CURV: Coherent Uncertainty-Aware Reasoning in Vision- Language Models for X-Ray Report Generation

Ziao Wang[1,2], Sixing Yan[1], Kejing Yin[1,*] Xiaofeng Zhang[3], William K. Cheung[1]

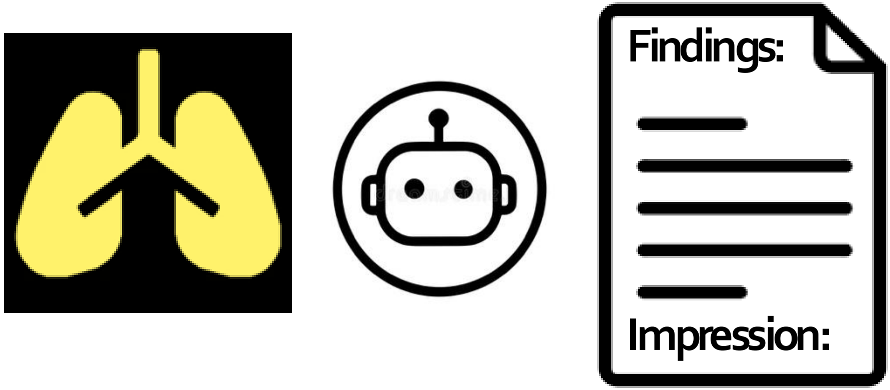
[1]Department of Computer Science, Hong Kong Baptist University
[2]Institute of Systems Medicine and Health Sciences, Hong Kong Baptist University
[3]Department of Computer Science, Harbin Institute of Technology

6/11/2025

香 港 浸 會 大 學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

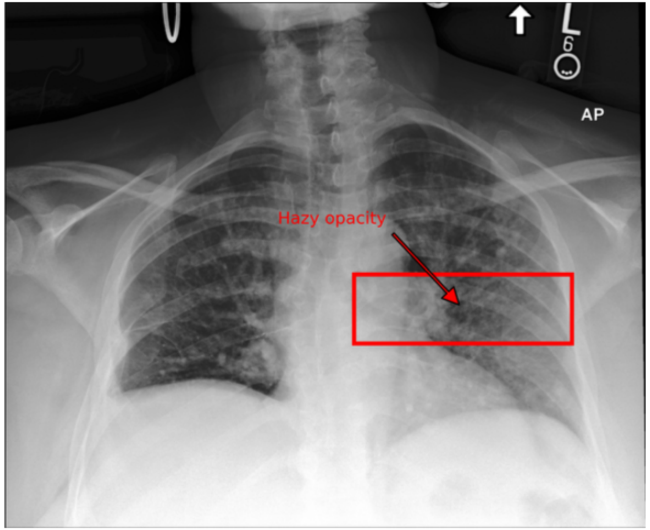
Vision Language Models For X-Ray Report Generation



Key Challenges:

- **Lacks Uncertainty:** Models don’t say "likely” or "possible".
- **“Black Box” Reasoning:** No clear logic from finding to impression.
- **Reduced Clinical Trust:** Clinicians can’t trust what they can’t understand.

Documenting Uncertainties in X-Ray Reports



(a) Chest X-ray image

Structural Uncertainty (Findings): *"Pulmonary nodules in the left upper lobe are also* ***not completely characterized*** *on this study. However, in addition, there is a more hazy widespread opacity projecting over the left mid upper lung which* ***could be compatible with*** *a coinciding pneumonia."*

