

# NTCIR-18 Hidden-RAD Overview:

## Hidden Causality Inclusion in Radiology Report Generation

Key-Sun Choi\*, You-Sang Cho, J.H. Hahn, K.M. Chae (Konyang University)  
Young-Gyun Hahm, Ye-Jee Kang (Teddysum)  
So-Yun Lee (KYU Hospital)

<https://github.com/hidden-rad/>

NTCIR18 Conference  
2025.6.11

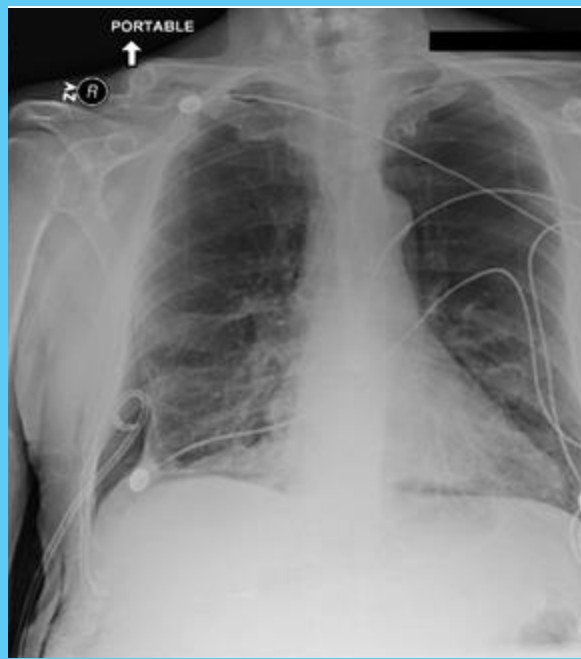


# Task Overview

To Recover the Hidden Causality why such Impression and Findings are drawn in the Radiology Report

■ Task input  
■ Task Output

## Radiography (option)



## Radiology report

### Impression:

Pneumothorax

### Finding:

The pneumothorax in this case may be attributed to a combination of factors, including trauma and anatomical location. The right pneumothorax **observed at** the T8-11 **thoracic spine level** in the right pleural space indicates a localized issue in the upper to middle 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**.

# What is 'Hidden Causality'?

## Traditional radiology reports focus

- on the final diagnosis (what)
- often omitting the causal reasoning (why) leading to it.

## Hidden-RAD Task requires

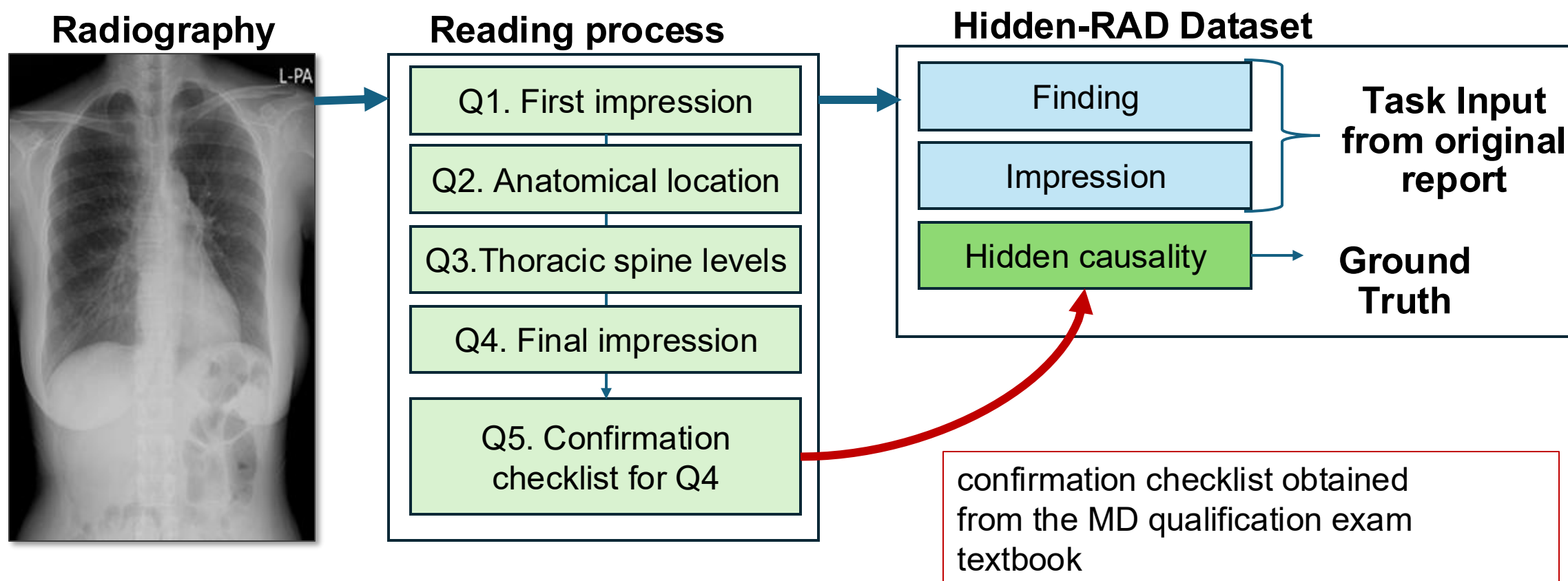
- linking observations to impressions with logical justifications

