Cognitive-Level Prefix-MoE Reward Modeling for Curriculum Alignment of Medical Language Models

Anonymous ACL submission

Abstract

Medical language models are increasingly evaluated on open-ended clinical reasoning, where good responses must combine factual recall, evidence integration, decision making, and synthesis. However, standard reward models compress these different standards into one reward score, while medical RL with answer keys mainly rewards multiple-choice correctness. This makes it difficult to improve non-multiple-choice medical reasoning without losing performance on exam-style benchmarks. In this paper, we propose a Prefix Mixture-of-Experts Reward Model (Prefix-MoE RM), which routes each query-response pair across recall, analysis, decision, and synthesis prefix experts, softly fuses their trainable prefixes, and then predicts one reward score. The MoE structure allows different abilities to use different scoring experts to reduce reward conflicts, while continuous prefixes provide a more expressive conditioning signal than fixed hard prompts. The reward model is trained on non-multiple-choice medical preferences and then used in mixed-data curriculum GRPO, where multiple-choice prompts provide additional answer-key feedback. On held-out reward-model evaluation, Prefix-MoE RM improves pairwise accuracy from 79.8 to 93.0 over a shared RM; with the RL pipeline fixed, replacing it with a shared RM or soft expert-head MoE lowers HealthBench from 43.8 to 24.8 and 34.6, respectively. With Qwen3-8B as the policy backbone, the full system outperforms the larger Qwen3-32B base on MedXpertQA and HealthBench, while the Qwen3-32B version reaches 50.1 HealthBench and 4.55 LLMEval-Med, remaining competitive with stronger closed general models and retaining an advantage on non-multiple-choice metrics.

1 Introduction

Medical language-model alignment must reward different kinds of medical reasoning within a single policy. Some questions ask for factual recall, while others ask the model to combine evidence, choose a safe next step, or synthesize a broader medical answer. These cases should not be judged by exactly the same standard: recall focuses on correctness and completeness, analysis on evidence use, decision making on risk and safety, and synthesis on balanced reasoning. If all feedback is merged into one reward, the training signal becomes unclear about which skill the policy is improving. This creates reward conflict: a scoring rule that favors concise factual recall can be a poor fit for longer differential diagnosis, risk-benefit discussion, or synthesis.

The standard RLHF pipeline trains a reward model from pairwise preferences and uses it to optimize a policy with reinforcement learning (Ouyang et al., 2022). For medical alignment, this design mixes easy factual questions with harder clinical reasoning questions. A single reward head can fit common factual patterns but miss the standards needed for decision or synthesis cases. Hand-written prompts can name the intended standard, but fixed text has limited expressive capacity, and hard expert selection still forces each example into one discrete level even when it lies between two levels. Prior structured reward models have attempted to reduce such conflicts by separating preference signals with hierarchical reward modeling, multi-objective preference optimization, or mixture-of-experts reward heads (Lai et al., 2024; Zhou et al., 2024; Quan, 2024; Wang et al., 2024a); we build on this idea with a simple ordered difficulty structure for medicine.

We propose Prefix-MoE RM, a cognitive-level reward model that adapts the revised Bloom taxonomy to four medical levels: recall, analysis, decision, and synthesis (Anderson and Krathwohl, 2001). A frozen cognitive router estimates a soft-level distribution for each query-response pair during reward scoring. The reward model then uses these distribution to fuse trainable level-specific

prefix embeddings before scalar reward prediction. The fused prefix is prepended to the query-response representation, so the model can adjust how it scores an answer without replacing the shared backbone or value head. The mixture separates scoring standards for different abilities, and the prefixes represent each standard as a learned continuous context rather than a fixed instruction. This design handles boundary cases by scoring an example with a blend of nearby levels instead of forcing it into one hard prompt or expert.

We use the learned reward model in mixed-data GRPO, with non-multiple-choice medical response quality as the main target. Open-ended preference data train the model to score long medical answers, while MedKgQA-style multiple-choice questions provide answer-key feedback during RL. Because the reward model is trained on non-multiple-choice preferences, we expect its most direct gains to appear on open-answer benchmarks rather than only on exam-style multiple-choice tasks. The multiple-choice data are mixed in during RL to prevent degradation on standard medical QA benchmarks and to provide low-variance feedback where an answer key is available. The training data are also organized by cognitive level so that policy optimization gradually expands from recall and analysis to decision and synthesis examples; we treat this curriculum as an empirical training heuristic and compare it with simpler schedules (Bengio et al., 2009).

Our main contributions are:

- We formulate medical reward modeling as level-aware preference learning, focusing on how recall, analysis, decision, and synthesis require different scoring standards.
- We introduce Prefix-MoE RM, which uses frozen router probabilities to softly fuse trainable level prefixes, allowing different abilities to use different scoring experts while producing one reward score for RL.
- We provide controlled reward-model and policy-level evidence that the approach improves non-multiple-choice medical responses while answer-key feedback preserves multiple-choice performance.

2 Method

Figure 1 shows the two places where cognitive levels enter the pipeline. During reward-model training, the router turns a query-response pair into a soft level distribution, which determines how level prefixes are fused before scalar scoring. During policy optimization, the trained reward model scores open-ended prompts while MCQ prompts use answer-key reward, so the RL stage can improve clinical response quality without treating answer extraction and long-form reasoning as the same supervision signal. This separation matters because the method is meant to identify where the open-ended supervision enters the alignment process. It avoids treating all gains as a generic effect of medical-domain RL.

2.1 Problem Setup

We train medical models from pairwise preference data and reinforcement learning prompts. Let x denote a medical query and y denote a candidate response. A reward model assigns a scalar score $R_\psi(x, y)$, where higher scores should correspond to preferred responses. Unlike a standard reward model, we assume each query is associated with a cognitive level $c \in \mathcal{C} = \{1, 2, 3, 4\}$. The four levels correspond to **factual recall**, **clinical analysis**, **clinical decision making**, and **biomedical synthesis**. The goal is to avoid one scoring rule for all abilities: the reward remains scalar, but its context changes with the level. Here c denotes the reasoning standard for evaluation rather than a disease specialty: two questions about the same condition may require different rewards if one asks for a definition and the other asks for a management plan.

2.2 Cognitive-Level Router

The cognitive router estimates which reasoning level is needed for a medical input. We instantiate the router with a Qwen3-8B backbone and a linear classification head over the final non-padding token. The router is trained with cross-entropy loss on labeled medical questions:

$$\mathcal{L}_{\text{router}} = -\log q_\phi(c^\star \mid x), \tag{1}$$

where $c^\star$ is the cognitive-level label. During reward-model training and inference, the router is frozen. Before training the router, we use an LLM classifier to label each query from its question stem and options, using a prompt grounded in Anderson and Krathwohl's revised Bloom taxonomy (Anderson and Krathwohl, 2001) and adapted to the four medical levels; the full prompt is shown

