Knowledge-Augmented Multimodal Clinical Rationale Generation for Disease Diagnosis with Small Language Models

Anonymous ACL submission

Abstract

Interpretation is critical for disease diagnosis, but existing models struggle to balance predictive accuracy with human-understandable rationales. While large language models (LLMs) offer strong reasoning abilities, their clinical use is limited by high computational costs and restricted multimodal reasoning ability. Small language models (SLMs) are efficient but lack advanced reasoning for integrating multimodal medical data. In addition, both LLMs and SLMs lack of domain knowledge for trustworthy reasoning. Therefore, we propose ClinRaGen, enhancing SLMs by leveraging LLM-derived reasoning ability via rationale distillation and domain knowledge injection for trustworthy multimodal rationale generation. Key innovations include a sequential rationale distillation framework that equips SLMs with LLM-comparable mutlimodal reasoning abilities, and a knowledge-augmented attention mechanism that jointly unifies multimodal representation from time series and textual data in a same encoding space, enabling it naturally interpreted by SLMs while incorporating domain knowledge for reliable rationale generation. Experiments on real-world medical datasets show that ClinRaGen achieves state-of-the-art performance in disease diagnosis and rationale generation, demonstrating the effectiveness of combining LLM-driven reasoning with knowledge augmentation for improved interpretability.

1 Introduction

The widespread adoption of electronic health records (EHRs) has transformed deep learning applications in healthcare by providing diverse data modalities, including medical notes, laboratory (lab) test results, and clinical events. These multimodal inputs are crucial for disease diagnosis, mortality prediction, and drug discovery (Niu et al., 2024; Laghuvarapu et al., 2024). Large language models (LLMs) have recently demonstrated strong diagnostic performance and reasoning capabilities through techniques such as prompt learning and Chain-of-Thought (CoT) reasoning (Wei et al., 2022; Singhal et al., 2023; Chen et al., 2023). However, despite these advancements, LLMs face significant challenges in real-world clinical deployment due to high computational costs, the need for external domain-specific data integration, and difficulties in processing multimodal inputs—particularly numerical time-series lab test. More critically, LLMs lack the ability to generate clinically grounded multimodal rationales, limiting their interpretability in medical decision-making.

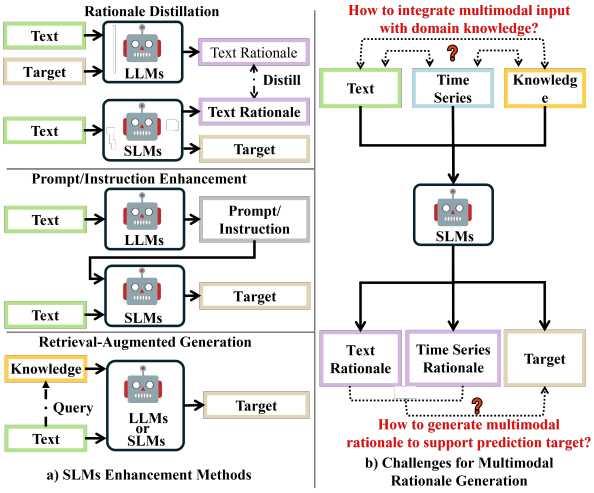

Figure 1: Existing SLM enhancement methods and challenges in multimodal rationale generation.

Small language models (SLMs) have emerged as a computationally efficient alternative, benefiting from recent advancements in rationale distillation, prompt learning, and retrieval-augmented generation (RAG) (Hsieh et al., 2023; Kang et al., 2024; Kwon et al., 2024). As shown in Figure 1a, these methods enable SLMs to inherit LLM-driven reasoning abilities, improve generalization through instruction-based adaptation, or leverage RAG for

more reliable outputs. However, as illustrated in Figure 1b, these approaches still suffer from two fundamental challenges. The first challenge is that they struggle to effectively integrate multimodal inputs with structured domain knowledge, as most methods focus on single-modality data (e.g., text-based rationales) rather than jointly processing textual and time series EHR data (Shi et al., 2024; Sohn et al., 2024). The second challenge is that they fail to provide coherent multimodal rationales that align with clinical decision-making, as rationale generation often remains text-centric and lacks interpretability across different data modalities.

To bring the best of both worlds, we propose ClinRaGen, a knowledge-augmented framework for multimodal clinical rationale generation. ClinRaGen enhances SLMs' trustworthy mutlimodal reasoning capabilities from two aspects. First, it transfers LLM-derived reasoning to SLMs through a sequential rationale distillation paradigm. Second, unlike approaches that rely solely on LLM-generated rationales (Kwon et al., 2024) or resource-intensive RAG (Kang et al., 2024), we propose a knowledge-augmented attention mechanism that achieves dual functionality: Efficient integration of external medical knowledge to enable multimodal rationale generation grounded in clinical validity, ensuring the production of clinically meaningful explanations; Unification of time-series and textual EHRs within a shared encoding space, thereby enhancing multimodal representation learning and facilitates interpretable decision-making.

The main contributions of this paper are:

- We propose ClinRaGen, a multimodal framework that transfer LLM reasoning capabilities into SLMs for disease diagnosis and clinical rationale generation, achieving both accuracy and interpretability.
- We introduce a knowledge-augmented attention mechanism that jointly encodes time-series EHRs into clinical textual representations while injecting domain knowledge, significantly improving mutlimodal rationale reliability and accuracy.
- State-of-the-art performance in disease diagnosis and rationale generation, validated through extensive experiments on benchmark EHR datasets (Johnson et al., 2016, 2023).

2 Related Work

Recent advancements in large-scale high-quality datasets and computational resources have enabled significant progress in Natural Language Processing (NLP), with improved training methodologies fueling the development of LLMs (Touvron et al., 2023; Achiam et al., 2023). In healthcare, LLMs have been applied to clinical question answering and diagnostic reasoning (Singhal et al., 2023; Yang et al., 2022). While effective in text-based tasks, these models struggle to generate clinically grounded multimodal rationales. Medical-specific LLMs (Chen et al., 2023; Zhang et al., 2023) mitigate this issue through domain adaptation, but their high computational costs and reliance on large-scale training data limit scalability.

To improve efficiency, rationale distillation transfers LLM-derived reasoning ability to SLMs, reducing computational overhead while preserving interpretability (Hsieh et al., 2023; Ho et al., 2023; Kang et al., 2024). Chain-of-thought prompting further enhances SLM reasoning capabilities (Wei et al., 2022). However, most distillation approaches remain text-centric and lack robust multimodal EHRs integration (Kang et al., 2024; Ho et al., 2023). RAG has been explored to improve rationale reliability by incorporating external knowledge, yet retrieval latency and adaptability remain key challenges (Jiang et al., 2025). Despite these advancements, multimodal rationale generation remains an open challenge. Current models struggle to fuse textual, time-series, and structured medical knowledge into coherent clinical rationales.

