FedCoT: Communication-Efficient Federated Reasoning Enhancement for Large Language Models

Anonymous ACL submission

Abstract

Enhancing LLM reasoning in federated settings is nontrivial due to stringent computational, communication, and privacy constraints—especially in healthcare, where clinically consequential decisions require not only accuracy but also interpretable, auditable rationales to meet safety, accountability, and regulatory requirements. Conventional federated fine-tuning largely imitates final answers rather than cultivating step-by-step reasoning, often relying on privacy-sensitive centralized distillation and still incurring substantial communication overhead. We address this gap with **FedCoT**, a federated reasoning framework that combines lightweight chain-of-thought resampling with a compact discriminator for selection, and client-aware LoRA stacking with weighted classifier aggregation to accommodate heterogeneity while reducing aggregation noise and communication; clients generate candidate chains and supervision locally, and only lightweight modules are aggregated on the server. Experiments on medical reasoning benchmarks show consistent gains under tight resource budgets while keeping data local and respecting privacy, offering an interpretable and resource-efficient solution.

1 Introduction

The development of large language models (LLMs) achieve advancing performance in complex reasoning tasks (Touvron et al., 2023; Bai et al., 2023; Guo et al., 2025a; Team et al., 2025; Chen et al., 2024b), which improves both the effectiveness and explainability based on the thought chains (Wei et al., 2022). The promising performance is attributed to the reinforcement learning (RL) algorithms (Christiano et al., 2017; Schulman et al., 2017; Shao et al., 2024; Zhou et al., 2025). However, the training paradigms of RL for reasoning models rely heavily on computational resources (Tian et al., 2023; Havrilla et al., 2024), which render them impractical for distributed edge environments, especially under the privacy constraint, e.g., the training data cannot be directly shared across different nodes/institutions in medical scenarios (Chen et al., 2023).

Training-free techniques (Wang et al., 2022; Xie et al., 2023; Wu et al., 2024) with prompting engineering or test-time scaling can alleviate the issues of distributing the data and model during the training phase. Although they are easy to deploy, their performance gains are quite limited and cannot fully leverage the collaborative potential of distributed device networks. Intuitively, federated learning (FL) (McMahan et al., 2017) serves as an alternative for achieving a better trade-off between privacy preservation guarantee and model performance. However, existing FL-based LLMs training paradigms (Wu et al., 2025; Wei et al., 2025; Zhang et al., 2024a) predominantly rely on federated supervised fine-tuning or simply incorporate parameter-efficient techniques, which still encounter high communication overhead and thus yield suboptimal performance gains.

Reasoning with Chain-of-Thought (CoT) is indispensable in medical domains. For example, the decision for clients should be not only accurate but also reliable with traceable rationales. In addition, these privacy-sensitive scenarios demand more strict data and model usage to prevent sensitive information leakage. However, existing studies (Magister et al., 2022; Li et al., 2022; Wang et al., 2023; Chen et al., 2024d) usually obtain the data and rationales by distilling knowledge from proprietary models or direct sharing among different sources, which might expose sensitive information and violate data privacy principles.

To facilitate reasoning-based LLMs training under a privacy-preserving setting without sharing the whole model in distributed nodes, we propose **FedCoT**, a federated learning based framework to enhance the model's reasoning capac-

ity under CoT prompting without data leakage. The core of our FedCoT is a dynamic chain-of-thought discrimination mechanism to enable cross-client reasoning enhancement. Specifically, a lightweight discriminator is deployed to evaluate reasoning path candidates in real-time, which enables the discrimination of optimized reasoning trajectories. Building upon the FLoRA algorithm's principal mechanism of modular LoRA stacking for federated fine-tuning (Wang et al., 2024b), we adapt this approach to lightweight BERT (Devlin et al., 2019)) models. Crucially, we incorporate a task-specific predictor (a BERT-based classifier) dedicated to reasoning path discrimination. To aggregate these task-specific classifier, we employ a weighted aggregation scheme, ensuring robust discriminative capability across the federation. Our FedCoT target simultaneously prevents privacy risks and achieves robust performance in privacy-preserving scenarios. The experimental results on five medical domains datasets demonstrate the effectiveness of our methods by outperforming existing strong baselines.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to employ CoT resampling in federated learning settings to enhance the reasoning capabilities of large language models.
- We propose an end-to-end federated reasoning enhancement framework that integrates dynamic reasoning path discrimination. FedCoT extends modular stacking to BERT-based discriminators with a weighted aggregation scheme, effectively handling client heterogeneity while maintaining robust performance.
- We conduct comprehensive experiments across multiple medical QA benchmarks. The results demonstrate that FedCoT significantly enhances reasoning performance and efficiency compared to strong baselines, robustly validating the effectiveness of our federated reasoning enhancement framework.

2 Related Works

Reasoning Enhancement for Large Language Models CoT prompting, proposed by Wei et al. (2022), acts as an effective mechanism to augment LLMs' reasoning ability with provided instructions, which has spawned numerous training-free variants (Wang et al., 2022; Chen et al., 2024c; Li et al., 2025; Nair and Wang, 2024; Wan et al., 2024; Guo et al., 2025b). The follow-up studies integrate CoT with parameter updates by using CoT-generated rationales for model supervised fine-tuning (Kim et al., 2023; Magister et al., 2022; Li et al., 2022; Wang et al., 2023; Hsieh et al., 2023), or training with reinforcement learning (Team et al., 2025; Shao et al., 2024). However, the assumption of these methods is centralized data access, which neglects privacy constraints and computational burdens of generating rationales for federated clients (McMahan et al., 2017), and thus offering no tailored mechanism for CoT capability enhancement under distributed settings (Zhang et al., 2024b; Dritsas and Trigka, 2025; Wang et al., 2024a; Li et al., 2019). To this end, our method utilizes the information of distributed clients to enhance the CoT reasoning ability while strictly ensuring privacy and low resource consumption.

Federated Learning for Large Language Models FL is one of the key solutions for LLMs training under privacy-preserving settings (Wei et al., 2025; Wu et al., 2025; Chen et al., 2022; Wu et al., 2020; Tariq et al., 2023, 2024; Ye et al., 2023; Qian et al., 2024).The gradients and data aggregation during the training phase result in the communication-efficient needs and are usually achieved by parameter-efficient tuning. For example, low-rank adaptation (LoRA) techniques (Zhang et al., 2024a) and matrix factorization methods (Wang et al., 2024b) enable efficient aggregation under heterogeneity. Besides, knowledge distillation (Fan et al., 2024) and federated RL (Tian et al., 2023) provide other alternatives as optimization pathways. However, existing FL-LLM studies neither explicitly enhance CoT reasoning capabilities nor mitigate the privacy risks inherent in distilling rationales from centralized teacher models (Havrilla et al., 2024; Li et al., 2022). In addition, federated RL-based approaches incur prohibitive computational and communication overhead for resource-constrained clients (Qi et al., 2021; Krouka et al., 2021; Imteaj et al., 2022). Thus, previous studies remain a critical gap for developing lightweight, privacy-preserving frameworks tailored to federated CoT enhancement, which our studies provide the first exploration.

3 Preliminaries

3.1 Chain-of-Thought Prompting

Chain-of-Thought (CoT) prompting (Wei et al., 2022) enhances complex reasoning by generating intermediate reasoning paths τ between input x and output y as guidance and explanation. Formally, given a model parameterized by θ and prompt I, the CoT generation process is defined as:

$$[\tau, y] \sim p_\theta(x|I)$$

Although CoT is effective in a range of scenarios, utilizing it in distributed settings alone cannot perform well without leveraging collaborative devices for specific training.

3.2 LoRA Federated Aggregation

Low-Rank Adaptation (LoRA) (Hu et al., 2022; Mao et al., 2024) enables efficient fine-tuning via parameters metric decomposition $\Delta W = BA$ ($A \in \mathbb{R}^{r\times n}, B \in \mathbb{R}^{m\times r}$). Standard federated learning conducts client model updates by averaging aggregation as

$$\mathbf{A} = \sum_{i}^{N} u_i \mathbf{A}_i, \quad \mathbf{B} = \sum_{i}^{N} u_i \mathbf{B}_i \tag{1}$$

where N denotes the number of clients, u_i denotes the weight of the client derived from the data volume ratio. This aggregation schema introduces cross-client noise terms as

$$\begin{aligned}\Delta\mathbf{W} &= (u_0\mathbf{B}_0 + u_1\mathbf{B}_1)(u_0\mathbf{A}_0 + u_1\mathbf{A}_1) \\ &= u_0^2\mathbf{B}_0\mathbf{A}_0 + u_1^2\mathbf{B}_1\mathbf{A}_1 \\ &+ \underbrace{u_0u_1(\mathbf{B}_0\mathbf{A}_1 + \mathbf{B}_1\mathbf{A}_0)}_{\text{noise term}}\end{aligned} \tag{2}$$

However, the global updates would be deviated as the noise term grows quadratically with local clients. Besides, it would result in a dimension mismatch between heterogeneous ranks, i.e., $r_1 \neq r_2$, leading to parameter updating with average fails.