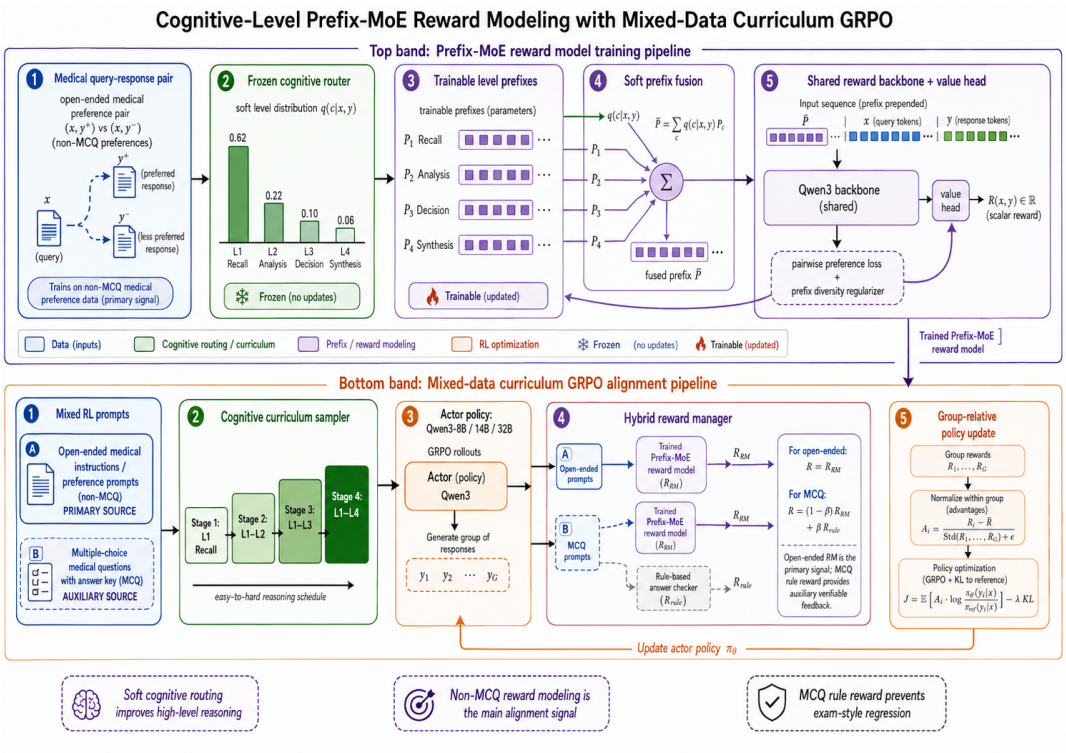

Figure 1: Overview of the proposed alignment pipeline. The top band trains a cognitive-level Prefix-MoE reward model on non-MCQ medical preference pairs, using a frozen router to softly fuse trainable level prefixes so different abilities can use different scoring contexts before scalar reward prediction. The bottom band uses the trained reward model as the primary signal for mixed-data curriculum GRPO, while MCQ answer checking helps preserve exam-style performance.

in Appendix C. Because the router is trained and selected on held-out LLM-labeled examples, its metrics measure consistency with the labeling protocol rather than manual annotation; after training, it is frozen so the reward model scores within fixed routed levels.

The router is used in two input formats: question-only labels group prompts for curriculum construction, while formatted query-response inputs produce level probabilities for reward scoring. We use a low-temperature softmax to obtain a sharp but non-degenerate distribution, with τ selected on the held-out router split:

$$\alpha_c(x, y) = \frac{\exp(z_c/\tau)}{\sum_{c'=1}^{4} \exp(z_{c'}/\tau)}, \quad \tau = 0.1, \quad (2)$$

where z_c is the router logit for level c. This soft distribution allows examples near a boundary to borrow information from multiple cognitive levels rather than being assigned to a single hard expert. The low temperature makes the largest level dominate while still leaving nonzero probability mass for boundary cases that sit between adjacent reasoning levels.

2.3 Prefix-MoE Reward Model

Prefix-MoE RM uses MoE to separate scoring standards through trainable level-specific prefixes rather than through separate reward heads. For each $c \in \mathcal{C}$, we define a prefix matrix $P_c \in \mathbb{R}^{m \times d}$, where m is the shared padded prefix length and d is the token embedding dimension of the backbone. Prefixes are initialized from the token embedding of natural-language level descriptions, padded to the shared length, and then treated as trainable parameters during reward-model training. Unlike a hard prompt, each prefix is a continuous parameter block, so pairwise preference training can adjust the scoring representation directly. Each row of P_c acts as a virtual token embedding; together, the rows define the scoring context for level c.

Given a query-response pair (x, y), the frozen router produces $\alpha_c(x, y)$ for each level. The model forms a fused prefix by weighted averaging:

$$\bar{P}(x, y) = \sum_{c=1}^{4} \alpha_c(x, y) P_c. \quad (3)$$

Let $E(x, y)$ be the token embedding sequence for the formatted pair. The reward backbone receives

the concatenated embedding sequence

$$\widetilde{E}(x,y) = [\bar{P}(x,y); E(x,y)]. \tag{4}$$

The final hidden state at the last non-padding token is passed through a PReLU activation and a linear value head to produce the scalar reward. The complete forward pass is therefore

$$\begin{aligned}(x,y) &\xrightarrow{q_\phi} \alpha(x,y) \xrightarrow{\{P_c\}_{c=1}^{4}} \bar{P}(x,y) \\ &\xrightarrow{[\cdot;E(x,y)]} \widetilde{E}(x,y) \xrightarrow{f_\psi} R_\psi(x,y),\end{aligned} \tag{5}$$

where f_ψ denotes the shared reward backbone and value head. This mixes prefix parameters before reward prediction: hard routing is recovered when α is one-hot, while soft routing blends cognitive-level contexts.

This architecture differs from hard-prompt reward conditioning in two ways. First, the router distribution is used continuously, so the reward model can interpolate between adjacent cognitive levels instead of committing to one label. Second, the prefixes are trainable parameters rather than fixed text embeddings, so pairwise preference supervision can directly shape the evaluation representation. Thus the model does not rely on a fixed prompt to both express the scoring rule and separate query types. Compared with separate expert heads or reward experts (Yuan et al., 2026), the fused-prefix design keeps a single shared backbone and value head while still changing the token-level representation seen by the reward model. We compare these design variants in the policy-level RM swap in Table 3.

2.4 Reward-Model Training Objective

The reward model is trained on pairwise preferences. Each example contains a query x, a chosen response y^+, a rejected response y^-, and ranks indicating that y^+ should receive higher reward. For each valid pair in a batch, we minimize the logistic ranking loss

$$\mathcal{L}_{\text{rank}} = -\log \sigma\left(R_\psi(x,y^+) - R_\psi(x,y^-)\right). \tag{6}$$

This loss treats the reward difference $R_\psi(x,y^+) - R_\psi(x,y^-)$ as the model's preference margin. Positive margins mean the chosen answer is scored higher than the rejected answer; negative margins are penalized.

To prevent the four prefixes from collapsing to the same representation, we add a diversity regularizer. Let $\bar{p}_c$ be the mean-pooled valid prefix embedding for level c. We penalize the average squared pairwise cosine similarity:

$$\mathcal{L}_{\text{div}} = \frac{2}{|\mathcal{C}|(|\mathcal{C}|-1)} \sum_{i<j} \cos^2(\bar{p}_i, \bar{p}_j). \tag{7}$$

The final reward-model objective is

$$\mathcal{L}_{\text{RM}} = \mathcal{L}_{\text{rank}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}, \quad \lambda_{\text{div}} = 0.1. \tag{8}$$

The first term learns preferences, while the second discourages prefix collapse.

