after heating) and the colony description point more strongly
    toward *Haemophilus influenzae*.
[...]
Therefore, the organism is *Haemophilus influenzae*, which is a
    Gram-negative coccobacillus.
```

Answer: (A) Gram-negative coccobacillus.
Analysis. Given the routed specialist traces, the generator correctly favors *H. influenzae* over *Neisseria* and maps the organism to its correct morphology. In doing so, it retains the useful organism-level hypothesis while rejecting the incorrect morphological characterization provided by OpenBioLLM-8B and the competing *Neisseria* interpretation proposed by MediX-R1-8B. The final answer is therefore correct even though the generator's mechanistic

explanation is not entirely precise: while heating does inactivate heat-labile inhibitors, the specific reference to "hemolysin" is unsupported. More precisely, heating lyses red blood cells to make X (hemin) and V (NAD) factors available and inactivates heat-labile NADases that would otherwise degrade factor V.

This case illustrates a key advantage of integrating specialist reasoning rather than relying solely on specialist final answers: even when every specialist selects an incorrect option, their intermediate traces can contain complementary diagnostic evidence. MedRouter enables the generator to exploit these partial signals while filtering contradictory or erroneous information, yielding a correct answer that neither the baseline generator nor any individual routed specialist produces independently.

G PROMPTS USED IN THIS WORK

Figure 4 presents the prompts used for answer evaluation and generation. The evaluation prompt asks the judge to compare a model response with the ground-truth answer and return a binary judgment. The generation prompt provides the medical question and, when specialists are selected, their responses in separate `<information>` blocks. It instructs the generator to consider these responses as potentially useful but fallible evidence, since specialists may be incorrect or disagree.

Answer Evaluation Prompt

Judge whether the model response gives the correct final answer to the question.

Question:

{question}

Model response:

{response}

Ground−truth answer(s):

{groundtruth}

Answer with only 'yes' or 'no'.

Answer Generation Prompt

Answer the medical question.

You may be given information from other medical LLMs. Use it as a reference, but be aware that some responses may be incorrect and different sources may conflict. Weigh the information against your own judgment rather than copying it.

Reply in the following format:
<reasoning>...</reasoning><answer>...</answer>.

Question:

{question}

<information>{specialist 1 response}</information>
<information>{specialist 2 response}</information>

...

Figure 4: Prompts used for answer evaluation and answer generation. Each selected specialist's response is placed in a separate information block.