Optimizing Causality-Based Radiology Reporting with Retrieval-Augmented and Structured Reasoning Approaches for the NTCIR-18 HIDDEN-RAD Task

Seung-Hoon Na(UNIST), Ju-Min Cho · Ho-Jin Yi · Myung-Kyu Kim · Se-Jin Jeong(Jeonbuk National University)

{ nash } @ unist.ac.kr, { properly59, dlghwls7889, rlaaudrb107, jeongsj } @ jbnu.ac.kr

I . Motivation

The Need for Explicit Causality in Radiology

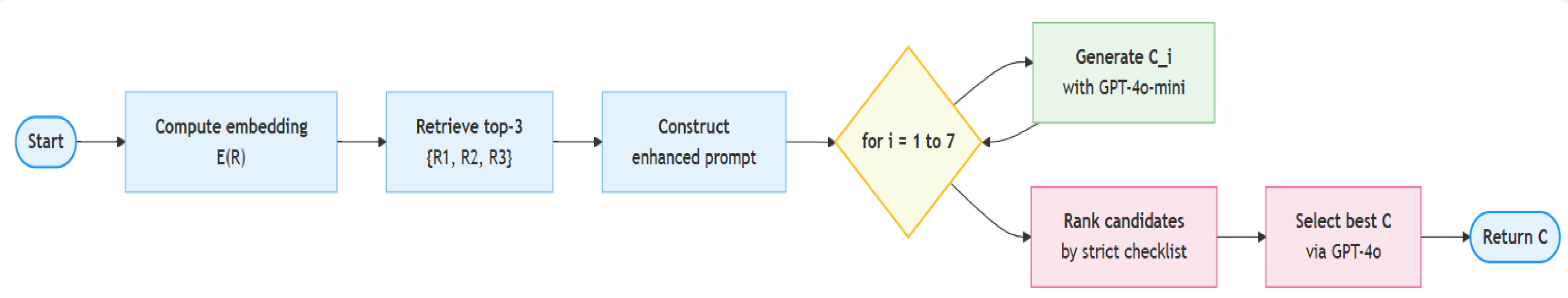
Radiology reports are crucial for diagnosis but often omit the explicit causal reasoning vital for transparency and trust in clinical decision-making by both humans and AI.

NTCIR-18 Hidden-RAD Task

The NTCIR-18 Hidden-RAD Task was established to address this, evaluating AI's ability to generate causality-based explanations by reconstructing radiologists' implicit reasoning processes. The "nash" team focused on the task's two facets: firstly, generating reports with diagnostic inferences by identifying hidden causalities in MIMIC-CXR reports, and secondly, simulating a radiologist's decision-making process to produce reports from initial impression to final conclusion using structured reasoning approaches.

II. Methods

Subtask 1: Retrieval-Augmented Causality Extraction



Algorithm 1: Retrieval-Enhanced Causal Report Generation Pipeline

For identifying hidden causalities, our "nash" team implemented a cost-efficient API-driven inference pipeline. This pipeline integrates few-shot in-context learning, retrieval-enhanced prompting, and a strict candidate selection process using an evaluation checklist. By dynamically retrieving the top-3 most similar cases from the training data, we enriched the prompt to improve contextual alignment. Our two-stage model approach utilized GPT-4o-mini to generate seven diverse candidate outputs per report, followed by the more powerful GPT-4o for final selection, ensuring high-quality causal explanations while optimizing computational costs.