2.5 Mixed-Data Curriculum GRPO

After reward-model training, mixed-data GRPO uses a prompt-type reward dispatch. The RL data combine two types of medical prompts. The first source is open-ended prompts from the UltraMedical open medical preference data (Zhang et al., 2024), where Prefix-MoE RM supplies the scalar reward. The second source is MedKgQA-style multiple-choice data, where generated responses are scored directly against an answer key. The reward used by GRPO is therefore

$$R(x,y) = \begin{cases} R_\psi(x,y), & x \in \mathcal{D}_{\text{open}}, \\ R_{\text{rule}}(x,y), & x \in \mathcal{D}_{\text{mcq}}, \end{cases} \tag{9}$$

where R_{rule} checks the extracted answer and assigns $+1$ to correct answers and -1 to incorrect or malformed answers. Mini-batches may contain both prompt types; reward assignment is per prompt before group-relative advantages are computed within each rollout group. This dispatch uses the learned RM only where open-ended scoring is needed and uses answer keys only where exact MCQ correctness is available.

We organize mixed-data RL with a cognitive-level curriculum. Let $\ell(t)$ be the maximum allowed level at training stage t. The curriculum sampler draws prompts from

$$\mathcal{D}_t = \{(x,y) : c(x) \le \ell(t)\}, \tag{10}$$

where $\ell(t)$ increases from L1 to L4. This schedule exposes the policy to factual and analytical examples before emphasizing higher-level decision and synthesis prompts. The curriculum is useful for mixed medical data because open-ended examples and MCQ answer checks differ in feedback source and difficulty. The reported runs use four fixed stages over the update budget: the first 15% samples L1 only, the next 25% samples L1–L2, the

next 30% samples L1–L3, and the final 30% samples all L1–L4 prompts. Within each stage, the sampler keeps a 60:40 open-ended-to-MCQ source ratio and samples uniformly over the active levels, with replacement when a level has fewer examples. Transitions are based on update count rather than validation performance, so the schedule is the same across seeds.

We optimize the policy with group relative policy optimization (GRPO) (Shao et al., 2024). For each prompt, the actor samples a group of responses $\{y_i\}_{i=1}^{G}$. Rewards are normalized within the group to estimate relative advantages:

$$\hat{A}_i = \frac{R(x, y_i) - \mathrm{mean}_j R(x, y_j)}{\mathrm{std}_j R(x, y_j) + \epsilon}. \quad (11)$$

Thus an MCQ prompt group and an open-ended prompt group can appear in the same optimization batch without sharing reward normalization. The advantage $\hat{A}_i$ compares responses sampled for the same prompt, which avoids directly comparing raw rule rewards and learned RM scores across prompt types. The policy is then updated with a clipped importance-ratio objective and a KL penalty against the reference model. This removes the need for a separately trained critic while retaining an RL objective that can use the learned medical reward model.

3 Experiments

3.1 Data Accounting and Evaluation Setup

Reward-model data. The reward-model training data use non-multiple-choice pairwise medical preferences with cognitive-level labels. UltraMedical-Preference samples (Zhang et al., 2024) are labeled by the LLM cognitive-level classifier prompt in Appendix C, then converted to the format $\{x, [y^+, y^-], [r^+, r^-], c\}$. The reward model is trained on a filtered UltraMedical preference subset with added cognitive-level labels; the larger prompt pool is reserved for RL.

RL data. The policy-optimization data mix open-ended prompts from the UltraMedical open medical preference data (Zhang et al., 2024) and MedKgQA multiple-choice prompts. Each example keeps its chat-formatted prompt, raw question text, reward metadata, and data-source identifier for prompt-type reward dispatch. MedKgQA examples include ground-truth answer letters for rule-based reward, while open-ended examples use the learned reward model as the training signal. Table 1 summarizes each source, its count, and where it enters the pipeline.

Table 1: Data accounting after filtering and deduplication. RM/router counts are preference pairs; RL counts are prompts.

Source	Count	Use
UltraMedical-Preference	9,742	level labeling; router/RM training
UltraMedical prompts	76,843	open-ended RL with RM reward
MedKgQA	51,229	MCQ RL with answer-key reward
Mixed RL total	128,072	curriculum RL

Evaluation benchmarks. We evaluate the policy on a text-only medical benchmark suite covering exam-style QA, biomedical literature QA, difficult scientific reasoning, expert-level clinical reasoning, and open-ended clinical communication. The closed-form suite includes MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), MMLU medical subsets (Hendrycks et al., 2020), MMLU-Pro health and biology tracks (Wang et al., 2024b), GPQA genetics and molecular biology tracks (Rein et al., 2023), and the text-only split of MedXpertQA (Zuo et al., 2025). The open-ended suite includes HealthBench (Arora et al., 2025) and LLMEval-Med (Zhang et al., 2025a). HealthBench contributes multi-turn healthcare conversations with physician-written rubrics, making it a useful complement to exam-style accuracy benchmarks. MedQA, MedMCQA, PubMedQA, MMLU, MMLU-Pro, GPQA, and MedXpertQA are scored through answer extraction and exact answer matching. HealthBench reports the percentage of physician-rubric points satisfied, while LLMEval-Med reports an external-judge score.

Baselines. The controlled comparison uses Qwen3-8B as the policy backbone and isolates the effects of supervision, reward design, reward conditioning, and curriculum scheduling. The base model is Qwen3-8B without task-specific medical alignment. The rule-only GRPO baseline uses MCQ answer-key rewards and omits learned rewards for open-ended examples. The shared-RM, hard-prompt RM, uniform-prefix, and prefix-MoE RM baselines use the same GRPO configuration but vary the reward model or conditioning signal. The final system adds cognitive-level curriculum scheduling. To evaluate scale, we repeat the full method with Qwen3-14B and Qwen3-32B. We also compare against representative medical-specialized

Table 2: Comparison with base, medical-specialized, and frontier models on representative exam-style and non-multiple-choice medical benchmarks.

Model	Type	MedQA	MedMCQA	PubMedQA	MedXpertQA	HealthBench	LLMEval-Med
Qwen3-8B	base open	61.8	56.4	66.3	14.8	23.8	3.02
Qwen3-32B	base open	78.6	68.5	77.4	24.1	31.6	3.58
HuatuoGPT-o1-8B (Chen et al., 2025b)	open medical	72.6	60.8	65.8	18.2	28.4	3.44
HuatuoGPT-o1-70B (Chen et al., 2025b)	open medical	88.1	75.4	80.9	33.5	40.2	4.12
GPT-5.5	closed general	94.8	84.2	87.9	40.8	45.3	4.30
Claude Opus 4.6	closed general	94.2	83.8	87.3	41.2	45.8	4.32
Gemini 3 Pro	closed general	**95.1**	**84.9**	**88.4**	40.5	44.9	4.27
DeepSeek-R1	reasoning model	92.4	79.6	82.8	37.8	39.5	4.12
Prefix-MoE RM + curriculum, Qwen3-8B	open	80.1	68.2	74.2	32.6	43.8	4.28
Prefix-MoE RM + curriculum, Qwen3-32B	open	94.0	82.4	87.8	**46.2**	**50.1**	**4.55**

and frontier models, including HuatuoGPT-o1, GPT-5.5, Claude Opus 4.6, Gemini 3 Pro, and DeepSeek-R1.