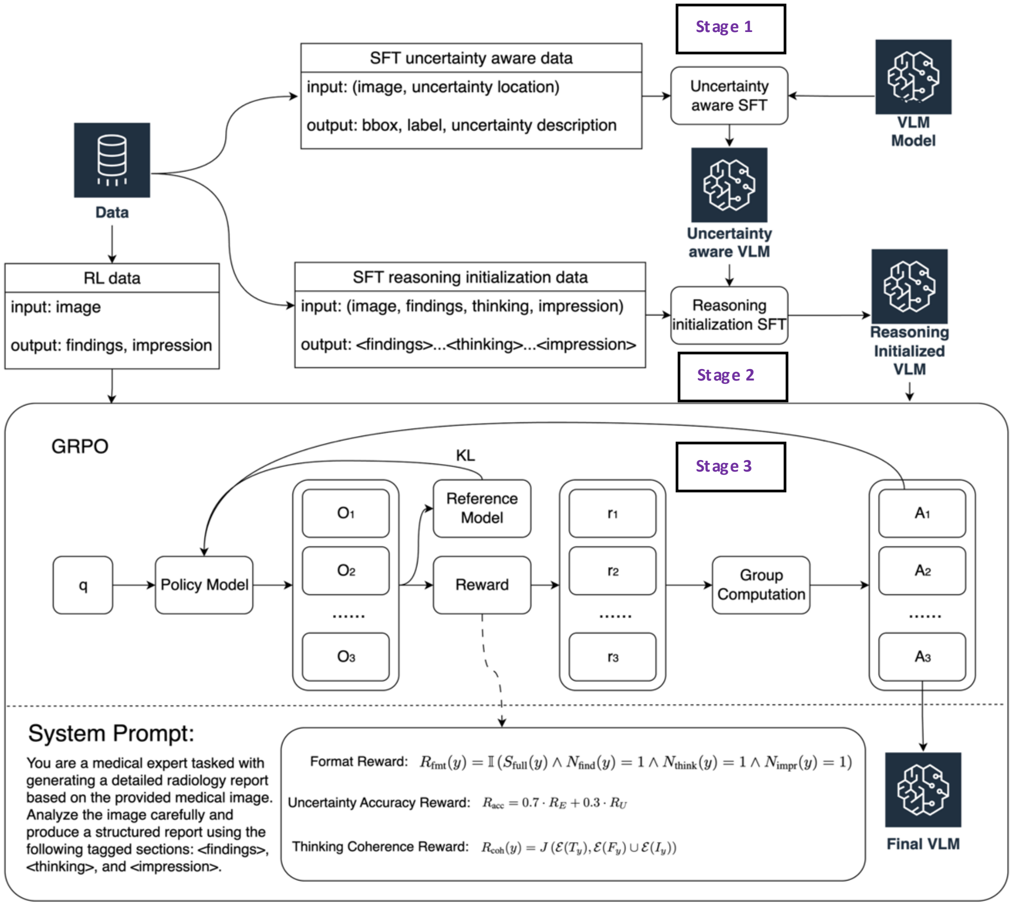
Semantic Uncertainty (Impression): *"Increasing left lung opacification which* ***may reflect*** *pneumonia superimposed on metastatic disease, although other etiologies such as lymphangitic pattern of metastatic spread* ***could be considered****. CT* ***may be helpful*** *to evaluate further if needed clinically."*

(b) Corresponding uncertain expressions from the radiology report.

DEPARTMENT OF COMPUTER SCIENCE
計算機科學系

香港浸會大學
HONG KONG BAPTIST UNIVERSITY

CURV Framework



Stage 1: Uncertainty Modeling

Stage 2: Reasoning Initialization

Stage 3: RL with Multi-component Rewards

Stage 1: Uncertainty Modeling

- Connect uncertainty phrase to specific bounding boxes
- Model learns to identify regions, assign anatomical labels, and express appropriate structural uncertainty

$$\mathcal{L}_{\text{uncertainty}} = - \sum_{(I,p,Y_{gt}) \in \mathcal{D}} \log \pi_\theta(\text{seq}(Y_{gt}) | I, p)$$


```
{
  "image": "cxr/02aa804e-bde0afdd-112c0b34-7bc16630-4e384014.jpg",
  "conversations": [
    {
      "from": "human",
      "value": "<image>\nDetect all anotomical objects in the image that most likely contain uncertainties and return their locations in the form of coordinates along with their organ label and uncertainty descriptions in JSON format."
    },
    {
      "from": "gpt",
      "value": "{\n"bbox_2d": [121, 104, 180, 162], “label” : “left lower lung zone” , "uncertainty":"Bilateral nodular opacities that most likely represent nipple shadows." "\n}"
    }
  ]
}
```

Stage 2: Reasoning Initialization & TRACE-CXR Dataset

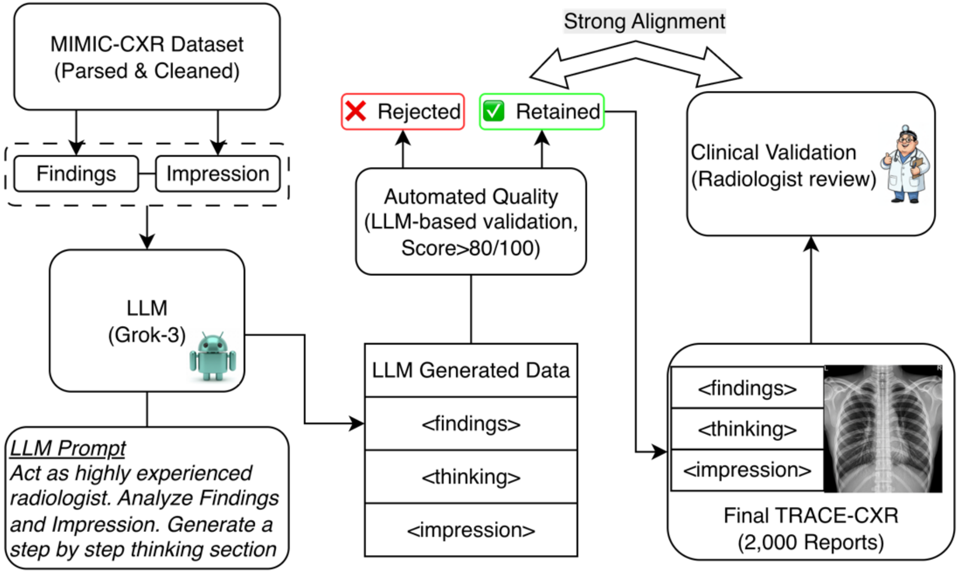
- **TRACE-CXR**: Novel dataset with 2,000 X-ray reports augmented with LLM-generated explicit 'Thinking' sections
- Model fine-tuned to produce structured tripartite output:

