SOCIOCOM at the NTCIR-18 RadNLP Main Task: Zero-Shot LLM Approaches for Lung Cancer Staging

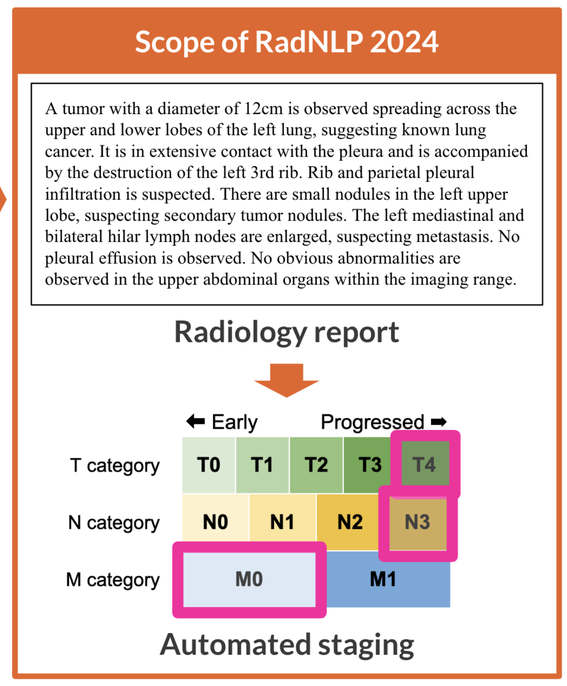
Yuki Tashiro[1] Yuta Nakamura[2] Eiji Aramaki[1]

1 Nara Institute of Science and Technology, Japan 2 The University of Tokyo, Japan

Introduction [Nakamura+2025]

- Writing radiology reports is one of the most critical workflows for radiologists
- Although radiologists communicate with physicians by providing the results of imaging studies, **radiology reports do not always specify the stage explicitly**
- This imposes extra workload on human experts for careful manual information extraction, which can be aided by automation
- **Task Objective: To classify the stage of lung cancer from radiology reports**

Lung cancer CT image



Dataset

- **Training Set**: Comprises 108 documents for 12 cases
- **Validation Set**: Includes 54 documents for 6 cases
- **Test Set**: Contains 216 documents for 24 cases

Our Method: Zero-shot LLM Approaches

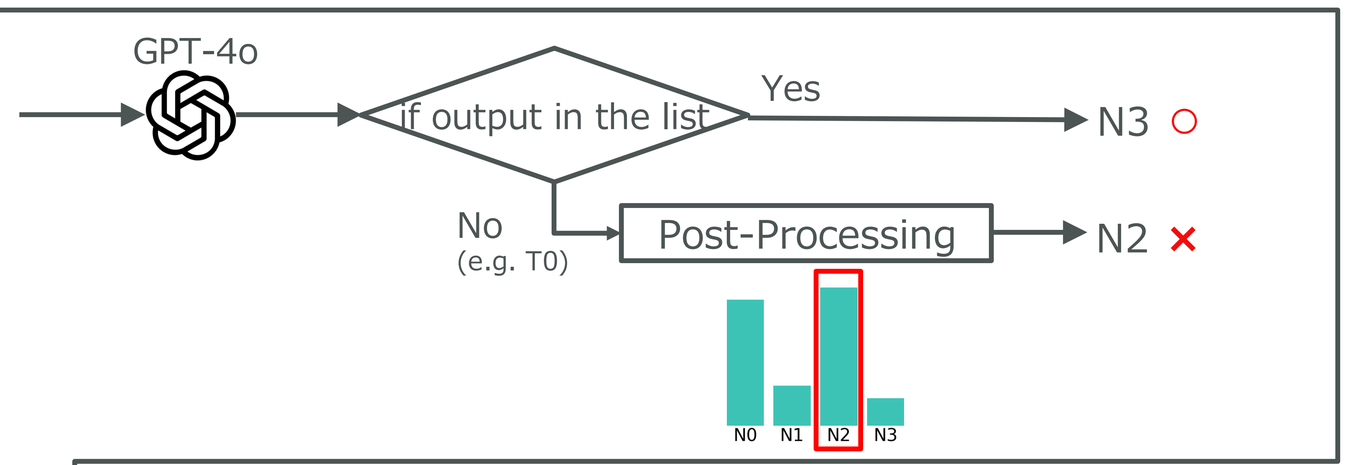
- This document was inferenced by GPT-4o in each category, such as T, N, M
- As post-processing, if a predicted label is not in the allowed list, we replace it with the most frequent label for the category

