Hidden Causality Inclusion in Radiology Report Generation (NTCIR-18 HIDDEN-RAD Task)

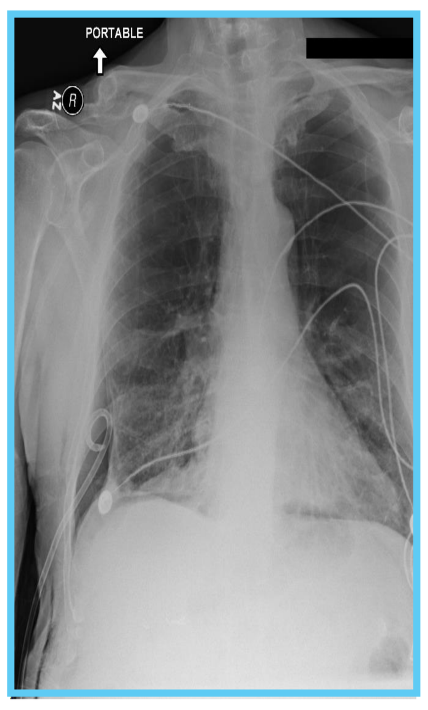
Key-Sun Choi
Konyang University, Republic of Korea
kschoi@konyang.ac.kr
kschoi@kaist.ac.kr

You-Sang Cho
Konyang University, Republic of Korea
davidecho@naver.com

Background & Motivation

Task input
Task Output

Radiography (option)



Radiology report

Impression:
Pneumothorax

Finding:
The pneumothorax in this case may be attributed to a combinatio n of factors, including trauma and anatomical location. The right pneumothorax observed at the T8-11 thoracic spine level in the ri ght pleural space indicates a localized issue in the upper to middl e region of the right lung.

Hidden causality

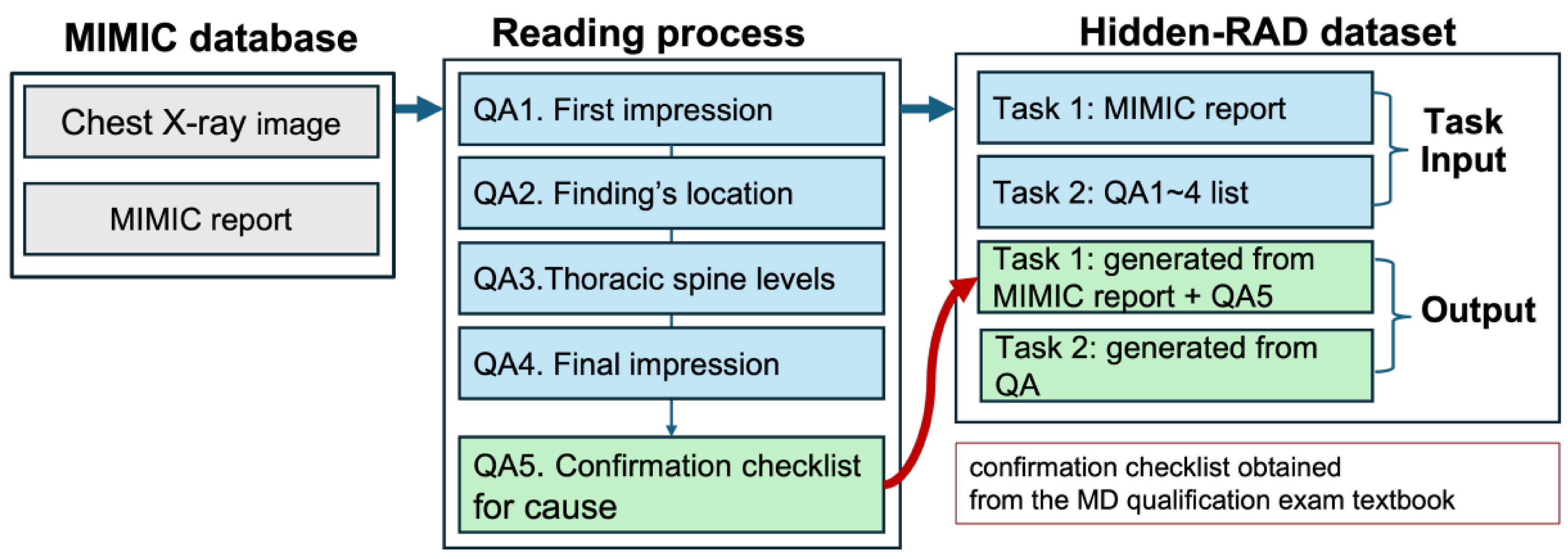
The lack of symmetry in the apical, upper, middle, and lower zones suggests an asymmetric distribution of air in the pleural space, further confirming the presence of pneumothorax.

