CECR: A Cloud-Edge Collaborative Reasoning Framework with Large and Small Language Models for Medical Question Answering

Anonymous ACL submission

Abstract

Large Language Models (LLMs) achieve outstanding performance in medical question answering, yet cloud-based inference poses severe privacy risks for sensitive patient data. While local Small Language Models (SLMs) preserve data confidentiality, they are limited by insufficient medical knowledge and weak reasoning, creating a pervasive privacy-utility trade-off in clinical applications. To address this issue, we propose CECR, a cloud-edge collaborative framework for privacy-preserving medical reasoning. CECR decouples generalized medical knowledge elicitation from privacy-sensitive patient inference, synergizing powerful cloud LLMs and privacy-aware edge SLMs. Specifically, the cloud LLM extracts sanitized medical knowledge without accessing raw patient data, while the local SLM integrates external knowledge for private on-device clinical reasoning. We further adopt a LoRA-based hybrid tuning strategy with Contextual Dropout to mitigate over-reliance on external knowledge and improve integration robustness. Evaluated on three standard medical benchmarks, our CECR framework enables lightweight SLMs to achieve performance competitive with trillion-parameter cloud LLMs, effectively balancing clinical reasoning quality and strict data privacy constraints.

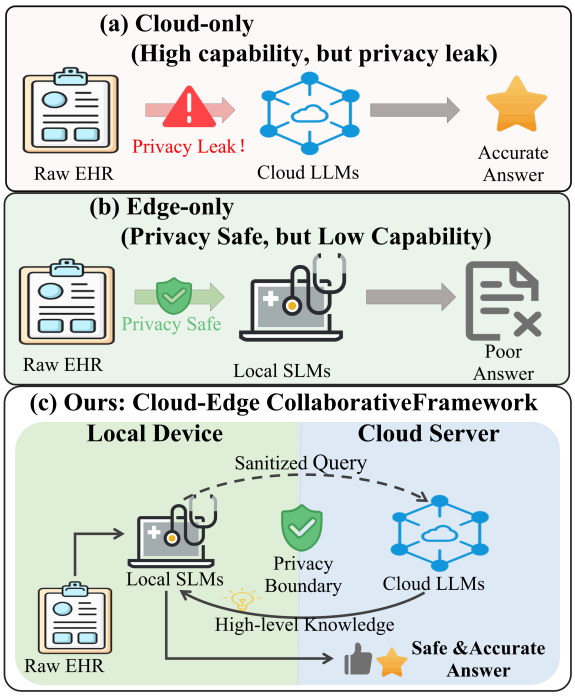


Figure 1: The core motivation of CECR. Unlike the **cloud-only** paradigm (a) which risks privacy leaks, or the **edge-only** paradigm (b) which exhibits limited reasoning capacity, our **cloud–edge collaborative framework** (c) enables lightweight local models to securely leverage cloud-based medical intelligence without exposing sensitive patient data.

1 Introduction

Large language models (LLMs) have demonstrated exceptional performance in natural language understanding and generation, showing great potential in medical question answering (MQA)(Thirunavukarasu et al., 2023). However, state-of-the-art LLMs typically require massive computational resources and are thus predominantly deployed on cloud servers(Shanmugam et al., 2025). In the healthcare domain, this deployment paradigm introduces significant privacy and compliance risks(Zhong et al., 2025). Medical queries often contain highly sensitive patient information, such as demographics, medical histories, and diagnostic records. Directly transmitting such data to cloud-based models violates strict data protection regulations and raises serious concerns regarding confidentiality and unauthorized data access (Kyu-Hwan, 2025; Ong et al., 2024), as illustrated in Fig. 1(a).

To circumvent these privacy barriers, recent studies have increasingly focused on deploying Small Language Models (SLMs)(Neveditsin et al., 2025)—typically characterized by parameters un-

der 10 billion(Kim et al., 2025)—directly on local clinical terminals (e.g., clinician workstations or edge devices). While SLMs ensure that sensitive data remains strictly on-device, their smaller parameter scale severely constrains their medical knowledge breadth and complex reasoning capacity(Wen et al., 2025; Merenda et al., 2026), as illustrated in Fig. 1(b).

This creates a fundamental trade-off in current medical AI deployment: powerful cloud-based LLMs offer superior reasoning at the cost of privacy, whereas local SLMs ensure data security but suffer from diminished clinical capabilities(Wang et al., 2026).

To this end, we propose a privacy–aware cloud–edge collaborative framework for MQA that effectively resolves this trade-off , as illustrated in Figure 1(c). Our core innovation lies in functionally isolating medical knowledge elicitation from patient-specific clinical reasoning. Specifically, the cloud-deployed LLM is isolated from raw patient data; it operates exclusively on privacy-sanitized queries to respond high-level, generalized medical background knowledge. Subsequently, the locally deployed SLM acts as a reasoning integrator, synthesizing the cloud-returned knowledge with the complete, unredacted patient records to generate a personalized and accurate answer. In this paradigm, sensitive information never leaves the local device, yet the system profoundly benefits from the rich knowledge capacity of cloud-based LLMs.

Overall, the primary contributions of our work are summarized as follows:

- We propose a novel cloud–edge collaborative medical reasoning framework that functionally decouples cloud-side knowledge elicitation from local clinical inference. This explicit task separation establishes a strict privacy boundary, where raw patient data is permanently confined to edge devices and only sanitized records are uploaded for cloud knowledge retrieval.

- We design a noise-resilient instruction tuning paradigm equipped with Contextual Dropout and contrastive noise training. Combined with LoRA-based parameter-efficient adaptation, our method strengthens the lightweight local SLM for robust knowledge integration and explicitly mitigates over-reliance on potentially unreliable external cloud knowledge.

- Empirically, we validate that our collaborative framework significantly narrows the performance gap between edge-resident SLMs and trillion-parameter cloud LLMs on both medical and general-domain benchmarks. Our work provides a practical privacy-preserving solution for reliable medical reasoning in resource-limited healthcare scenarios.