3 Methodology

We introduce ClinRaGen, a knowledge-augmented framework designed to enhance disease diagnosis and clinical rationale generation in SLMs by integrating LLM-derived reasoning and structured domain knowledge. ClinRaGen bridges the gap between large-scale medical knowledge and efficient multimodal reasoning, enabling the generation of two types of rationales: 1). medical note-based rationales ($\boldsymbol{R}^m$) and 2). lab test-based rationales ($\boldsymbol{R}^t$) from medical notes ($\boldsymbol{M}$), time-series lab test results ($\boldsymbol{T}$), and disease-specific knowledge ($\boldsymbol{K}$).

ClinRaGen consists of two key components: Knowledge Retrieval and LLM-Guided Rationale Generation (Section 3.1), which collects domain knowledge and generates LLM-derived rationales

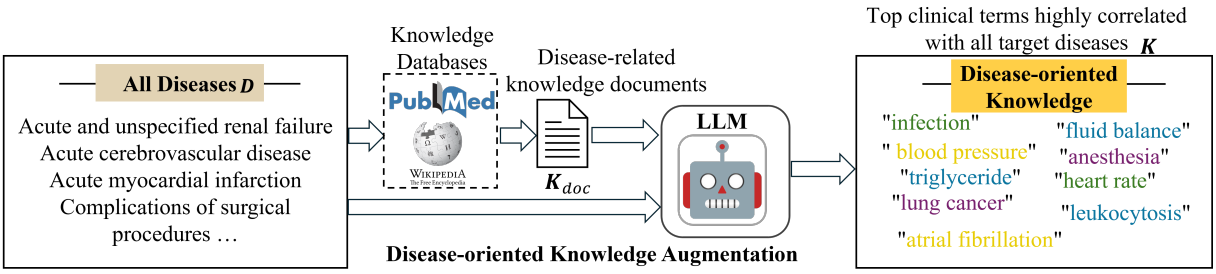

Figure 2: Knowledge augmentation in ClinRaGen. Given diagnosed diseases $\boldsymbol{D}$, relevant descriptions $\boldsymbol{K}_{doc}$ are retrieved and processed by an LLM to extract key clinical terms $\boldsymbol{K}$, enhancing multimodal rationale generation.

as distillation data for subsequent model training, and Knowledge-augmented Attention with Sequential Multimodal Rationale Distillation (Section 3.2), which progressively integrates structured knowledge to enhance multimodal reasoning in SLMs.

3.1 Knowledge Retrieval and LLM-Guided Rationale Generation

This step focuses on gathering domain knowledge and leveraging LLMs to generate structured rationales. The generated rationales serve as distillation targets for training SLMs in later stages. This ensures that SLMs receive high-quality, structured reasoning data to develop robust multimodal reasoning capabilities.

3.1.1 Collecting Domain-Specific Medical Knowledge

LLMs encode extensive medical knowledge but are computationally expensive and impractical for direct deployment. Meanwhile, SLMs such as Flan-T5 and Flan-PaLM (Chung et al., 2024) are computationally efficient but lack sufficient domain-specific expertise to perform complex medical reasoning (Kang et al., 2024; Ho et al., 2023). To bridge this gap, ClinRaGen retrieves relevant medical knowledge from external sources and structures it for integration into SLM training.

As shown in Figure 2, ClinRaGen collects disease-related documents $\boldsymbol{K}_{doc}$ from PubMed[1] and Wikipedia[2], extracting key medical terms using LLM-based processing to construct a structured knowledge base $\boldsymbol{K}$:

$$\boldsymbol{K} = \underset{\boldsymbol{K}'}{\text{argmax}} P_{LLM}(\boldsymbol{K}' \mid \boldsymbol{D}, \boldsymbol{K}_{doc}). \quad (1)$$

This structured knowledge base is not used directly by the SLM during inference but instead supports rationale generation in the next step. The retrieval and extraction process iterates until a stable set of key medical terms is obtained.

3.1.2 Generating Rationales for Distillation

ClinRaGen employs LLMs to generate structured rationales that serve as distillation targets for SLM training. Unlike direct knowledge retrieval(Kang et al., 2024; Jiang et al., 2025), this step synthesizes structured explanations that explicitly link medical knowledge with clinical decision-making, enabling SLMs to internalize complex reasoning patterns during later training stages. To construct high-quality rationale data, we collaborated with clinicians to curate representative EHR samples and formulate corresponding gold-standard rationales $\boldsymbol{O}$. These rationales guide the LLM in generating structured explanations, ensuring that the distilled knowledge supports multimodal reasoning. To improve LLM comprehension of numerical lab test data, we applied anomaly detection (Vinutha et al., 2018) and designed structured prompts that convert numerical values into interpretable textual explanations $\boldsymbol{T}^*$ (see Appendix A.1).

Figure 3 illustrates the multimodal rationale generation process. ClinRaGen first generates rationales ($\boldsymbol{R}^m$) based on medical notes:

$$\boldsymbol{R}^m = \underset{\boldsymbol{R}'}{\text{argmax}} P_{LLM}(\boldsymbol{R}' \mid \boldsymbol{M}, \boldsymbol{D}, \boldsymbol{O}). \quad (2)$$

Then, lab test-based rationales ($\boldsymbol{R}^t$) are generated using insights from both medical notes, time series anomalies, and the generate note-based rationales:

$$\boldsymbol{R}^t = \underset{\boldsymbol{R}'}{\text{argmax}} P_{LLM}(\boldsymbol{R}' \mid \boldsymbol{M}, \boldsymbol{T}^*, \boldsymbol{D}, \boldsymbol{O}, \boldsymbol{R}^m). \quad (3)$$

These LLM-generated rationales form the foundation of the subsequent distillation process (detailed in Section 3.2) and enable SLMs to learn structured, multimodal reasoning efficiently. For further

[1] https://pubmed.ncbi.nlm.nih.gov/
[2] https://www.wikipedia.org/

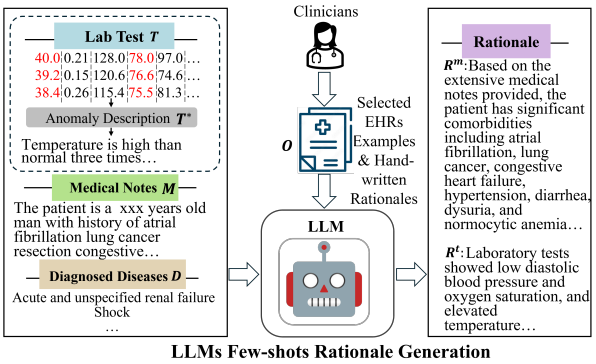

Figure 3: LLM-based clinical rationale generation. Medical notes ($\boldsymbol{M}$), lab test results ($\boldsymbol{T}$ and $\boldsymbol{T}^*$), diagnosis ($\boldsymbol{D}$), and clinicians provide examples ($\boldsymbol{O}$) are used to produce medical note-based ($\boldsymbol{R}^m$) and lab test-based ($\boldsymbol{R}^t$) rationales.

details on data processing and prompt engineering, refer to Appendix A.2.