## Ultimately, the goal is to

- Move beyond shallow summarization to structured diagnostic reasoning
- Enhance the interpretability and clinical trust of AI models.

# Hidden-RAD: dataset overview

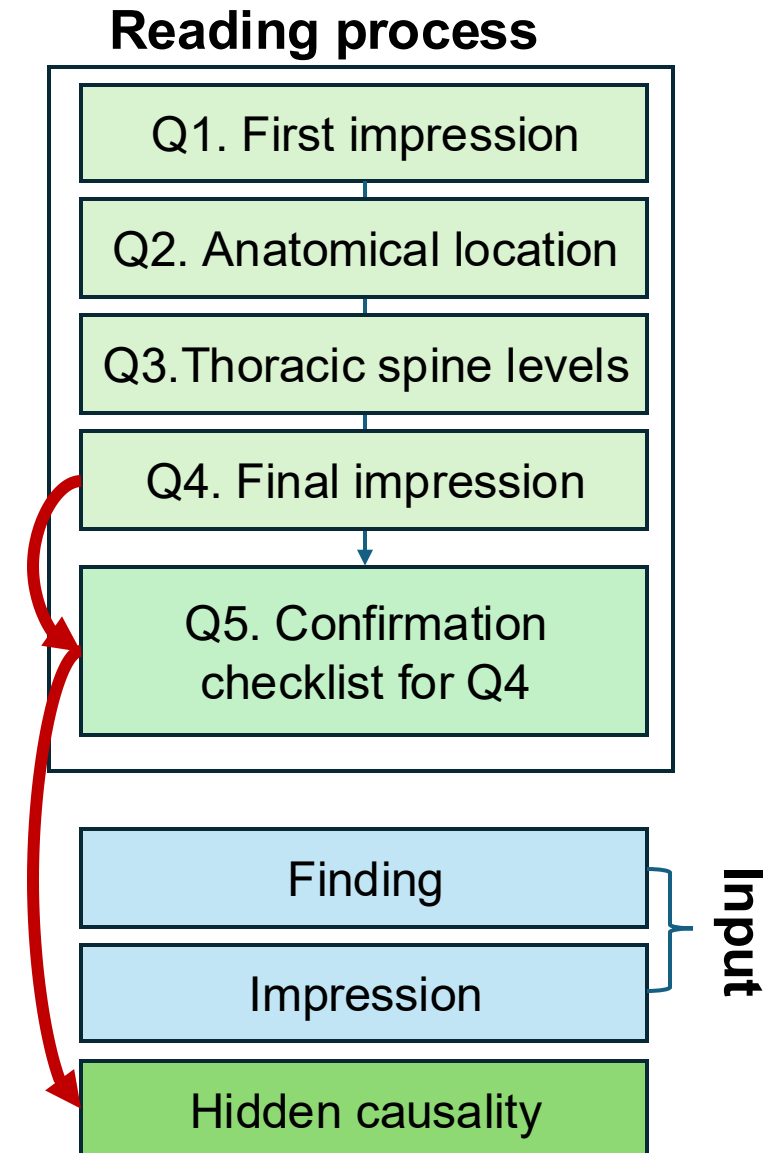
- **Data collection: Decision of Steps in Reading Process of real doctors.**
- **Hidden-RAD dataset: generated reports from the collected data**



The reading process was annotated with questions and answers to reduce costs, and GPT-4 was used to generate the dataset into a report format.