4 Methodology

We propose FedCoT, a federated learning-based framework to enhance the reasoning capability of LLMs under privacy constraints. The overview of FedCoT is presented in Figure 1 with dynamic path selection and parameter-efficient aggregation. With the candidate reasoning paths generated locally by client LLM, the signals derived from these paths supervise the training of lightweight discriminators, whose LoRA modules and classifier are then aggregated within the server-side to construct a global discriminator. This federated model subsequently enables clients to dynamically select optimal reasoning traces during inference, yielding privacy-preserving, CoT-enhanced answers.

4.1 Local Candidates Generation

Under the federated learning framework, each client should first generate candidate reasoning paths based on the questions of their associated local datasets $\{x_j, y_j\}$, which then serve as the basis for the subsequent discriminator training. Formally, K candidate reasoning paths are generated by diversity sampling of an LLM p_θ, which can also be analogously regarded as the actor model in a reinforcement-learning scenario, corresponding to the input x_j as:

$$[\tau_{j,k}, \hat{y}_{j,k}] \sim p_\theta(x_j|I)$$

where the local ground truth y_j is assigned with binary labels as Eq. 6 and then, the discrimination datasets for discriminator training are formed as Eq. 7.

$$z_{j,k} = \mathbf{1}(\hat{y}_{j,k} = y_j), \quad k = 1, 2, ..., K \tag{3}$$

$$\mathcal{D} = \{(\mathbf{h}_{j,k}, z_{j,k}) \mid \mathbf{h}_{j,k} = [x_j \parallel \tau_{j,k} \parallel \hat{y}_{j,k}]\} \tag{4}$$

To ensure high-quality supervision, we introduce a contrastive candidate selection mechanism to refine the dataset $\mathcal{D}$. We first discard queries with unanimously correct or incorrect paths, as they lack discriminative signals. For the remainder, we dynamically sample a balanced subset of positive and negative candidates, explicitly ensuring diverse error types (i.e., reasoning errors vs. missing answers) are represented.

4.2 Local Training for Candidates Discrimination

We formulate the reasoning path discrimination as a binary classification task motivated by Shi et al. (2024), where a lightweight discriminator at BERT-scale effectively evaluates candidate correctness.

In our FedCoT framework, clients initialize their local models from one of the following choices: (1) the server-provided global modules (for non-initial rounds) or (2) a base pre-trained model locally (for the first round). During each

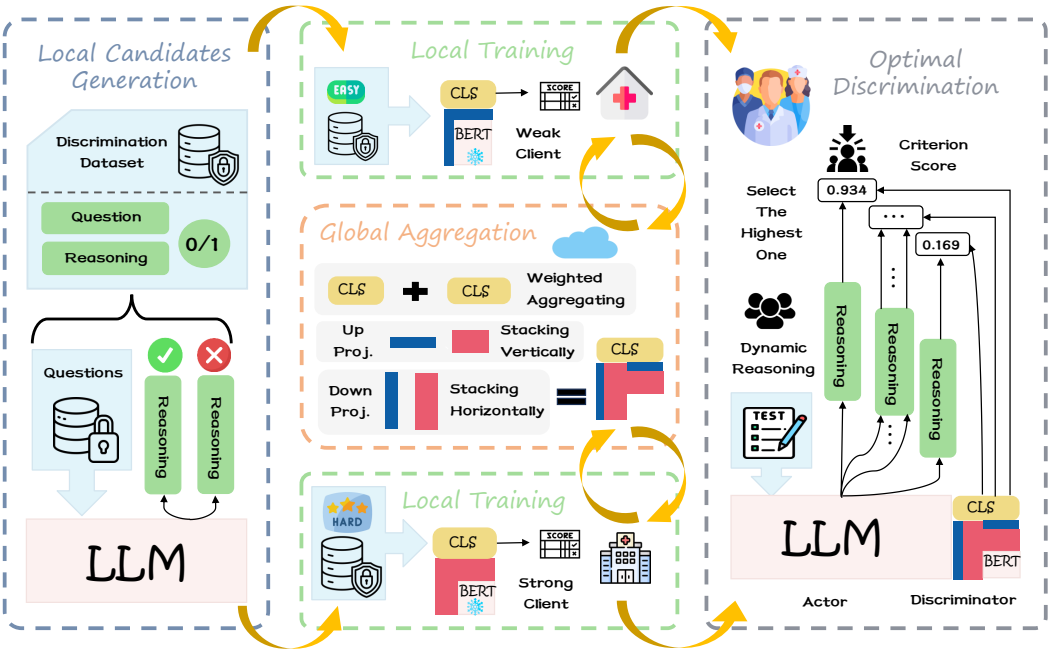

Figure 1: Overview of FedCoT framework. **Left:** The data preparation for training the discriminator. **Middle:** The federated fine-tuning of discriminator without the participation of local LLM which is only used in data preparation. **Right:** The Optimal Discrimination at test time. "Up Proj." corresponds to the $\mathbf{A}$ matrix in LoRA, and "Down Proj." corresponds to the $\mathbf{B}$ matrix. "CLS" here denotes the classifier module of discriminator.

global communication round, clients receive and initialize the model with the latest aggregated parameters, comprising both the LoRA matrices and classifier that encapsulate information from the entire federation while preventing the complete model drift from local domains.

Formally, given a question-reasoning pair $(x_j, \tau_{j,k})$, the discriminator $d_\theta : \mathcal{X} \times \mathcal{T} \rightarrow [0, 1]$ outputs a criterion score via sigmoid activation function, which is optimized to minimize the binary cross-entropy loss as:

$$\begin{aligned}\mathcal{L} = -\,[s_{j,k} \log d_\theta(\mathbf{h}_{j,k}) + \\ (1 - s_{j,k}) \log(1 - d_\theta(\mathbf{h}_{j,k}))]\end{aligned} \tag{5}$$

Instead of relying on harsh binary labels like (Shi et al., 2024), we design an soft label to evaluate the logical consistency between the reasoning path $\tau_{j,k}$ and the predicted answer $\hat{y}_{j,k}$. We utilize a Natural Language Inference model to compute the entailment probability $p_{\text{entail}} \in [0, 1]$. The continuous soft label $s_{j,k}$ is formulated as a combination of the outcome correctness and the reasoning validity:

$$s_{j,k} = (1 - \alpha) \cdot \mathbf{1}(\hat{y}_{j,k} = y_j) + \alpha \cdot p_{\text{entail}} \tag{6}$$

where $\alpha = 0.5$ is a balancing coefficient. This continuous label assigns partial credit to flawed answers with rigorous reasoning, while penalizing correct guesses lacking logical support. The discrimination dataset is then constructed as:

$$\mathcal{D} = \{(\mathbf{h}_{j,k}, s_{j,k}) \mid \mathbf{h}_{j,k} = [x_j \parallel \tau_{j,k} \parallel \hat{y}_{j,k}]\} \tag{7}$$

4.3 Modular Global Aggregation

We adopt and integrate FLoRA (Wang et al., 2024b) to achieve noise-free aggregation of LoRA matrix with protecting data privacy. When aggregating local LoRA modules, the global model update ΔW can be expressed as

$$\begin{aligned}\Delta \mathbf{W} &= \sum_{i=1}^{N} \mathbf{B}_i \mathbf{A}_i \\ &= (\mathbf{B}_1 \oplus \mathbf{B}_2 \oplus \cdots \oplus \mathbf{B}_N) \\ &\quad \cdot (\mathbf{A}_1 \oplus \mathbf{A}_2 \oplus \cdots \oplus \mathbf{A}_N)\end{aligned} \tag{8}$$

where "$\oplus$" represents the matrix stacking operation, i.e., stacking them vertically along the row direction for $\mathbf{A}_i$ and stacking horizontally along the column direction for $\mathbf{B}_i$, respectively. With the principle of block matrix multiplication, the product of these two global produced matrices, $\mathbf{B} \cdot \mathbf{A}$, is mathematically equivalent to the sum of the individual local updates, $\sum_{i=1}^{N} \mathbf{B}_i \mathbf{A}_i$.

This approach makes the globally aggregated discriminator more reliable and adaptable to heterogeneity, which arises from varying client capabilities (e.g., weaker clients using smaller LoRA ranks, stronger ones using larger ranks). We can also intentionally create heterogeneity by assigning smaller ranks to simpler tasks and larger ranks to complex ones. Regardless of the source, the stacking method integrates these diverse LoRA matrices through unified merging, ensuring smooth federated learning.

Besides, the classifier weights of each client are aggregated using a weighted average approach at each global round, to integrate information across downstream tasks as

$$\mathbf{W}^{cls} = \sum_{i}^{N} u_i \mathbf{W}_i^{cls} \tag{9}$$

4.4 Optimal Discrimination

During the inference stage, each client utilizes the final global discriminator model to score the multiple candidate reasoning paths and then selects the one with the highest score as the final output to achieve dynamic reasoning as:

$$r(h_{j,k}) = \sigma(d_\theta(h_{j,k})) \tag{10}$$

$$\hat{y_j} = \arg \max_{k \in \{1,\cdots,K\}} r(h_{j,k}) \tag{11}$$

The comprehensive process is described as an Algorithm provided in the Appendix A as the overall process of federated reasoning in our FedCoT.