3.2 Multimodal Rationale Distillation

Figure 4a presents the ClinRaGen framework, which comprises a Time Series Encoder for processing numerical lab test data, a Knowledge-Augmented Attention Module for integrating structured domain knowledge, and a SLM for generating disease diagnoses and structured clinical multimodal rationales. The framework enables progressive multimodal reasoning by leveraging structured knowledge and sequential learning mechanisms.

As illustrated in Figure 4b, ClinRaGen employs a three-phase rationale distillation paradigm that systematically integrates textual, numerical, and structured domain knowledge. The first phase distills medical note-based rationales, allowing the SLM to develop a foundational understanding of textual clinical information. The second phase introduces knowledge-augmented attention, aligning numerical lab test with structured medical knowledge to for distilling lab test-based rationales. The final phase fully integrates textual and numerical inputs, enabling the SLM to generate clinically coherent multimodal rationales to support disease diagnosis.

3.2.1 Phase 1: Rationale Distillation from Medical Notes

In the first phase, the SLM is trained exclusively on medical notes $\boldsymbol{M}$ to establish a foundational understanding of clinical reasoning. This stage enables the model to generate disease diagnoses $\boldsymbol{D}$ while also producing medical note-based rationales $\boldsymbol{R}^m$ and lab test-based rationales $\boldsymbol{R}^t$. By learning to extract meaningful insights from structured textual data, the SLM develops its initial ability to infer clinical relationships. The model is trained using a language model generation objective:

$$\mathcal{L}_{note}(\theta) = \mathbb{E}[-\log P_{SLM_\theta}(\boldsymbol{D}, \boldsymbol{R}^m, \boldsymbol{R}^t \mid \boldsymbol{M})], \tag{4}$$

where θ represents the trainable parameters of the SLM. This phase not only enables the model to internalize explicit diagnostic reasoning from medical notes but also allows it to implicitly capture latent patterns associated with lab test results, laying the groundwork for multimodal integration in subsequent phases.

3.2.2 Phase 2: Knowledge Injection and Time-Series Rationale Distillation

To enable the SLM to effectively interpret numerical lab test data and generate time-series-based rationales ($\boldsymbol{R}^t$) that support disease diagnosis ($\boldsymbol{D}$), we introduce a Knowledge-Augmented Attention (KA) Module. This mechanism integrates domain-specific medical knowledge into the reasoning process, enhancing the model's ability to produce clinically coherent and robust multimodal rationales.

We first use a Time Series Encoder (TSE) to encode raw lab test values $\boldsymbol{T}$ into structured hidden embeddings $\boldsymbol{T}^e$. To align domain knowledge with the SLM, we construct a domain-specific vocabulary $\boldsymbol{V}^k$ by filtering standard language vocabulary $\boldsymbol{V}$ based on structured medical knowledge $\boldsymbol{K}$:

$$\boldsymbol{V}^k = \{v_1, \ldots, v_n \mid v_1 \in \boldsymbol{K}, \boldsymbol{K} \subseteq \boldsymbol{V}\}. \tag{5}$$

A cross-attention mechanism is then applied to integrate knowledge-driven representations into the model. The lab test embeddings ($\boldsymbol{T}^e$) serve as the Query, while domain knowledge tokens ($\boldsymbol{V}^k$) act as the Key and Value:

$$\begin{aligned} \boldsymbol{H} &= f_\phi(\boldsymbol{T}, \boldsymbol{V}^k), \\ &= SoftMax\Big(\frac{(\boldsymbol{T}^e\boldsymbol{W}^q)(\boldsymbol{V}^{k\top}\boldsymbol{W}^k)}{\sqrt{d}}\Big)(\boldsymbol{V}^k\boldsymbol{W}^v), \end{aligned} \tag{6}$$

where d is the hidden dimension of the SLM, and $\boldsymbol{W}^q$, $\boldsymbol{W}^k$, $\boldsymbol{W}^v$ are learnable attention weight matrices, f indicates the encoding function of the TSE and attention, and ϕ represents the trainable parameters of f. The resulting knowledge-enhanced embeddings $\boldsymbol{H}$ are then fed into the frozen distilled SLM to refine its reasoning and generate lab test-based rationales ($\boldsymbol{R}^t$) and diagnosis ($\boldsymbol{D}$).

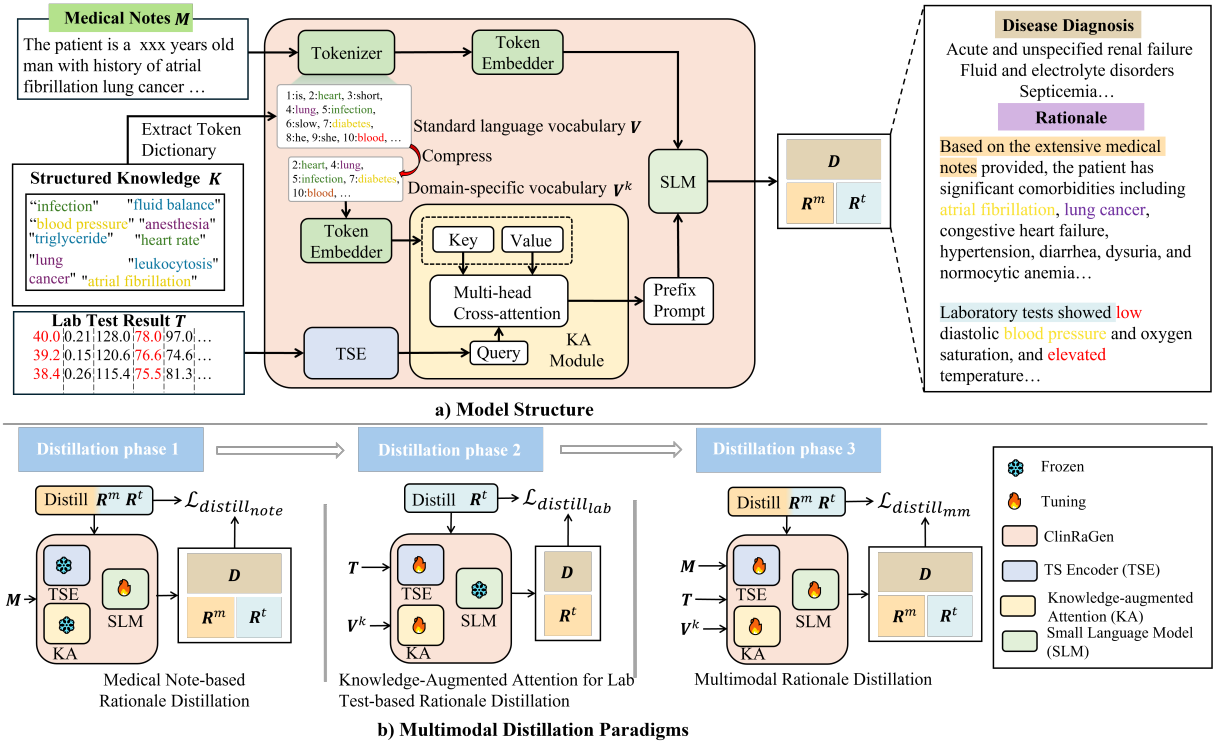

Figure 4: Overview of ClinRaGen. (a) Model structure comprising a time series encoder, knowledge-augmented attention module, and a SLM. (b) Three-phase rationale distillation: Medical Note-based Rationale Distillation, Knowledge-Augmented Attention for Lab Test-based Rationale Distillation, and Multimodal Rationale Distillation.

