DP-HHEM: A Fine-Tuning and Alignment Framework Based on a Dual Pool Heterogeneous Hybrid Expert Model

Anonymous ACL submission

Abstract

The hybrid expert model shows significant potential in the field of auxiliary diagnosis and treatment, but in the application scenario of multi department consultation in Internet hospitals, the existing model lacks the ability to take into account the breadth of general knowledge and the depth of specialized knowledge, and it is difficult to transform their comprehensive advantages into effective diagnosis and treatment support. To solve the above problems, this paper proposes a fine tuning and alignment framework based on Dual Pool Heterogeneous Hybrid Expert Model (DP-HHEM). The experimental results on MedDialog-X and KaMed-X show that the performance of the proposed method is significantly better than the baseline model.Extensive experiments on the MedDialog-X and KaMed-X datasets demonstrate that DP-HHEM outperforms the strongest baseline, achieving improvements of 13.79% and 9.21% on the E-F metric, respectively.

1 Introduction

In real clinical practice, patients often have multiple conditions at oncea diabetic might also have heart disease, or a hypertensive patient could have kidney problems. These complex cases usually require specialists from different departments to collaborate. Internet hospitals make this easier by letting doctors across regions weigh in on the same case. But this puts pressure on AI systems: they need both broad general knowledge and deep specialty expertise to be useful in these multi-disciplinary settings. Most current models struggle to deliver both.

While instruction fine-tuning and alignment enable Large Language Models(LLMs) to follow complex instructions (Li et al., 2025a), building reliable online medical dialogue systems remains challenging. Simulating professional doctor-patient interactions (Zhi et al., 2025) requires balancing fluent general conversation with strict adherence to evidence-based medical standardsa core difficulty in open medical alignment frameworks (Munnangi and Savage, 2026).Although dense models perform well on static medical knowledge benchmarks, they exhibit fatal flaws in dynamic, real-world scenarios (Zhi et al., 2025; John Snow Labs, 2026). Generated replies often prioritize "probabilistic rationality" over "clinical accuracy," leading to verbose, unfocused content and severe medical hallucinations (John Snow Labs, 2026; MedCity News, 2026). Additionally, deploying large-parameter models for high-concurrency online requests incurs unacceptable latency and memory pressure. To expand model capacity while controlling costs, the Mixture-of-Experts (MoE) architecture offers a promising solution via sparse activation, enabling sub-linear parameter scaling with constant inference cost (IBM, 2026; NVIDIA, 2026). However, migrating homogeneous MoE to vertical healthcare faces three bottlenecks: undifferentiated routing fails to balance general language ability with specialized medical logic; continual fine-tuning triggers catastrophic forgetting of base capabilities (IBM, 2026); and traditional alignment relies on Proximal Policy Optimization (PPO), which requires loading a value model alongside policy and reference models, causing memory overflow in MoE scenarios. Moreover, reward functions dependent on external LLM judges are prone to "reward hacking," encouraging lengthy but clinically hollow responses (Wang et al., 2025b; Mullappilly et al., 2026).

To address these issues, we propose a novel domain-adaptive heterogeneous hybrid expert architecture and alignment paradigm (Figure 1). The core innovation of our approach is the Dual Pool Heterogeneous Expert Routing Mechanism, which decouples the expert networks into two physically isolated yet cooperative pools:

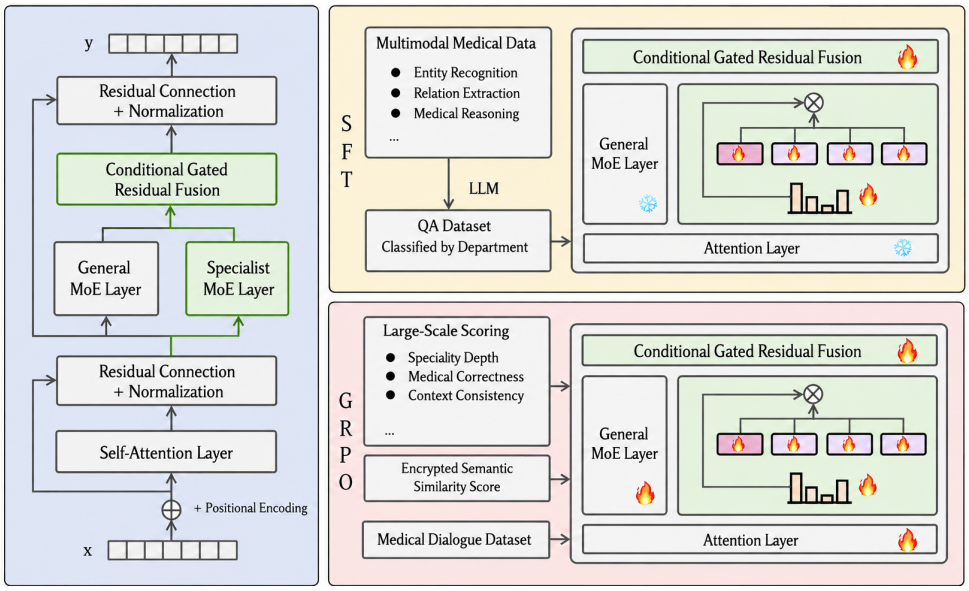

Figure 1: The Overall Architecture Of MoE Medical Dialogue System

- **General Expert Pool**: Captures broad medical knowledge and conversational context. Each token in the input is routed dynamically to the top-k general experts, ensuring that the model maintains fluent general language ability and can process everyday patient input effectively.

- **Department-Specialized Expert Pool**: Focuses on vertical domain knowledge. Unlike the general pool, this pool uses low-frequency, semantic-unit-level routing to maintain consistency within a conversation turn. It guarantees that professional medical reasoning and specialist knowledge are applied cohesively, preventing dilution of expertise.

The two expert pools are fused using a conditional gated residual mechanism, which dynamically integrates general semantic features with specialized medical knowledge. The gating coefficient adapts based on the context: high for specialist-heavy interactions and low for general conversational content. This design ensures that the model preserves general abilities while selectively amplifying specialized knowledge.