5 Experiments

5.1 Experimental Setup

Datasets We evaluate our method on five biomedical Question-Answering (QA) datasets served as privacy-preserving benchmarks following previous studies (Chen et al., 2024b; Song and Lee, 2025; Chen, 2024; Zhang et al., 2023), including BioASQ(Tsatsaronis et al., 2015), MedMCQA(Pal et al., 2022), MedQA(Jin et al., 2021), MMLU-MED(Hendrycks et al., 2020), and PubMedQA(Jin et al., 2019). The statistics are provided in Table 1. These datasets span diverse task categories, ranging from medical examination questions to literature-based QA, enabling us to evaluate the models' performance across complex reasoning tasks comprehensively. The detailed information is provided in Appendix G.

Datasets	Train	Test	Source
PubMedQA (2019)	450	500	Experts
BioASQ (2015)	494	124	Articles
MMLU (2020)	1299	163	Examination
MedMCQA (2022)	3000	4183	Examination
MedQA (2021)	10178	1273	Examination

Table 1: The size and source of the medical QA datasets used in the experiment.

Cross-silos Setting The five datasets are regarded as independent clients respectively, and the privacy of each client's data is strictly protected during the training process. This cross-silo setting reflects the model's reasoning capability on different tasks and verifies its robustness under data distribution heterogeneity, which aligns with the data islands (Huang et al., 2021, 2023; Tang and Wong, 2021; Liu et al., 2022) in the real-world setting.

Prompt Templates Our designed CoT template adheres to a standardized structure characterized by concise instructions to mitigate verbosity, and a structured, itemized format for requirements. We also incorporate a one-shot CoT demonstration in the template, where the completed information is provided in Appendix H.

Hyperparameter Settings We generate 8 candidate responses for each query sample with a maximum length of 512 tokens. The LoRA ranks for each client model in BioASQ, MedMCQA, MedQA, MMLU, and PubMedQA datasets are set as 4, 32, 32, 16, and 4, respectively. During supervised fine-tuning of the federated LLMs, a uniform LoRA with a rank of 32 was used for model training. The global round is set to 2, and the local training epoch is set to 1 with a batch size of 2 in SFT of LLMs as baselines. The global round is set to 3, and the local training epoch is set to 1 with a batch size of 16 in our discriminator training.

Baselines and Evaluation Metrics Our experiments utilize different models for evaluation, including `Qwen2.5-7B-Instruct` (Bai et al., 2023), `LLaMA-3-8B-Instruct` (Touvron et al., 2023), as core LLMs for main evaluation, and the `Longformer-base-4096` (Beltagy et al., 2020) as discriminator model, following Shi et al. (2024). We compare with both the training-free and training-based baselines under both federated and non-federated scenarios to evaluate our **FedCoT** as follows: We evaluate **FedCoT** against a diverse set of strong baselines, including **Self-Consistency** (Wang et al., 2022) as

Method	BioASQ		MedMCQA		MedQA		MMLU		PubMedQA		Avg.	
	Acc. (%)	Δ(%)	Acc. (%)	Δ(%)	Acc. (%)	Δ(%)	Acc. (%)	Δ(%)	Acc. (%)	Δ(%)	Acc. (%)	Δ(%)
LLaMA-3-8B-Instruct	37.90	—	29.80	—	27.20	—	38.70	—	9.20	—	28.56	—
+Self-Consistency	40.30	+2.40	31.50	+1.70	24.70	-2.50	41.10	+2.40	2.80	-6.40	28.08	-0.48
+FedIT	42.74	+4.84	**47.29**	**+17.49**	53.73	+26.53	**71.17**	**+32.47**	13.20	+4.00	45.63	+17.07
+FedFFA-LoRA	25.00	-12.90	39.47	+9.67	37.78	+10.58	57.67	+18.97	4.60	-4.60	32.90	+4.34
+FLORA	37.10	-0.80	42.05	+12.25	45.56	+18.36	61.96	+23.26	7.00	-2.20	38.73	+10.17
+FedPrompt	52.42	+14.52	45.80	+16.00	49.96	+22.76	42.94	+4.24	39.40	+30.20	46.10	+17.54
+FedCoT (Ours)	**68.50**	**+30.60**	45.20	+15.40	**54.10**	**+26.90**	68.70	+30.00	**41.00**	**+31.80**	**55.50**	**+26.94**
Qwen2.5-7B-Instruct	73.40	—	43.70	—	29.50	—	50.30	—	38.80	—	47.14	—
+Self-Consistency	86.30	+12.90	47.10	+3.40	28.00	-1.50	57.10	+6.80	39.80	+1.00	51.66	+4.52
+FedIT	82.26	+8.86	48.48	+4.78	44.30	+14.80	68.71	+18.41	47.20	+8.40	58.19	+11.05
+FedFFA-LoRA	85.48	+12.08	48.74	+5.04	44.93	+15.43	70.55	+20.25	53.00	+14.20	60.54	+13.40
+FLORA	83.87	+10.47	51.30	+7.60	45.88	+16.38	73.62	+23.32	57.40	+18.60	62.41	+15.27
+FedPrompt	66.13	-7.27	46.62	+2.92	44.23	+14.73	68.71	+18.41	50.40	+11.60	55.22	+8.08
+FedCoT (Ours)	**96.00**	**+22.60**	**51.70**	**+8.00**	**53.40**	**+23.90**	**74.20**	**+23.90**	**68.20**	**+29.40**	**68.70**	**+21.56**

Table 2: Performance of different methods across five privacy-preserving medical datasets on top of two backbone LLMs under different settings. The best results are in **Bold** and the second-highest results are indicated with an underline.

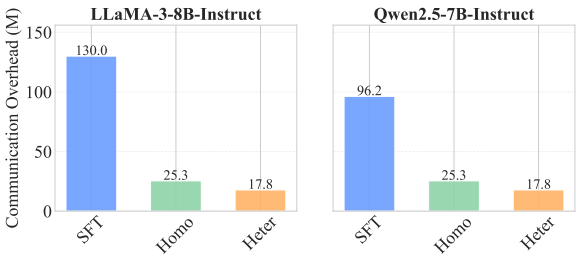

Figure 2: Analysis of communication efficiency in federated SFT and our FedCoT. "SFT" represents FedSFT/FedIT, "Homo" represents FedCoT with lora rank of 32, "Heter" represents FedCoT with lora rank of 4, 32, 32, 16, 4.

a training-free inference strategy via diverse sampling and majority voting. For federated learning baselines, we compare with **FedIT** (Zhang et al., 2024a) utilizing weighted averaging, **FedFFA-LoRA** (Sun et al., 2024) for parameter-efficient tuning, **FLORA** (Wang et al., 2024b) designed for heterogeneous LoRA rank adaptations, and **FedPrompt** (Zhao et al., 2023) for communication-efficient prompt tuning. Accuracy is adopted as the primary evaluation metric to be consistent with previous studies (Chen et al., 2024b). All evaluations measure accuracy under CoT prompting, which not only quantifies performance but also aligns with real-world medical demands for interpretability and response safety via inherent step-by-step reasoning.

5.2 Main Results

Overall Performance The overall results are presented in Table 2. FedCoT significantly and consistently outperforms all baseline methods across the five medical datasets on both backbone LLMs, demonstrating the superior efficacy of our framework.

Specifically, FedCoT achieves substantial average absolute improvements of 26.94% and 21.56% over the base `LLaMA-3-8B-Instruct` and `Qwen2.5-7B-Instruct` models with CoT prompting. Furthermore, our framework establishes a clear state-of-the-art by surpassing the strongest advanced federated fine-tuning baselines, outperforming FedPrompt by 9.40% on LLaMA-3 and FLORA by 6.29% on Qwen2.5.

While other parameter-efficient federated methods (e.g., FedIT, FedFFA-LoRA, and FLORA) yield moderate gains over the base models, they struggle to generalize evenly across diverse medical reasoning tasks. In stark contrast, FedCoT demonstrates remarkable stability and robust improvements across highly varied domains, from specialized literature (BioASQ, PubMedQA) to comprehensive clinical examinations (MedQA, MedMCQA). This validates that our dynamic reasoning path discrimination mechanism is strictly superior to generic federated fine-tuning for unlocking LLM reasoning capabilities under strict privacy constraints.

Efficiency Comparison The efficiency comparison is shown in Figure 2, which compares the

Method	**BioASQ**	**MedMCQA**	**MedQA**	**MMLU**	**PubMedQA**	**Avg.**
Best-of-N (Logits-based)	40.32	27.76	26.24	38.04	5.60	27.59
Best-of-N (Qwen3-0.6B-Reranker)	49.20	26.50	29.60	36.20	12.40	30.78
FedCoT (Local)	57.30	42.60	40.80	56.40	29.20	45.26
FedCoT (Qwen3-0.6B-Base)	50.00	43.70	37.00	49.70	13.20	38.72
FedCOT (ours)	68.50	45.20	54.10	68.70	41.00	55.50

Table 3: Performance of different filtering method on 8 candidates. FedCoT (ours) denotes the full federated framework. FedCoT (Local) represents models trained only on individual client datasets. FedCoT (Qwen3-0.6B-Base) replaces the default discriminator with a generative backbone. All results are accuracy (%).