# Dataset: Representing and Ensuring Causality's Reliability

- Based on MIMIC-CXR:
  - radiology reports + QA1–QA5 structure
  - QA4: Final Impressions, QA5: Checklist with 28 questions
- **Causal links** reconstructed via expert mapping between findings and diagnoses
- Data reviewed by radiologists to ensure coherence and reliability



# Confirmation checklist: ABCDE approach

- 32 checklists,
- **Chest x-ray review** is a key competency for medical students, junior doctors and other allied health professionals.
- Using A, B, C, D, E is a helpful and systematic method for [chest x-ray review](#):

- **A: airways – 5 checklist**
- **B: breathing (the lungs and pleural spaces) - 11 checklist**
- **C: circulation (cardiomediastinal contour) - 5 checklist**
- **D: disability (bones - especially fractures) - 6 checklist**
- **E: everything else, e.g. pneumoperitoneum - 5 checklist**

Q5-1. “Trace down the trachea to the carina. Is there tracheal deviation?” (in checklist)

## 1. Anatomical tracing:

- < “Trachea (body structure)” 44567001 **is\_connected\_to** “Carina of trachea (body structure)” 297171001 >
- < “Carina of trachea (body structure)” 297171001 **is\_inferior\_to** “Trachea (body structure)” 44567001 >

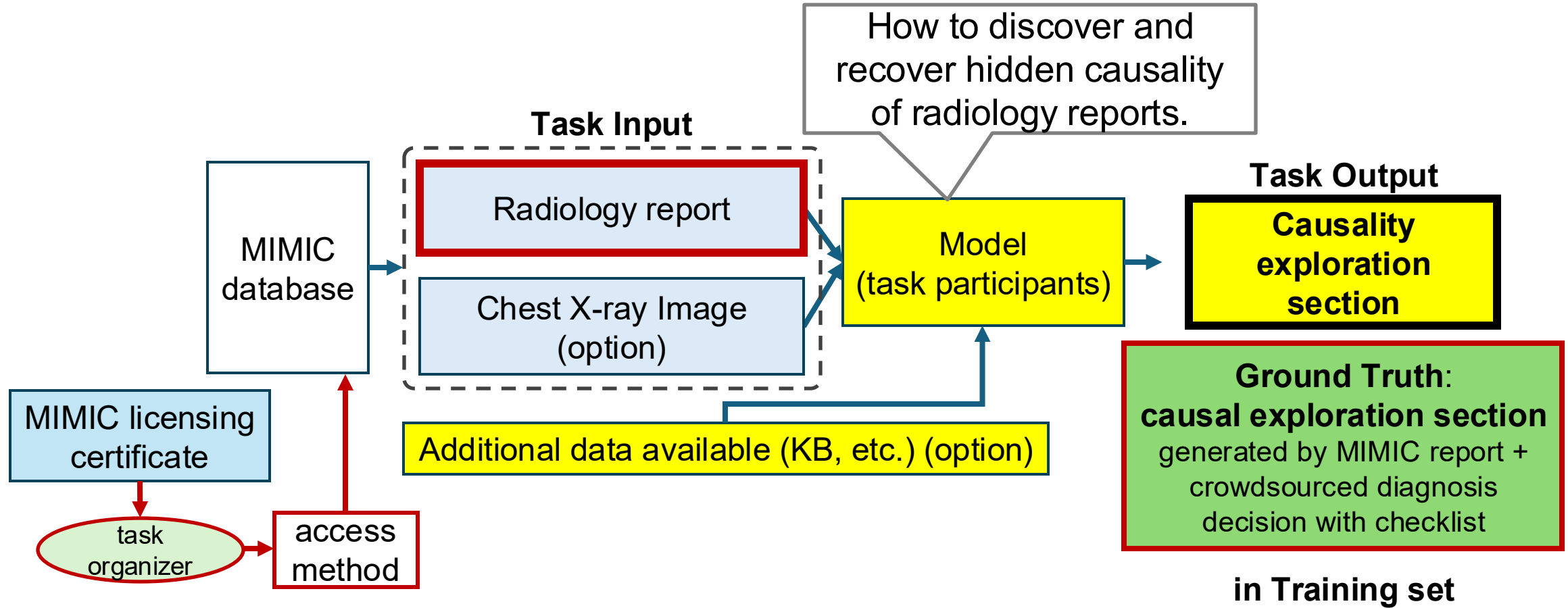
## 2. Finding (Tracheal deviation):