Building on this foundation, we further introduce two supporting contributions:

- **Anti-forgetting strategy based on parameter expansion and absolute freezing**: Freezes all pre-trained base parameters while only training newly added department experts, achieving immunity to catastrophic forgetting.

- **Memory-efficient GRPO and dual-track hybrid reward function**: Replaces PPO with a value-model-free Group Relative Policy Optimization(GRPO) algorithm and a hybrid reward system combining semantic similarity with structured LLM referees, effectively suppressing reward hacking and reducing training cost.

Moreover, considering the inherent mismatch between traditional classification or sequence labeling tasks and the generative paradigm of LLMs, we construct a training dataset by reorganizing existing public medical corpora into natural language QA pairs. Separately, we annotate two additional public datasets with Chain-of-Thought(CoT) for reinforcement learning and evaluation.

2 Related Work

The LLM shows excellent potential in simulating clinical dialogues and answering medical questions, but its implementation in the actual telemedicine scene still faces severe challenges. Existing studies have shown that although the general dense model performs well in the static medical benchmark test, its essence is a statistical language model based on probability, which lacks real clinical judgment. As a result, the output is often unable to directly drive high-risk clinical decisions (John Snow Labs, 2026). In addition, real medical practice not only requires the accuracy of facts, but also involves complex contextual reasoning and patient emotion perception (Munnangi and Savage, 2026). Recently

released ThreadMed-QA and other rounds of dialogue evaluation benchmarks further revealed that the large model is prone to context consistency collapse and error accumulation when confronted with continuous questioning from real patients (Cao et al., 2026). The "8% problem" raised by intelligence factory also pointedly points out that even the most advanced large-scale model still shows significant limitations in high-risk precision medical tasks (IBM, 2026; NVIDIA, 2026). This shows that it is difficult to bridge the huge gap between the general model and rigorous clinical practice by relying solely on the cue word project or a simple single round of fine-tuning (MedCity News, 2026).

In order to control the computational overhead while expanding the model parameter capacity, hybrid expert architecture is widely used in various LLMs. Through the sparse activation mechanism, this architecture only calls part of the "expert" network to process the input, thus realizing the sub linear expansion of the model capability at a constant computational cost (Sanseviero et al., 2023). However, when the pre trained MoE model is adapted to a specific vertical medical field, the traditional full parameter fine-tuning can easily lead to "catastrophic forgetting", that is, the model quickly loses its original common language ability when learning the knowledge of a new Department.

In order to solve this problem, recent studies have explored a variety of strategies, such as using forgetting aware pruning or incremental moment matching to slow down the loss of knowledge (Xiao, 2026; Lee et al., 2017). Under the specific background of MoE, how to design an efficient routing mechanism and isolate domain specific parameters has become a core breakthrough to alleviate cross domain interference and catastrophic forgetting (Google Research, 2026; Simonds et al., 2024). For example, the latest MoE-CT framework proposes a new method for LLM training with the ability to resist catastrophic forgetting (Li et al., 2024).

In order to make the output of large models more in line with human preferences and security standards in specific fields, reinforcement learning based on human feedback (RLHF) has become a standard post training alignment paradigm (Shao et al., 2024). However, the traditional RLHF algorithm (such as PPO) not only needs to load a huge reward model and reference model, but also needs an additional value model (critical), which is easy to cause video memory overflow when fine-tuning the MoE model with a large number of parameters. In addition, in high-risk medical conversations, overly simplified reward functions often lead the model to learn "reward hacking", that is, to cheat high scores by generating lengthy, professional but empty content(Liu et al., 2025; Wang et al., 2025b) . In order to reduce the computational cost of reinforcement learning and improve the reliability of medical scenarios, a group relative strategy optimization algorithm without value model has been proposed in recent research. The algorithm replaces the absolute value judgment with the relative reward output within the group, and combines the compound semantic similarity reward to significantly improve the stability and training efficiency of the model in complex reasoning tasks(hlfshell, 2026). At the same time, in order to evaluate the generated content more accurately, the evaluation mechanism based on semantic similarity and llm-as-a-judge has also been widely introduced into alignment training(Confident AI, 2026; Yan, 2024).

3 Methodology

3.1 Dual Pool Heterogeneous Expert routing strategy

In order to solve the problem of long winded, unfocused and unprofessional model responses, this study broke the traditional paradigm of MoE single sharing expert pool,and creatively decoupled the expert network from the two physically isolated sub pools of General Expert Pool and Department-Specialized Expert Pool. Here we take $\mathrm{N_{gen}}$ a collection of general experts $\mathrm{A_{gen}}$ and $\mathrm{N_{spec}}$ experts from departments $\mathbf{A}_{\mathrm{spec}}$,when entering $x$to reach MoE, the system executes two sets of gating logic in parallel(Equation 1):

$$\mathrm{A_t} = \mathrm{TopK}(\mathrm{s^t}, \mathrm{K_t} = \mathrm{N_t}), \quad t \in \{gen, spec\} \tag{1}$$

Where s is the output of the input x activated by the softmax of the corresponding module, from which we can obtain the comprehensive output of the heterogeneous MoE layer(Equation 2):

$$h_{HMoE} = \left(\sum_{i \in A_{\mathrm{gen}}} s_i^{\mathrm{gen}} \, E_i^{\mathrm{gen}}(\mathbf{x}) \right) + \left(\sum_{j \in A_{\mathrm{spec}}} s_j^{\mathrm{spec}} \, E_j^{\mathrm{spec}}(\mathbf{x}) \right) \tag{2}$$

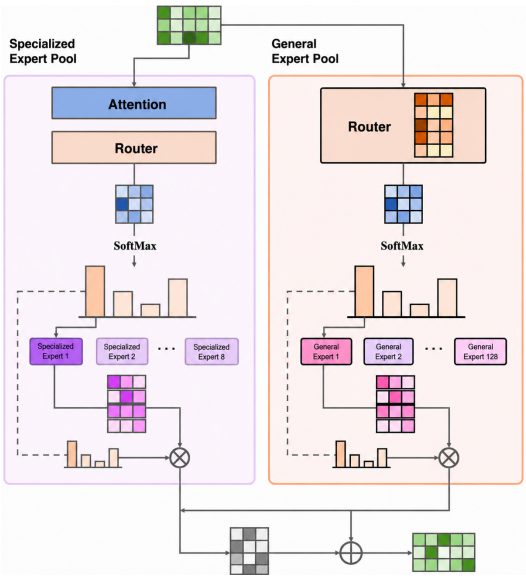

Figure 2: Model Architecture (The subject connects the general and specialized routers through a conditional gated residual mechanism, selecting 8 general experts from 128 and 1 specialized expert from 8.)