Method	BioASQ	MedMCQA	MedQA	MMLU	PubMedQA	Avg.(%)
CoT	37.90	29.80	27.20	38.70	9.20	28.56
FedCoT (r=4,4,4,4,4)	67.70	44.50	52.60	65.00	40.80	54.12
FedCoT (r=8,8,8,8,8)	68.50	45.40	53.40	65.60	41.00	54.78
FedCoT (r=16,16,16,16,16)	69.40	45.00	54.30	65.60	41.00	55.06
FedCoT (r=32,32,32,32,32)	66.10	45.70	54.00	66.30	41.00	54.62
FedCoT (r=4,32,32,16,4)	68.50	45.20	54.10	68.70	41.00	55.50

Table 4: The different performances of FedCoT under different LoRA configurations. “r" represents the LoRA rank of different clients, corresponding to the clients of the datasets BioASQ, MedMCQA, MedQA, MMLU, and PubMedQA in sequence. The best results are in **Bold**.

communication efficiency between federated SFT methods (Fed-SFT/FedIT) and our FedCoT. During the federated learning process, the Federated SFT method performs LoRA fine-tuning on the LLM actor `LLaMA-3-8B-Instruct` and `Qwen2.5-7B-Instruct`. Here, the values 130M and 96M refer to the total number of parameters that need to be transmitted across all clients in the federated system during one global round of training. However, even with LoRA fine-tuning, such a volume of parameters still incurs large computational and communication overheads for low-resource clients and the whole federated learning system. In contrast, our FedCoT greatly reduces the training and communication overheads by fine-tuning a lightweight model. Thus, the parameter quantity of FedCoT during the federated learning process only accounts for 25.3M and 17.8M, which is much more efficient than the compared existing SFT methods, demonstrating its efficiency in low-resource environments.

5.3 Ablation Study

Candidate Processing To evaluate our data processing components, we conducted ablation studies by independently removing Candidate Selection and Soft Labeling (Table 5). Results show that omitting Selection introduces uninformative, imbalanced samples that degrade the discriminator's decision boundaries. Similarly, replacing Soft Labeling with binary labels fails to distinguish between rigorous reasoning and “lucky guesses." The full FedCoT framework achieves the highest accuracy, particularly on the MMLU dataset, confirming that refined, logically smoothed supervision is essential for complex multi-disciplinary reasoning.

Selection Strategies Table 3 compares the filtering performance of our proposed FedCoT against various reward modeling strategies. FedCoT consistently achieves the highest accuracy (55.50% on average), significantly outperforming both the vanilla Logits-based baseline and specialized reranking models like Qwen3-0.6B-Reranker. Notably, our BERT-based federated approach surpasses its local counterpart, demonstrating the necessity of cross-client knowledge aggregation. Furthermore, although larger generative backbones (e.g., Qwen3-0.6B) or SOTA open-source reward models (detailed in Appendix X) were tested, they failed to match our performance. This highlights that in low-resource scenarios, evaluating the complex deep semantics of

Method	BioASQ		MedMCQA		MedQA		MMLU		PubMedQA		Avg.
	Acc. (%)	Size	Acc. (%)	Size	Acc. (%)	Size	Acc. (%)	Size	Acc. (%)	Size	Acc. (%)
FedCoT (Full)	68.50	364	45.20	1420	54.10	7278	68.70	915	41.00	205	55.50
w/o Candidate Selection	69.40	494	44.90	2000	53.70	10178	62.90	1299	41.00	450	54.24
w/o Soft Labeling	67.70	364	45.10	1420	54.50	7278	60.30	915	41.00	205	53.10
w/o CS & SL	65.30	494	45.20	2000	56.10	10178	54.00	1299	41.00	450	52.32

Table 5: Ablation study of FedCoT components. "Size" denotes the number of candidate paths retained for local discriminator training.

small models may be less effective than assessing their relatively shallow semantics.

5.4 Analysis of Heterogeneous LoRA Configurations

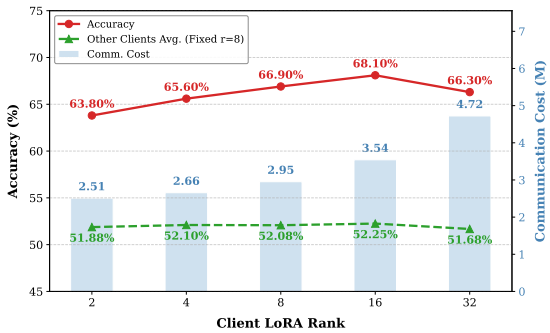

Figure 3: Impact of varying the MMLU client's LoRA rank on local accuracy, communication cost, and the average performance of other fixed-rank clients ($r = 8$).

In real-world medical federated networks, clients often face extreme heterogeneity not only in local computational budgets and network bandwidths but also in the intrinsic complexity of their respective clinical tasks. Forcing a homogeneous LoRA rank across all clients is heavily suboptimal, which wastes communication bandwidth and risks overfitting on simpler tasks, while creating severe capacity bottlenecks for complex medical reasoning tasks.

To investigate how the system dynamically responds to asymmetric resource constraints, we conduct a controlled stress test (Figure 3). We fix the LoRA rank of four clients at $r = 8$ while drastically varying the rank of the MMLU client from $r = 2$ to $r = 32$. As the results illustrate, reducing the MMLU rank from 16 down to 2 significantly cuts its local communication cost from 3.54M to 2.51M parameters per round. Despite this aggressive downscaling, the client suffers only a smooth, acceptable degradation in accuracy from 68.10% to 63.80%. More importantly, throughout this compression phase, the average performance of the other four clients remains remarkably stable at approximately 52.10%. This stability empirically proves that the adopted modular stacking mechanism effectively isolates local capability bottlenecks, preventing severe low-rank truncation noise from poisoning the global aggregated knowledge. Conversely, naively amplifying the MMLU rank to 32 maximizes the communication burden (4.72M) but triggers a counterintuitive overfitting penalty. Rather than improving reasoning capabilities, this over-parameterization not only degrades MMLU's local accuracy to 66.30% but also inadvertently drags down the average performance of the remaining network to 51.68%.

Guided by these systemic insights, a practical federated framework must flexibly accommodate heterogeneous ranks to adapt to diverse real-world conditions, such as varying computational budgets and task complexities. Enforcing a uniform rank across all clients inevitably forces a rigid compromise between communication efficiency and local model capacity. As shown in Table 4, by allowing clients to independently operate with varied ranks (e.g., $r \in \{4, 16, 32\}$) tailored to their specific local constraints, FedCoT seamlessly integrates these asymmetric updates. Remarkably, this flexible heterogeneous configuration eliminates the unrealistic need for strict system-wide uniformity. This demonstrates that FedCoT elegantly adapts to diverse and constrained client environments while maximizing global reasoning stability.

6 Conclusion

We propose FedCoT, a federated reasoning enhancement framework that optimizes LLM performance under strict privacy and resource constraints. By combining a lightweight discriminator for candidate selection with LoRA-based fine-tuning and weighted server aggregation, FedCoT addresses the challenges of limited reasoning capability and high communication overhead.

Limitations and Future Work

This study is constrained by available compute and deployment resources: we evaluate FedCoT mainly in cross-silo biomedical QA with moderate client scales and model sizes, rather than in broader or more demanding settings (e.g., large-scale or cross-device deployments). In addition, our discriminator is trained with final-answer supervision, which effectively selects high-quality reasoning paths but does not provide a fully granular, process-level assessment; incorporating step-wise evaluation and supervision is left for future work.

Ethics Statements

Our framework preserves data locality and shares only aggregated updates. Real-world deployment should incorporate differential privacy and regulatory compliance. This is a research prototype and not a clinical decision tool; any medical application requires rigorous validation and professional oversight. All experiments use publicly available data following their usage agreements, and no patient privacy or personal information is used throughout this process.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Alice Chen. 2024. *Policy-Based Access Control in Federated Clinical Question Answering*. Massachusetts Institute of Technology.

Cen Chen, Tiandi Ye, Li Wang, and Ming Gao. 2022. Learning to generalize in heterogeneous federated networks. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 159–168.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2024a. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024b. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2023. A customized text sanitization mechanism with differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5747–5758.

Wenqing Chen, Weicheng Wang, Zhixuan Chu, Kui Ren, Zibin Zheng, and Zhichao Lu. 2024c. Self-para-consistency: Improving reasoning tasks at low cost for large language models. In *62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pages 14162–14167. Association for Computational Linguistics.

Xin Chen, Hanxian Huang, Yanjun Gao, Yi Wang, Jishen Zhao, and Ke Ding. 2024d. Learning to maximize mutual information for chain-of-thought distillation. *arXiv preprint arXiv:2403.03348*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Elias Dritsas and Maria Trigka. 2025. Federated learning for iot: A survey of techniques, challenges, and applications. *Journal of Sensor and Actuator Networks*, 14(1):9.