3.2 Main Model Comparison

Table 2 compares base open models, medical-specialized systems, closed general models, and our aligned Qwen3 policies. Rows specify the model family and type, while columns report three representative exam-style benchmarks and three non-multiple-choice medical reasoning benchmarks. Against the unaligned Qwen3-8B base, Prefix-MoE RM improves MedQA by 18.3 points, MedMCQA by 11.8 points, PubMedQA by 7.9 points, MedXpertQA by 17.8 points, HealthBench by 20.0 points, and LLMEval-Med by 1.26. The same trend holds at larger scale: relative to the Qwen3-32B base, the 32B aligned policy gains 15.4 points on MedQA, 13.9 on MedMCQA, 10.4 on PubMedQA, 22.1 on MedXpertQA, 18.5 on HealthBench, and 0.97 on LLMEval-Med. These gains are largest on non-multiple-choice metrics, but the exam-style benchmarks also improve, showing that the mixed RL data preserve answer-key performance while the reward model improves open-ended responses. The 8B aligned policy is therefore competitive with substantially larger public models on the metrics targeted by the reward model. It exceeds the Qwen3-32B base on MedXpertQA, HealthBench, and LLMEval-Med, and compared with HuatuoGPT-o1-70B it is lower on exam-style benchmarks but higher on HealthBench and LLMEval-Med. After scaling to Qwen3-32B, the policy is comparable to the strongest closed models on exam-style benchmarks and higher on MedXpertQA, HealthBench, and LLMEval-Med.

3.3 Reward-Model Design Ablation

Table 3 isolates reward-model design while keeping the policy backbone, RL prompts, 60:40 data mix, prompt-type reward dispatch, four-stage curriculum, and optimizer fixed. Rows differ only in how the open-ended reward model is conditioned, and columns report both held-out reward-model accuracy and downstream policy metrics. The hard-prompt baseline is the key contrast because it exposes level information as text, while Prefix-MoE RM learns continuous scoring contexts.

The results show that reward-model accuracy transfers to policy quality only when the conditioning mechanism is well matched to the task. Uniform prefix averaging has reasonable RM accuracy but weak downstream transfer, suggesting that simply adding prefix capacity is not enough. Prefix-MoE RM improves over hard-prompt conditioning by 5.0 MedQA points, 6.2 MedMCQA points, 9.3 PubMedQA points, 11.8 MedXpertQA points, 13.4 HealthBench points, and 0.86 LLMEval-Med, showing the benefit of learned continuous scoring contexts. It also improves over the soft expert-head MoE by 5.4 RM-accuracy points, 8.4 MedXpertQA points, 9.2 HealthBench points, and 0.56 LLMEval-Med, indicating that prefix-level fusion is more effective here than moving mixture structure only to the scalar head.

3.4 Reward-to-Policy Transfer

Held-out preference accuracy is useful but incomplete, because policy optimization can weaken, amplify, or ignore differences between reward models. The transfer analysis keeps the RL data, curriculum, answer-key reward, and optimizer fixed, and changes only the reward model used for open-ended rollouts. This probes whether a better level-conditioned reward model produces better downstream medical responses.

Figure 2 tests whether better preference modeling predicts better policy behavior under the same RL pipeline. Across the reward-model variants,

Table 3: Policy-level RM swap with the full RL pipeline fixed. All rows use the same Qwen3-8B actor, RL data, RM training data, answer-key MCQ reward, and curriculum.

Reward model	RM Acc	MedQA	MedMCQA	PubMedQA	MedXpertQA	HealthBench	LLMEval-Med
Shared RM	79.8	72.8±0.6	58.6±0.7	61.0±0.7	15.9±0.8	24.8±0.9	3.05±0.07
Hard-prompt RM	83.9	75.1±0.5	62.0±0.6	64.9±0.6	20.8±0.7	30.4±0.8	3.42±0.06
Uniform prefix average	84.7	73.6±0.6	59.8±0.7	61.8±0.7	16.8±0.8	25.9±0.9	3.11±0.07
Soft expert-head MoE	87.6	76.9±0.5	63.6±0.6	67.0±0.6	24.2±0.7	34.6±0.7	3.72±0.06
Prefix-MoE RM	**93.0**	**80.1±0.3**	**68.2±0.4**	**74.2±0.4**	**32.6±0.4**	**43.8±0.5**	**4.28±0.03**

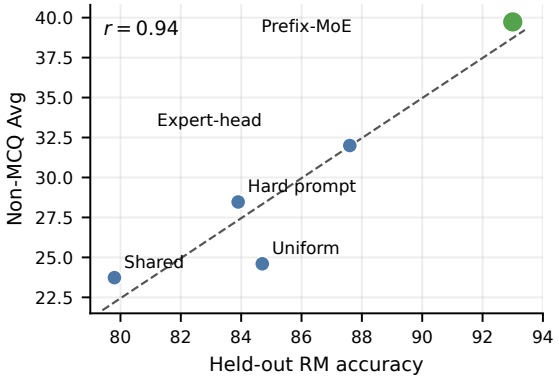

Figure 2: Transfer from held-out reward-model accuracy to downstream Non-MCQ Avg across the fixed-pipeline reward-model variants in Table 3. Non-MCQ Avg averages MedXpertQA, HealthBench, and 10× LLMEval-Med.

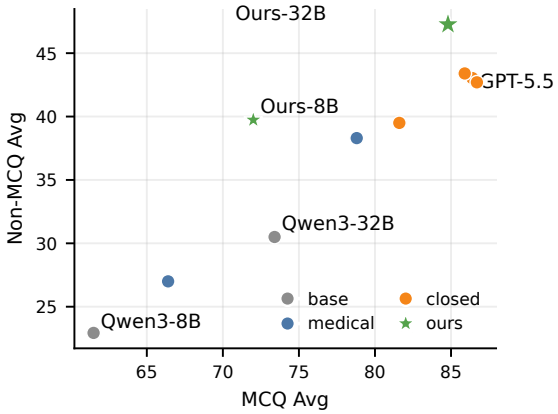

Figure 3: Model-level benchmark trade-off. MCQ Avg averages the answer-key MCQ suite, and Non-MCQ Avg averages MedXpertQA, HealthBench, and 10× LLMEval-Med.

held-out RM accuracy has a strong positive association with downstream Non-MCQ Avg ($r = 0.94$), and Prefix-MoE RM sits at the high-accuracy, high-policy-quality end of the curve. The uniform-prefix point remains below the trend, matching Table 3: extra prefix capacity without router-weighted fusion does not transfer reliably. The plot therefore supports two claims at once: reward-model quality matters for policy behavior, and the proposed soft prefix fusion is a better way to use that capacity than uniform prefix averaging.

3.5 Training-Signal Ablation

Table 4 measures how the training signals compose under the Qwen3-8B backbone. Rows move from the unaligned model through MCQ-only, mixed-data, no-curriculum, and full-curriculum variants; columns match Table 2.

MCQ-only RL strongly improves MedQA, MedMCQA, and PubMedQA but leaves open-ended metrics close to the base model, confirming that answer-key reward alone does not train the behavior targeted by the reward model. Adding open-ended preference feedback through a shared RM improves HealthBench but still trails level-aware conditioning by 16.8 HealthBench points, while uniform prefixes perform worst among the mixed-data variants because averaging all level prefixes weakens the conditioning signal. Prefix-MoE RM without curriculum already gives strong non-MCQ performance, and the curriculum adds a further 2.1 MedQA points, 3.2 MedMCQA points, 7.1 MedXpertQA points, 7.3 HealthBench points, and 0.50 LLMEval-Med. The full system is therefore not just a result of adding medical RL data; it depends on the combination of prompt-type rewards, prefix-conditioned scoring, and staged difficulty.