Figure 2 shows the model architecture. This mandatory allocation ratio of "8 general+1 specialist" has a far-reaching theoretical basis. In multiple rounds of online medical dialogues, a patient's input is often "half vernacular, half symptom description". By setting $K_{gen}= 8$ We ensure that the model has enough universal representation bandwidth to capture the emotional tone and context reference of the conversation, which directly eliminates the mechanization and ice cold feeling caused by the over fitting of medical terms in the traditional medical model(Haider et al., 2025). And force $K_{spec}= 1$ ensures that no matter how oral the input text is, there must be a gradient and information flow through the specialized medical knowledge network. This mechanism physically blocks the possibility of the model to "talk in general" key medical clues to general experts on the computational graph, thus solving the core pain point of unfocused generated content and lack of professionalism at the root.

3.2 Anti forgetting strategy based on parameter expansion and absolute freezing

After the establishment of heterogeneous routing architecture, how to train on the huge multi department medical data set and completely eliminate catastrophic forgetting has become another core difficulty in the implementation of the system. This study abandoned the inefficient regularization penalty or data replay, and instead adopted an evolutionary learning strategy of "parameter expansion and absolute freezing", which was inspired by the progressive neural network in continuous learning and MoE-CT architecture.

Taking an open source MoE model as a starting point, we define all the initial parameters of the model as the pre training parameter set $_{pre}$. Before entering the continuous fine-tuning (SFT and RLHP) phase in the medical field, we freeze the general branch parameters, specifically for all the self attention layer parameters (W_q,W_k,W_v,W_o Layer normalization parameters, general expert set A_{gen} and the corresponding router W_{gen} Turn off all gradient updates; At the same time, we are interested in the new K_{spec}feedforward neural network modules, that is, untrained Department expert set A_{spec} and the corresponding router W_{spec} for parameter training. Through this mechanism, when the model learns new clinical department knowledge, its underlying common language manifold is preserved intact and lossless](Li et al., 2025b). Department experts are actually fitting the residual vector between "general expert output" and "professional medical diagnosis goal". Compared with Lora and other methods, this absolute parameter isolation scheme not only provides a large enough parameter capacity to accommodate the complex medical guidelines, but also isolates the gradient interference, so that the knowledge of different departments can be independently deposited in the corresponding expert neurons, achieving immunity to catastrophic forgetting.

3.3 Alignment and hybrid reward function based on GRPO

While PPO remains the industry standard for aligning LLMs via reinforcement learning, it suffers from challenges such as training instability and high computational overhead. However, the core of PPO algorithm is that it uses an actor critical architecture to estimate the dominance function. This requires that in addition to retaining the policy model and reference model in the video memory, a value model equivalent to the parameter size of the policy model must be additionally trained and loaded during training(Yugen.ai, 2026). For the MoE model, which already contains a large number of expert parameters, this doubling of memory usage can easily lead to the depletion of the memory of the computing node, which seriously hinders

the concurrency and throughput of large-scale online training.

For this reason, the system fully uses GRPO algorithm to replace PPO for policy alignment. GRPO is an efficient RL paradigm proposed by Deepseek when exploring the mathematical logic reasoning of large models. Its biggest breakthrough is to completely remove the value model and calculate the advantage value by comparing multiple generated samples within a group under the same prompt word, thus reducing the display memory overhead by nearly half.

This section focuses on the dual track hybrid reward function we designed. The success or failure of reinforcement learning ultimately depends on the gradient signal quality provided by the reward function. If the scalar score given by the referee based on GPT-4o and other general large models is directly used, it will often lead to disastrous reward hacking phenomenon. The research shows that even if the professionalism and conciseness are emphasized in the prompts, LLM judges still have inherent "length deviation" and "format preference" in the bottom alignment preference(Huang et al., 2026). If we only rely on the big model to score, the trained big medical model will soon find that "as long as we generate extremely lengthy and complex medical disclaimers and all-round gossip, we can cheat the judges' high scores", which completely deviates from our original intention of solving the problem of "verbosity and non focus".

In order to impose constraints on the model that are in line with both clinical evidence-based medicine and online interaction norms, this architecture proposes an innovative dual track hybrid reward system, which checks and balances each other by integrating the semantic similarity at the bottom and the multi-dimensional external referee scoring at the top. The specific formula is as follows(Equation 3):

$$R_{\text{total}}(o_i, q, y_{\text{ref}}) = W_{\text{sem}} \cdot R_{\text{sem}}(o_i, y_{\text{ref}}) + W_{\text{judge}} \cdot R_{\text{judge}}(o_i, q) \quad (3)$$

Then, the meaning of the above parameters is introduced from the reward function of two different dimensions.

Semantic similarity reward based on dense vector R_{sem}

In order to ensure the accuracy of the clinical facts of the generated content, we avoid using the surface coincidence index (such as BLEU or ROUGE) which is not sensitive to medical synonyms, and instead introduce the deep semantic similarity reward. This mechanism uses a lightweight encoder (such as BGE-m3 or Med sentence transformer) that has been specially fine tuned by medical literature, and the currently generated reply o_i and gold standard reference answer y_{ref}(desensitization interaction records from real senior clinicians) are mapped to high-dimensional embedding spaces respectively, and their cosine similarity is calculated(Akarajaradwong et al., 2025).

Semantic reward plays the role of "factual gravity" in reinforcement learning. If the policy model generates a large amount of redundant text that appears professional but deviates from the target or has no clinical significance to cater to the LLM's criteria, its dense vector representation in the feature space will inevitably be diluted and deviate from the true diagnostic benchmark, resulting in a sharp drop in R_{sem}. This fundamentally locks the space of the model to generate empty long text through the underlying distance measurement.