Tao Fan, Guoqiang Ma, Yan Kang, Hanlin Gu, Yuanfeng Song, Lixin Fan, Kai Chen, and Qiang Yang. 2024. Fedmkt: Federated mutual knowledge transfer for large and small language models. *arXiv preprint arXiv:2406.02224*.

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

YiQiu Guo, Yuchen Yang, Zhe Chen, Pingjie Wang, Yusheng Liao, Ya Zhang, Yanfeng Wang, and Yu Wang. 2025b. Dsvd: Dynamic self-verify decoding for faithful generation in large language models. *arXiv preprint arXiv:2503.03149*.

Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. 2024. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*.

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*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Chao Huang, Ming Tang, Qian Ma, Jianwei Huang, and Xin Liu. 2023. Promoting collaboration in cross-silo federated learning: Challenges and opportunities. *IEEE Communications Magazine*, 62(4):82–88.

Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. 2021. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7865–7873.

Ahmed Imteaj, Khandaker Mamun Ahmed, Urmish Thakker, Shiqiang Wang, Jian Li, and M Hadi Amini. 2022. Federated learning for resource-constrained iot devices: Panoramas and state of the art. *Federated and transfer learning*, pages 7–27.

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 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 (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv:2305.14045*.

Mounssif Krouka, Anis Elgabli, Chaouki Ben Issaid, and Mehdi Bennis. 2021. Communication-efficient and federated multi-agent reinforcement learning. *IEEE Transactions on Cognitive Communications and Networking*, 8(1):311–320.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2019. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*.

Yiwei Li, Ji Zhang, Shaoxiong Feng, Peiwen Yuan, Xinglin Wang, Jiayi Shi, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, et al. 2025. Revisiting self-consistency from dynamic distributional alignment perspective on answer aggregation. *arXiv preprint arXiv:2502.19830*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. 2025. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*.

Ken Liu, Shengyuan Hu, Steven Z Wu, and Virginia Smith. 2022. On privacy and personalization in cross-silo federated learning. *Advances in neural information processing systems*, 35:5925–5940.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

Yulong Mao, Kaiyu Huang, Changhao Guan, Ganglin Bao, Fengran Mo, and Jinan Xu. 2024. Dora: Enhancing parameter-efficient fine-tuning with dynamic rank distribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Inderjeet Nair and Lu Wang. 2024. Midgard: Self-consistency using minimum description length for structured commonsense reasoning. *arXiv preprint arXiv:2405.05189*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Jiaju Qi, Qihao Zhou, Lei Lei, and Kan Zheng. 2021. Federated reinforcement learning: Techniques, applications, and open challenges. *arXiv preprint arXiv:2108.11887*.

Yuwen Qian, Luping Rao, Chuan Ma, Kang Wei, Ming Ding, and Long Shi. 2024. Toward efficient and secure object detection with sparse federated training over internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 25(10):14507–14520.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

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, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Haotian Sun, Hang Wu, Carl Yang, and May D Wang. 2024. Medadapter: Efficient test-time adaptation of large language models towards medical reasoning. *arXiv preprint arXiv:2405.03000*.

InSeo Song and KangYoon Lee. 2025. Optimizing communication and performance in federated learning for large language models. In *2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 0964–0967. IEEE.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. 2024. Improving lora in privacy-preserving federated learning. *arXiv preprint arXiv:2403.12313*.

Ming Tang and Vincent WS Wong. 2021. An incentive mechanism for cross-silo federated learning: A public goods perspective. In *IEEE INFOCOM 2021-IEEE conference on computer communications*, pages 1–10. IEEE.

Asadullah Tariq, Mohamed Adel Serhani, Farag Sallabi, Tariq Qayyum, Ezedin S Barka, and Khaled A Shuaib. 2023. Trustworthy federated learning: A survey. *arXiv preprint arXiv:2305.11537*.

Asadullah Tariq, Mohamed Adel Serhani, Farag M Sallabi, Ezedin S Barka, Tariq Qayyum, Heba M Khater, and Khaled A Shuaib. 2024. Trustworthy federated learning: A comprehensive review, architecture, key challenges, and future research prospects. *IEEE Open Journal of the Communications Society*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Chunlin Tian, Zhan Shi, and Li Li. 2023. Learn to select: Efficient cross-device federated learning via reinforcement learning.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. 2024. Reasoning aware self-consistency: Leveraging reasoning paths for efficient llm sampling. *arXiv preprint arXiv:2408.17017*.

Haozhao Wang, Yabo Jia, Meng Zhang, Qinghao Hu, Hao Ren, Peng Sun, Yonggang Wen, and Tianwei Zhang. 2024a. Feddse: Distribution-aware sub-model extraction for federated learning over resource-constrained devices. In *Proceedings of the ACM Web Conference 2024*, pages 2902–2913.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. Scott: Self-consistent chain-of-thought distillation. *arXiv preprint arXiv:2305.01879*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. 2024b. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *arXiv preprint arXiv:2409.05976*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shuyue Wei, Yongxin Tong, Zimu Zhou, Yi Xu, Jingkai Gao, Tongyu Wei, Tianran He, and Weifeng Lv. 2025. Federated reasoning llms: a survey. *Frontiers of Computer Science*, 19(12):1912613.

Huiwen Wu, Cen Chen, and Li Wang. 2020. A theoretical perspective on differentially private federated multi-task learning. *arXiv preprint arXiv:2011.07179*.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*.

Yebo Wu, Chunlin Tian, Jingguang Li, He Sun, Kahou Tam, Li Li, and Chengzhong Xu. 2025. A survey on federated fine-tuning of large language models. *arXiv preprint arXiv:2503.12016*.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Xie. 2023. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 36:41618–41650.

Tiandi Ye, Senhui Wei, Jamie Cui, Cen Chen, Yingnan Fu, and Ming Gao. 2023. Robust clustered federated learning. In *International Conference on Database Systems for Advanced Applications*, pages 677–692. Springer.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. 2024a. Towards building the federatedgpt: Federated instruction tuning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6915–6919. IEEE.

Jingbo Zhang, Qiong Wu, Pingyi Fan, and Qiang Fan. 2024b. A comprehensive survey on joint resource allocation strategies in federated edge learning. *arXiv preprint arXiv:2410.07881*.

Xinlu Zhang, Shiyang Li, Xianjun Yang, Chenxin Tian, Yao Qin, and Linda Ruth Petzold. 2023. Enhancing small medical learners with privacy-preserving contextual prompting. *arXiv preprint arXiv:2305.12723*.

Haodong Zhao, Wei Du, Fangqi Li, Peixuan Li, and Gongshen Liu. 2023. Fedprompt: Communication-efficient and privacy-preserving prompt tuning in federated learning. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1–5. IEEE.

Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. 2025. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. *arXiv preprint arXiv:2504.21277*.

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 Algorithm

Here, we present the pseudo-code of the entire FedCoT method, see Algorithm 1:

Algorithm 1: FedCoT Algorithm

Input: Total rounds T; Local training epochs E; Client datasets $\{\mathcal{D}_i\}_{i=1}^{N}$; Pretrained model d_θ; Number of candidates K

for $t = 1$ **to** T **do**
  foreach *client* i **do**
    Receive last aggregated adapter $\mathbf{W}^{t-1}$ $(\mathbf{A}_i^t, \mathbf{B}_i^t, \mathbf{W}_i^{cls}) \leftarrow$ *LocalUpdates*$(i, \mathbf{W}^{t-1})$;
  end
  Aggregate LoRA modules and classifier using Equation 8, 9 to get adapter $\mathbf{W}^t$;
end
Function *LocalUpdates(i,* $\mathbf{W}$*):*
  for $e = 1$ **to** E **do**
    Setup local discriminator from d_θ with $\mathbf{W}$ and apply local LoRA modules $\mathbf{A}_i, \mathbf{B}_i$ and classifier ;
    Generate reasoning paths $\{(\tau_{j,k}, \hat{y}_{j,k})\}_{k=1}^{K}$;
    Concatenate feature vector $\mathbf{h}_{j,k}$;
    Predict using Equation 10 ;
    Update LoRA modules and classifier using Equation 5;
  end
  Return updated LoRA modules and classifier to server;
foreach *test sample* x_j **do**
  Generate reasoning paths $\{(\tau_{j,k}, \hat{y}_{j,k})\}_{k=1}^{K}$;
  Concatenate feature vector $\mathbf{h}_{j,k} = [x_j \parallel \tau_{j,k} \parallel \hat{y}_{j,k}]$;
  Select answer via Equations 10, 11;
end
Output: Final answer $\{\hat{y}_j\}$

B Analysis of Different Context Length

To address potential concerns regarding truncation effects under the 512-token constraint (e.g., premature termination before generating complete answer), we extended the maximum generation length to 1024 tokens. This intervention effectively eliminated truncation issues, as evidenced by the near-zero truncation rates in Table 6.