Subtask 2: PRISMA-Guided Structured Diagnostic Reasoning

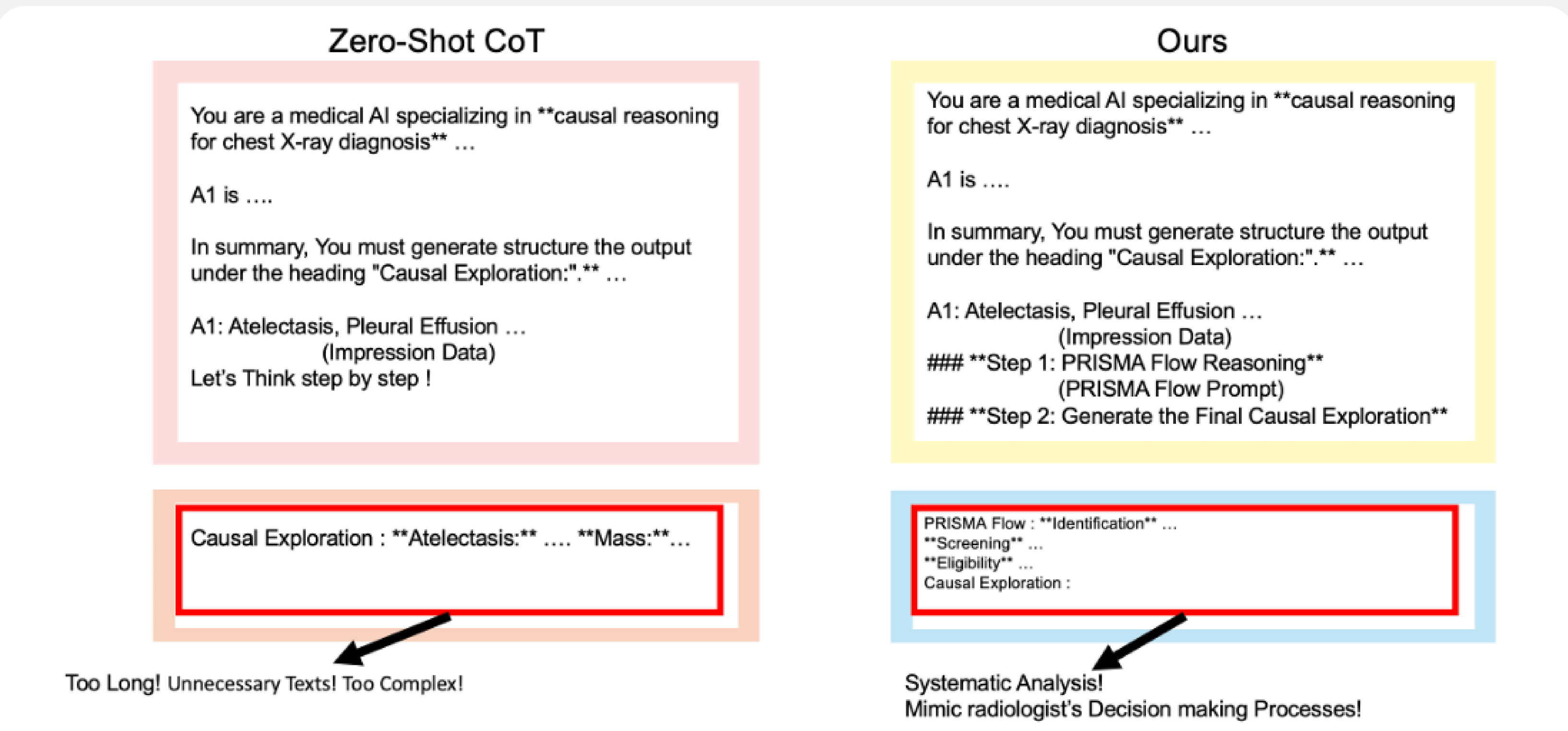


Figure 1: Comparison of PRISMA-Guided Reasoning and Chain-of-Thought Prompting

To overcome the limitations of standard Chain-of-Thought (CoT) prompting, we implemented a structured reasoning approach inspired by the PRISMA methodology, which is widely used in systematic reviews. This framework guides the LLM to mimic an expert radiologist's workflow through four key stages: Identification, Screening, Eligibility, and Inclusion. By systematically evaluating diagnostic possibilities, this process significantly enhances the transparency, reliability, and clinical trustworthiness of the final conclusions. We also explored an alternative approach combining LoRA fine-tuning with domain-specific CoT prompting to improve model adaptability.