3.6 Benchmark Trade-off Analysis

The benchmark trade-off analysis asks whether improvements on open-ended medical reasoning are obtained by sacrificing answer-key MCQ performance. We use aggregate axes only for this visualization: MCQ Avg summarizes the answer-key MCQ suite, and Non-MCQ Avg summarizes MedXpertQA, HealthBench, and 10× LLMEval-Med. In this plot, a vertical gain means better open-ended medical reasoning at similar MCQ accuracy,

Table 4: Controlled Qwen3-8B ablation over three seeds. All non-base rows use GRPO; "Mixed" denotes the 60:40 open-ended/MCQ RL data, and rows without "curriculum" use non-curriculum sampling.

Method	MedQA	MedMCQA	PubMedQA	MedXpertQA	HealthBench	LLMEval-Med
Qwen3-8B base	61.8±0.2	56.4±0.3	66.3±0.3	14.8±0.3	23.8±0.4	3.02±0.03
MCQ-only rule reward	78.6±0.5	67.0±0.5	74.0±0.5	15.7±0.8	23.4±0.9	2.98±0.07
Mixed + shared RM	74.8±0.6	60.5±0.7	66.8±0.7	18.2±0.8	27.0±0.9	3.20±0.07
Mixed + hard-prompt RM	76.0±0.5	62.4±0.6	69.3±0.6	22.0±0.7	31.6±0.8	3.47±0.06
Mixed + uniform prefixes	73.5±0.6	58.7±0.7	64.0±0.7	16.5±0.8	25.0±0.9	3.08±0.07
Prefix-MoE RM w/o curriculum	78.0±0.4	65.0±0.5	71.0±0.5	25.5±0.7	36.5±0.7	3.78±0.05
Prefix-MoE RM + curriculum	**80.1±0.3**	**68.2±0.4**	**74.2±0.4**	**32.6±0.4**	**43.8±0.5**	**4.28±0.03**

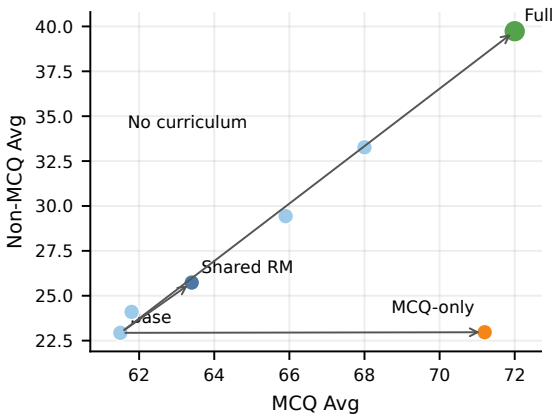

Figure 4: Training-signal benchmark trade-off under the Qwen3-8B backbone.

while a top-right movement means the system improves both kinds of evaluation.

Figure 3 shows that the 8B policy moves far upward on the Non-MCQ axis while retaining competitive MCQ accuracy, and the 32B policy further improves both axes. This pattern supports the main claim that the reward model primarily improves open-ended reasoning while mixed MCQ feedback preserves exam-style performance. The comparison is also useful for interpreting scale: the 8B model reaches the open-ended region of much larger open and closed models, while the 32B model becomes the strongest point on the Non-MCQ axis. Thus, the gains in Table 2 are not only separate benchmark improvements; they move the model into a different trade-off region.

We next analyze the same trade-off within the controlled Qwen3-8B ablations to separate the roles of answer-key reward, learned open-ended reward, and the full mixed pipeline. This complements Table 4 by showing each training signal as a direction in metric space.

Figure 4 shows why both reward sources are needed: MCQ-only training mainly moves right, shared-RM training mainly moves upward, and the full system reaches the best balance. The full system is not just an average of those two directions; it reaches a better top-right point because prefix-conditioned reward improves the open-ended signal and the curriculum further raises Non-MCQ performance without erasing the MCQ gain. This makes the trade-off plot a compact check on the intended division of labor between answer-key rewards and the learned medical reward model.

4 Conclusion

We presented Prefix-MoE RM, a cognitive-level prefix-MoE reward model for medical alignment with both open-ended preferences and MCQ answer-key feedback. The core idea is to keep a single scalar reward while conditioning its scoring context on soft routing over recall, analysis, decision, and synthesis levels. Instead of forcing all answers through one scoring rule, the router softly selects among level-specific contexts, and trainable prefixes make these contexts more flexible than fixed prompts. The same reward model guides mixed-data GRPO, where open-ended prompts receive learned rewards and MCQ prompts keep answer-key feedback. Across fixed-pipeline RM swaps and policy ablations, prefix conditioning improves reward-to-policy transfer, especially on non-multiple-choice benchmarks, while answer-key feedback preserves exam-style performance. Scaling to Qwen3-32B further suggests that the recipe transfers within the same model family. Together, these results support level-aware reward conditioning as a practical way to align medical reasoning without abandoning scalar RL rewards. Future work should add broader expert evaluation and test whether learned prefixes capture interpretable clinical reasoning differences.

5 Limitations

This work has five limitations. First, cognitive-level labels are generated by an LLM classifier rather than manual annotation, so label noise can affect both reward scoring and curriculum order. Second, the mixed-data RL pipeline uses answer-key rule rewards for MCQ prompts, so answer-extraction errors may reward answer-format behavior rather than medical reasoning quality. Third, the targeted human evaluation is intentionally small and does not replace live clinician-patient interaction studies, long-term safety analysis, or uncertainty estimates. Extending the evaluation to expert human review and controlled multi-turn clinical dialogue is therefore important before deployment-oriented claims can be made. Fourth, our taxonomy controls show that cognitive levels outperform several alternatives, but they do not prove that this is the best grouping for all medical alignment data. Finally, comparisons to proprietary systems may involve different access conditions or reported-number provenance when direct access is unavailable; they should be interpreted as contextual rather than strict head-to-head claims.

References

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. 2025. Medical large language model benchmarks should prioritize construct validity. *arXiv preprint arXiv:2503.10694*.

Lorin W. Anderson and David R. Krathwohl. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quinonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, and 1 others. 2025. MedHELM: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*.

Yoshua Bengio, Jerome Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48.

Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. 2025a. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *The Thirteenth International Conference on Learning Representations*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2025b. HuatuoGPT-o1, towards medical complex reasoning with LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Auke Elfrink, Iacopo Vagliano, Ameen Abu-Hanna, and Iacer Calixto. 2023. Soft-prompt tuning to predict lung cancer using primary care free-text dutch medical notes. In *International Conference on Artificial Intelligence in Medicine*, pages 193–198. Springer.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. 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*, pages 2567–2577.

Yongnan Jin, Xurui Li, Feng Cao, Liucun Gao, and Juanjuan Yao. 2025. Multidimensional rubric-oriented reward model learning via geometric projection reference constraints. *arXiv preprint arXiv:2511.16139*.