Max Length	Actor Model	Avg.(%)
512	LLaMA-3-8B-Instruct	5.93
	Qwen2.5-7B-Instruct	26.28
1024	LLaMA-3-8B-Instruct	0.02
	Qwen2.5-7B-Instruct	0.04

Table 6: The mean truncation rates on all datasets across maximum generation lengths.

Performance comparisons in Table 7 reveal consistent accuracy improvements across most configurations after length extension. Our FedCoT method achieved gains of +3.04% and +1.00% for `LLaMA-3-8B-Instruct` and `Qwen2.5-7B-Instruct`, respectively, while FedIT showed a +0.43% improvement for `LLaMA-3-8B-Instruct`. The marginal decline observed for FedIT on `Qwen2.5-7B-Instruct` (-0.27%) may be attributable to increased noise in extended reasoning chains.

Critically, FedCoT demonstrates robust effectiveness when controlling for truncation effects, delivering substantial improvements of +9.30% and +9.16% over FedIT baselines for `LLaMA-3-8B-Instruct` and `Qwen2.5-7B-Instruct`, respectively token length with 1024. These gains exceed those observed under 512-token constraints, validating both the efficacy and robustness of our approach across generation length parameters.

C Discussion on Fine-grained Process-oriented Discrimination

While our primary framework relies on outcome-based labels for path discrimination, we investigate whether process-oriented evaluation could provide more fine-grained signals. Inspired by process reward models (Lightman et al., 2023) in reinforcement learning, we design a step-wise self-evaluation mechanism, where each client model assigns confidence scores to its intermediate reasoning steps.

Surprisingly, as shown in Table 8, models exhibit strong positivity bias, with self-rated step accuracy ranging from 81.63% to 99.98% across all datasets. This overconfidence persists even in cases where the final answer accuracy is critically low (e.g., only 9.20% on PubMedQA), indicating that self-assessment fails to distinguish correct

Model	Method	BioASQ	MedMCQA	MedQA	MMLU	PubMedQA	Avg.(%)
LLaMA-3-8B-Instruct	FedIT-512	42.74	47.29	53.73	71.17	13.20	45.63
	FedCoT-512	65.30	45.20	56.10	54.00	41.00	52.32
	FedIT-1024	45.16	46.45	54.36	70.55	13.80	46.06
	FedCoT-1024	78.20	44.40	53.70	58.30	42.20	**55.36**
Qwen2.5-7B-Instruct	FedIT-512	82.26	48.48	44.30	68.71	47.20	58.19
	FedCoT-512	96.80	50.00	52.50	66.30	64.80	66.08
	FedIT-1024	82.26	48.55	43.28	68.10	47.40	57.92
	FedCoT-1024	97.60	51.60	56.50	59.50	70.20	**67.08**

Table 7: Performance comparison (%) across different maximum generation token lengths. The best results are in **Bold** and the second-highest results are indicated with an underline.

Dataset	Positive	Negative	Ratio(%)
BioASQ	5,740	145	97.54
MedMcQA	47,290	5,408	89.74
MedQA	118,719	26	99.98
MMLU	36,676	979	97.40
PubMedQA	3,288	740	81.63

Table 8: Step-wise self-evaluation performance across medical QA benchmarks. Positive: count of reasoning steps judged correct by the model; Negative: count of steps judged incorrect; Ratio: proportion of correct self-evaluation.

from flawed reasoning paths. This may be because models lack reliable internal uncertainty estimation for intermediate steps, and the self-evaluation task, being trained on the same data distribution, inherits the model's existing biases.

D Performance of General Open-source Reward Model

We further verify the applicability of general-domain reward models (Liu et al., 2025) in medical task scenarios-specifically focusing on their ability to filter and select high-quality reasoning traces in medical scenarios, and the results are shown in Table 9. It can be seen that both the general reward models with small parameter counts and those with large parameter counts performed very poorly, achieving even worse results than direct CoT prompting. This reflects the special nature of medical scenarios; crudely and directly utilizing these models cannot transfer their excellent performance from general scenarios.

E Case Study

This case in Table 10 involves a 29-year-old man with burning urination (urethritis), acute asymmetric joint pain (right ankle, left knee), bilateral conjunctivitis, and recent antibiotic-treated bloody diarrhea.

The question asks for the most likely additional finding among four options. The correct answer is B (Tenderness at the insertion of the Achilles tendon), indicative of reactive arthritis triggered by enteric infection (e.g., Shigella/Salmonella).

Generation 1 incorrectly prioritized finger joint pain despite the patient's ankle pain being the critical clue. Reactive arthritis typically involves lower extremities (e.g., Achilles enthesitis), not finger joints. Generation 2 ignored the 2-week latency between diarrhea and joint symptoms—a hallmark of reactive (not septic) arthritis.

Generation 3 correctly diagnosed reactive arthritis and prioritized Achilles tenderness (B) as the key additional finding, aligning with the patient's ankle pain

Other generation all made wrong answers.

The discriminator model correctly assigned the highest score (0.834) to generation 3 for its precise pathophysiology: linking Achilles tenderness to reactive arthritis.

Other generations scored 0.62–0.79 for plausible-but-incorrect "abdominal infection" theories, yet the model still ranked them below the correct answer.

F Improvement Potential

Table 11 shows the improvement of the test set of multiple models with three different parameter levels under multiple sampling, where the models with less than 3B parameters can hardly

Model	BioASQ	MedMCQA	MedQA	MMLU	PubMedQA
Baseline	37.90	29.80	27.20	38.70	9.20
Skywork-Reward-V2-Qwen3-0.6B	31.50	24.50	20.20	32.50	4.00
Skywork-Reward-V2-Llama-3.2-1B	41.90	23.90	25.80	28.80	6.40
Skywork-Reward-V2-Qwen3-1.7B	25.00	24.80	21.20	26.40	1.00
Skywork-Reward-V2-Llama-3.1-8B	16.10	12.80	7.60	16.00	0.60

Table 9: Performance of open-source SOTA reward models in the medical domain across different benchmarks (in %).

achieve performance improvement through multi-sampling. The model with 3B parameters shows a significantly higher improvement under multiple samplings than models with less than 3B parameters, and its improvement performance is relatively close compared to that of the model with 7B parameters.

In addition, we can see that in the case of less than 3B, the Qwen series of models can still achieve a slight improvement, while the LLaMA series of models basically have no improvement at all. This may be due to different treatments during pre-training.

Overall, whether for the Qwen series or the LLaMA series, the larger the model parameter count, the better the basic performance under the corresponding dataset conditions, and the greater the improvement obtained through sampling. This reflects that the basic capabilities of the model are a very crucial part for further subsequent improvement. If the basic capabilities of the base model are too low, then even more sampling cannot achieve significant improvement.

G Datasets Information

We will showcase the medical datasets used for generating CoT candidates during the training phase and the medical datasets used for evaluation during the testing phase here.

- PubMedQA (Jin et al., 2019): A biomedical question-answering dataset based on PubMed abstracts, containing 1,000 expert-annotated multiple-choice (yes/no/maybe) biomedical question-answering samples. Its knowledge source involves 211,300 PubMed papers. This task requires judging the answer to the research question based on the abstract content. The average lengths of the question stem and context are 14.4 and 238.9 tokens respectively. It is distributed under MIT license.

- BioASQ (Tsatsaronis et al., 2015): A large-scale biomedical semantic indexing and question-answering dataset, including two tasks: information retrieval (Task A) and machine reading comprehension (Task B). This paper focuses on 618 binary judgment (yes/no) reading comprehension questions in Task B in the recent five years (2019 - 2023). The average length of the question tokens is 17. Similar to PubMedQA, this dataset supports machine reading comprehension research by providing standard answer text segments. It is distributed under the CC-BY-4.0 license.

- MMLU-Med (Hendrycks et al., 2020): A medical reasoning specialized dataset extracted from the multi-task language understanding benchmark MMLU (hereinafter referred to as MMLU for short). In this paper's experiments, seven medical-related fields are selected: clinical knowledge, college biology, college medicine, high school biology, medical genetics, professional medicine, and virology. The reasoning ability of the model in professional medical scenarios is mainly investigated. It is distributed under MIT license.