- < “Trachea (body structure)” 44567001 **has\_finding** “Displacement of trachea (finding)” 29857009 >

ref: SNOMED-CT


# Task-1 Definition

Generating causality section by discovering hidden (missing) causality in radiology reports.





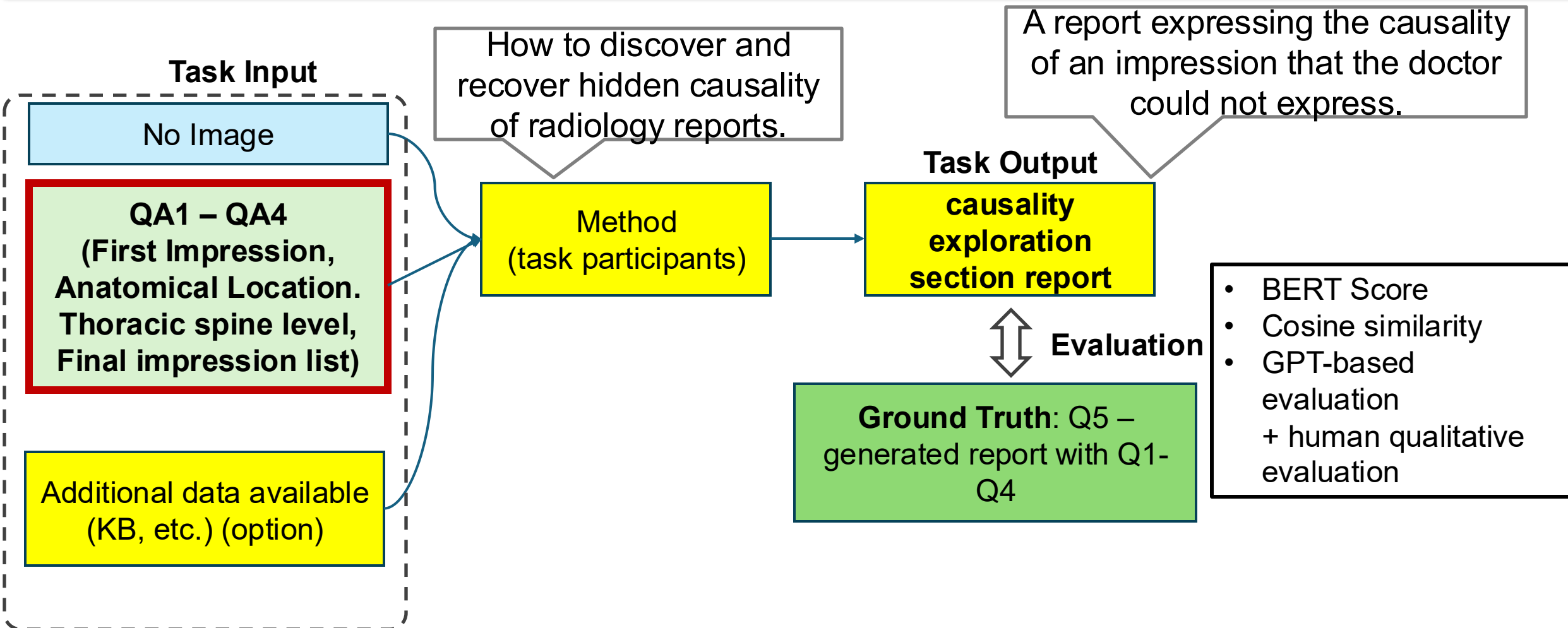
# Task1 Example: Generating causality exploration section from a radiology report