❖ Problem
- Traditional radiology reports state only the final diagnosis, omitting the underlying causal reasoning

❖ Significance
- The Hidden-Rad task aims to enhance interpretability and trust in AI by requiring models to explicitly explain the rationale behind clinical interpretations

❖ Hidden-Rad Task Overview
- **Goal**: Generate causal explanations from radiology reports and structured questionnaire responses
- **Task 1**: Report → Causal Exploration section
- **Task 2**: QA1–QA4 responses → Causal Exploration text

Dataset



❖ Dataset
- The Hidden-Rad dataset is derived from MIMIC-CXR, comprising paired chest X-ray images and radiology reports

❖ Annotation Process
- Conducted via a structured questionnaire reflecting radiologists' reading workflow (QA1–QA5)

❖ Data Distribution by Task
- **Task 1**: Training 1,219 cases / Evaluation 314 cases
- **Task 2**: Training 804 cases / Evaluation 216 cases

❖ Common Diagnoses(in Task 1 Training Set)
- Normal(366), Pleural Effusion(324), Cardiomegaly(187), Atelectasis(172), Pneumonia(143), Edema(80), Mass(44), Pneumothorax(36), Congestion(28), Emphysema (24)

Evaluation Metrics

❖ Quantitative Metrics (80%)
- **BERTScore (5%):** Assesses contextual semantic similarity between generated explanations and ground-truth reports using pre-trained BERT embeddings
- **Cosine Similarity (5%):** Measures structural and semantic alignment via cosine similarity between report vector representations
- **BioSentVec (20%):** Captures domain-specific similarity with biomedical sentence embeddings trained on PubMed and MIMIC-III
- **GPT-White (25%):** Calculates scores based on contextual similarity referencing an external evaluation scheme (For full criteria, scan the QR code below to view on GitHub)
- **GPT-Black (25%):** Evaluates completeness, accuracy, and logical consistency of generated explanations using internal bonus and penalty criteria (For full criteria, scan the QR code below to view on GitHub)

❖ Qualitative Evaluation by Experts (20%)
- Expert review of 18 (Task 1) and 10 (Task 2) system runs selected from top-5 of each quantitative metric, after duplicate removal
- Comprehensive assessment of clinical validity, readability, and causal fidelity

Methods in official runs

❖ Team Approaches in Official Runs
- **Teddysum**: Applied CoT, RAG, and ToT prompting on a large Blossom LLM (70B)
- **RADPHI3**: Fine-tuned a Rad-Phi-3.5-Vision-CXR (4.2B) model with LoRA and data augmentation in both text-only and multimodal settings
- **Nash**: Built an optimized pipeline using GPT-4o APIs with retrieval augmentation and strict candidate selection

❖ Key Techniques Compared
- CoT+RAG+ToT vs. LoRA fine-tuning vs. API-based optimization
- Image handling: separate VLM (Teddysum), integrated multimodal model (RADPHI3), text-only (Nash)

Main Results

❖ Task 1 Final Rankings & Scores

Team (Model Name)	BERTScore	COS Sim	BioSentVec	GPT (W)	GPT (B)	Qual. Score	Final Score
Nash (nasher-002)	**0.281**	**0.570**	**0.785**	**0.696**	**0.715**	0.689	**0.69**
RADPHI3 (CARE-v6)[a]	0.236	0.522	0.770	0.691	0.713	**0.694**	0.68
RADPHI3 (CARE-v2.32)[b]	0.256	0.541	0.766	0.680	0.700	0.690	0.68
RADPHI3 (CARE)[c]	0.259	0.538	0.767	0.683	0.696	0.682	0.68
Teddysum (Blossom)	0.179	0.571	0.765	0.633	0.689	0.694	0.66

[a] RADPHI3's GPT-4o multimodal baseline submission.
[b] RADPHI3's Rad-Phi-3.5-Vision-CXR text-only submission.
[c] RADPHI3's Rad-Phi-3.5-Vision-CXR multimodal submission.

- **Nash** (1st) : 0.694
- **RADPHI3** (2nd) : 0.682
- **Teddysum** (3rd) : 0.662

❖ Task 2 Final Rankings & Scores

Team (Model Name)	BERTScore	COS Sim	BioSentVec	GPT (W)	GPT (B)	Qual. Score	Final Score
Teddysum (bllossom)[f]	0.099	**0.669**	**0.827**	**0.827**	**0.859**	**0.816**	**0.79**
Nash (Prisma-zero-shot)	0.123	0.590	0.762	0.798	0.788	0.780	0.74
Nash (Joh-3B)[g]	**0.224**	0.634	0.778	0.740	0.723	0.783	0.72

[f] Teddysum's **Blossom** model, achieved **1st place**. Spelled bllossom on leaderboard.
[g] Nash's fine-tuned Llama-3.2-3B model.

- **Teddysum** (1st) : 0.792
- **Nash** (2nd) : 0.735

❖ Key Insights
- Retrieval-augmentation with strict candidate selection excels in Task 1
- Combined CoT, RAG, and ToT pipeline demonstrates strong performance in Task 2
- Domain-specialized smaller models (RADPHI3) achieve results close to large LLM approaches

Discussion & Future Work

❖ Discussion
- **Retrieval-Augmentation** performs well in Task 1 but degrades on rare cases due to limited similar literature
- **CoT+RAG+ToT** pipeline effectively captures deep causal relations in Task 2, yet incurs higher computational cost and latency
- **Data Limitations**: The Task 1 training set is skewed toward a few common findings (e.g., Pleural Effusion, Normal), making generalization to rare conditions challenging

❖ Future Work
- **Deepening Multimodal Integration**
  Enhance causal reasoning via advanced vision-language models
- **Combining Prompting & Fine-tuning Strategies**
  Maximize RAG effectiveness using specialized medical knowledge sources and optimized query strategies
- **Output Control & Evaluation**
  Develop output management methods and clinically aligned evaluation metrics for readability and style control
- **Scalability & Clinical Validation**
  Apply methods to large-scale datasets and validate in real hospital workflows