- MedMCQA (Pal et al., 2022): A large-scale multiple-choice medical question-answering dataset that integrates real question resources from the All India Institute of Medical Sciences (AIIMS) and the National Eligibility cum Entrance Test for Postgraduate (NEET-PG) in India. This dataset contains over 194,000 high-quality medical questions, covering 2,400 health topics and 21 medical discipline areas, featuring significant topic di-

Question:A previously healthy 29-year-old man comes to the emergency department because of burning with urination for several days. He has also had pain in the right ankle for 3 days and pain and swelling in the left knee for 1 day. Two weeks ago, he had several days of fever and bloody diarrhea, for which he was treated with antibiotics. Examination shows a small left knee effusion and bilateral conjunctival injection. Which of the following is the most likely additional finding in this patient?\n A. Palpable mass in the right lower quadrant\n B. Tenderness at the insertion of the Achilles tendon\n C. Pain on passive extension of the fingers\n D. Circular erythematous rash with central clearing **Ground-Truth Answer**: **B**;

Generation 1: Here is the step-by-step reasoning to answer the question:\n\nThe patient presents with a combination of symptoms, including burning with urination, pain in the right ankle, and pain and swelling in the left knee. These symptoms suggest a systemic infection, rather than a localized issue.\n\nThe patient's recent history of fever and bloody diarrhea, which was treated with antibiotics, is likely related to an underlying infection. The development of new symptoms, such as joint pain and swelling, suggests that the infection may have spread to other parts of the body.\n\nThe patient's conjunctival injection (inflammation of the conjunctiva) is also a significant finding, as it is a common manifestation of disseminated infection.\n\nGiven the patient's symptoms and medical history, the most likely diagnosis is reactive arthritis, which is a type of spondyloarthropathy that occurs in response to an underlying infection, often gastrointestinal.\n\nReactive arthritis typically affects the joints, eyes, and skin, and can cause symptoms such as joint pain, swelling, and inflammation. The patient's bilateral conjunctival injection and left knee effusion are consistent with this diagnosis.\n\nThe most likely additional finding in this patient would be:\n\n#### C. Pain on passive extension of the fingers\n\nThis symptom is a common manifestation of reactive arthritis, as the condition can cause inflammation and stiffness in the hands and fingers. Passive extension of the fingers would likely cause pain and stiffness, which would be consistent with the patient's underlying diagnosis.\n\nThe other options are not as likely:\n\nA. Palpable mass in the right lower quadrant: While the patient has symptoms in the right ankle, there is no indication of a palpable mass in this area.\n\nB. Tenderness at the insertion of the Achilles tendon: Achilles tendonitis is a possible diagnosis, but it is not the most likely explanation for the patient's symptoms given the patient's medical history and conjunctival injection.\n\nD. Circular erythematous rash with central clearing: While a rash could be a symptom of reactive arthritis, it is not the most likely finding in this patient given the patient's other symptoms and medical history.
Prediction 1: **C**; **Score 1:** 0.77344

Generation 2: Here is the step-by-step reasoning to answer the question:\n\nThe patient presents with a complex set of symptoms, including burning with urination, pain in the right ankle, and pain and swelling in the left knee. The patient also reports a history of fever and bloody diarrhea two weeks ago, which was treated with antibiotics.\n\nStep 1: Identify the possible underlying causes of the patient's symptoms.\n\nThe patient's symptoms suggest a possible infection, given the presence of fever, bloody diarrhea, and pain and swelling in the joints. The antibiotics given two weeks ago may have contributed to the development of a resistant infection.\n\nStep 2: Consider the possible sources of the infection.\n\nGiven the patient's history of bloody diarrhea, it is possible that the infection originated in the gastrointestinal tract. The patient's symptoms of burning with urination and bilateral conjunctival injection suggest that the infection may have spread to other parts of the body, such as the urinary tract and eyes.\n\nStep 3: Evaluate the patient's symptoms in relation to the possible sources of infection.\n\nThe patient's symptoms of pain and swelling in the left knee and right ankle suggest that the infection may have spread to the joints. The patient's small left knee effusion and bilateral conjunctival injection support this possibility.\n\nStep 4: Consider the possible causes of the patient's symptoms in the context of the possible sources of infection.\n\nGiven the patient's history of bloody diarrhea and the presence of symptoms in the joints and eyes, it is possible that the patient has developed a septic arthritis, which is an infection of the joint that is caused by bacteria that have spread from another part of the body.\n\nStep 5: Evaluate the patient's symptoms in relation to the possible causes of septic arthritis.\n\nThe patient's symptoms of pain and swelling in the joints, as well as the presence of a small left knee effusion, are consistent with septic arthritis. The patient's bilateral conjunctival injection also suggests that the infection may have spread to the eyes.\n\nStep 6: Consider the most likely additional finding in the patient.\n\nGiven the patient's symptoms and the possible causes of septic arthritis, the most likely additional finding is a palpable mass in the right lower quadrant, which could indicate an abscess or a collection of pus in the right lower quadrant of the abdomen.\n\n#### A. Palpable mass in the right lower quadrant
Prediction 2: **A**; **Score 2:** 0.717285

Generation 3: Here is the step-by-step reasoning:\n\nThe patient presents with a combination of symptoms, including burning with urination, pain in the right ankle, and pain and swelling in the left knee. This suggests that the patient may have a systemic infection that is affecting multiple parts of the body.\n\nThe patient had a recent episode of bloody diarrhea, which is a common symptom of gastrointestinal infections. The fact that he was treated with antibiotics suggests that the infection was likely bacterial in nature.\n\nThe patient's symptoms are consistent with a urinary tract infection (UTI), as burning with urination is a common symptom of UTIs. The presence of a small left knee effusion and bilateral conjunctival injection suggests that the infection may have spread to the joints and eyes, respectively.\n\nThe patient's symptoms are also consistent with reactive arthritis, a condition that can occur as a complication of certain bacterial infections, such as Salmonella or Shigella. Reactive arthritis is characterized by joint pain and swelling, as well as conjunctivitis and urethritis.\n\nGiven the patient's symptoms and medical history, the most likely additional finding is:\n\nB. Tenderness at the insertion of the Achilles tendon\n\nThis is because reactive arthritis often affects the joints, including the ankles, and the Achilles tendon is a common site of tenderness in this condition.\n\n#### B. Tenderness at the insertion of the Achilles tendon
Prediction 3: **B**; **Score 3:** 0.833984

Generation 4: .....
Prediction 4: **A**; **Score 4:** 0.621094

Generation 5: .....
Prediction 5: **A**; **Score 5:** 0.785645

Generation 6: .....
Prediction 6: **A**; **Score 6:** 0.781250

Generation 7: .....
Prediction 7: **A**; **Score 7:** 0.777344

Generation 8: .....
Prediction 8: **C**; **Score 8:** 0.717285

Table 10: A case study on MedQA dataset.

Model	bioasq		medmcqa		medqa		mmlu		pubmedqa		average		Δ
	p@1	p@k	p@1	p@k	p@1	p@k	p@1	p@k	p@1	p@k	p@1	p@k	
LLaMA-3-8B	37.90	70.97	29.84	68.18	27.18	70.46	38.65	71.78	9.20	42.00	28.55	64.68	36.12
LLaMA-3.1-8B	25.81	64.52	35.07	72.22	32.05	73.53	39.26	77.30	15.20	54.80	29.48	68.47	38.99
Qwen2.5-7B	73.39	98.39	43.75	70.98	29.46	66.54	50.31	78.53	38.80	73.40	47.14	77.57	30.43
Qwen2.5-3B	10.48	46.77	18.98	55.41	5.03	23.80	26.38	65.64	1.20	7.40	12.41	39.81	27.39
LLaMA3.2-3B	31.45	70.16	31.41	65.57	20.03	59.62	32.52	70.55	4.80	24.20	24.04	58.02	33.98
Qwen-1.5B	0.81	9.68	1.65	11.09	1.73	15.48	0.61	10.43	0.00	0.40	0.96	9.41	8.46
LLaMA3.2-1B	0.00	0.00	0.10	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.07	0.05
Qwen2.5-0.5B	0.00	0.81	1.63	8.92	0.47	5.89	1.23	7.98	0.60	3.20	0.78	5.36	4.57

Table 11: Model performance (in %) on biomedical test sets demonstrating accuracy potential through multiple sampling. The table shows pass rates at first sample (p@1) and after k samples (p@k), with Δ = p@k − p@1 indicating accuracy improvement potential.

versity. The average token length of the questions is 12.77. It is distributed under Apache-2.0 license.

- MedQA (Jin et al., 2021): A clinical medicine question bank based on the United States Medical Licensing Examination (USMLE), containing 12,723 questions sourced from 18 widely used authoritative clinical medicine textbooks. The questions cover a rich variety of clinical medicine topics and require professional-level reasoning by integrating multi-source evidence. The average length of the questions is 116.6 tokens, and the average length of the options is 3.5 tokens. It is distributed under the CC-BY-4.0 license.

- GPQA (Medical) (Rein et al., 2024): As the biomedical domain subset derived from the Graduate-Level QA dataset, GPQA (Medical) is characterized by expert-curated multiple-choice questions of high difficulty. These questions are specifically designed to necessitate advanced reasoning capabilities, with an intentional design objective to avoid being answerable through simple search-engine queries—thereby eliminating over-reliance on surface-level information retrieval and emphasizing the need for in-depth domain understanding.

- MedBullets(Chen et al., 2024a): This dataset comprises practice questions curated from the MedBullets medical education platform, a specialized resource for medical learning and skill assessment. Owing to their alignment with high-difficulty medical knowledge application scenarios, these subsets function as rigorous evaluation benchmarks for assessing the performance of models in the biomedical domain.

- MedXpertQA (Zuo et al., 2025): This dataset consists of 50 medical reasoning questions, all of which are authored by domain experts and structured to require multi-step analytical processes. The expert-authored nature of the questions ensures high validity and authenticity of the scenarios, while the multi-step reasoning framework aligns with the complexity of real-world clinical decision-making, thereby enabling rigorous, nuanced assessments of model performance in biomedical reasoning tasks. It is distributed under MIT license.