Prompt Example, Category: N

```
The following is a task to predict the diagnostic outcome of lung cancer based
on CT reading reports. Please provide the prediction results in the following
JSON format, adhering to diagnostic guidelines.
Example:
{
   "output": "N3",
   "explanation": "Determined as N3 due to enlargement of contralateral
mediastinal lymph nodes. "
}
You must output one of the words from the following list                  allowed list
['N0', 'N1', 'N2', 'N3']

Complete atelectasis of the left upper lobe.
74 mm mass with poor contrast enhancement, obstructing …                  content
Normal appearance of the visualized upper abdomen.

{                                                                         definition of label
  "N0": "No regional lymph node metastasis",
  "N1": "Metastasis in ipsilateral peribronchial …",
  "N2": "Metastasis in ipsilateral mediastinal and/or subcarinal lymph node(s)",
  "N3": "Metastasis in contralateral mediastinal, contralateral hilar …"
}
```



Baseline

- Fine-tuned bi-encoder models: BERT, JMEDRoBERTa
- Data augmentation: Shuffled and rearranged sentences on a per-sentence basis

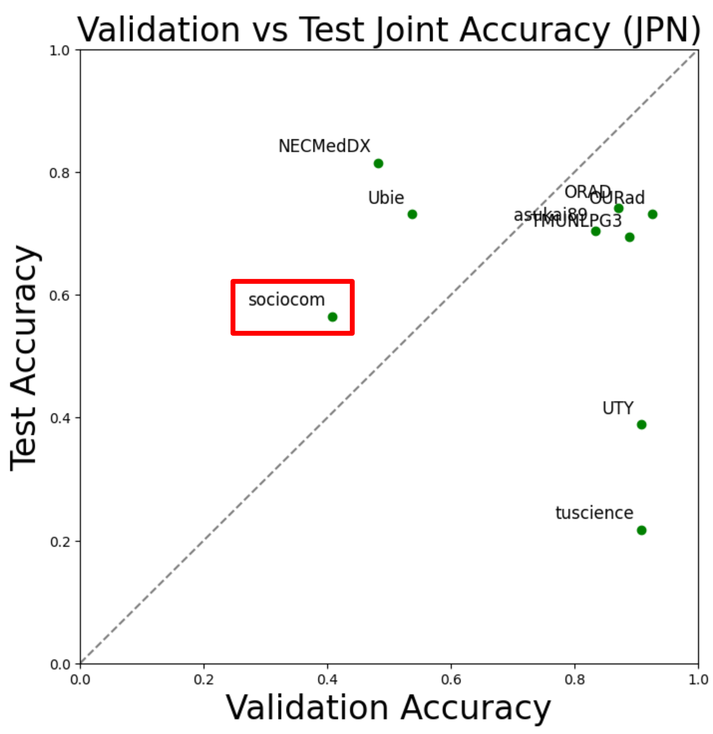
Results

- GPT-4o achieved the highest scores on the joint task
- GPT-4o is not necessary to post-processing
- Demonstrated performance that significantly improved bi-encoder models

Model Name	Joint		T		N		M	
	fine	coarse	fine	coarse	fine	coarse	fine	coarse
Baseline (most frequency label)	0.0	0.0	0.3333	0.3333	0.3704	0.3704	0.5	0.5
BERT	0.0741	0.2407	0.2593	0.4259	**0.7963**	**0.7963**	0.7037	0.7778
BERT (DA)	0.1852	0.3889	0.4074	0.5185	0.7222	0.7222	0.7407	0.8519
JMEDRoBERTa (DA, WordPiece)	0.0926	0.2037	0.4074	0.5741	0.5000	0.5000	0.5926	0.8519
JMEDRoBERTa (DA, SentencePiece)	0.2037	0.4630	0.4074	0.6111	0.7222	0.7222	0.6111	0.8889
GPT-4o-mini	0.2407	0.4815	0.5556	0.7778	0.6111	0.6111	0.8148	0.8889
GPT-4o-mini (Post Processing)	0.3148	0.5556	0.5556	0.7778	0.6852	0.6852	0.8148	0.8889
GPT-4o	**0.4259**	**0.5926**	**0.6667**	**0.8333**	0.7593	0.7593	**0.9444**	**0.9630**
GPT-4o (Post Processing)	**0.4259**	**0.5926**	**0.6667**	**0.8333**	0.7593	0.7593	**0.9444**	**0.9630**

* Bold font indicates the best score for each evaluation metric.
DA: Data Augmentation

Our model shows less overfitting than other teams' systems



Conclusions

- Our results demonstrated that GPT-4o achieved higher accuracy compared to traditional bi-encoder models
- Shared a simple, yet robust, GPT-4-mini baseline method with the community
- These findings highlight the potential of LLMs in medical natural language processing tasks

Nakamura et al. NTCIR-18 RadNLP 2024 Overview: Dataset and solutions for automated lung cancer staging. In Proceedings of the NTCIR-18 Conference, 2025