2 Related Work

LLMs in Biomedicine. Recent advances in LLMs have unlocked unprecedented performance in medical natural language processing (NLP) tasks, from clinical question answering to diagnostic reasoning. General-purpose models such as GPT-5 (OpenAI, 2025) and Qwen3-VL-Thinking (Alibaba Qwen Team, 2025) have demonstrated remarkable capabilities in complex clinical scenarios including differential diagnosis and pathophysiological reasoning, even outperforming human clinicians in specific tasks (Hoppe et al., 2024). Parallel efforts in medical-specialized LLMs, such as Med-PaLM and Med-PaLM 2 (Singhal et al., 2023, 2025), have further advanced this field by fine-tuning on vast biomedical corpora, achieving state-of-the-art results in handling professional medical consultation and reasoning problems. Despite these advances, a critical bottleneck remains: these models require cloud-based deployment due to their massive parameter scale. In real-world healthcare settings, however, the direct transmission of protected health information (PHI) to external cloud servers is strictly constrained by stringent data privacy regulations, such as HIPAA and the GDPR (Moor et al., 2023; Meskó and Topol, 2023). This creates a fundamental tension between leveraging high-performance medical reasoning and maintaining data confidentiality, significantly impeding the direct integration of proprietary, cloud-reliant LLMs into clinical practice.

SLMs for Edge Clinical Deployment. To ensure patient data remains strictly on-device, research has increasingly focused on domain-specific SLMs. Early representation models like ClinicalBERT (Rumshisky et al., 2020) and PubMedBERT (Gu et al., 2021) have paved the way for recent generative SLMs, ranging from Me-LLaMA (Xie et al., 2025) and the Meditron series (Chen et al., 2023) to the recently introduced Meerkat models (Kim et al., 2025). While these compact models can be securely deployed in edge environments (e.g.,

hospital intranets or clinician workstations), their restricted parameter scale (typically < 10B) inevitably introduces a representational "knowledge bottleneck." Consequently, compared to state-of-the-art cloud LLMs, local SLMs frequently struggle with comprehensive medical knowledge recall and complex multi-hop diagnostic reasoning (Liévin et al., 2024; Li et al., 2025). Our work addresses this fundamental capability gap not by naively scaling up local parameters, but by proposing a novel cloud–edge collaborative framework (Jayaweera et al., 2025; Corradini et al., 2025). This paradigm enables lightweight local models to indirectly harness cloud-scale intelligence, achieving advanced clinical reasoning while adhering to the strict data isolation protocols required in modern healthcare.

Knowledge Distillation and Transfer. Knowledge distillation (KD) has emerged as a predominant paradigm for transferring complex reasoning capabilities from powerful LLMs to compact SLMs (Yang et al., 2025). Conventional approaches typically leverage teacher LLMs to generate step-by-step rationales or structured synthetic instructions, which are subsequently used to fine-tune smaller student models (Zhu et al., 2024; Kang et al., 2026). However, this "static" distillation permanently encodes knowledge into the SLM's rigid parametric memory. Consequently, these models often exhibit degraded generalization on rare, long-tail clinical scenarios and inherently fail to accommodate the rapid, dynamic updates of medical guidelines (Clusmann et al., 2023; Zakka et al., 2024). In contrast to relying solely on static parametric memorization, our CECR framework introduces a novel "dynamic knowledge mediation" mechanism. By leveraging the cloud LLM as an on-demand knowledge engine to supply generalized, case-specific medical context, the edge SLM can dynamically synthesize this external intelligence to perform robust and personalized diagnostic reasoning.

Augmenting Reasoning with External Context. To supplement the internal knowledge of language models, Retrieval-Augmented Generation (RAG) is widely employed to fetch relevant documents from static external corpora (Ma et al., 2025). Parallel advancements have also explored utilizing LLMs as implicit knowledge bases to elicit internal reasoning chains via advanced prompting strategies (Chu et al., 2024; Zhou et al., 2024). While our framework shares the overarching goal of context augmentation, it differs fundamentally in execution. Instead of relying on passive document retrieval—which is often constrained by corpus coverage—or simple rationale prompting, we leverage a cloud LLM to actively generate structured, pathophysiological contexts tailored to sanitized clinical queries (Jin and Wu, 2025). Furthermore, recent studies reveal that small models are highly susceptible to the "contextual distraction curse," where irrelevant or noisy retrieved context severely degrades their reasoning performance (Wang et al., 2025; Wu et al., 2025). To overcome this critical vulnerability, our framework implements a targeted noise-resilient tuning strategy, ensuring that the local SLM integrator maintains strict diagnostic autonomy and robustly filters out cloud-generated noise.

3 Cloud-Edge Collaborative Reasoning Framework

3.1 Overview of CECR

As illustrated in Fig. 2, our CECR framework comprises two functional entities, namely a cloud server and a local edge device, which collaboratively conduct medical reasoning while strictly complying with patient privacy constraints.

- **Cloud Server:** The cloud server deploys an LLM with parameters θ_l, which is empowered by comprehensive medical knowledge and augmented with a professional medical knowledge base. Serving as a centralized *Knowledge Provider*, the cloud LLM exclusively processes sanitized medical data and delivers domain-specific background knowledge (K) without accessing raw patient records.

- **Local Device:** The edge device (e.g., clinician workstations) maintains a lightweight SLM parameterized by θ_s. The local SLM undertakes dual functional roles: (1) a *Sanitization Engine* that abstracts raw private data before cloud transmission; (2) a *Reasoning Integrator* that aggregates cloud-derived knowledge K with complete raw patient context to yield clinically accurate responses.

In this paradigm, only sanitized data and elicited medical knowledge cross the privacy boundary. Local operations guarantee rigorous privacy protection and personalized inference, while cloud operations serve as a privacy-isolated knowledge provider. We detail the implementation of each module in the following sections.