Multidimensional scoring of large model based on structured scale R_{judge}

In order to effectively avoid the subjective bias and hallucination of the referee model under a single coarse-grained score, this paper designs a multidimensional scale evaluation system based on strict thinking chain, which requires the referee model to give quantitative scores in five independent dimensions (the score interval is set to [0,1]). The relevant prompt word templates are provided in the Appendix A.

4 Experiments and results

4.1 Experiments Setup

1) Hyperparameter: Experiments are conducted on the Qwen3-30B-A3B model, a large-scale sparse activation language model with a hybrid expert architecture. Compared to dense models of similar parameter scale, Qwen3-30B-A3B achieves better performance with higher computational efficiency, making it an ideal testbed for medical reasoning under practical deployment constraintsthis also motivates our choice. For fine-tuning, we use LoRA with rank r = 64 and scaling factor = 16. Training is performed on 8 NVIDIA A100 GPUs with mixed precision, using the AdamW optimizer, a learning rate of 1e-5, batch size of 4, and 3 training epochs, along with linear learning rate scheduling and a warmup strategy to stabilize training.

Table 1: Method Comparisons on KaMed-X and MedDialog-X

	KaMed			MedDialog-X		
	BLEU-1	ROUGE-1	E-F	BLEU-1	ROUGE-1	E-F
Qwen3-30B-A3B(MoE)	0.0367	0.0823	13.86	0.0373	0.0901	13.15
SFT	0.1258	0.1844	19.88	0.1300	0.1954	18.05
DPO	0.1267	0.1872	20.34	0.1383	0.2005	19.19
GRPO	0.1422	0.1906	21.40	0.1476	0.2029	20.16
Ours	0.1669	0.2118	23.37	0.1677	0.2213	22.94

Table 2: Ablation Study Results

Module Fusion	BLEU-1	ROUGE-1	E-F
w/o Specialist routing	0.1534	0.2047	21.78
w/o Efficient fine tuning	0.1558	0.2062	21.05
w/o Reinforcement learning	0.1609	0.2145	21.57
Ours	0.1677	0.2213	22.94

*2) Datasets:*During the fine-tuning stage, we curated datasets spanning 6 major medical NLP tasks which are shown in AppendixB. These datasets were uniformly converted into natural language question-answering pair format and supplemented with department labels, aiming to enhance the model's capability in multi-task medical language understanding and specialty expert routing. In addition, MedDialog and KaMed annotated with CoT following another work(construction method is given in the appendix **??**) are employed for GRPO training and evaluation, denoted as MedDialog-X and KaMed-X respectively.

3) Metrics: In order to comprehensively evaluate the effectiveness of this work in the task of medical dialogue generation, this section selects evaluation indicators covering multiple dimensions of text quality and medical accuracy. First, the BLEU-1/BLEU-4 index is used to measure the language fluency and basic semantic coincidence of the generated reply by calculating the exact matching degree of n-gram between the generated text and the reference text. Secondly, the ROUGE-1/ROUGE-l index is introduced to evaluate the information recall rate and the similarity of the overall text structure between the generated content and the doctor's manual reference answer from the perspective of word overlap rate and the longest common subsequence. In view of the strict requirements for professionalism in the field of medical dialogue, this chapter specifically introduces Entity-F1 (E-F) as the core evaluation index; Through the harmonic average of accuracy rate and recall rate, this index penetrates the diversity of natural language expression, accurately quantifies the capture and generation ability of the model for key medical entities, so as to objectively reflect the clinical fidelity of the model output and the availability of the system in the actual medical scene.

4) Baselines: This section compares the proposed method with the following baseline methods:

(1) SFT: Supervised fine-tuning is the foundational alignment stage for enabling LLMs to follow instructions and infuse domain-specific knowledge. This paradigm constructs high-quality "instruction-response" paired datasets (such as standardized specialist doctor-patient dialogues and medical record reasoning data) and updates the parameters of the pre-trained model using maximum likelihood estimation as the objective function.

(2) DPO: Direct Preference Optimization is a lightweight alignment algorithm designed to simplify the reinforcement learning from human feedback (RLHF) pipeline. By constructing an implicit contrastive loss on "preferred-rejected" response pairs, DPO can directly adjust the model policy without requiring an additional value baseline network, efficiently widening the log probability gap between high-quality responses that adhere to clinical facts and low-quality ones containing hallucinations or verbose, unfocused content.

(3) GRPO: Group Relative Policy Optimization is a memory-friendly reinforcement learning algorithm designed for aligning extremely large-scale language models. GRPO samples a group of candidate responses in parallel for the same input prompt and calculates a relative advantage function by normalizing reward scores within that group.

Table 3: Fine Tuning Contrast Experiment

method	MedDialog-X		
	BLEU-1	ROUGE-1	E-F
Qwen3-30B-A3B(MoE)	0.0373	0.0901	13.15
Adapter	0.1389	0.2022	19.25
LoRA	0.1343	0.1966	18.53
X-LoRA	0.1365	0.2001	19.07
Ours	0.1677	0.2213	22.94

4.2 Result analysis

Comparative experiment

Table 1 shows the effect comparison of this work with other training methods in the overall method. The method proposed in this paper is superior to the existing mainstream algorithm in all evaluation dimensions. On the KaMed-X dataset, the BLEU-1 and ROUGE-1 of this method are improved by 17.3% and 11.1% respectively compared with the benchmark GRPO algorithm; On the E-F index reflecting the depth of medical logic, the highest score was 23.37. It is proved that by isolating the general ability pool and the specialized ability pool, the pain point of medical expertise diluted by the general neuron can be effectively solved, and the model can maintain the general ability and have the accuracy of medical diagnosis.

The efficient fine-tuning construction method of parameters introduced in this framework successfully realizes the balance between knowledge injection and ability maintenance by freezing the parameters of the general expert pool and only fitting the residuals of the specialist expert pool. The data shows that while this method improves the professional logic score, the general language indicators such as BLEU-1 do not decline due to catastrophic forgetting, but achieve synchronous growth due to the synergistic effect of routing mechanism.