The model is trained using the following objective function:

$$\mathcal{L}_{lab}(\phi) = \mathbb{E}[-\log P_{SLM_\theta}(\boldsymbol{D}, \boldsymbol{R}^t \mid \boldsymbol{H})]. \quad (7)$$

This phase ensures that the SLM can naturally interpret lab test while effectively leveraging medical knowledge to enhance its reasoning capabilities.

3.2.3 Phase 3: Full Multimodal Rationale Distillation

In the final phase, ClinRaGen is trained to generate full multimodal clinical rationales by integrating medical notes, lab test, and structured domain knowledge. To ensure effective multimodal reasoning, lab test $\boldsymbol{T}$ is formatted as prefix prompts (Niu et al., 2024), allowing the SLM seamlessly incorporates it with textual EHRs.

During this stage, the model is optimized to generate both medical note-based rationales ($\boldsymbol{R}^m$) and lab test-based rationales ($\boldsymbol{R}^t$), ensuring that all available information contributes to clinically coherent and interpretable decision-making. The multimodal rationale distillation objective is formulated as follows:

$$\mathcal{L}_{mm}(\theta, \phi) = \mathbb{E}[-\log P_{SLM_\theta}(\boldsymbol{D}, \boldsymbol{R}^m, \boldsymbol{R}^t \mid \boldsymbol{M}, f_\phi(\boldsymbol{T}, \boldsymbol{V}^k))]. \quad (8)$$

The fine-tuning of all ClinRaGen components, ensuring that multimodal EHRs are effectively integrated, enhances diagnostic accuracy and produces modality-consistent rationales.

4 Experiments

4.1 Experimental Settings

Dataset: We evaluate ClinRaGen on two public EHR datasets: MIMIC-III (Johnson et al., 2016) (28,456 EHRs include medical notes and time series lab tests) and MIMIC-IV (Johnson et al., 2023) (28,900 EHRs). Both datasets use benchmark tools (Harutyunyan et al., 2019) for time series processing, with missing values filled by nearest available data. We target 25 disease phenotypes and follow a 4:1 training-to-testing split (Harutyunyan et al., 2019). Our model is available at github[3].

Baseline Methods: To evaluate the effectiveness of ClinRaGen for disease diagnosis generation, we compared it with following baselines: Flan-T5 (Chung et al., 2024), PROMPTEHR (Wang and Sun, 2022), FROZEN (Tsimpoukelli et al., 2021), EHR-KnowGen (Niu et al., 2024), Clinical CoT

[3]https://anonymous.4open.science/r/ClinRaGen-6C9D/

Models	Size	Modality		Micro			Macro		
		Lab	Note	Precision	Recall	F1	Precision	Recall	F1
MIMIC-III									
Flan-T5	60M		✓	$0.5812_{(0.11)}$	$0.6623_{(0.07)}$	$0.6203_{(0.05)}$	$0.5656_{(0.10)}$	$0.6247_{(0.08)}$	$0.5887_{(0.07)}$
PROMPTEHR	75.2M		✓	$0.5929_{(0.11)}$	$0.6553_{(0.07)}$	$0.6224_{(0.02)}$	$0.5744_{(0.10)}$	$0.6287_{(0.06)}$	$0.5910_{(0.03)}$
LLaMA-ft	7B	✓	✓	$0.6142_{(0.21)}$	$0.6598_{(0.15)}$	$0.6364_{(0.04)}$	$0.6108_{(0.15)}$	$0.6164_{(0.13)}$	$0.6055_{(0.04)}$
FROZEN	265M	✓	✓	$0.6102_{(0.18)}$	$0.6401_{(0.16)}$	$0.6231_{(0.03)}$	$0.5976_{(0.16)}$	$0.6001_{(0.17)}$	$0.5915_{(0.03)}$
EHR-KnowGen	77M	✓	✓	$0.6001_{(0.03)}$	$0.6551_{(0.02)}$	$0.6262_{(0.01)}$	$0.5834_{(0.04}$	$0.6181_{(0.03)}$	$0.5944_{(0.01)}$
Clinical CoT									
-w/o TSE	60M		✓	$0.6115_{(0.03)}$	$0.6402_{(0.04)}$	$0.6311_{(0.03)}$	$0.6024_{(0.04)}$	$0.5989_{(0.06)}$	$0.5969_{(0.03)}$
-w/ TSE	85M	✓	✓	$0.5967_{(0.05)}$	$0.6607_{(0.06)}$	$0.6328_{(0.03)}$	$0.5924_{(0.06)}$	$0.6092_{(0.07)}$	$0.5975_{(0.05)}$
LLM Zero-shot									
-LLaMA	7B	✓	✓	$0.1227_{(0.08)}$	$0.0421_{(0.06)}$	$0.0627_{(0.06)}$	$0.0392_{(0.06)}$	$0.0622_{(0.06)}$	$0.0438_{(0.05)}$
-ChatGPT	175B	✓	✓	$0.4474_{(0.07)}$	$0.1405_{(0.05)}$	$0.2139_{(0.05)}$	$0.4883_{(0.08)}$	$0.1872_{(0.05)}$	$0.2188_{(0.04)}$
ClinRaGen	87M	✓	✓	$0.6104_{(0.02)}$	$0.6751_{(0.02)}$	$\underline{0.6410}_{(0.01)}$	$0.5991_{(0.03)}$	$0.6311_{(0.04)}$	$\underline{0.6113}_{(0.02)}$
ClinRaGen*	793M	✓	✓	$0.6047_{(0.03)}$	$0.6875_{(0.03)}$	$\mathbf{0.6501}_{(0.02)}$	$0.5943_{(0.04)}$	$0.6531_{(0.03)}$	$\mathbf{0.6196}_{(0.03)}$
MIMIC-IV									
Flan-T5	60M		✓	$0.6624_{(0.05)}$	$0.6953_{(0.02)}$	$0.6792_{(0.04)}$	$0.6428_{(0.06)}$	$0.6601_{(0.05)}$	$0.6479_{(0.04)}$
PROMPTEHR	75.2M		✓	$0.6524_{(0.07)}$	$0.7031_{(0.06)}$	$0.6802_{(0.02)}$	$0.6353_{(0.05)}$	$0.6702_{(0.07)}$	$0.6501_{(0.03)}$
LLaMA-ft	7B	✓	✓	$0.6854_{(0.11)}$	$0.6954_{(0.07)}$	$0.6929_{(0.03)}$	$0.6753_{(0.09)}$	$0.6624_{(0.11)}$	$0.6621_{(0.06)}$
FROZEN	265M	✓	✓	$0.6781_{(0.08)}$	$0.6908_{(0.09)}$	$0.6842_{(0.01)}$	$0.6627_{(0.10)}$	$0.6521_{(0.10)}$	$0.6530_{(0.02)}$
EHR-KnowGen	77M	✓	✓	$0.6580_{(0.06)}$	$0.7085_{(0.05)}$	$0.6816_{(0.02)}$	$0.6382_{(0.05)}$	$0.6724_{(0.06)}$	$0.6511_{(0.02)}$
Clinical CoT									
-w/o TSE	60M		✓	$0.6751_{(0.05)}$	$0.7069_{(0.03)}$	$0.6905_{(0.03)}$	$0.6607_{(0.04)}$	$0.6796_{(0.06)}$	$0.6612_{(0.02)}$
-w/ TSE	85M	✓	✓	$0.7011_{(0.04)}$	$0.6808_{(0.06)}$	$0.6917_{(0.04)}$	$0.6971_{(0.05)}$	$0.6354_{(0.03)}$	$0.6577_{(0.03)}$
LLM Zero-shot									
-LLaMA	7B	✓	✓	$0.1357_{(0.11)}$	$0.0997_{(0.07)}$	$0.1150_{(0.06)}$	$0.0435_{(0.09)}$	$0.1466_{(0.07)}$	$0.0619_{(0.05)}$
-ChatGPT	175B	✓	✓	$0.4536_{(0.07)}$	$0.1458_{(0.05)}$	$0.2207_{(0.04)}$	$0.4532_{(0.06)}$	$0.1831_{(0.06)}$	$0.2147_{(0.05)}$
ClinRaGen	87M	✓	✓	$0.7009_{(0.01)}$	$0.6963_{(0.02)}$	$\underline{0.6989}_{(0.01)}$	$0.6868_{(0.03)}$	$0.6603_{(0.01)}$	$\underline{0.6685}_{(0.02)}$
ClinRaGen*	793M	✓	✓	$0.6848_{(0.04)}$	$0.7429_{(0.02)}$	$\mathbf{0.7127}_{(0.02)}$	$0.6779_{(0.02)}$	$0.7087_{(0.01)}$	$\mathbf{0.6893}_{(0.01)}$