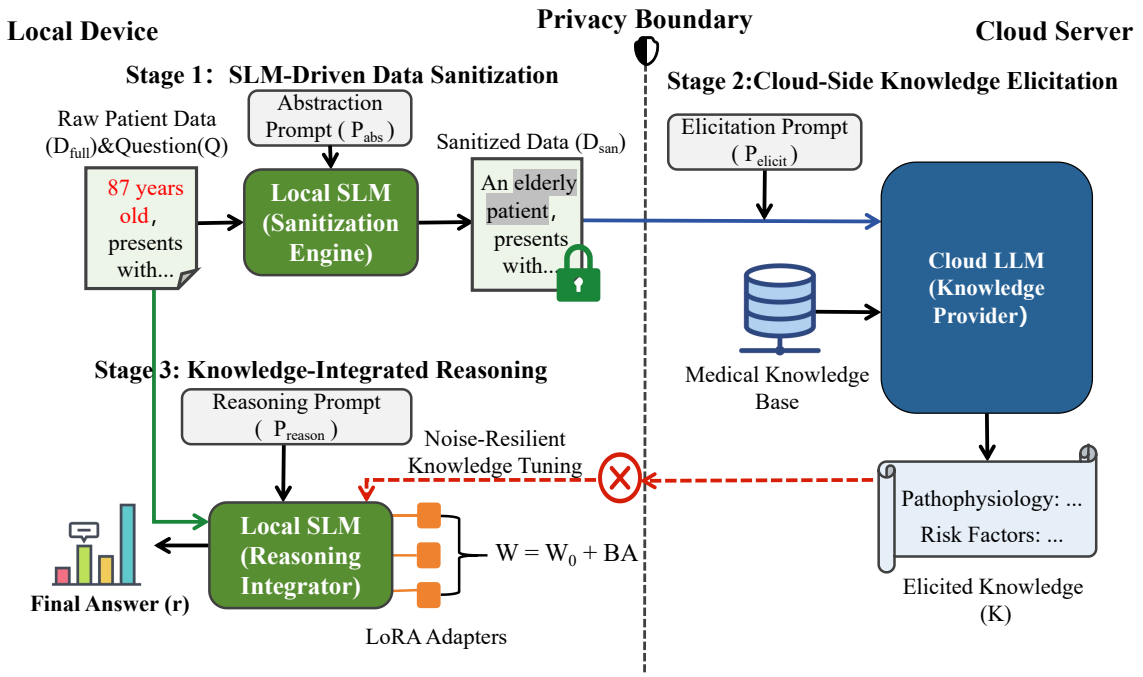


Figure 2: The privacy-preserving cloud-edge collaborative medical reasoning framework.

3.2 SLM-Driven Privacy Sanitization

To mitigate privacy leakage during cloud transmission, we formulate medical data sanitization as an end-to-end text-to-text rewriting task, as illustrated in Stage 1 of Fig. 2. Different from conventional rule-based NER pipelines with unstable extraction performance, we exploit the instruction-following capability of the local SLM to generate privacy-compliant medical records D_{san}. Acting as a semantic information bottleneck, this module minimizes protected health information (PHI) exposure while retaining clinically critical semantics.

Specifically, we design a dedicated abstraction prompt $\mathcal{P}_{\text{abs}}$ to guide the rewriting procedure:

$$D_{\text{san}} = \text{SLM}_{\theta_s}\left([\mathcal{P}_{\text{abs}}; D_{\text{full}}]\right), \tag{1}$$

where $[\cdot;\cdot]$ denotes text concatenation. The rewriting process jointly optimizes two complementary objectives:

- **Privacy Abstraction:** Explicit identifiers (e.g., names, IDs) are erased, while implicit quasi-identifiers are generalized into coarse categories. For instance, specific age figures are abstracted into "an elderly patient".
- **Semantic Preservation:** Essential clinical information, including symptoms and laboratory indicators, is retained to provide sufficient contextual cues for cloud-side knowledge elicitation.

3.3 Cloud-Side Knowledge Elicitation

As illustrated in Fig. 2, the sanitized data D_{san} is delivered to the cloud module located on the right. Within this module, the cloud-hosted LLM θ_l serves as a centralized Knowledge Provider. Leveraging its built-in medical knowledge base, the LLM takes the sanitized input and the elicitation prompt $\mathcal{P}_{\text{elicit}}$ as inputs, and finally generates structured medical background knowledge:

$$K = \text{LLM}_{\theta_l}\left([\mathcal{P}_{\text{elicit}}; D_{\text{san}}]\right). \tag{2}$$

Notably, the cloud model cannot access any raw patient identifiers. It solely processes privacy-abstracted content to generate high-level medical knowledge for downstream local reasoning.

3.4 Knowledge-Integrated Reasoning

As illustrated in Fig. 2, the local inference module at the bottom executes the collaborative reasoning procedure. Upon receiving external medical knowledge K, the local SLM serves as a dedicated Reasoning Integrator.

The model aggregates the task query Q, cloud-derived knowledge K, and complete raw patient records D_{full} via a reasoning prompt $\mathcal{P}_{\text{reason}}$ to produce the final diagnosis response r:

$$r = \text{SLM}_{\theta_s}\left([\mathcal{P}_{\text{reason}}; Q; D_{\text{full}}; K]\right). \tag{3}$$

Notably, D_{full} is strictly confined within the secure local pathway and never leaves the edge device. This on-device reasoning paradigm ensures privacy safety and enables seamless fusion of cloud medical intelligence and raw clinical information.

3.5 Knowledge-Guided Hybrid Optimization

To enhance the local SLM's knowledge aggregation capability and improve its robustness against irrelevant cloud-side information, we develop a hybrid optimization paradigm for the local integration module.

3.5.1 Noise-Resilient Knowledge Tuning

To prevent the SLM from developing a "blind reliance" on external knowledge, we introduce a noise-resilient tuning paradigm, corresponding to the designated module and operation in Fig. 2. During instruction tuning, we construct three complementary training scenarios to enhance alignment, offline robustness and noise immunity:

- **Gold Correspondence:** We adopt perfectly matched knowledge-query pairs for training to strengthen semantic alignment between clinical queries and medical background knowledge.
- **Contextual Dropout:** External knowledge is replaced with null prompts in partial training cases, empowering the model to maintain autonomous reasoning capability even when cloud services are unavailable.
- **Contrastive Noise:** Irrelevant distracting background knowledge is randomly injected into training samples. The SLM is explicitly trained to generate a Rejection Rationale before conducting independent clinical reasoning.

3.5.2 Hybrid Adaptation Strategy

To accommodate edge hardware limitations, we adopt Low-Rank Adaptation (LoRA) for parameter-efficient tuning. Trainable low-rank matrices A and B are injected into the frozen backbone weights W_0:

$$W = W_0 + BA, \tag{4}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ denote low-rank projection matrices. As indicated by the adapter modules in Fig. 2, only lightweight low-rank layers are updated. This design enables efficient clinical reasoning deployment on resource-constrained edge workstations.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our proposed CECR framework on three widely recognized medical question-answering (MQA) benchmarks: (1) **MedQA-USMLE** (Jin et al., 2021), which comprises professional-grade questions from the United States Medical Licensing Examination; (2) **HEADQA** (Vilares and Gómez-Rodríguez, 2019), a multi-disciplinary dataset encompassing biology, pharmacology, and clinical medicine; and (3) **MedMCQA** (Pal et al., 2022), a large-scale collection of entrance examination questions from Indian medical institutes. All datasets are processed in a multiple-choice format to ensure a standardized quantitative assessment of reasoning performance.

