Teddysum at the NTCIR-18 HIDDEN-RAD Task: Using RAG and Tree-of-Thought for Causal Explanation Generation

Youngseob Won (Kyung Hee University), Younggyun Hahm (Teddysum),
Chanhyuk Yoon (Teddysum), Seong Tae Kim† (Kyung Hee University)

1. Introduction

Radiology reports are essential for clinical decision-making. However, they often focus on findings and diagnoses while omitting **explicit causal explanations**—how and why certain observations relate to specific conditions.

The **NTCIR-18 HIDDEN-RAD task** was designed to address this gap. It challenges participants to develop AI systems capable of **extracting and generating causal explanations** in radiology, across two subtasks:
- **Task 1**: Recovering causal relationships from radiology reports and optional chest X-ray images.
- **Task 2**: Generating a causal explanation based on structured text responses from radiologists.

We, the **Teddysum team**, present a **training-free, LLM-based causal reasoning framework** that combines:
- **Chain-of-Thought (CoT) prompting**: for structured stepwise reasoning
- **Retrieval-Augmented Generation (RAG):** to ground explanations in RadGraph medical knowledge
- **Tree-of-Thought (ToT) evaluation**: for iterative self-evaluation and refinement

Our system uses a fine-tuned **LLaMA 70B model, Blossom**, for both subtasks. In Task 1, we apply **KG-LLaVA** to convert chest X-rays into text before reasoning. In Task 2, we apply the same reasoning pipeline directly to text input.

Our method ranked **1st in Task 2**, demonstrating the effectiveness of structured reasoning in medical report generation and causal inference.

2. Method

Our framework is a training-free causal reasoning pipeline built on a fine-tuned LLaMA-70B model (Blossom) using efficient prompting strategies. It is applied to both image-augmented (Task 1) and text-only (Task 2) radiology report generation.We decompose the process into three core modules, designed to work sequentially:

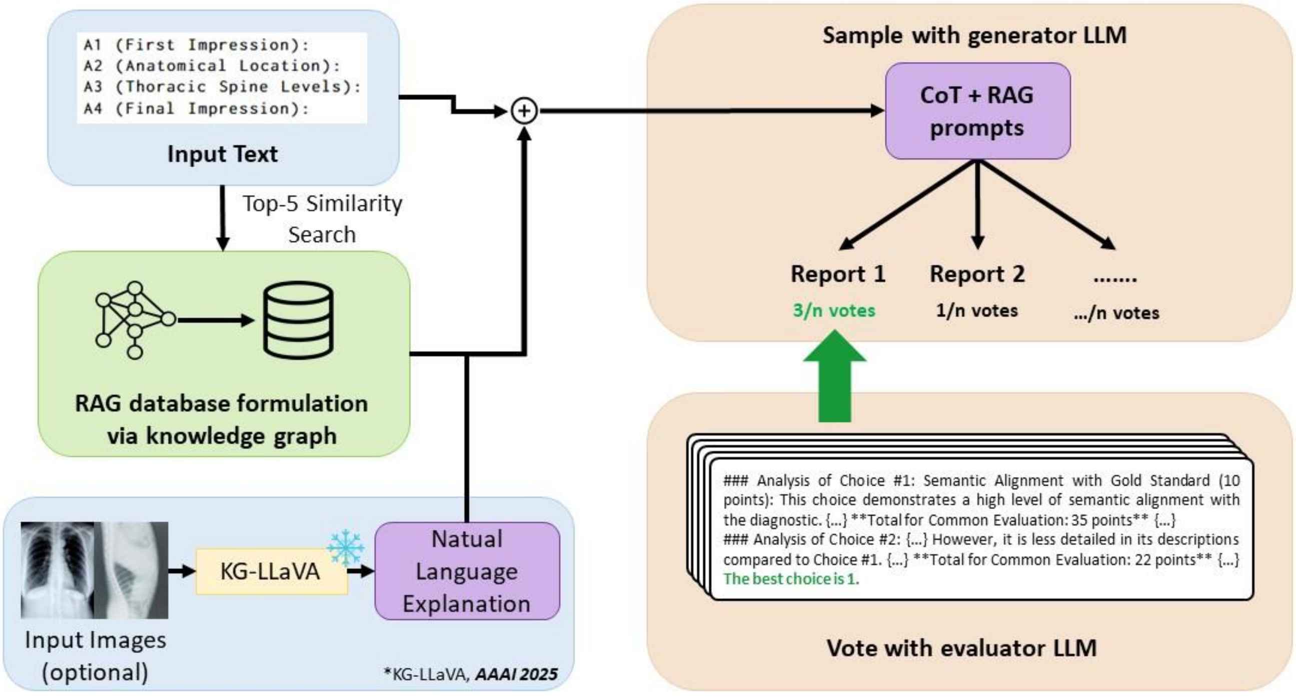
01 Chain-of-Thought (CoT) Prompting

We use structured CoT prompts to guide stepwise diagnostic reasoning based on the radiologist-provided fields A1-A4:
- **A1**: First Impression
- **A2**: Anatomical Location
- **A3**: Thoracic Spine Level
- **A4**: Final Impression

The prompt explicitly asks the model to explain how the findings in A1 and A2, localized via A3, lead to A4 using standard radiological reasoning. We further reinforce medical style by asking the model to rewrite the explanation to match professional report formatting.

02 Retrieval-Augmented Generation (RAG)

To enrich factual grounding, we retrieve relevant clinical knowledge from **RadGraph**, a knowledge graph of radiology entities and relations.



Analysis of Choice #1: Semantic Alignment with Gold Standard (10 points): This choice demonstrates a high level of semantic alignment with the diagnostic. {...} **Total for Common Evaluation: 35 points** {...}
Analysis of Choice #2: {...} However, it is less detailed in its descriptions compared to Choice #1. {...} **Total for Common Evaluation: 22 points** {...}
The best choice is 1.

(Figure 1) Overall architecture of the causal reasoning framework.

- The Final Impression (A4) is encoded and used to query a **FAISS index** built from RadGraph data.
- The top-5 retrieved passages are embedded into the prompt, improving completeness and reducing hallucinations.

03 Tree-of-Thought (ToT) Evaluation

To improve robustness, we implement **ToT-style self-evaluation**, where the model:
1. Samples multiple explanation candidates
2. Scores them based on completeness, clarity, and factual consistency
3. Selects the best explanation via an internal voting mechanism

This iterative process reduces incoherent or contradictory outputs and enhances causal accuracy. The overall architecture is shown at Figure 1.

3. Results

We evaluated our system on both subtasks of the **NTCIR-18 HIDDEN-RAD** challenge using official metrics for semantic accuracy, clinical relevance, and explanation quality.

Task 2: Text-Based Causal Reasoning (Ranked 1st Place)
Our method achieved the highest overall score in Task 2, demonstrating strong performance in generating causal explanations from structured radiologist input. (Table 1)

4. Conclusion

Our training-free pipeline—combining CoT prompting, RAG retrieval, and ToT evaluation—achieved 1st place in Task 2, validating the power of structured reasoning in medical report generation.

In contrast, image-based inputs in Task 1 showed potential but were limited by insufficient alignment between visual features and causal reasoning.

Going forward, we aim to reduce verbosity and improve multimodal grounding for more clinically usable outputs.

Task	BERTScore	COS Similarity	BioSentVec	GPT Base Score (White)	GPT Base Score (Black)	Qualitative Score
Task 1	0.179	0.571	0.765	0.633	0.689	0.694
Task 2 (without style matching)	0.099	**0.669**	**0.827**	0.827	**0.859**	0.816
Task 2 (with style matching)	**0.157**	0.664	0.825	**0.830**	0.841	-

(Table 1) Evaluation results for Task 1 (image + text) and Task 2 (text-only)