Table 1: The performance of comparative methods in the disease diagnosis tasks on MIMIC-III and MIMIC-IV. The best results are highlighted in bold, and the second-best results are marked with an underline.

(with/without TSE) (Kwon et al., 2024), and LLM-based models LLaMA-7B (Touvron et al., 2023) (zero-shot and fine-tuning) and ChatGPT (Open, 2023) (zero-shot). Baseline and implementation details are provided in Appendices A.3 and A.4. For a fair comparison, all baselines (except LLaMA) use Flan-T5-Small as the backbone; our model employs Flan-T5-Small (ClinRaGen) and Flan-T5-Large (ClinRaGen*) for evaluate effect of varying scales. ChatGPT (GPT-3.5-turbo) serves as our teacher LLM. Results are averaged over five runs with statistical significance determined at $p < 0.05$ by t-test.

4.2 Disease Diagnosis Performance

Comparison with Baselines: We evaluate disease diagnosis using micro and macro precision, recall, and F1 scores. Table 1 shows that multimodal models outperform single-modality models, confirming the value of lab test results. Clinical CoT surpasses other baseline models, highlighting rationale distillation's effectiveness. ClinRaGen (80M) achieves the best performance, with an average F1 score improvement of over 1.1% across all baselines, even outperforming LLaMA-7B-ft. Furthermore, ClinRaGen* (793M) improves by over 1.5%, significantly exceeding other baselines. The weak performance of zero-shot LLMs confirms the absence of data leakage. These results demonstrate ClinRaGen's ability to match or surpass LLMs in clinical tasks through multimodal rationale distillation and the knowledge-augmented attention mechanism.

Ablation Study: We assess the impact of key components in ClinRaGen: (1) *w/o* LAB&KNOW removes lab tests and knowledge input, (2) *w/o* KNOW replaces the knowledge-based vocabulary with a standard one, and (3) *w/o* REASONING excludes rationale distillation while maintaining model structure. Table 2 shows that *w/o* REASONING performs worst, highlighting the importance of rationale distillation. The drop in F1 scores for *w/o* LAB&KNOW confirms the value of multimodal integration, while *w/o* KNOW shows the KA module's contribution to diagnostic accuracy.

Model Efficiency: We evaluate ClinRaGen's efficiency by comparing model parameters, micro F1 scores (Figure 5), and training times (Table 3).

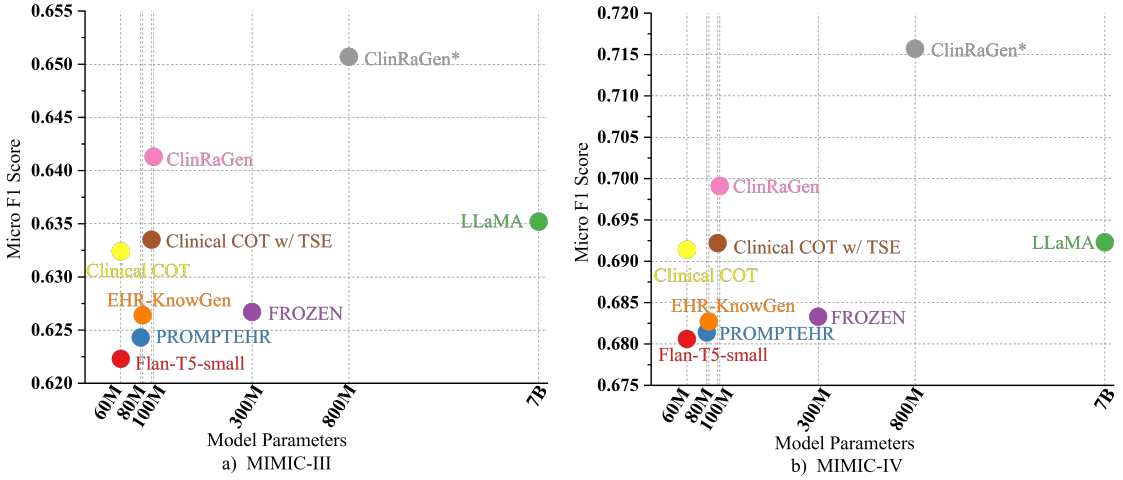

Figure 5: Model Parameter Counts and Micro F1 Scores

Models	Micro F1	Macro F1
MIMIC-III		
ClinRaGen	**0.6410**	**0.6113**
w/o LAB&KNOW	0.6323	0.6021
w/o KNOW	0.6349	0.6042
w/o REASONING	0.6255	0.5915
MIMIC-IV		
ClinRaGen	**0.6989**	**0.6685**
w/o LAB&KNOW	0.6925	0.6643
w/o KNOW	0.6936	0.6644
w/o REASONING	0.6828	0.6541

Table 2: Ablation studies on disease diagnosis.