Implementation Details. All local training and inference are conducted on a single NVIDIA RTX 5090 GPU (32GB) under BFloat16 precision to ensure numerical stability. We implement a strict label masking strategy during the supervised fine-tuning phase: the training objective focuses exclusively on the model's final response tokens, effectively mitigates overfitting on the input prompt and background context. Consistent with the design of the optimization module, Contextual Dropout is integrated into the training pipeline to enhance the model's autonomous reasoning robustness. All experimental results are reported as the average of three independent runs.We present detailed experimental configurations and parameters in Appendix B.

Model Backbones and Baselines. To demonstrate the model-agnostic nature of our framework, we employ several SLMs as local backbones: BioLinkBERT-Base/Large (Yasunaga et al., 2022), BioMedLM (Bolton et al., 2022), Meditron3-Gemma2-2B (OpenMeditron Initiative, 2025), Phi-3.5-mini-instruct (Microsoft, 2024), and Qwen3.5-4B-Base (Qwen Team, 2026).

Context Generation with Cloud LLM. For the cloud-side knowledge elicitation, we utilize the Qwen3.5-122B-A10B model (Qwen Team, 2026) as the centralized knowledge provider. We employ greedy decoding during the elicitation process to ensure deterministic pathophysiological rationales. The elicited knowledge is structured into predefined clinical categories to maintain high signal-to-noise ratio for downstream local integration.

Table 1: Performance of various SLM backbones under the CECR framework. We compare standalone local models against those enhanced with cloud-elicited knowledge. ↑ denotes absolute improvement. FT and ZS indicate full-parameter fine-tuning and zero-shot inference, respectively. † denotes results cited from original papers. MedGemma has two variants with different parameter sizes; FTC is a knowledge-enhanced fine-tuning method for medical models from prior work.

Model / Method	Params	MedQA-USMLE (%)	HEADQA (%)	MedMCQA (%)
Cloud LLM Upper Bound				
GPT-5 (OpenAI, 2025)	N/A	93.70†	88.50†	84.20†
MedGemma (Google, 2025)	27B	87.70†	-	74.20†
Medical Specialized LLMs				
Me-LLaMA (Xie et al., 2024)	13B	42.70†	-	44.90†
MedGemma (Sellergren et al., 2026)	4B	53.30†	-	55.70†
BioMistral (Labrak et al., 2024)	7B	46.60†	-	45.70†
FTC (Li et al., 2024)				
+BioLinkBERT-base	110M	48.60	60.16	48.32
+BioLinkBERT-large	340M	49.69	61.23	49.40
+BioMedLM	2.7B	52.58	62.03	50.09
Standalone Local Baselines (Edge-Only/ No Cloud Knowledge)				
BioLinkBERT-Base (FT)	110M	41.22	39.14	32.15
BioLinkBERT-Large (FT)	340M	44.60	39.53	35.86
BioMedLM (FT)	2.7B	46.46	48.40	57.30
Phi-3.5-mini-instruct (ZS)	3.8B	58.65	69.72	56.25
CECR: Cloud-Edge Collaborative Reasoning (Ours)				
BioLinkBERT-Base (FT)	110M	51.10 ↑ 9.88	62.20 ↑ 23.06	53.17 ↑ 21.02
BioLinkBERT-Large (FT)	340M	55.64 ↑ 11.04	67.15 ↑ 27.62	57.15 ↑ 21.29
BioMedLM (FT)	2.7B	63.72 ↑ 17.26	68.59 ↑ 20.19	68.13 ↑ 10.83
Phi-3.5-mini-instruct (ZS)	3.8B	68.73 ↑ 10.08	74.34 ↑ 4.62	69.04 ↑ 12.79

Fine-tuning Strategies. To demonstrate the versatility of the CECR framework across varied hardware constraints, we adopt size-dependent optimization strategies tailored for resource-limited edge devices. For compact legacy backbones such as BioLinkBERT-Base/Large (Yasunaga et al., 2022) and BioMedLM (Bolton et al., 2022), we perform full-parameter fine-tuning. For larger-scale SLMs including Qwen3.5-4B-Base, Meditron3-Gemma2-2B, and Phi-3.5-mini-instruct, we employ Low-Rank Adaptation (Hu et al., 2022). This method freezes backbone weights and only inserts trainable low-rank matrices into attention and feed-forward layers, which cuts peak memory usage during training. Our bifurcated optimization scheme enables efficient parameter updates and ensures the reasoning modules can be deployed on clinical workstations.

Evaluation Metrics. Following the standard evaluation protocol for medical multiple-choice tasks, we adopt prediction Accuracy (%) as the primary metric for reasoning performance. To ensure a fair and consistent comparison, we maintain identical hyperparameters and prompting templates across all experimental settings. During inference, we employ greedy decoding and a regular expression-based parsing strategy to robustly extract predicted options for standardized evaluation.

4.2 Quantitative Performance Analysis

We evaluate the performance of the CECR framework through a two-phase analysis. Table 1 (Phase I) investigates the inherent utility of cloud-elicited knowledge for off-the-shelf models, while Table 2 (Phase II) examines the impact of our robust tuning strategy on state-of-the-art SLM backbones.

Efficacy of Knowledge Elicitation. As illustrated in Table 1, the injection of high-level pathophysiological contexts (K) brings steady performance improvement for standalone SLMs with different parameter scales. Even for high-performance models like Phi-3.5-mini-instruct, CECR achieves zero-shot accuracy gain up to 12.79% on MedMCQA. These steady gains indicate that insufficient internal medical knowledge constitutes a major performance bottleneck for edge-side SLMs.

Impact of Collaborative Fine-Tuning. Table 2 reflects the combined effect of external knowledge and local optimization strategy. Standalone small models tend to reach performance limita-

Table 2: Overall accuracy (%) on three medical benchmarks (Part 2). The best performance among local methods is bolded. Blue upward arrows denote absolute improvements over standalone baselines. ZS = Zero-Shot, FT = Fine-Tuned.