Yuhang Lai, Siyuan Wang, Shujun Liu, Xuan-Jing Huang, and Zhongyu Wei. 2024. ALARM: Align language models via hierarchical rewards modeling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7817–7831.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yusheng Liao, Shuyang Jiang, Yu Wang, and Yanfeng Wang. 2024. MING-MOE: Enhancing medical multi-task learning in large language models with sparse mixture of low-rank adapter experts. *arXiv preprint arXiv:2404.09027*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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

Cheng Peng, Xi Yang, Kaleb E. Smith, Zehao Yu, Aokun Chen, Jiang Bian, and Yonghui Wu. 2024. Model tuning or prompt tuning? a study of large language models for clinical concept and relation extraction. *Journal of Biomedical Informatics*, 153:104630.

Shanghaoran Quan. 2024. DMoERM: Recipes of mixture-of-experts for effective reward modeling. *arXiv preprint arXiv:2403.01197*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, and 1 others. 2024. Capabilities of Gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024a. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592.

Yuan Wang, Shujian Gao, Jiaxiang Liu, Songtao Jiang, Haoxiang Xia, Xiaotian Zhang, Zhaolu Kang, Yemin Wang, and Zuozhu Liu. 2026. Beyond n-grams: A hierarchical reward learning framework for clinically-aware medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33719–33727.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024b. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*.

Zihao Wang, Chirag Nagpal, Jonathan Berant, Jacob Eisenstein, Alexander Nicholas D'Amour, Sanmi Koyejo, and Victor Veitch. 2024c. Transforming and combining rewards for aligning large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 51161–51176. PMLR.

Lingyan Wu, Xiang Zheng, Weiqi Zhai, Wei Wang, Xuan Ren, Zifan Zhang, Hu Wei, and Bing Zhao. 2026. MedPRMBench: A fine-grained benchmark for process reward models in medical reasoning. *arXiv preprint arXiv:2604.17282*.

Ziqin Yuan, Ruiqi Wang, Dezhong Zhao, Baijian Yang, and Byung-Cheol Min. 2026. PrefMoE: Robust preference modeling with mixture-of-experts reward learning. *arXiv preprint arXiv:2605.00384*.

Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Yanjun Shao, Yonghoe Koo, Minhyeok Ko, Qingyu Chen, Mark Gerstein, Michael Moor, and Jaewoo Kang. 2025. Med-PRM: Medical reasoning models with stepwise, guideline-verified process rewards. *arXiv preprint arXiv:2506.11474*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, and 1 others. 2024. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081.

Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, and 1 others. 2025a. LLMEval-Med: A real-world clinical benchmark for medical LLMs with physician validation. *arXiv preprint arXiv:2506.04078*.

Xiaotian Zhang, Yuan Wang, Zhaopeng Feng, Ruizhe Chen, Zhijie Zhou, Yan Zhang, Hongxia Xu, Jian Wu, and Zuozhu Liu. 2025b. Med-U1: Incentivizing unified medical reasoning in LLMs via large-scale reinforcement learning. *arXiv preprint arXiv:2506.12307*.

Yiyang Zhao, Huiyu Bai, and Xuejiao Zhao. 2026. GEM: Generative entropy-guided preference modeling for few-shot alignment of LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. 2024. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10586–10613.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Implementation Details

The main experiments use Qwen3-8B as actor, router, and reward backbone, and repeat the full recipe with Qwen3-14B/32B actors to test scaling (Qwen Team, 2025). The router is trained for three epochs with batch size 8, learning rate 10^{-5}, bfloat16 precision, and a 9:1 train-validation split. The reward model is trained for three epochs with batch size 1 per device, learning rate 5×10^{-5}, cosine scheduling, gradient clipping at 1.0, bfloat16 precision, and best-checkpoint selection by pairwise accuracy. GRPO uses four rollouts per prompt, actor learning rate 5×10^{-7}, KL coefficient 0.02, prompt length 512, response length 2048, and eight GPUs for the 8B runs; the 32B scaling runs use tensor parallelism with the same optimization hyperparameters. For MCQ evaluation and rule rewards, we first extract a final option with a regex over final-answer markers and option letters; if no marker is found, we fall back to the last standalone option letter and mark the response malformed when no valid option can be extracted. Unless otherwise stated, controlled ablations report the mean and standard deviation over three random seeds.

B Related Work

Medical LLMs and medical reasoning alignment. Medical LLM research has moved from exam-oriented QA toward broader clinical reasoning and communication: Med-Gemini expands medical capabilities with multimodal and long-context modeling (Saab et al., 2024), and open systems such as UltraMedical and HuatuoGPT-o1 use domain data and RL to improve instruction following and complex reasoning (Zhang et al., 2024; Chen et al., 2025b). Recent medical RL work such as Med-U1 explores unified reward design across MCQ, computation, and generation tasks (Zhang et al., 2025b). Hierarchical medical reward learning has also been used for clinically aware report generation, where rewards progress from fluency to factual grounding and diagnostic consistency (Wang et al., 2026). Our work is complementary: rather than optimizing report generation or making exam-style tasks the main objective, we train a cognitive-level reward model on open-ended medical preferences and use MCQ data only as added RL feedback.

Reward modeling, preference optimization, and conditioning. RLHF trains a reward model from preferences and optimizes a policy with online RL (Ouyang et al., 2022; Schulman et al., 2017), while DPO avoids explicit RL through pairwise preference likelihoods (Rafailov et al., 2023); reasoning systems increasingly combine group-relative objectives with verifiable rewards (Shao et al., 2024; Guo et al., 2025). At the reward-model level, RewardBench highlights the limits of using one all-purpose reward score (Lambert et al., 2024), and recent work decomposes rewards through hierarchical objectives, pluralistic or personalized preference models, multi-objective preference optimization, reward transformation and combination, MoE reward heads, interpretable routing, and medical process verification (Lai et al., 2024; Chen et al., 2025a; Zhou et al., 2024; Wang et al., 2024c; Quan, 2024; Wang et al., 2024a; Yun et al., 2025; Wu et al., 2026). PrefMoE is especially related because it uses soft routing over reward experts for mixed preference data (Yuan et al., 2026); medical rubric-oriented reward learning instead decomposes clinical quality into multidimensional rubrics and geometric constraints (Jin et al., 2025), while GEM studies entropy-guided preference modeling for few-shot alignment in specialized domains (Zhao et al., 2026). Our design conditions a language-model reward backbone through fused cognitive-level prefixes, so the routed signal changes token representations before a shared scalar head rather than learning separate expert heads, rubric dimensions, or user-personalized reward prototypes. Medical MoE adaptation methods such as MING-MOE route inputs to sparse low-rank adapter experts for multi-task medical generation (Liao et al., 2024); in contrast, our experts are not generator adapters but reward-side continuous prefixes tied to an ordered reasoning taxonomy. Prefix tuning and prompt tuning provide lightweight continuous conditioning mechanisms (Li and Liang, 2021; Liu et al., 2022), with clinical NLP evidence that soft prompts can help under data scarcity (Peng et al., 2024; Elfrink et al., 2023). Prefix-MoE RM uses this idea to replace fixed hard prompts with trainable reward-side scoring contexts conditioned on ordered medical cognitive levels.