H Prompt Template

We present the one-shot CoT prompt for generating candidate reasoning paths in Figure 4. This carefully designed template guides the model in rigorous reasoning and follows required formatting for easy answer extraction and evaluation. We also presented the prompt template used by LLM-as-Judge to evaluate the factual fidelity, explainability, and overall quality of reasoning paths, as shown in Figure 5, with evaluation scores ranging from 1 to 10.

The Prompt Template for one-shot CoT

To answer the following question, provide a highly detailed and comprehensive step-by-step reasoning.
Your explanation should break down the problem into clear, logical steps, with each step on a new line.
The goal is to construct a robust and self-contained argument that fully justifies the final conclusion.
After the reasoning, present the final answer.
Please model the overall structure on the provided example.

Example:
Q:
Primary Colonization of plaque is dominated by
A. Facultative Aerobic Gram+ve rods
B. Facultative Anaerobic Gram-ve rods
C. Facultative Aerobic Gram-ve cocci
D. Facultative Anaerobic Gram+ve cocci

A:
The oral cavity is a complex ecosystem where various microorganisms coexist, and plaque is a biofilm that forms on tooth surfaces.
Primary colonization of plaque is dominated by Streptococcus species, which are Gram-positive, facultative anaerobic cocci.
These Streptococcus species are able to adhere to the tooth surface and multiply, creating a matrix for other microorganisms to colonize.
D. Facultative Anaerobic Gram+ve cocci

Here is the question:
Q:
`{question}`
`{options}`

Figure 4: Template for Multiple-choice Questions (MCQs). Placeholders `{question}` and `{options}` denote the MCQ stem and options.

The Prompt Template for 3 dimensional judgments.

You are a medical expert evaluating AI responses to medical questions. Please evaluate the following response based on three criteria and provide scores from 1-10 (10 being best).
Medical Question: `{question}`
AI Response: `{response}`

Evaluation Criteria:
1. FACTUAL FAITHFULNESS (1-10): Assess whether the medical information is accurate, evidence-based, and consistent with established medical knowledge. Penalize hallucinations, unsupported claims, or factual inaccuracies.
2. EXPLANATORY QUALITY (1-10): Evaluate how well the response explains medical concepts, provides reasoning, and makes complex information understandable. Consider clarity, logical flow, and educational value.
3. OVERALL QUALITY (1-10): Judge the comprehensive quality including completeness, relevance to the question, professionalism, and practical usefulness for medical understanding.

Output Format - Provide your evaluation in this exact JSON format:
{{
'factual_faithfulness': {{
'score': 7,
'reasoning': 'Brief explanation of strengths and weaknesses in factual accuracy' }},
'explanatory_quality': {{
'score': 8,
'reasoning': 'Brief analysis of explanatory effectiveness'
}},
'overall_quality': {{
'score': 7,
'reasoning': 'Overall assessment of the response's medical value'
}}
}}
Ensure scores are integers from 1-10. Be strict but fair in medical accuracy assessment.

Figure 5: Template for 3 dimensional judgments. Placeholders `{question}` and `{response}` denote the medical question and response.

I Analysis of Candidate Sampling Numbers

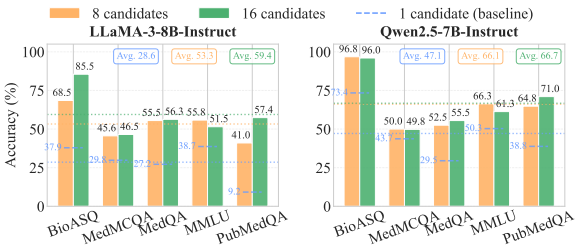

Figure 6: Performance improvement on difference candidate numbers of FedCoT. Different shapes represent different datasets.

In candidates generation, the sampling number of candidates is sensitive to model performance. Our experimental analysis, illustrated in Figure 6, demonstrates that increasing candidate samples from 8 to 16 consistently enhances performance across models. This improvement is particularly pronounced for `LLaMa-3-8B-Instruct`, where average accuracy increases from 52.32% to 59.44%. Notably, on the BioASQ dataset, its accuracy rises substantially from 65.30% to 85.50%.

Although `Qwen2.5-7B-Instruct` also exhibits improvement gains (from 66.08% to 66.72%), the marginal improvement suggests model-dependent sensitivity to candidate quantity. In particular, the performance gap between 8 and 16 candidates is significantly smaller than that between 1 and 8 candidates for all backbone models, indicating that 8 candidates sufficiently demonstrate methodological efficacy.

J Assessing via LLM-as-Judge: Factuality, Explainability, Quality

We assess response quality using an LLM-as-Judge protocol with a strong judge model (`DeepSeek-V3.2-Exp`), which scores generated solutions along three axes-medical factuality, explainability/coherence, and overall quality-on a 0-10 scale; prompt details and rubric are in Appendix H. Figure 7 summarizes results on PubMedQA, MedMCQA, MedQA, and MMLU. FedCoT consistently improves over direct CoT across all datasets and axes. The largest improvements appear on MMLU (+0.31 factuality, +0.25 quality) and MedMCQA (+0.18 factuality, +0.16 explainability), while PubMedQA shows smaller but uniformly positive gains (+0.04 factuality, +0.01 explainability, +0.03 quality). These findings align with the accuracy improvements in Section 5.2 and indicate that, despite using only final correctness as supervision during federated training, FedCoT effectively steers inference-time selection toward higher-quality reasoning chains.

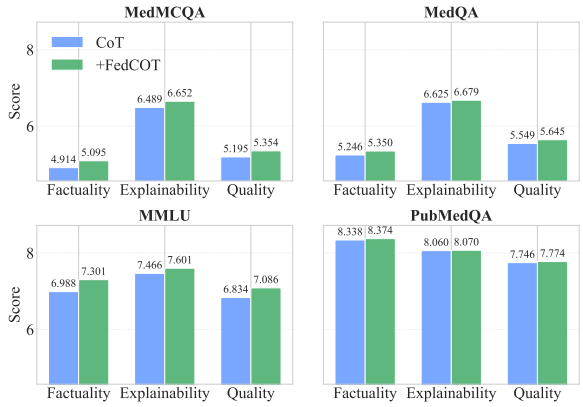

Figure 7: LLM-as-Judge evaluation on four biomedical QA datasets across three axes (higher is better): medical factuality, explainability/coherence, and overall quality.

K Performance on Smaller Size LLMs

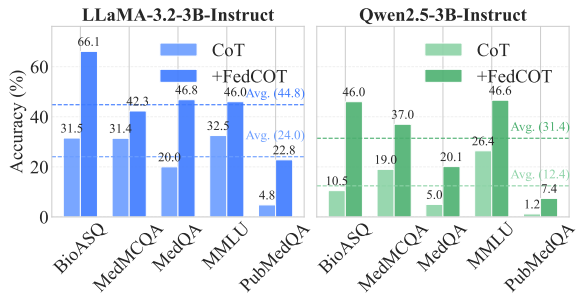

Figure 8: Performance improvement on 3B LLMs via federated reasoning fine-tuning on top of our FedCoT.

We further investigate the performance on smaller size LLMs, and the results are presented in Figure 8. We can observe that our FedCoT exhibit moderate yet consistent gains on smaller models. Specifically, FedCoT significantly outperforms the CoT method with specific training on our federated learning framework and achieves an average performance improvement of 20.8% on `LLaMA-3.2-3B-Instruct` and 19.00% on `Qwen2.5-3B-Instruct` These results underscore our methods' robust generalization and adaptability across a spectrum of model sizes.

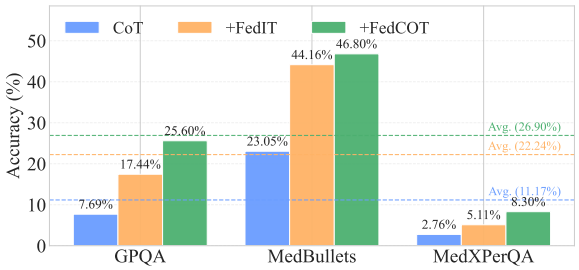

Figure 9: The validation in out-of-distribution situation.

L OOD Generalization

To assess generalization beyond the training distribution, we evaluate on three out-of-distribution (OOD) benchmarks-GPQA (biomedical track), MedBullets, and MedXPertQA. FedCoT maintains strong performance and consistently surpasses both the CoT and FedIT baseline on all benchmarks, as shown in Figure 9. Specifically, FedCoT attains 25.60/46.80/8.30% on GPQA/MedBullets/MedXPerQA, compared with FedIT's 17.44/44.16/5.11% and CoT's 7.69/23.05/2.76%. Averaged over the three datasets, FedCoT reaches 26.90%, yielding a +4.66% improvement over FedIT and a +15.73% gain over CoT; the gains are most pronounced on GPQA (+8.16% vs. FedIT). These results indicate that FedCoT offers better OOD robustness and transfers more effectively to other biomedical scenario under distribution shift.