Model / Method	Params	MedQA-USMLE (%)	HEADQA (%)	MedMCQA (%)
Standalone Local Baselines (Edge-Only Deployment)				
Meditron3-Gemma2-2B	2B	41.12	56.75	43.39
Qwen3.5-4B-Base	4B	62.89	76.00	56.83
Ours: Cloud-Edge Collaborative Reasoning (CECR)				
Meditron3-Gemma2-2B (ZS)	2B	55.40 ↑ 14.28	59.50 ↑ 2.75	52.09 ↑ 8.70
Meditron3-Gemma2-2B (FT)	2B	73.11 ↑ 31.99	74.55 ↑ 17.80	72.31 ↑ 28.92
Qwen3.5-4B-Base (ZS)	4B	63.42 ↑ 0.53	79.45 ↑ 3.45	64.62 ↑ 7.79
Qwen3.5-4B-Base (FT)	4B	**83.57** ↑ 20.68	**85.63** ↑ 9.63	**81.54** ↑ 24.71

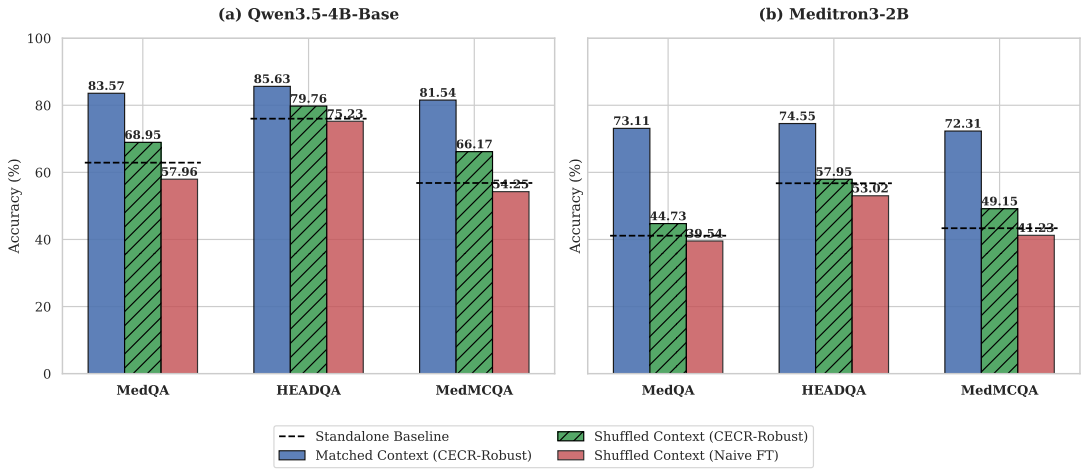


Figure 3: Robustness analysis under knowledge interference (Noise Attack). We compare optimized CECR scheme against naive fine-tuning and standalone baselines. The dashed lines represent the standalone reasoning performance. Naive fine-tuned models suffer noticeable performance degradation under interference, while CECR maintains stable reasoning capacity near the standalone baseline level.

tion easily, while CECR optimized models substantially narrow the performance gap between 4B-scale lightweight models and large cloud-based LLMs. For instance, Qwen3.5-4B-Base enhanced by CECR obtains 83.57% accuracy on MedQA, effectively reducing the disparity with high-capacity cloud models.

4.3 Resilience against Contextual Interference

To explore whether local SLMs generate excessive dependence on the external cloud context, we conduct shuffled context attack experiments as shown in Figure 3.

Experimental results reveal common limitations of conventional knowledge enhanced models. When matched cloud information is replaced with mismatched content, naive fine-tuned models suffer noticeable performance degradation, with prediction accuracy dropping below a standalone baseline. Taking Qwen3.5-4B-Base on MedQA-USMLE as an example, the naive fine-tuned model drops to 57.96%, revealing a marked degradation in independent reasoning ability.

Compared with naive schemes, our CECR framework equipped with multi-strategy noise-resilient tuning shows better anti-interference ability. Benefiting from Contextual Dropout and contrastive noise training adopted in optimization stage, the model can effectively weaken disturbance from unreliable external information. In most test cases, CECR achieves higher accuracy than naive fine-tuning method and keeps stable reasoning capacity comparable to or even exceeding the standalone baseline.

For example, the Qwen3.5-4B-Base model maintains 68.95% accuracy on MedQA under interference attack. The result verifies that our framework helps local SLMs reasonably integrate multi-source information and prioritize credible local clinical content when external knowledge lacks reliability.

4.4 Cross-Domain Generalization

Privacy and data confidentiality requirements exist not only in clinical scenarios, but also in diverse open-domain reasoning tasks. We select two mainstream public benchmarks to verify the cross-domain applicability of CECR: (1) **CommonsenseQA (CSQA)** (Talmor et al., 2018), focusing on daily world knowledge reasoning; (2) **OpenBookQA (OBQA)** (Mihaylov et al., 2018), designed for multi-step scientific reasoning.

We adopt publicly available T5-base (Raffel et al., 2020) as local lightweight backbone for cross-domain tests. To verify the inherent effect of cloud extracted knowledge without extra task adaptation, we adopt zero-shot integration setting, where external knowledge is directly appended to original query without downstream fine-tuning. We compare our zero-shot CECR with standard supervised fine-tuning baseline, which is fully optimized on target datasets without external auxiliary knowledge.

As shown in Table 3, CECR still achieves prominent accuracy improvement when migrated to non-medical backbone and open-domain tasks. Specifically, our method raises average accuracy by 14.67% on OBQA (56.20% to 70.87%) and 8.87% on CSQA (59.90% to 68.77%).

The results prove that CECR is not limited to medical reasoning tasks. As a universal knowledge supplement paradigm, it can effectively address knowledge deficiency of SLMs and realize flexible plug-and-play reasoning across different domains.

Table 3: Zero-shot generalization performance (Accuracy %) on open-domain benchmarks using a T5-base backbone. CECR achieves substantial gains over the SFT baseline without task-specific fine-tuning.

Method	OBQA (%)	CSQA (%)
SFT Baseline	56.20	59.90
CECR (Zero-shot)	**70.87** ↑ 14.67	**68.77** ↑ 8.87

5 Conclusion

The deployment of artificial intelligence in clinical settings faces a fundamental tension: privacy regulations and patient confidentiality requirements restrict access to sensitive health data, while high-performance medical reasoning often depends on the rich domain knowledge and computational power of LLMs. To address this challenge, we present CECR, a privacy-preserving cloud-edge collaborative reasoning framework designed to reconcile secure data handling with the capabilities of powerful models.