Medical evaluation and curricula. MedQA, MedMCQA, PubMedQA, and MMLU medical subsets remain standard exam-style benchmarks (Jin et al., 2021; Pal et al., 2022; Jin et al., 2019; Hendrycks et al., 2020), while MMLU-Pro and GPQA add harder scientific reasoning stress tests

(Wang et al., 2024b; Rein et al., 2023). Newer benchmarks emphasize expert-level and open-ended evaluation: MedXpertQA targets difficult clinical reasoning (Zuo et al., 2025), HealthBench evaluates physician-rubric healthcare conversations (Arora et al., 2025), LLMEval-Med uses physician-validated real-world scenarios (Zhang et al., 2025a), and MedHELM argues for broader task-taxonomy coverage (Bedi et al., 2025). Construct-validity arguments further caution that medical benchmarks should measure explicit clinical constructs rather than only leaderboard aggregates (Alaa et al., 2025), while MedPRMBench shows how process-level medical reward models can be evaluated with fine-grained reasoning errors (Wu et al., 2026). Curriculum learning and Bloom-style taxonomies motivate ordering examples by difficulty and reasoning demand (Bengio et al., 2009; Anderson and Krathwohl, 2001); we use this structure as one practical way to condition the non-multiple-choice reward model, while treating MCQ performance as a check that exam-style accuracy is preserved.

C LLM Cognitive-Level Labeling Prompt

Before router training, we label UltraMedical-Preference queries with an LLM classifier at temperature 0. The classifier sees only the question stem and options when present, and mixed cases are assigned to the highest required level. The labeling prompt is:

> **Prompt.** You are a cognitive-level classifier for medical questions. Based only on the question stem and options, without introducing external assumptions, classify the question into one of four levels. This framework is grounded in Anderson and Krathwohl's (2001) revision of Bloom's Taxonomy, adapted for the medical domain.
>
> **L1 – Medical Knowledge Recall (Remember / Understand):** direct retrieval of factual medical knowledge, definitions, mechanisms, classifications, or anatomical structures; single-concept understanding.
>
> **L2 – Clinical Analysis and Application (Apply / Analyze):** application of medical knowledge to clinical scenarios; integration of symptoms, labs, or imaging for diagnosis or short-chain treatment selection; no full management plan.
>
> **L3 – Clinical Reasoning and Decision-Making (Evaluate):** multi-step clinical reasoning with explicit planning or decision-making; treatment planning, therapeutic-option evaluation, comorbidity management, or decisions under uncertainty.
>
> **L4 – Biomedical Synthesis and Innovation (Create):** cross-domain synthesis, research protocol design, literature review, critical evaluation, ethics, policy, public health analysis, novel hypotheses, or comprehensive reports.
>
> **Decision rules:** single fact or concept suffices → L1; clinical data integration for diagnosis or short-chain application → L2; multi-step treatment planning or patient management → L3; cross-domain synthesis, research design, ethical analysis, or creative generation → L4. For mixed cases, assign the highest required level: L4 > L3 > L2 > L1.
>
> **Output format:** first line outputs only the label in a LaTeX box, `\boxed{L1/L2/L3/L4}`. Second line is optional and gives a short English reason of at most 25 words.
>
> **Medical question:** `{question_text}`

D Additional Diagnostics

Table 5: Curriculum and reward-source diagnostics. Non-MCQ Avg uses the same definition as Figure 3; Parse is the percentage of MCQ outputs with an extractable final option.

Probe	Variant	MedQA	MedMCQA	PubMedQA	Non-MCQ Avg	Parse %
Schedule	Random mix	78.0±0.4	65.0±0.5	71.0±0.5	33.3±0.5	98.6
Schedule	Difficulty-balanced	78.4±0.4	65.7±0.4	72.0±0.4	34.5±0.5	98.7
Schedule	Reverse curriculum	76.8±0.5	63.9±0.5	69.6±0.5	31.4±0.6	98.3
Schedule	Four-stage curriculum	**80.1±0.3**	**68.2±0.4**	**74.2±0.4**	**39.7±0.4**	98.8
Reward	MCQ rule only	78.6±0.5	67.0±0.5	74.0±0.5	23.0±0.6	98.7
Reward	Non-MCQ RM only	72.4±0.4	61.2±0.5	68.5±0.5	34.9±0.4	94.2
Reward	Prompt-type dispatch	**80.1±0.3**	**68.2±0.4**	**74.2±0.4**	**39.7±0.4**	98.8
Reward	Format-only MCQ rule	74.0±0.5	61.5±0.5	68.0±0.6	25.7±0.7	99.5

Table 5 tests whether the gains require both the staged data schedule and prompt-type-specific rewards. This diagnostic is useful because mixed medical RL can improve MCQ accuracy while silently degrading open-ended responses, so we vary one design factor at a time under the same Qwen3-8B pipeline. The schedule rows compare random mixing, difficulty balancing, reverse ordering, and the proposed four-stage curriculum; the reward rows compare MCQ-only, Non-MCQ-only, dispatch, and a format-only MCQ control. The four-stage curriculum gives the strongest Non-MCQ Avg, and prompt-type dispatch preserves MCQ performance while avoiding the large open-ended drop of MCQ-only reward. The format-only control keeps parse rate high but performs worse, indicating that the MCQ gains are not explained by answer-format compliance alone. Together, the table supports treating curriculum order and reward dispatch as separate contributors rather than attributing the full result to mixed-data RL.

Table 6 tests whether the routed cognitive signal is actually used by the policy rather than serving as a cosmetic label. We rerun the full Qwen3-8B pipeline after changing the routing rule: hard argmax routing removes uncertainty, larger temperatures smooth the distribution, and random flips

Table 6: Router sensitivity under the full Qwen3-8B pipeline. Non-MCQ Avg uses the same definition as Figure 3.

Routing setting	MedQA	MedMCQA	PubMedQA	Non-MCQ Avg
Soft routing, $\tau = 0.1$	80.1±0.3	68.2±0.4	74.2±0.4	39.7±0.4
Hard argmax routing	78.8±0.4	67.0±0.4	73.0±0.5	37.8±0.5
Soft routing, $\tau = 0.5$	79.4±0.4	67.6±0.4	73.8±0.4	38.5±0.5
Soft routing, $\tau = 1.0$	78.5±0.4	66.8±0.4	72.7±0.5	37.1±0.5
10% random route flips	78.1±0.5	66.2±0.5	72.3±0.5	36.8±0.6
20% random route flips	77.2±0.5	65.4±0.5	71.4±0.6	35.0±0.7

inject controlled routing noise. Every perturbation lowers Non-MCQ Avg relative to soft routing with $\tau = 0.1$, with the largest degradation under 20% random route flips. This pattern indicates that downstream performance depends on preserving informative level probabilities, not merely on attaching any level tag to the reward model. The hard-routing drop also suggests that uncertainty over neighboring cognitive levels is useful, especially when prompts combine factual, analytic, and decision-making cues.

Table 7: Diversity-regularizer sweep. Prefix cosine is the mean pairwise cosine similarity between level prefixes; Non-MCQ Avg uses the same definition as Figure 3.

λ_{div}	RM Acc	Prefix cosine	Non-MCQ Avg
0	91.8±0.3	0.61	38.1±0.5
0.03	92.5±0.2	0.48	38.8±0.4
0.10	**93.0±0.2**	0.32	**39.7±0.4**
0.30	92.4±0.3	0.21	38.3±0.5

Table 7 examines whether prefix diversity improves specialization or simply over-regularizes the reward model. We sweep λ_{div} and track three linked quantities: reward-model accuracy, mean pairwise prefix cosine, and downstream Non-MCQ Avg after RL. Without the regularizer, prefixes remain highly similar and downstream performance is lower, suggesting insufficient specialization across cognitive levels. A moderate value, $\lambda_{\text{div}} = 0.10$ gives the best RM accuracy and Non-MCQ Avg while reducing prefix cosine substantially. Increasing the weight to 0.30 further separates prefixes but hurts downstream performance, so diversity is useful as a calibration term rather than an objective to maximize by itself.