ClinRaGen-Small achieves superior diagnostic performance with 80× fewer parameters and less than half the training time of LLaMA. These results highlight the effectiveness of our sequential multimodal distillation paradigm and KA mechanism in enabling efficient and accurate clinical reasoning.

4.3 Rationale Generation Performance

Evaluation Methods: To assess the quality of generated multimodal rationales and maximize the potential of SLMs, we evaluate ClinRaGen (80M) using five evaluation criteria—*Correctness*, *Readability*, *Soundness*, *Consistency*, and *Persuasiveness*—based on clinicians and prior research (Lin et al., 2024; Kwon et al., 2024). Scores range from 1 to 5 on a Likert scale (details of criteria defined in Appendix A.5). We conduct both LLM-based and human evaluations. For LLM comparisons, we use Mistral-7B, LLaMA2-7B, and LLaMA3-8B with five-shot prompting. Distilled rationales from ChatGPT serve as ground truth (GT). Comparative LLMs receive time series anomalies and medical notes, while ClinRaGen directly processes numerical lab test and medical notes. Following Lin et al. (2024); Chiang and Lee (2023), we use GPT-4 to evaluate 1000 randomly selected samples. For human assessment, 15 professional postgraduates rate 100 samples, achieving moderate intra-class (0.637) and inter-class (0.608) agreement, indicating reasonable consistency despite the task's subjectivity.

Models	Time Cost (Seconds)
Knowledge Retrieval	12,636
LLM-Guided Rationale Generation	604,715
LlaMA – 7B Tuning	259,113
ClinRaGen – 84M Tuning	94,623

Table 3: Time Cost Evaluation.

Evaluation Results: Figures 6(a) and (b) show GPT-4 and human evaluations across five criteria, with closely aligned results. LLaMA3 performs best, benefiting from its large scale and pre-training. ClinRaGen ranks second, matching LLaMA3 in readability and correctness while surpassing LLaMA2 and Mistral, which often generate incoherent rationales. Unlike other LLMs relying on anomaly captions, ClinRaGen achieves the second-highest consistency score, demonstrating the KA mechanism's effectiveness in consist multimodal reasoning. ClinRaGen also outperforms LLaMA2 and Mistral in soundness and persuasiveness, further underscoring our method's effectiveness. Appendix A.6 further validates rationale quality using BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2019).

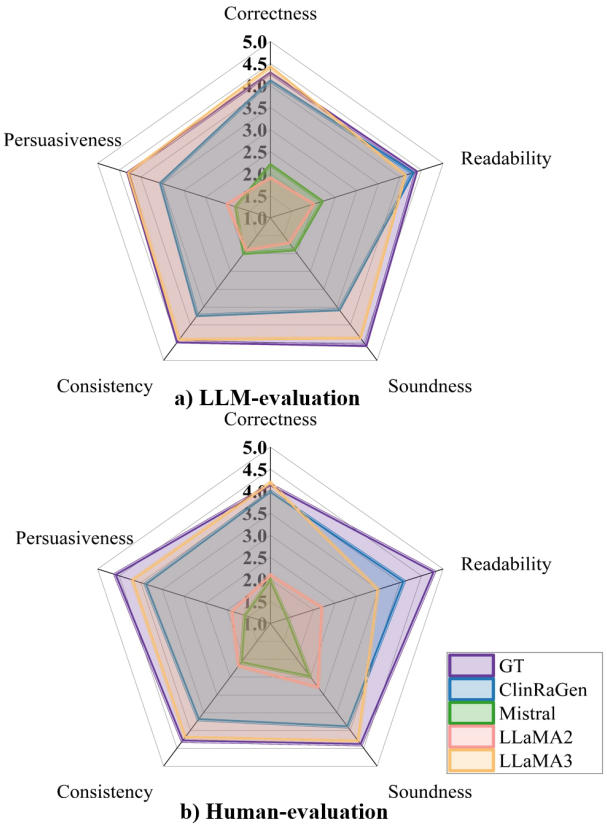

Figure 6: Clinical rationale generation evaluation

Example 1
Disease Diagnosed:
Acute cerebrovascular disease, Complications of surgical procedures or medical care, Disorders of lipid metabolism
LLM Rationale:
Based on the medical notes, the patient... The presence of basal ganglia hemorrhage raises suspicion for acute cerebrovascular disease... Lab test shows repeated elevated diastolic and mean blood pressures indicating acute cerebrovascular disease post craniotomy...
ClinRaGen Rationale:
Based on the medical notes provided, the patient... The patient exhibited signs of acute cerebrovascular disease, including weakness, lethargy, and basal ganglia hemorrhage. Management included heparin drip, antiplatelet therapy, and blood pressure control... Lab test shows transiently elevated diastolic blood pressure and mean blood pressure post craniotomy for aneurysm clipping...

Example 2
Disease Diagnosed:
Chronic obstructive pulmonary disease and bronchiectasis, Coronary atherosclerosis and other heart disease, Disorders of lipid metabolism, Essential hypertension, Cardiac dysrhythmias
LLM Rationale:
Based on the medical notes provided, the patient … hypertension, or cardiac dysrhythmias in the provided medical notes. Therefore, the rationale for diagnosis is "no disease was diagnosed" due to the focus on the acute neurosurgical intervention and recovery… Lab test shows episodes of significantly elevated fraction inspired oxygen, supportive of hypoxemia following…
ClinRaGen Rationale:
Based on the medical notes provided, the patient …. Therefore, the comprehensive rationale for the diagnosed diseases includes acute cerebrovascular disease, disorders of lipid metabolism, and essential hypertension… Lab test shows transiently elevated fraction inspired oxygen and increased mean blood pressure, consistent…

Figure 7: Case studies on disease diagnosis and clinical rationale generation compared with teacher LLM.

Case Studies: As illustrated in Figure 7, our model ClinRaGen can produce both medical note-based rationales (e.g., "*Based on the medical notes...*") and lab test-based rationales (e.g., "*Lab test shows...*"), akin to the outputs of teacher LLM. For medical note-based rationale generation, ClinRaGen effectively extracts key medical terms essential for disease diagnosis (highlighted in green). Additionally, for lab test-based rationales, our model accurately identifies abnormal lab test features (highlighted in blue), demonstrating its capability to understand numerical time series lab test data effectively. These results indicate that ClinRaGen competently produces clinically relevant multimodal rationales to support disease diagnosis.