The core design of CECR separates general medical knowledge acquisition from privacy-sensitive clinical inference. By keeping all protected health information (PHI) confined to local edge devices, the framework effectively safeguards confidential patient data, while still allowing lightweight SLMs to leverage abundant, de-identified medical knowledge extracted from cloud-based LLMs. This structure avoids the need to transmit sensitive data to external servers, a key requirement for compliance with common healthcare privacy standards.

Extensive experiments conducted on three medical reasoning datasets and two open-domain benchmarks yield three key findings that support the effectiveness of the proposed framework. First, high-quality medical knowledge elicited from powerful LLMs such as Qwen3.5-122B effectively compensates for the insufficient inherent domain knowledge of lightweight SLMs, addressing a major limitation of small models in complex clinical reasoning tasks. Second, the proposed knowledge-guided instruction tuning and Contextual Dropout mechanism help local SLMs properly fuse external medical information with task-specific inputs, mitigating over-reliance on biased knowledge shortcuts and encouraging more robust, context-aware reasoning pathways. Third, CECR exhibits strong model adaptability and cross-domain generalization ability, achieving consistent performance gains across diverse SLM architectures (from BERT-based models to modern decoder-only models) and reasoning tasks.

Overall, this work presents a feasible and promising solution for secure AI deployment in privacy-sensitive clinical scenarios. It demonstrates how collaborative cloud-edge architectures can unlock the potential of medical LLMs without compromising patient privacy. Looking ahead, future work will focus on validating the framework in real-world clinical workflows, evaluating its performance on more diverse patient populations and rare disease cases, and exploring additional privacy-preserving mechanisms to further strengthen its security guarantees.

Limitations

Although the CECR framework demonstrates competitive reasoning performance, several limitations warrant further investigation in future research.

First, our evaluation relies primarily on reference-matching metrics such as Accuracy. While standard for multiple-choice benchmarks, these metrics are constrained by lexical matching and may not fully capture the qualitative nuances of clinical reasoning logic. Responses that are medically valid but follow different linguistic styles or provide generalized interpretations may be penalized under stringent matching protocols, potentially leading to an underestimation of the model's actual diagnostic capability.

Second, while the noise-resilient knowledge tuning strategy significantly reduces over-reliance on external context, the framework remains susceptible to highly complex or deceptive cloud-side misinformation. If the centralized knowledge provider generates factually incorrect but internally consistent pathophysiological hallucinations, the local integrator may still experience performance fluctuations. Developing more sophisticated semantic gating mechanisms and adaptive filtering remains essential to further enhance the resilience of the system in safety-critical clinical environments.

Third, the multi-stage pipeline—comprising data sanitization, cloud-side elicitation, and local integration—introduces higher inference latency and increased token consumption compared to standalone local prompting. This architectural trade-off may pose challenges in urgent clinical settings, such as emergency triage, where real-time response is critical. Future work should explore techniques such as knowledge caching and prompt compression to optimize the balance between reasoning accuracy and computational overhead for practical healthcare deployment.

References

Alibaba Qwen Team. 2025. Qwen3-vl-thinking. Hugging Face model card, accessed 2026-05-16.

Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. 2022. Biomedlm. Official project page, accessed 2026-05-20.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, and 1 others. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

Flavio Corradini, Matteo Leonesi, and Marco Piangerelli. 2025. State of the art and future directions of small language models: a systematic review. *Big Data and Cognitive Computing*, 9(7):189.

Google. 2025. google/medgemma-27b-text-it. Hugging Face biomedical pre-trained model.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

John Michael Hoppe, Matthias K Auer, Anna Strüven, Steffen Massberg, and Christopher Stremmel. 2024. Chatgpt with gpt-4 outperforms emergency department physicians in diagnostic accuracy: retrospective analysis. *Journal of medical Internet research*, 26:e56110.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Rasanga Jayaweera, Himanshu Agrawal, and Nickson M Karie. 2025. Federated security for privacy preservation of healthcare data in edge-cloud environments. *Sensors*, 25(16):5108.

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.

Hongpeng Jin and Yanzhao Wu. 2025. Ce-collm: Efficient and adaptive large language models through cloud-edge collaboration. In *2025 IEEE International Conference on Web Services (ICWS)*, pages 316–323. IEEE.

Minki Kang, Jongwon Jeong, Seanie Lee, Jaewoong Cho, and Sung Ju Hwang. 2026. Distilling llm agent into small models with retrieval and code tools. *Advances in Neural Information Processing Systems*, 38:106501–106538.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, and 1 others. 2025. Small language models learn enhanced reasoning skills from medical textbooks. *NPJ digital medicine*, 8(1):240.

Jung Kyu-Hwan. 2025. Large language models in medicine: Clinical applications, technical challenges, and ethical considerations. *hir*, 31(2):114–124.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Shiyang Li, Xianjun Yang, Chenxin Tian, Yao Qin, Linda Petzold, and 1 others. 2024. Enhancing small medical learners with privacy-preserving contextual prompting. In *International Conference on Learning Representations*, volume 2024, pages 28361–28389.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025. Small models struggle to learn from strong reasoners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25366–25394.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yusong Ma, Hongxuan Nie, Chao Chen, Jiujie Zhang, Jiali Jiang, Bisheng Wang, and Yuqin Xia. 2025. A survey of retrieval-augmented generation (rag) for large language models. In *2025 International Conference on Trustworthy Big Data and Artificial Intelligence (ICTBAI)*, pages 7–13.

Flavio Merenda, Jose Manuel Gomez-Perez, and German Rigau. 2026. Can LLMs reason like doctors? exploring the limits of large language models in complex medical reasoning. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2432–2452, Rabat, Morocco. Association for Computational Linguistics.

Bertalan Meskó and Eric J Topol. 2023. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine*, 6(1):120.

Microsoft. 2024. Phi-3.5-mini-instruct. Hugging Face model card.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391. Association for Computational Linguistics.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Nikita Neveditsin, Pawan Lingras, and Vijay Mago. 2025. Clinical insights: A comprehensive review of language models in medicine. *PLOS Digital Health*, 4(5):e0000800.

Jasmine Chiat Ling Ong, Shelley Yin-Hsi Chang, Wasswa William, Atul J Butte, Nigam H Shah, Lita Sui Tjien Chew, Nan Liu, Finale Doshi-Velez, Wei Lu, Julian Savulescu, and 1 others. 2024. Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*, 6(6):e428–e432.

OpenAI. 2025. Chatgpt (gpt-5). Large language model, released August 7, 2025.