  Findings → Observations from the X-ray

  Thinking → Reasoning pathway

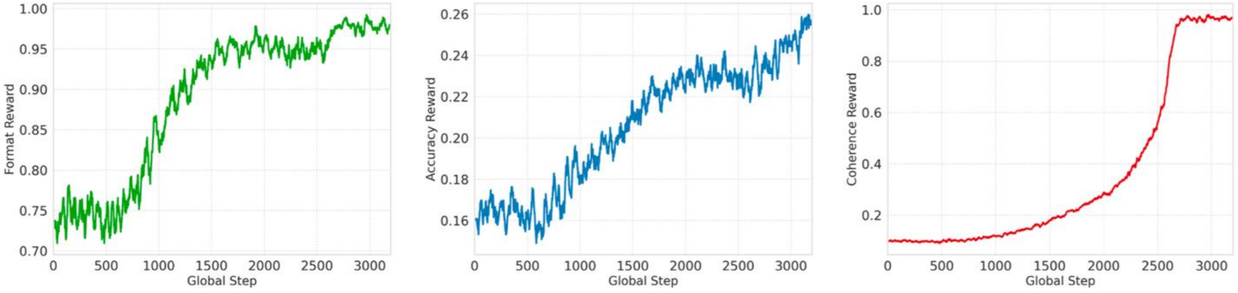
  Impression → Clinical conclusions

$$\mathcal{L}_{\text{reasoning}} = -\sum_{(I,p,Y_{\text{structured}})\in\mathcal{D}_{\text{reason}}} \log \pi_\theta(\text{seq}(Y_{\text{structured}})|I,p)$$



Stage 3: RL with Multi-Component Rewards

- Based on only <findings> and <impression> but not <thinking>
- **Discover a better reasoning path** guided by the reward signals.
  - R_{fmt} **(Format Adherence):** Ensures the tripartite structure.
  - R_{acc} **(Accuracy):** Measures medical accuracy and uncertainty.
  - R_{coh} **(Coherence):** Rewards logical coherence.



(a) Format Reward (b) Accuracy Reward (c) Coherence Reward

Experimental Setup & Datasets

- Model & Training: Qwen-2.5-VL-3B (3B parameters)
  - Trained on 4xA100 GPUs (~100 hours)
  - Batch size: 16, Learning rate: 1×10-6
- Datasets:
  - MIMIC-CXR base (227,835 reports)
  - Uncertainty-annotated set (112,111 samples)
  - TRACE-CXR reasoning set (2,000 samples)
- Evaluation:
  - Text: BLEU-1/2/3/4, METEOR, ROUGE-L
  - Clinical: CheXbert, RadGraph F1
  - LLM-based: Reasoning & uncertainty scoring

Performance Evaluation

- ✅ CURV achieves state-of-the-art on generation and clinical metrics
- ✅ CURV outperforms Gemini-2.5 pro
- ✅ CURV (3B) outperforms 7B-models
- ✅ OOD experiments on IU X-ray