Further Discussions: LLMs may introduce bias into distilled clinical rationales. To evaluate ClinRaGen's correctness, we assess the relevance of key medical terms to diagnosed diseases. In one case, our model identifies *weakness*, *lethargy*, and *basal ganglia hemorrhage* as evidence for *acute cerebrovascular*, while teacher LLM captures only *basal ganglia hemorrhage*, missing relevant symptoms (Unnithan et al., 2023). In another case, while the teacher LLM reports *no disease*, ClinRaGen correctly identifies conditions like *disorders of lipid metabolism* and *essential hypertension*. These results highlight ClinRaGen's ability to mitigate LLM biases by capturing time-series variations and integrating structured knowledge.

5 Conclusion and Future Work

We present ClinRaGen, a knowledge-augmented framework that enhances SLMs with LLM-derived reasoning and structured medical knowledge for disease diagnosis and multimodal rationale generation. It introduces a knowledge-augmented attention module that jointly unifies time-series and textual EHRs in the same encoding space while injecting domain knowledge for reliable rationale generation and a sequential multimodal distillation paradigm for transferring LLMs' reasoning capabilities to SLMs. Extensive evaluations on real world datasets, including quantitative and qualitative analyses, show that ClinRaGen enables SLMs to achieve LLM-comparable performance in disease diagnosis and multimodal rationale generation. This work bridges the performance gap between LLMs and SLMs in clinical tasks. Future research will extend ClinRaGen to a broader range of SLM architectures, datasets, and medical applications.

Limitations

While ClinRaGen effectively enhances multimodal clinical reasoning, certain limitations remain:

- First, although rationale distillation transfers reasoning capabilities from LLMs to SLMs, potential biases in LLM-generated rationales may persist.
- Second, the effectiveness of the knowledge-augmented attention module depends on the quality and coverage of external knowledge sources.
- Lastly, ClinRaGen is evaluated on structured EHR datasets, and its applicability to unstructured clinical text or other medical modalities requires further exploration.

Future work will refine knowledge integration, enhance bias mitigation strategies, and extend evaluations to diverse clinical settings.

Ethics Statement

Data Privacy: The datasets utilized in our research are publicly accessible and feature de-identified patient data, accessing these datasets still requires passing the CITI Exam[4] and download from Physionet[5]. In addition, this study used the Azure OpenAI service and completed the "opting out of the review process" agreement.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Cheng-Han Chiang and Hung-Yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882.

Cheng-Yu Hsieh, Chun-Liang Li, CHIH-KUAN YEH, Hootan Nakhost, Yasuhisa Fujii, Alex Jason 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. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. 2025. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. *ICLR*.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2024. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *Advances in Neural Information Processing Systems*, 36.

Taeyoon Kwon, Kai Tzu-iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Beomseok Sohn, Yongsik Sim, Dongha Lee, and Jinyoung Yeo. 2024. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18417–18425.

Siddhartha Laghuvarapu, Zhen Lin, and Jimeng Sun. 2024. Codrug: Conformal drug property prediction with density estimation under covariate shift. *Advances in Neural Information Processing Systems*, 36.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 2359–2370.

Shuai Niu, Jing Ma, Liang Bai, Zhihua Wang, Li Guo, and Xian Yang. 2024. Ehr-knowgen: Knowledge-

[4] https://about.citiprogram.org/

[5] https://physionet.org/

enhanced multimodal learning for disease diagnosis generation. *Information Fusion*, 102:102069.

AI Open. 2023. Chatgpt (mar 14 version)[large language model].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2024. Rationale-guided retrieval augmented generation for medical question answering. *arXiv preprint arXiv:2411.00300*.

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*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

A K A Unnithan, J M Das, and P Mehta. 2023. Hemorrhagic stroke. In *StatPearls*. StatPearls Publishing.

HP Vinutha, B Poornima, and BM Sagar. 2018. Detection of outliers using interquartile range technique from intrusion dataset. In *Information and Decision Sciences: Proceedings of the 6th International Conference on FICTA*, pages 511–518. Springer.

Zifeng Wang and Jimeng Sun. 2022. Promptehr: Conditional electronic healthcare records generation with prompt learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2873 – 2885. Association for Computational Linguistics.

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.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, et al. 2022. Gatortron: a large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.

A Appendix

A.1 Lab Test Anomaly Caption

To caption lab test results into textual descriptions, we initially employ the Inter Quartile Range (IQR) anomaly detection method (Vinutha et al., 2018) to identify anomalous lab test features. Subsequently, we craft multiple text templates to caption these anomalies. These templates are delineated in Table 4.

A.2 Prompts for Multimodal Rationale Generation Via ChatGPT

The overall procedure for ChatGPT generating clinical rationale is illustrated in Figure 3. The specific medical note-based rationale prompt and lab test-based rationale prompt are detailed as follows. **Medical note-based rationale prompt** for ChatGPT,

"*Below is an instruction that describes examples of generating the rationale of disease diagnosis; please refer to the examples style to generate the Output from the Input:*

Instruction:

There are some examples please to refer:

Condition: If the lab test value is not an abnormal value: Prompt: {Lab features} is normal all the time.
Condition: If the lab test value is an abnormal value higher than the standard: Prompt: {Lab features} is higher than normal {number of times} times.
Condition: If the lab test value is an abnormal value lower than the standard: Prompt: {Lab features} is lower than normal {number of times} times.
Condition: If the lab test value is abnormal, it includes both higher and lower than the standard value: Prompt: {Lab features} is higher than normal {number of times} times and lower than normal {number of times} times.

Table 4: Lab test anomaly caption template.

Example 1, Example 2, Example ...

Input:

Medical note: [$\boldsymbol{M}$]

Diagnosed diseases: [$\boldsymbol{D}$]

Please review the patient's medical records. Adhere to the provided format to craft a succinct 100-word rationale for diagnosing these conditions (Start with "Based on the medical notes..."). If the diagnosis indicates "no disease was diagnosed," the rationale must state "no disease was diagnosed." Otherwise, provide a comprehensive rationale for the diagnosis.

Response:

Output:

Medical note-based Rationale: [$\boldsymbol{R}^n$]

Lab test-based rationale prompt for ChatGPT, we denote the lab test anomalies as $\boldsymbol{T}^*$:

"*Below is an instruction that describes examples of generating the rationale of disease diagnosis, please refer to the examples style to generate the Output from the Input:*

Instruction:

There are some examples please to refer:

Example 1, Example 2, Example ...

Input:

Medical note: [$\boldsymbol{M}$]

Descriptions of lab test abnormalities: [$\boldsymbol{T}^*$]

Diagnosed diseases: [$\boldsymbol{D}$]

Medical note-based rationale: [$\boldsymbol{R}^n$]

Please review the patient's medical notes, laboratory test anomaly results, and existing rationales in the medical record. Construct a concise, one-sentence rationale, limited to max 50 words, that accurately describes a diagnosed condition based on descriptions of laboratory test abnormalities (Start with "Lab test shows..."). Pay close attention to potential inaccuracies in the lab descriptions.