OpenMeditron Initiative. 2025. Meditron3-gemma2-2b. Biomedical language model based on Gemma 2.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann. 2020. Proceedings of the 3rd clinical natural language processing workshop. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2026. Medgemma technical report. In *Proceedings of the Medical Imaging with Deep Learning Conference (MIDL)*.

Divya Shanmugam, Monica Agrawal, Rajiv Movva, Irene Y Chen, Marzyeh Ghassemi, Maia Jacobs, and Emma Pierson. 2025. Generative artificial intelligence in medicine. *Annual Review of Biomedical Data Science*, 8.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: Answering commonsense questions leveraging massive semantic knowledge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4149–4160. Association for Computational Linguistics.

Anand J. Thirunavukarasu, David S.J. Ting, K. Elangovan, and 1 others. 2023. Large language models in medicine. *Nature Medicine*, 29:1930–1940.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Yanbo Wang, Zixiang Xu, Yue Huang, Chujie Gao, Siyuan Wu, Jiayi Ye, Xiuying Chen, Pin-Yu Chen, and Xiangliang Zhang. 2025. Breaking focus: Contextual distraction curse in large language models. In *Will Synthetic Data Finally Solve the Data Access Problem?*

Zian Wang, Ziyi Wang, Haonan Jin, Jie Xing, and Lanshan Zhang. 2026. Tailorllm: Collaborative end-cloud inference of large and small language models based on low-rank adaptation. In *Proceedings of the 21st European Conference on Computer Systems*, pages 1144–1159.

Andrew Wen, Qiuhao Lu, Yu-Neng Chuang, Guanchu Wang, Jiayi Yuan, Jiamu Zhang, Liwei Wang, Sunyang Fu, Kurt D Miller, Heling Jia, and 1 others. 2025. Context matching is not reasoning when performing generalized clinical evaluation of generative language models. *npj Digital Medicine*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jinyang Wu, Shuai Zhang, Feihu Che, Mingkuan Feng, Pengpeng Shao, and Jianhua Tao. 2025. Pandora's box or aladdin's lamp: A comprehensive analysis revealing the role of rag noise in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5019–5039.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, and 1 others. 2024. Me-llama: Foundation large language models for medical applications. *Research square*, pages rs–3.

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

Chuanpeng Yang, Yao Zhu, Wang Lu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. 2025. Survey on knowledge distillation for large language models: Methods, evaluation, and application. *ACM Trans. Intell. Syst. Technol.*, 16(6).

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, and 1 others. 2024. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068.

Xiaoying Zhong, Siyi Li, Zhao Chen, Long Ge, Dongdong Yu, Shijia Wang, Liangzhen You, and Hongcai Shang. 2025. Considerations for patient privacy of large language models in health care: Scoping review. *Journal of Medical Internet Research*, 27.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V Le, Denny Zhou, Swaroop Mishra, Huaixiu S Zheng, and 1 others. 2024. Self-discover: Large language models self-compose reasoning structures. *Advances in Neural Information Processing Systems*, 37:126032–126058.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024. Distilling mathematical reasoning capabilities into small language models. *Neural Networks*, 179:106594.

A Detailed Dataset Profiles and Licensing

In this section, we provide detailed descriptions and licensing information for all datasets utilized in our experiments, covering both the medical and open-domain reasoning benchmarks.

A.1 Medical Reasoning Benchmarks

MedQA-USMLE. This dataset is designed to evaluate professional clinical reasoning using questions from the United States Medical Licensing Examination. We utilize the English subset (4-option version), which contains 10,178 training samples, 1,273 development samples, and 1,273 test samples.

MedMCQA. A large-scale multiple-choice dataset containing over 194k questions derived from Indian medical entrance exams (AIIMS and NEET-PG). It covers 2,400+ healthcare topics. We utilize the standard train and development splits provided by the original authors.

HEADQA. A multi-disciplinary dataset originating from the Spanish healthcare system specialization exams (FIR, BIR, QIR, PIR, EIR, MIR). We utilize the English version, which encompasses six categories: biology, chemistry, pharmacology, psychology, nursing, and medicine. The English test set contains 2,742 professional questions.

A.2 Open-Domain Reasoning Benchmarks

CommonsenseQA (CSQA). CSQA is a challenging dataset designed to evaluate commonsense reasoning that goes beyond surface-level pattern matching. It requires the model to utilize diverse world knowledge to answer multiple-choice questions. It contains 12,102 questions.

OpenBookQA (OBQA). Modeled after open-book elementary science exams, this dataset focuses on multi-step reasoning with a core set of 1,326 facts. It requires combining the provided facts with broad world knowledge to solve 5,957 questions.

All datasets (MedQA, MedMCQA, HEADQA, CSQA, and OBQA) and models (Qwen, Meditron) used in this study are publicly available for academic research. They are utilized in strict accordance with their respective open-source licenses, including CC BY-SA 4.0, Apache 2.0, and MIT. Our implementation will also be released under the Apache 2.0 license

B SLM IMPLEMENTATION AND TRAINING DETAILS

We implement the CECR framework using the HuggingFace Transformers (Wolf et al., 2020) and PEFT libraries.All local experiments are conducted on a single NVIDIA RTX 5090 GPU (32GB). To accommodate the diverse architectural properties of the selected SLMs, we employ tailored optimization and precision settings.