Tables 8 and 9 check two failure modes of the frozen router before it is used inside the reward model. Table 8 asks whether the trained router reproduces the LLM labeling protocol on held-out labeled questions, rather than drifting toward a different taxonomy. Its errors are concentrated between adjacent levels, especially L3 and L4, which is the expected ambiguity boundary between clinical decision-making and broader synthesis. Table 9 asks whether routing based only on the question is close to routing based on the full query-response pair, since the policy-time router observes the prompt rather than a candidate answer. The 87.6% same-argmax rate and low Jensen-Shannon divergence indicate that question-only routing is usually stable, while the non-adjacent shift rate remains small. These results justify using the frozen router as an operational approximation of the labeling protocol, but they also delimit the claim: the router is a reproducible training component, not a manually validated clinical taxonomy.

Table 8: Router validation confusion matrix on a held-out LLM-labeled split.

Label \ pred.	L1	L2	L3	L4	Acc.
L1 recall	270	24	5	1	90.0
L2 analysis	22	250	24	4	83.3
L3 decision	4	28	251	17	83.7
L4 synthesis	1	7	48	244	81.3

Table 9: Router mode-consistency check on held-out query-response pairs.

Question-only vs. query-response routing	Rate
Same argmax level	87.6
Adjacent-level shift	10.1
Non-adjacent shift	2.3
Mean Jensen-Shannon divergence	0.074
Pair input routes higher than question-only	6.5
Pair input routes lower than question-only	5.9

E Qualitative Error Analysis

Table 10: Targeted human evaluation on 200 held-out L3/L4 prompts. Clinical adequacy is a 1–5 rubric score; safety flags and full-system win/tie/loss are percentages.

Policy	Adequacy	Safety flag %	Full win/tie/loss
Shared-RM fixed pipeline	3.71±0.08	9.8	58.5 / 21.0 / 20.5
Prefix-MoE RM w/o curriculum	3.89±0.07	7.2	51.5 / 28.0 / 20.5
Full Prefix-MoE RM	**4.12±0.06**	**4.6**	–

Table 10 tests whether the automatic Non-MCQ improvements correspond to clinically meaningful differences on the hardest prompt types. This check is needed because open-ended medical benchmarks can reward surface fluency, while L3/L4 use cases require adequate reasoning, appropriate caution,

Table 11: Representative L3/L4 qualitative cases from the targeted human evaluation. The table summarizes behavior rather than reproducing full model outputs.

Case type	Prompt summary	Improvement in full Prefix-MoE RM	Remaining issue
L3 decision	Chest pain after phosphodiesterase inhibitor use, with diaphoresis and risk factors	Prioritizes emergency evaluation, ECG/troponin testing, and nitrate avoidance rather than giving generic reflux advice	Does not quantify pre-test probability or local protocol variation
L3 decision	Fever and spreading erythema around a diabetic foot ulcer	Flags possible sepsis and need for urgent in-person assessment; distinguishes outpatient wound care from admission-level risk	Antibiotic discussion remains broad without culture or renal-function details
L4 synthesis	Critique an observational study claiming a drug prevents dementia	Identifies confounding by indication, immortal-time bias, and need for sensitivity analyses before causal claims	Misses a possible negative-control analysis in one output
L4 synthesis failure	Compare antiseizure medication options during pregnancy with incomplete patient history	More cautious than baselines and recommends specialist coordination, but the answer becomes verbose and incomplete on dose adjustment	Still needs stronger handling of missing constraints and shared decision making

Table 12: RM-only ablation on held-out pairwise preference data over three seeds. Overall pairwise accuracy and level-wise pairwise accuracy are reported; higher is better.

Reward model	Overall	L1 recall	L2 analysis	L3 decision	L4 synthesis
Shared RM	79.8 ± 0.5	82.4 ± 0.5	80.5 ± 0.5	77.1 ± 0.7	74.6 ± 0.8
Hard-prompt RM	83.9 ± 0.4	85.8 ± 0.4	84.2 ± 0.5	82.0 ± 0.6	78.9 ± 0.7
Uniform prefix average	84.7 ± 0.4	86.7 ± 0.4	85.3 ± 0.5	82.8 ± 0.6	80.1 ± 0.7
Router probabilities as features	85.8 ± 0.4	87.6 ± 0.4	86.1 ± 0.5	83.7 ± 0.6	81.0 ± 0.7
Embedding-cluster prefixes	86.6 ± 0.4	88.4 ± 0.4	86.9 ± 0.5	84.6 ± 0.6	82.2 ± 0.6
Body-system prefixes	86.9 ± 0.4	88.8 ± 0.4	87.2 ± 0.5	85.0 ± 0.6	82.5 ± 0.6
Soft expert-head MoE	87.6 ± 0.3	89.1 ± 0.4	87.8 ± 0.4	85.8 ± 0.5	83.7 ± 0.6
Prefix-MoE RM	$\mathbf{93.0\pm0.2}$	$\mathbf{94.2\pm0.3}$	$\mathbf{93.3\pm0.3}$	$\mathbf{92.1\pm0.3}$	$\mathbf{90.3\pm0.4}$

and avoidance of unsafe advice. Human evaluators score 200 held-out L3/L4 prompts for clinical adequacy and safety, and compare full Prefix-MoE RM against the shared-RM and no-curriculum variants. The full system receives the highest adequacy score and the lowest safety-flag rate, and it wins more often than it loses against the shared-RM fixed pipeline. The evaluation is targeted rather than deployment-level, but it supports the interpretation that the main benchmark gains reflect better clinical response quality rather than only metric-specific improvements.

Table 11 illustrates what changes qualitatively in the same L3/L4 setting. We include this table because aggregate adequacy and safety scores do not show which clinical behaviors improved or which errors remain. The cases are selected from the targeted human evaluation and summarized rather than reproduced verbatim, so the table focuses on error categories and response behavior. The full system more consistently recognizes urgent evaluation needs, critiques causal claims, and remains cautious when patient information is incomplete. At the same time, the remaining issues show that the model still needs better calibration, protocol specificity, and shared-decision framing. These examples therefore explain the human-evaluation trend while keeping the limitation visible: cognitive routing improves difficult responses but does not remove the need for clinical oversight.

Table 12 isolates the reward-modeling contribution before policy optimization. The purpose is to distinguish cognitive-level prefix fusion from simpler explanations such as extra parameters, prompt conditioning, embedding clusters, or body-system specialization. We evaluate all variants on the same held-out preference pairs over three seeds and report both overall and level-wise pairwise accuracy. Prefix-MoE RM reaches 93.0 overall accuracy and has its largest margins on L3 decision and L4 synthesis, where preference judgments require more than factual correctness. The capacity-matched alternatives improve over a shared RM but remain well below the fused cognitive-level prefixes, showing that the gain is not explained by conditioning capacity alone. This level-wise pattern supports the paper's central claim that cognitive routing is most valuable where a single shared scoring rule is least adequate.