Ablation Experiment

The experimental results are shown in Table 2.From the table results, we can see that the complete framework has achieved the best performance in all evaluation dimensions, which verifies the overall effectiveness of the two-stage training paradigm and the underlying architecture design proposed in this paper. The specific ablation analysis of each core module is as follows:

When the specialist channel is removed and the traditional route is used, the performance of the model decreases significantly. BLEU-1 and ROUGE-1 decreased to 0.1534 and 0.2047, respectively, which were the lowest values in all ablation groups, and E-F index also decreased to 21.78. Under the single shared expert pool architecture, the deep-seated medical differential diagnosis information is easy to be diluted by universal neurons. Due to the lack of independent specialist gating and expert mechanism, the model is difficult to effectively model specific pathological features, resulting in weak performance in the accurate expression of medical text, which affects the professionalism and pertinence of the inquiry.

Without efficient fine-tuning settings, the E-F index of the model decreased from 22.94 to 21.05, while BLEU-1 and ROUGE-1 also decreased to 0.1558 and 0.2062, respectively. This shows that when dealing with complex cross specialty knowledge, the fine-tuning method of conventional adapter or non isolated feature space is difficult to improve the model expression ability, and may even cause catastrophic forgetting.

After removing the GRPO based reinforcement learning alignment phase, the BLEU-1 and ROUGE-1 of the model were slightly higher than those of the previous two ablation experiments, but there was still a significant gap compared with the complete framework, and the E-F index decreased to 21.57. In the absence of dense vector semantic rewards and multidimensional scale constraints, the model is prone to degenerate and generate redundant text that lacks substantive diagnostic logic. Because this kind of lengthy reply has a certain confusion in the degree of literal overlap, the decline of its surface text index is relatively slow; However, the obvious decrease of E-F essentially reflects the performance decline of the model in the identification and screening efficiency and the accuracy of medical facts. The hybrid reward system in this paper successfully guides the model to return to the screening logic of the medical core through dense semantic scoring and multidimensional large model scale constraints.

Analytical experiment

In order to further verify the effectiveness of the method, this section introduces the following methods for efficient fine-tuning of parameters for comparison:

(1) Adapter(Yan et al., 2023): a classical and efficient parameter tuning method, which usually realizes the injection of domain specific knowledge by inserting a small trainable feedforward network into the transformer architecture

(2) LoRA(Devalal and Karthikeyan, 2018): an efficient parameter tuning method. By introducing

Table 4: Experimental Results Of Model Parameter Analysis

Added parameters	KaMed-X			MedDialog-X		
	BLEU-1	ROUGE-1	E-F	BLEU-1	ROUGE-1	E-F
Attention Layer	0.1242	0.1827	19.88	0.1327	0.1986	19.37
General MoE Layer	0.1279	0.1903	19.95	0.1346	0.2014	19.62
Specialized MoE (Ours)	0.1669	0.2118	23.37	0.1677	0.2213	22.94

low rank matrix approximate full parameter tuning in key modules such as attention, it can achieve downstream task adaptation while reducing memory overhead.

(3) X-LoRA(Buehler and Buehler, 2024): an advanced variant of LoRA, which usually combines the idea of mixed experts and introduces a routing mechanism in low rank space to enhance the expression ability of the model in multi task or complex knowledge domain.

The experimental results are shown in Table 3. The dual pool heterogeneous expert and strong chemical alignment framework proposed in this study achieves the optimal performance on all core indicators, which is better than the basic model and the traditional efficient fine-tuning baseline.

General large models perform poorly on medical dialogue benchmarks due to their lack of professional medical focus. While parameter-efficient tuning methods like adapters, LoRA, and X-LoRA bring improvements, their low-rank feature spaces face bottlenecks in complex, cross-disciplinary medical scenarios. The proposed method achieves the best BLEU-1 and ROUGE-1 scores, owing to a frozen general ability pool and dense semantic constraints that enhance text naturalness and reduce redundancy. Crucially, it attains an entity F1 score of 22.94, roughly 19% higher than the suboptimal adapter, demonstrating that its specialized ability pool with isolation and fine-tuning more accurately captures key medical entities and pathological features, thereby strengthening clinical differential diagnosis.

Model parameter experiment

From an experimental design perspective, this study strictly controlled the scale of newly added parameters, injecting equal amounts into three structures: attention layers, standard MoE layers, and the proposed independent specialist MoE with fusion mechanism. This comparison isolates the effect of parameter allocation and architecture from mere scale expansion. Table 4 show that simply adding parameters to attention or standard MoE layers yields limited gains. In contrast, the proposed Dual Pool Heterogeneous Expert architecture shows superior improvement across metrics, especially achieving over 16% relative gain on the E-F score, highlighting its advantage in medical knowledge modeling and specialized semantic expression.

These results demonstrate that performance gains stem not from parameter expansion but from rational structural organization and semantic alignment with the task. The sentence-level specialist routing and conditional gating mechanism enable the decoupling and synergy of general and specialized capabilities, improving specialized modeling while maintaining generalization.

5 Conclusion

We proposes a fine-tuning and alignment framework based on a Dual Pool Heterogeneous Expert model to address key challenges in adapting LLMs for clinical applications, such as the coupling of general and specialized knowledge, lack of long-term coherence, and reliance on single reward signals. The framework decouples these capabilities into independent expert pools and dynamically fuses them via a conditional gating residual mechanism. During fine-tuning, it employs a branch decoupling strategy for knowledge injection, and during alignment, it integrates the GRPO algorithm with a multi-dimensional reward mechanism. Experiments demonstrate that our method outperforms the sub-optimal baseline by 13.79% and 9.21% on MedDialog-X and KaMed-X, validating its effectiveness in the medical vertical and thereby laying a foundation for developing safe and controllable clinical decision-making agents.

Ethical Considerations

This study, based on the DP-HHEM framework, strictly adheres to medical research ethics and copyright laws in the use of all datasets. Fine-tuning is conducted on the multi-task medical language understanding corpus constructed by reorganizing existing public medical NLP datasets (including

NER, RE, NORM, TS, TC, and NLI tasks) into natural language QA pairs, while evaluation is performed on MedDialog-X and KaMed-X. All data has undergone thorough anonymization during collection and usage, ensuring no patient privacy information is involved. The data sources comply with open-source licenses or have obtained proper authorization. Any manual annotation or evaluation involved in the dataset construction process was conducted solely by the authors of this paper, with no ethical concerns.