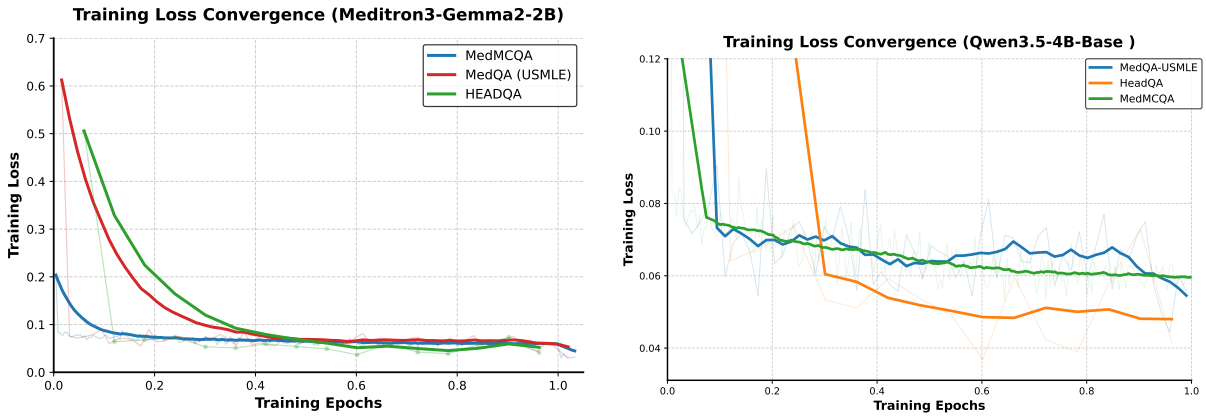
Optimization and Precision. For the Qwen3.5-4B-Base, we manually inject a standard ChatML template during fine-tuning to imbue it with instruction-following capabilities. To facilitate efficient edge deployment, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) to all linear layers (target_modules="all-linear") with a rank $r = 16$ and an alpha value $\alpha = 32$. For the BioMedLM backbone, we utilize the Adafactor optimizer (Shazeer and Stern, 2018) with a warmup ratio of 0.06 and a weight decay of 0.05. For all other models, including BioLinkBERT, Qwen3.5, and Meditron3, we adopt the AdamW optimizer (Loshchilov and Hutter, 2019). Regarding numerical precision, we employ BFloat16 for BioMedLM, Qwen3.5, and Meditron3 to ensure training stability, while FP16 is used for BioLinkBERT-Base/Large following original implementation standards.

Hyperparameters. We adopt size-dependent tuning strategies as follows:

- **BioLinkBERT-Base/Large:** We perform full-parameter fine-tuning with learning rates of 2×10^{-5} and 3×10^{-5}, respectively. The training lasts for 1 to 3 epochs depending on the specific benchmark.

- **BioMedLM:** We set a learning rate of 5×10^{-6} for full-parameter fine-tuning. The total effective batch size is set to 32 for MedQA and 16 for MedMCQA, achieved through gradient accumulation with a per-device batch size of 1.

- **Other SLMs (Qwen3.5-4B-Base and Meditron3-2B):** We apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) with a rank $r = 16$, $\alpha = 32$, and a learning rate of 1×10^{-4}.



(a) Meditron3-Gemma2-2B

(b) Qwen3.5-4B-Base

Figure 4: Training loss convergence across three medical benchmarks. Bold lines represent exponentially smoothed loss, while faded lines indicate raw batch-level updates. The plots illustrate rapid structural alignment during the initial 0.2 epoch and stable convergence below 0.05 across all datasets.

Training Dynamics. The maximum sequence length is set to 512 tokens for standalone SFT baselines and extended to 1536 tokens for CECR-integrated models to accommodate the elicited medical context. For modern SLMs and BioMedLM, training is conducted for 3 epochs. For the CECR-Robust tuning phase, we apply a Knowledge Dropout probability of $p = 0.2$ and incorporate negative sampling. We implement a strict label masking strategy where the loss is calculated exclusively on the response tokens (e.g., "The correct answer is [X]").

Evaluation and Reproducibility. All reported results are based on deterministic greedy decoding. For the BioLinkBERT-Base backbone, we report the mean performance across three random seeds $\{42, 43, 44\}$. For larger-scale backbones, we report results from a single stable run due to computational cost, while maintaining consistent prompting templates across all experimental settings to ensure a fair comparison.

C Quantitative Privacy Evaluation

Since standard MQA benchmarks (e.g., MedQA) are pre-de-identified for public distribution, they do not contain real-world Protected Health Information (PHI). To substantiate the privacy-preserving claims of the CECR framework, we conducted a "Privacy Stress Test" by synthetically augmenting the dataset with sensitive identifiers.

C.1 Experimental Methodology

We randomly sampled 200 cases from the MedQA test set. Using the Faker library, we injected synthetic PHI into the clinical vignettes, including patient names, specific birth dates, and names of medical facilities. For example, a raw vignette "A 62-year-old man..." was augmented to "Patient [James Miller], born on [1962-05-14], was admitted to [Mayo Clinic]. This 62-year-old man...".

We then utilized our local SLM-driven Sanitization Engine (Stage 1) to rewrite these augmented records. To audit the de-identification results, we conducted manual inspection to check for residual PHI contained in the sanitized output (D_{san}).

C.2 Metrics and Results

We report two primary safety metrics in Table 4:

- **PHI Removal Recall**: The percentage of injected PHI elements successfully removed or masked by the SLM.
- **Re-identification (Re-ID) Risk**: We simulated a *linkage attack* where an adversary uses a dense retriever (*all-MiniLM-L6-v2*) to match D_{san} back to the original D_{full} based on semantic similarity.

The results demonstrate that our generative sanitization effectively erases direct identifiers and significantly lowers the re-identification risk by altering the linguistic signatures of the records, reducing the adversary's Top-1 matching success from 91.5% to 5.4%.

Table 4: Results of the Privacy Stress Test on 200 MedQA samples.

Metric	**Raw** D_{full}	**CECR** D_{san}
Name/ID Leakage (↓)	100%	0.8%
Date/Facility Leakage (↓)	100%	4.2%
PHI Removal Recall (↑)	-	**97.2%**
Top-1 Re-ID Risk (↓)	91.5%	**5.4%**

D Training Convergence Analysis

To provide further insights into the optimization process of the CECR framework, we present the training loss convergence curves for our two primary backbones: Meditron3-Gemma2-2B and Qwen3.5-4B-Base.

As shown in Figure 4, both models experience a rapid drop in training loss within the first 0.2 epoch. This elbow-shaped convergence trend indicates that the SLMs can quickly adapt to the structured reasoning templates adopted in our tuning strategy. Empirically, MedQA (USMLE) yields relatively higher training loss values than MedMCQA, which presumably arises from the higher cognitive complexity of professional clinical cases compared to routine examination questions. The favorable low steady loss, especially achieved by the Qwen3.5 backbone, is largely attributed to our strict label-masking design. This design primarily constrains gradient updates toward critical diagnostic response tokens during parameter optimization.

To secure satisfactory generalization ability and mitigate overfitting risks, we adopt an early stopping mechanism throughout fine-tuning. We monitor validation loss at fixed intervals and halt training automatically if no performance gain is detected across three consecutive evaluation rounds. This mechanism helps the derived model parameters maintain a reasonable trade-off between convergence degree and reasoning generalization capability. The convergence curves demonstrate that the CECR-Robust tuning scheme supports effective knowledge assimilation, with no obvious optimization instability observed during training.