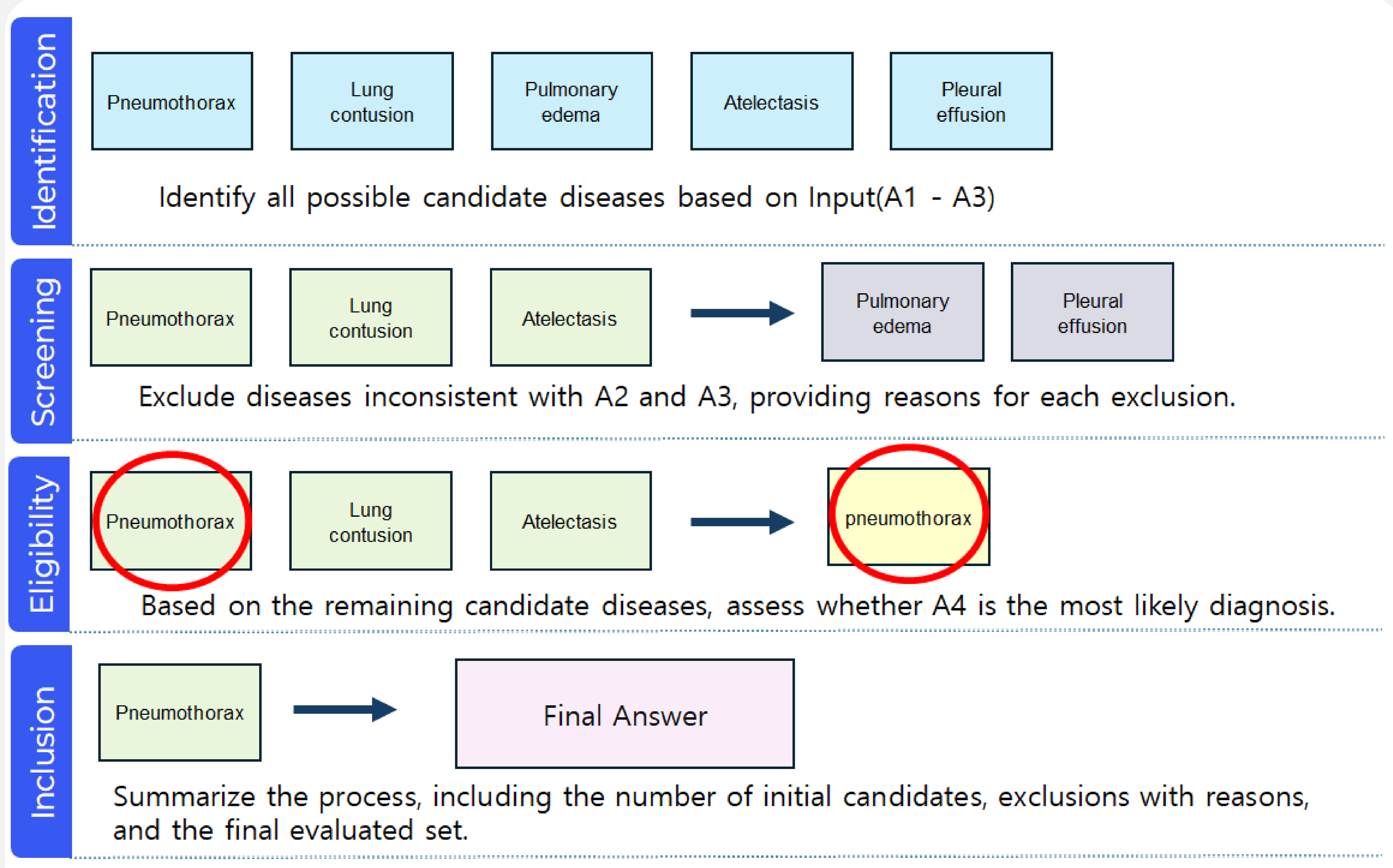


Figure 2: The PRISMA-Inspired Four-Stage Diagnostic Reasoning Framework

III. Results

Subtask 1: Top Performance in Hidden Causality Identification,

Our model, nasher-002, achieved the highest ranking (1st place) in the official evaluation for Subtask 1. This approach yielded a final score of 69.00, demonstrating the effectiveness of leveraging retrieved similar cases to dynamically enrich prompts and employing strict candidate selection.

Model	BERTScore	COS Sim	SentVec	GPT (W)	GPT (B)	Qual. Score	Final Score
nasher-002	0.281	0.570	0.785	0.696	0.715	0.689	69.00
CARE-v6	0.236	0.522	0.770	0.691	0.713	0.694	68.19
blissom	0.179	0.571	0.765	0.633	0.689	0.694	65.98

Table 1: Final Evaluation Results for Subtask 1

Subtask 2: Strong Performance in Causal Explanation Generation

For Subtask 2, our Prisma-zero-shot model, which utilized structured PRISMA flow with large language models, secured 2nd place in the official evaluation with a final score of 74.07. This highlighted the benefits of PRISMA-guided systematic reasoning in enhancing interpretability. Our alternative approach, Joh-3B, which combined fine-tuning and domain-specific prompting, was not included in the final ranking but demonstrated considerable potential in enhancing domain-specific model interpretability and achieved competitive

Model	BERTScore	COS Sim	SentVec	GPT (W)	GPT (B)	Qual. Score	Final Score
blissom	0.099	0.669	0.827	0.827	0.859	0.8158	78.84
Prisma-zero-shot	0.123	0.590	0.762	0.798	0.788	0.7804	74.07
Joh-3B	0.224	0.634	0.778	0.740	0.723	-	-

Table 2: Final Evaluation Results for Subtask 2

IV. Conclusion

Advancing Explainable AI in Radiology

Our work successfully demonstrates the efficacy of retrieval-enhanced prompting and PRISMA-guided structured reasoning in generating causality-based diagnostic inferences, achieving 1st place in Subtask 1 and 2nd place in Subtask 2, respectively. These findings significantly contribute to advancing explainable AI (XAI) in radiology by bridging the critical gap between automated systems and human expert decision-making. Future work will focus on integrating multimodal (text and image) data to improve causal inference and exploring hybrid methods that combine our structured reasoning framework with adaptive fine-tuning.