Despite these precautions, potential risks remain in deploying LLMs for medical dialogue. The proposed DP-HHEM framework may still generate medically inaccurate, unsafe, or hallucinated responses, particularly in complex clinical scenarios. Therefore, the system is intended for research purposes only and should not replace professional medical judgment or real-world clinical decision-making. In addition, biases inherited from training data may affect the fairness and reliability of generated responses across different patient groups and medical specialties.

Limitations

The studys experimental validation relies primarily on the KaMed and MedDialog datasets, which are derived from online consultation platforms and thus exhibit a conversational style and biases in department and disease distribution. This limits their ability to reflect complex clinical decision-making scenarios and may overestimate the models robustness in real-world medical settings. At the architectural level, the 8 general + 1 specialist static routing strategy lacks adaptability to varying query complexity, leading to computational redundancy in simple cases and potential reasoning limitations in complex, interdisciplinary scenarios. Additionally, fully freezing the underlying parameters prevents catastrophic forgetting but restricts the models ability to evolve and adapt to new medical knowledge and paradigms.

References

Pawitsapak Akarajaradwong, Chompakorn Chaksangchaichot, Pirat Pothavorn, Ekapol Chuangsuwanich, Attapol Rutherford, and Sarana Nutanong. 2025. Aligning LLMs for Thai legal question answering with efficient semantic-similarity rewards. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 304–316, Suzhou, China. Association for Computational Linguistics.

Eric L. Buehler and Markus J. Buehler. 2024. X-lora: Mixture of low-rank adapter experts, a flexible framework for large language models with applications in protein mechanics and molecular design. *APL Machine Learning*, 2(2):026119.

Ruike Cao, Shaojie Bai, Fugen Yao, Liang Dong, Jian Xu, and Li Xiao. 2026. ATPO: ADAPTIVE TREE POLICY OPTIMIZATION FOR MULTI-TURN MEDICAL DIALOGUE. In *The Fourteenth International Conference on Learning Representations*.

Confident AI. 2026. G-Eval Simply Explained: LLM-as-a-Judge for LLM Evaluation. https://www.confident-ai.com/blog/g-eval-the-definitive-guide. Accessed: 2026-03-15.

Shilpa Devalal and A. Karthikeyan. 2018. Lora technology - an overview. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 284–290.

Google Research. 2026. Mixture-of-Experts with Expert Choice Routing. https://research.google/blog/mixture-of-experts-with-expert-choice-routing/. Accessed: 2026-03-15.

Syed Ali Haider, Srinivasagam Prabha, Cesar Abraham Gomez-Cabello, Sahar Borna, Ariana Genovese, Maissa Trabilsy, Bernardo G Collaco, Nadia G Wood, Sanjay Bagaria, Cui Tao, and Antonio Jorge Forte. 2025. Synthetic PatientPhysician Conversations Simulated by Large Language Models: A Multi-Dimensional Evaluation. *Sensors (Basel)*, 25(14):4305. PMID: 40732431; PMCID: PMC12298718; Accessed: 2026-03-15.

hlfshell. 2026. GRPO in DeepSeek-R1. https://hlfshell.ai/posts/grpo/. Accessed: 2026-03-15.

Zixuan Huang, Xin Xia, Yuxi Ren, Jianbin Zheng, Xuefeng Xiao, Hongyan Xie, Li Huaqiu, Songshi Liang, Zhongxiang Dai, Fuzhen Zhuang, Jianxin Li, Yikun Ban, and Deqing Wang. 2026. Real-time aligned reward model beyond semantics. *Preprint*, arXiv:2601.22664.

IBM. 2026. What is Catastrophic Forgetting? https://www.ibm.com/think/topics/catastrophic-forgetting. Accessed: 2026-03-15.

John Snow Labs. 2026. Why LLM Output Alone Cannot Drive Clinical Decisions: Lessons from Production Deployments. https://www.johnsnowlabs.com/why-llm-output-alone-cannot-drive-clinical-decisions-lessons-from-production-deployments/. Accessed: 2026-03-15.

Sang-Woo Lee, Jin-Hwa Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Overcoming catastrophic forgetting by incremental moment matching.

Junzhuo Li, Bo Wang, Xiuze Zhou, and Xuming Hu. 2025a. Dynamic expert specialization: Towards catastrophic forgetting-free multi-domain MoE adaptation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18478–18493, Suzhou, China. Association for Computational Linguistics.

Raymond Li, Charles Guille-Escuret, Denis Kocetkov, Joel Lamy-Poirier, Luke Kumar, Max Tian, Sebastien Paquet, Christopher Pal, and Torsten Scholak. 2025b. Mitigating forgetting in continually pretraining moe-LLMs by adding and chilling experts.

Tianhao Li, Shangjie Li, Binbin Xie, Deyi Xiong, and Baosong Yang. 2024. Moe-ct: A novel approach for large language models training with resistance to catastrophic forgetting. *Preprint*, arXiv:2407.00875.

Yizhou Liu, Dingkang Yang, Zizhi Chen, Minghao Han, Xukun Zhang, Keliang Liu, Jingwei Wei, and Lihua Zhang. 2025. Adaptive reinforcement for open-ended medical reasoning via semantic-guided reward collapse mitigation.

MedCity News. 2026. Prompting is Not Clinical Practice: The Limits of General LLMs in Healthcare. https://medcitynews.com/2026/03/prompting-is-not-clinical-practice-the-limits-of-general-llms-in-healthcare/. Accessed: 2026-03-15.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Omair Mohamed, Mohamed Zidan, Fahad Khan, Salman Khan, Rao Anwer, and Hisham Cholakkal. 2026. Medix-r1: Open ended medical reinforcement learning. *Preprint*, arXiv:2602.23363.