Input: Radiology report (MIMIC database) + optional data		Output: text for causality
Report	<p>FINAL REPORT EXAMINATION: CHEST (AP upright AND LAT) INDICATION: ___M with AMS and R hip pain s/p recent fall COMPARISON: CT chest ___ FINDINGS: The imaging findings indicate signs of compromised lung integrity, with evidence of pneumothorax and pleural effusion, likely stemming from trauma or an underlying lung pathology. Additionally, the presence of subcutaneous emphysema suggests air leakage into the subcutaneous tissue, possibly due to rib fractures or a breach in the lung tissue. The stable cardiomediastinal contours, along with the placement of a right chest tube, support the current management approach for these conditions. IMPRESSION: Large mass in the right upper lung. Otherwise, unremarkable.</p>	<p><b>Causal Exploration:</b></p> <p>The presence of pneumothorax and pleural effusion indicates compromised lung integrity, likely resulting from trauma or underlying lung pathology.</p> <p>The subcutaneous emphysema suggests air leakage into the subcutaneous tissue, possibly due to rib fractures or a breach in lung tissue.</p> <p>The stable cardiomediastinal contours and the right chest tube in place support ongoing management of these conditions.</p>
Image (Option)		
Additional resources from task participant (Option)		

# Task-2 Definition (w/o MIMIC-licensing)

input: Crowdsourced data about Q1-Q4 (first/final impression, anatomic loc.)

output: text generated from Q5 (checklist answer for each case of Q4 decision)



## Task2: from Radiologist questionnaire for Chest X-ray image to causal exploration text generation

Input		Output
A1 (First Impression)	Emphysema pleural effusion Pneumothorax	<b>Causal Exploration:</b>  1. Emphysema: The presence of subcutaneous emphysema as indicated in A5 can be caused by a rupture in the airways or lung tissue, which allows air to escape into the subcutaneous tissues. This can be linked back to A1~A4 as the condition was identified through these processes.  2. Pleural Effusion: The absence of certain findings in A5 such as lack of blunted costophrenic angles or no abnormalities in the hilar vessels could indicate that there is no pleural effusion. However, further investigations are needed as these signs are not definitive.  3. Pneumothorax: The fact that both lungs are well expanded and similar in volume as stated in A5, alongside other normal findings, suggests there is no pneumothorax present. This is in accordance with the findings from A1~A4.  Thus, the Causal Explorations for these conditions are derived from the findings in A1~A4 and are further supported by the details in A5.
A2 (Anatomical Location)	'Subcutaneous tissue', 'Parenchyme', 'RUL(Right Upper Lobe)', 'Parenchyme', 'RML(Right Middle Lobe)', 'Parenchyme', 'RLL(Right Lower Lobe)', 'Parenchyme', 'LUL(Left Upper Lobe)', 'Parenchyme', 'LLL(Left Lower Lobe)', 'Pleural', 'Right Pleural'	
A3 (Thoracic Spin Level)	{'begin': 1, 'end': 12}	
A4 (Final Impression)	'emphysema' 'pleural effusion' 'pneumothorax'	

# Evaluation Methods and Validity

Three major evaluation types used in HIDDEN-RAD:

- **Similarity-based (vector embeddings):** BERTScore, Cosine Similarity, BioSentVec
- **LLM-based:** GPT-White (rubric scoring), GPT-Black (bonus/penalty logic)
- **Human experts** : judgment of diagnostic plausibility and completeness

LLM-based scores (esp. GPT-Black) aligned closely with expert assessments

# Evaluation Process Overview

---

## 1. Quantitative Evaluation (80 points)

---

- BERTScore (5%) – Ensures **causal explanations** align with the **original report**.
  - COS Similarity (5%) – Measures **semantic coherence** in generated explanations.
  - BioSentVec (20%) – Validates **medical accuracy** using **MIMIC embeddings**.
  - GPT-based Score (White) (25%) – Rewards **structured** and **logical explanations**.
  - GPT-based Score (Black) (25%) – Penalizes **inconsistencies** to enhance **reliability**.
- 

## 2. Qualitative Evaluation (20 points)

---

- Top 5 models from each metric
  - Human evaluation based on predefined criteria
-