Response:

Output:

Lab test-based rationale: [$\boldsymbol{R}^t$]

A.3 Baseline Details

- **Flan-T5**: Flan-T5 is introduced in the scaling instruction-fine-tuning method for language models (Chung et al., 2024). It is trained on comprehensive datasets designed for tasks like summarization, question answering, and reasoning, enhancing its chain-of-thought capabilities.
- **PROMPTEHR**: PROMPTEHR (Wang and Sun, 2022) innovates generative modelling for EHRs through conditional prompt learning; in this experiment, we focus on applying it, particularly on disease diagnosis.
- **LLaMA**: The LLaMA-7B model (Touvron et al., 2023), a prominent large language model, employs Reinforcement Learning with Human Feedback (RLHF) and instructional tuning, showcasing its adaptability across diverse NLP tasks. This study applied both zero-shot and fine tuning for disease diagnosis.
- **FROZEN**: The FROZEN framework (Tsimpoukelli et al., 2021) stands out in multimodal vision-language modeling for few-shot learning. Here, it's tailored to disease diagnosis,

analyzing both lab test results and medical notes.

- **EHR-KnowGen**: As a leading model in EHR multimodal learning, EHR-KnowGen (Niu et al., 2024) specializes in generating disease diagnoses. This study excludes external knowledge to maintain a balanced evaluation.
- **Clinical CoT**: Clinical CoT (Kwon et al., 2024) integrates clinical reasoning into a diagnostic framework for EHRs using prompt-based learning methods distilled from GPT. To ensure a fair comparison, we incorporate the same time series encoder (TSE) as used in our model for multimodal processing.
- **ChatGPT**: ChatGPT (Open, 2023) is a state-of-the-art LLM optimized for conversational applications, such as dialogue, summarization, and text completion.

A.4 Implementation Details

For our experiments, we utilized version 2.0.1 of the PyTorch framework, running on a CUDA 11.7 setup. The training processes were conducted using the DeepSpeed[6] framework. We opted for the AdamW optimizer, starting with a learning rate of $1e^{-5}$ and incorporating a weight decay of 0.05. We implemented a warm-up phase that spanned 10% of the training period. The experimental setup included two NVIDIA A100 GPUs, each with 80 GB of memory. To process time series data consistently, we padded all lab test results to a standard length of 1,000 time steps, dividing the data into 125 patches, where each patch included 8-time steps.

A.5 Rationale Evaluation Metrics

We defined the rationale evaluation metrics for the LLM and human evaluation as follows: 1). *Correctness:* how medically accurate the rationale supports the diagnosis results. 2). *Readability:* the extent to which a clinical rationale adheres to proper grammar and structural rules. 3). *Soundness:* the logical coherence and insight provided by the clinical rationale. 4). *Consistency:* the degree of alignment between the clinical rationale derived from medical notes and lab test results. 5). *Persuasiveness:* the effectiveness of the clinical rationale in convincing the reader of its validity.

Evaluation scores based on Likert scale:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

A.6 Rationale Evaluation with BLUE and BERTScore.

In addition to the criteria defined for evaluating rationale performance, Table 5 presents the performance of our model, ClinRaGen, alongside various baselines, using both BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2019) on the MIMIC-III and MIMIC-IV datasets. The results show that ClinRaGen outperformed all other models in both metrics across the datasets. The latest open-source LLM, LLaMA3, ranked second, while Mistral exhibited the poorest performance. These results are consistent with those from LLM and human evaluations.

Models	MIMIC-III		MIMIC-IV	
	BLEU	BERTScore	BLEU	BERTScore
Mistral	0.0163	0.7348	0.0532	0.7625
LLaMA2	0.1441	0.8714	0.2357	0.8808
LLaMA3	0.1641	0.8804	0.2568	0.8919
ClinRaGen	**0.2689**	**0.8972**	**0.2963**	**0.9044**

Table 5: Rationale evaluation with BLEU and BERTScore.

A.7 Discussion on SLMs Selection

In this section, we discuss our choice of Flan-T5 as the base SLM for our research, focusing on the following aspects: 1). **Required CoT ability**: Flan-T5 (Chung et al., 2024) has been extensively instruction-tuned on numerous datasets and hundreds of tasks, endowing it with strong zero-shot, few-shot, and Chain-of-Thought (CoT) abilities that outperform the original T5 (Raffel et al., 2020). In contrast, other SLMs, such as OPT (Zhang et al., 2022) and GPT-2 (Radford et al., 2019), lack these robust CoT capabilities, which is crucial as a initialization ability for further clinical reasoning distillation. 2). **Maximizing SLM potential for practical usage**: Although other instruction-finetuned SLMs (e.g., Flan-PaLM) exist, they have substantially larger parameter counts (ranging from 8B to 540B), which is not practical in real world clinical applications and not our target SLMs to investigate.

[6] https://github.com/microsoft/DeepSpeed

We selected Flan-T5-Small (80M) and Flan-T5-Large (780M) as our base models to maximize the potential of SLMs for accurate disease diagnosis and LLM-comparable multimodal reasoning, while maintaining cost-effectiveness in practical applications. Although we currently use Flan-T5, future work will explore a broader range of SLM architectures to further enhance accuracy.

A.8 Discussion on Teacher LLMs Selection

In this section, we discuss our choice of ChatGPT (GPT-3.5-turbo) as the teacher LLM for our research, focusing on the following aspects: 1). **High quality clinical rationales** : Although ChatGPT is known for its strong language modeling capabilities, its generated rationales may still contain noise and bias—issues that are critical in precision medicine. To address this, we incorporate external medical domain knowledge and introduce a novel knowledge-augmented attention mechanism during multimodal clinical rationale generation. Our extensive experiments (Section 4.3) show that ClinRaGen effectively mitigates incomplete or incorrect diagnoses and rationales distilled from the teacher LLM, thereby reducing the impact of bias in LLM-generated outputs for SLMs distillation. 2) **Test set leakage** : The PhysioNet Credentialed Data Use Agreement prohibits the use of MIMIC-series data in public LLMs' training and applications[7], ensuring that test set leakage is not an issue with ChatGPT. Furthermore, the poor performance of ChatGPT under zero-shot prompting (as shown in Table 1) indicates that MIMIC-III and MIMIC-IV data were not used in its training. 3) **More powerful LLM for evaluation**: While we did not choose the most powerful LLM as our teacher, our current teacher LLM sufficiently enhances SLM capabilities in disease diagnosis and clinical rationale generation. Our evaluations (Sections 4.2 and 4.3), supported by quantitative metrics, comparisons with a superior LLM, and human assessments, confirm the effectiveness of using ChatGPT as the teacher LLM through our multimodal rationale distillation paradigm and knowledge-augmented attention mechanism to improve SLMs' accuracy in disease diagnosis and modality-consistent rationale generation. In future work, more powerful LLMs can be seamlessly integrated with our method to further enhance evaluation accuracy and robustness.

[7] https://physionet.org/news/post/gpt-responsible-use