Monica Munnangi and Saiph Savage. 2026. Threadmed-qa: A multi-turn medical dialogue benchmark from real patient questions. *Preprint*, arXiv:2603.11281.

NVIDIA. 2026. What Is Mixture of Experts (MoE) and How It Works? https://www.nvidia.com/en-us/glossary/mixture-of-experts/. Accessed: 2026-03-15.

Omar Sanseviero, Lewis Tunstall, Philipp Schmid, Sourab Mangrulkar, Younes Belkada, and Pedro Cuenca. 2023. Mixture of experts explained.

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. *Preprint*, arXiv:2402.03300.

Toby Simonds, Kemal Kurniawan, and Jey Han Lau. 2024. MoDEM: Mixture of domain expert models. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 75–88, Canberra, Australia. Association for Computational Linguistics.

Guandong Wang, Yating Wang, Yanhao Li, and Jie Zhai. 2025a. Fmt: Simulating clinical reasoning via forward-inference framework in medical dialogue systems. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 7657–7664.

Pengkai Wang, Linus, Pengwei Liu, Zhijie Sang, Congkai Xie, and Hongxia Yang. 2025b. Infimed-orbit: Aligning llms on open-ended complex tasks via rubric-based incremental training. *Preprint*, arXiv:2510.15859.

Baicen Xiao. 2026. Avoiding Amnesia: Some Practical Guides to Mitigate Catastrophic Forgetting in LLMs Post-training. https://medium.com/@baicenxiao/avoiding-amnesia-some-practical-guides-to-mitigate-catastrophic-forgetting-in-llms-post-training-6a23e4f064cb. Accessed: 2026-03-15.

Yongyu Yan, Kui Xue, Xiaoming Shi, Qi Ye, Jingping Liu, and Tong Ruan. 2023. Af adapter: Continual pretraining for building chinese biomedical language model. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 953–957.

Ziyou Yan. 2024. Evaluating the Effectiveness of LLM-Evaluators (aka LLM-as-Judge). https://eugeneyan.com/writing/llm-evaluators/. Accessed: 2026-03-15.

Yugen.ai. 2026. Understanding the Math Behind GRPO — DeepSeek-R1-Zero. https://medium.com/yugen-ai-technology-blog/understanding-the-math-behind-grpo-deepseek-r1-zero-9fb15e103a0a. Accessed: 2026-03-15.

Xiaoquan Zhi, Hongke Zhao, Likang Wu, Chuang Zhao, and Hengshu Zhu. 2025. Reinventing clinical dialogue: Agentic paradigms for llm enabled healthcare communication. *Preprint*, arXiv:2512.01453.

A Referee model scoring prompt

Figure A1 shows the scoring prompt. The referee model quantifies scoring based on five independent dimensions: Consistency of Specialty Depth and Domain, Medical Accuracy, Consistency with the Patient Query, Personalization and Appropriateness, and Expression Clarity.

B Fine tuning training data construction

In this work, we construct a multi-task medical language understanding corpus to build the model's fundamental medical capabilities in entity recognition, semantic relation modeling, and reasoning. Specifically, we curate datasets spanning 6 major medical NLP tasks (as shown in Table B1), including Named Entity Recognition (NER), Relation

Role Definition

You are an attending physician with extensive experience at a tertiary hospital. You are also an expert in applying artificial intelligence to evaluation in medical domains. Your task is to conduct a rigorous, objective, and fair evaluation of the quality of responses generated by large language models in real-world medical consultation scenarios.

Task Objective

Please evaluate the response generated by an [AI model] based on the provided [Patient Query], [User Prompts], and [Clinical Reference Answers]. Perform a multidimensional assessment. You need to score based on the five criteria listed below. Each score must be on a scale from [0.0, 1.0] (one decimal place), where 0.0 indicates extremely poor or highly risky, and 1.0 indicates completely excellent and fully aligned with clinical gold standards.

Evaluation Criteria and Scoring Rubric

1. Consistency of Specialty Depth and Domain

- [0.0–0.3]: Severely off-topic or irrelevant to the specialty perspective; even critical details (e.g., recommendations for internal medicine related issues) are missing.
- [0.4–0.7]: Demonstrates basic knowledge of the specialty, but the depth of diagnostic reasoning or pathophysiological explanations is insufficient.
- [0.8–1.0]: Highly accurate. Demonstrates strong specialty-specific clinical reasoning and recommends targeted evaluations (e.g., recommending coronary angiography for suspected coronary artery disease rather than vague tests like general blood work). Exhibits high-level clinical thinking.

2. Medical Accuracy

- [0.0–0.3]: Contains serious factual errors, clinical hallucinations, or potentially harmful/risks-inducing treatment suggestions.
- [0.4–0.7]: Fundamentally harmless, but contains minor inaccuracies or does not fully comply with the latest clinical guidelines/evidence.
- [0.8–1.0]: Scientifically accurate, rigorous, and complete. Diagnostic and treatment recommendations fully align with current authoritative clinical guidelines; safe and reliable.

3. Consistency with the Patient Query

- [0.0–0.3]: Fails to capture key information in the patient's query (e.g., ignores stated drug allergies, symptoms, etc.).
- [0.4–0.7]: Addresses part of the query, but ovorlooks key historical details.
- [0.8–1.0]: Fully addresses all key issues through multi-turn dialogue, with coherent and cohesive responses.

4. Personalization and Appropriateness

- [0.0–0.3]: Provides general information without individualization; fails to consider patient-specific factors (e.g., pregnancy, older age).
- [0.4–0.7]: Includes basic patient information and makes surface-level adjustments, but lacks depth in personalized treatment.
- [0.8–1.0]: Accurately captures individual differences and provides highly personalized prevention or treatment plans based on age, sex, and medical history.

5. Expression Clarity

- [0.0–0.3]: Extremely confusing (e.g., medical jargon, illogical structure), making it impossible for the patient to understand.
- [0.4–0.7]: Generally clear, but contains low-information-density statements or partially unclear explanations.
- [0.8–1.0]: Clearly structured, concise, professional, and easy to understand, effectively reducing the patient's cognitive burden.