# Task 1 Ranking

Ranking	TeamName	ModelName	BERTScore	COS similarity	BioSentVec	GPT base score (White)	GPT base score (Black)	Qualitative Score	Final Score
1	nash	nasher-002	0.281	0.57	0.785	0.696	0.715	0.689	<b>69</b>
2	RADPHI3	CARE-v6	0.236	0.522	0.77	0.691	0.713	0.694	<b>68.19</b>
3	Teddysum	bllossom	0.179	0.571	0.765	0.633	0.689	0.694	<b>65.98</b>

- A total of **40 models** were submitted for evaluation.
- A total of **18 models** were selected (for qualitative scoring by human experts)  
based on the **top 5 models from each evaluation metric**.
- This leaderboard presents the top-performing model from each team, **comparing only the best submission per team** rather than all submitted models.

# Task 1: Key findings based on the evaluation results

1. Score in the phase after the evaluation criteria (scoring rubric) were provided for GPT-White,
  - a shift from simply listing information to clearly describing causal explanations
    - Before opening the evaluation criteria:
      - tendency for only input data-based explanations to be output
    - After opening the evaluation criteria:
      - model's output became more organized
      - more tendency to include major causal relationships
  - But the best score was found before opening the evaluation criteria
2. Overall, many models fit well with the contents of the input report and maintains contextual similarity.
3. However, many results omitted causal relationships, or unnecessary content was added, even when causal relationships were restored.

# Evaluation Results: Highlights (Task 2)

Model	BERT	CosSim	BioSent	GPT-W	GPT-B	Expert	Wtd / Eq
A	0.099	0.669	0.827	0.827	0.859	0.816	0.790 / 0.683
B	0.123	0.590	0.762	0.798	0.788	0.780	0.740 / 0.640
C	0.224	0.634	0.778	0.740	0.723	0.783	0.720 / 0.647

- Model A scored highest on GPT and expert metrics.
- **GPT-Black** was the most discriminative: up to 0.136 score gap.
- GPT-White showed strong **alignment with expert judgment**
  - e.g., 0.827 vs 0.816.
- Surface metrics (**BERTScore**) diverged from clinical quality
  - e.g., Model C's 0.224 not matched by GPT/Expert.
- **Task-weighted evaluation** yielded more reliable rankings.
  - e.g., 25% GPT, 20% Expert



# Evaluation Insights & LLM Issues

- **LLM hallucination** reduced through structured prompting: CoT, RAG, ToT, PRISMA.
- Surface similarity fails to capture causal logic
- Prompt quality and consistency critically affect evaluation.
- GPT-White/Black offer **scalable, rubric-aligned evaluation**, but need careful tuning.
- **Expert reviews** remain essential but are resource-intensive.
- Ranking shifts under different metric weight schemes
- **Future direction**: hybrid methods combining LLMs, clinical knowledge graphs (e.g., SNOMED-CT), and structured reviewer rubrics.

# Conclusion: Key Findings

- Proof of Causality Modeling Feasibility
  - Confirmed effectiveness of structured reasoning (78.84% in Task 2)
- Methodological Diversity and Effectiveness
  - Complex pipelines vs single specialized models
- Evaluation framework validity and limitations
  - Effectiveness of multi-dimensional evaluation
  - Confirmed consistency of GPT-based evaluation
  - Persistent subjectivity and scalability issues
- Critical role of Evaluation design
  - GPT-Black: Superior discriminative power for evaluation
  - Surface metrics: Limited alignment with clinical quality
  - Weighting schemes: Significant impact on model rankings

# Data, paper and PRICAI workshop

## Dataset for NTCIR-18 (Sample data)

- Task1: <https://github.com/hidden-rad/Task1> \* Task2: <https://github.com/hidden-rad/Task2>

## NTCIR-19 Hidden-Rad tasks: more data and more refined tasks

## Paper: (in LLM4Eval workshop@SIGIR2025, July/17/2025)

- "Evaluating Causal Explanation in Medical Reports with LLM-Based and Human-Aligned Metrics"

## Workshop accepted in PRICAI2025 (will open CFP homepage soon)

- **Workshop HIDDEN-RAD: Unlocking Causal Explanations in Medical AI and Beyond**

\* Hidden-Rad session (9:30-10:30, 12/June) and Round table (16:15-17:15).

\* Thank you very much to all the participating teams and NTCIR Co-Chairs and organizers!