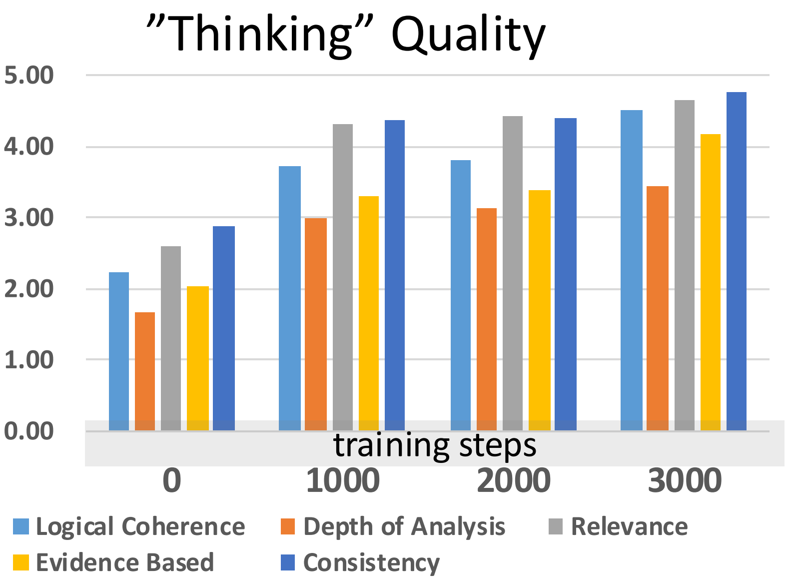
Results on **MIMIC-CXR** dataset

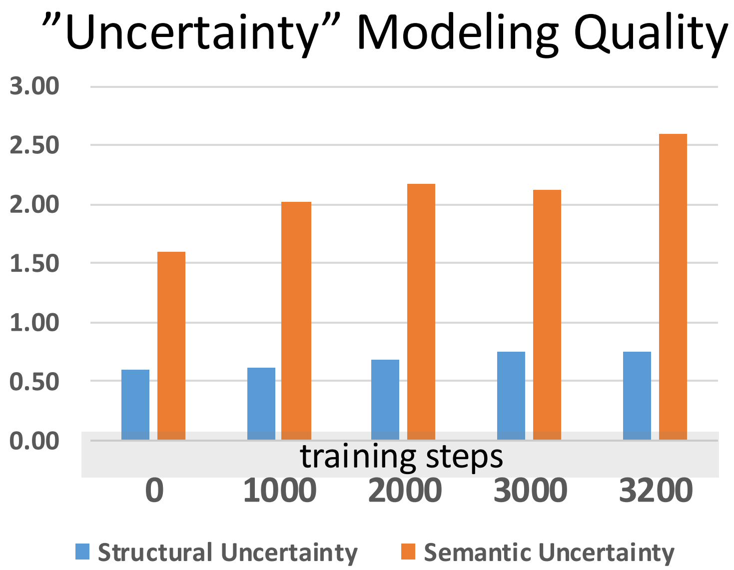
Model	BLEU2	BLEU3	BLEU4	METEOR	ROUGEL	gritlm	chexbertf1	radgraphf1
llava-1.5-7b	7.46	2.81	1.25	19.16	18.36	44.25	38.54	4.95
llava-1.5-7b-sft-cxr	15.06	9.43	6.13	25.71	28.09	50.28	51.51	13.06
Huatuo-GPT-Vision-7B	9.42	4.64	1.93	26.01	20.78	47.32	48.62	9.06
Maira2	14.12	9.01	6.14	26.78	28.65	47.48	46.53	17.05
Qwen2.5-VL-3B	5.42	2.08	0.89	20.81	15.23	44.57	37.66	4.66
gemini2.5-pro	5.20	2.25	1.05	21.19	15.01	40.41	48.45	7.71
CURV_stage1	7.08	3.33	1.61	23.47	19.07	45.15	30.75	9.95
CURV_stage2	5.22	2.43	1.10	18.57	14.59	42.76	26.83	6.03
CURV	**15.58**	**9.85**	**6.18**	**30.43**	**31.19**	**50.48**	**57.12**	**19.54**

Results on **IU X-ray** dataset

Model	BLEU2	BLEU3	BLEU4	METEOR	ROUGEL	gritlm	chexbertf1	radgraphf1
llava-1.5-7b	6.60	3.00	1.40	19.65	17.48	45.78	46.81	8.76
llava-1.5-7b-sft-cxr	12.95	8.03	5.20	23.24	26.40	46.21	53.33	10.31
Huatuo-GPT-Vision-7B	10.70	6.28	2.81	31.02	23.42	50.22	67.07	13.96
Maira2	15.60	9.64	6.03	25.52	31.18	54.18	70.75	24.01
Qwen2.5-VL-3B	5.05	2.28	1.05	21.65	15.01	45.95	49.47	6.26
CURV_stage1	6.24	3.18	1.57	23.64	18.18	46.76	40.30	12.13
CURV_stage2	5.32	2.75	1.27	18.64	13.93	44.46	33.67	7.28
CURV	**18.76**	**12.08**	**6.86**	**38.30**	**39.08**	**54.89**	**74.36**	**25.65**

LLM-based Evaluation





✅ Improvement on thinking and uncertainty modeling capabilities.

Conclusion & Impact

- CURV generates more trustworthy X-Ray reports via **Reasoning** with **Uncertainty Awareness**
- CURV 3B model outperforms other 7B models, proving effectiveness
- **TRACE-CXR** dataset will be released to the public

- *Limitations*
  - Performance relying on the quality of the initial curated datasets
  - Generalization to other medical imaging requires further investigation
  - A large-scale clinical validation are still needed.

Thank You!

Code and Dataset

Contact Me on Telegram