Output Format (Strict Compliance)

You must output the evaluation result in pure JSON format. Before finalizing the score, you must first provide your reasoning process in the "reasoning" field. The length of the reasoning must be controlled within 50–100 words.
Please strictly use the following JSON format:

```
{
  "Consistency of Specialty Depth and Domain": { "reasoning": "...", "score": 0.0 },
  "Medical Accuracy": { "reasoning": "...", "score": 0.0 },
  "Consistency with the Patient Query": { "reasoning": "...", "score": 0.0 },
  "Personalization and Appropriateness": { "reasoning": "...", "score": 0.0 },
  "Expression Clarity": { "reasoning": "...", "score": 0.0 }
}
```

Figure A1: Example of Scoring Prompt For the Referee Model

Table B1: Table of Training Dataset

Dataset	Type	Total Size
CMeEE	NER	23,000
CMeIE	RE	22,406
CHIP-CDN	NORM	18,000
CHIP-STS	TS	30,000
CHIP-CTC	TC	40,836
KUAKE-QIC	TC	10,880
KUAKE-QTR	NLI	32,552
KUAKE-QQR	NLI	18,195
PubMedQA	NLI	213,000
MedMCQA	NLI	190,400

Extraction (RE), Normalization (NORM), Textual Similarity (TS), Text Classification (TC), and Natural Language Inference (NLI). These datasets are uniformly converted into natural language question-answering (QA) pair format and supplemented with department labels, considering the inherent mismatch between traditional classification or sequence labeling tasks and the generative paradigm of large language models. This enables the model to enhance multi-task medical language understanding and specialty expert routing.

C CoT construction method

This section describes how we supplement MedDialog and KaMed with CoT annotations.This method employs a closed-loop mechanism of "reasoning-revision-reflection" to generate prior diagnostic chains without relying on the doctors' original responses, thereby more authentically simulating the clinical reasoning process(Wang et al., 2025a).The specific process is shown in the figure C1. MedDialog covers 29 macro-level clinical departments (e.g., internal medicine, surgery, pediatrics and gynecology) and 172 sub-specialties (e.g., cardiovascular medicine, neurology, oncology), ensuring broad domain coverage. In addition, KaMed contains over 60,000 high-quality medical dialogues with an average of 11.62 turns, posing higher demands on long-range logical consistency and multi-step clinical reasoning.

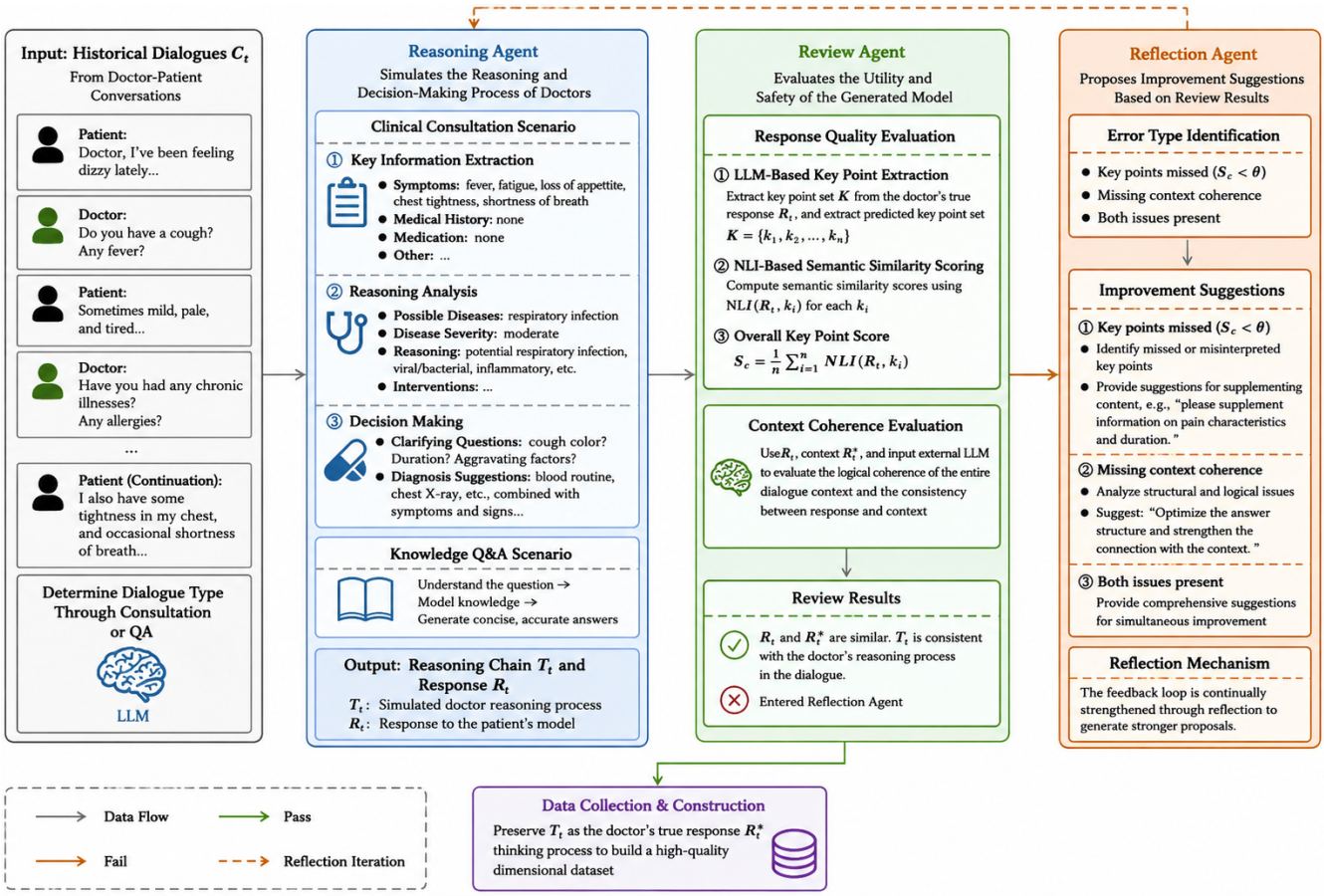

Figure C1: The Forward Reasoning Framework of Constructing CoT