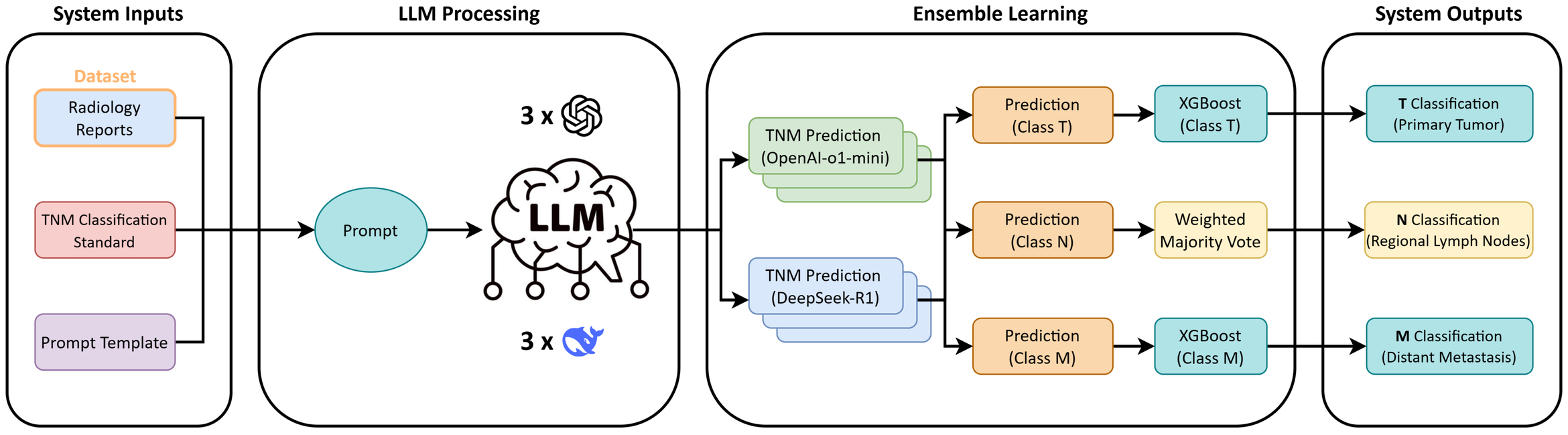
From Divergent LLM Predictions to Reliable Lung Cancer Staging with Ensemble Fusion: CYUT at the NTCIR-18 RadNLP Main Task

Tsz-Yeung Lau
Chaoyang University of Technology Taiwan (R.O.C)
s11327605@gm.cyut.edu.tw

Shih -Hung Wu
Chaoyang University of Technology Taiwan (R.O.C)
shwu@cyut.edu.tw

System Overview

This study investigates the application of Large Language Models (LLMs) for automated lung cancer staging based on radiology reports, as part of the CYUT team's participation in the NTCIR-18 RadNLP Main Task. Our system presents a comprehensive framework for automating lung cancer staging through the analysis of radiological reports, as illustrated in the following Figure:



Data Preprocessing

We discover that using "mm" units typically converts measurements to integers (e.g., "37 mm" vs. "3.7 cm"), which can improve model performance. Integers offer a consistent format, simplifying numerical comparison for the LLM. They are often processed as single tokens (e.g., "37"), potentially making them easier to compare directly than decimals, which can be tokenized into multiple parts (e.g., "3", ".", "7").

Report ID	Unit	Text
923073	cm	A mass with a maximum diameter of 3.7 cm in the left upper lobe of the lung.
923073	mm	A mass with a maximum diameter of 37 mm in the left upper lobe of the lung.

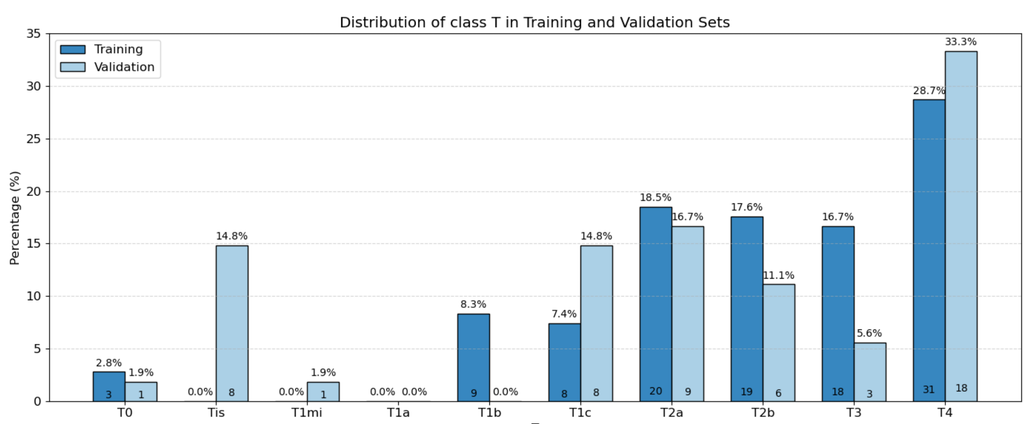
Table 2: DeepSeek-R1 Tokenization Examples.

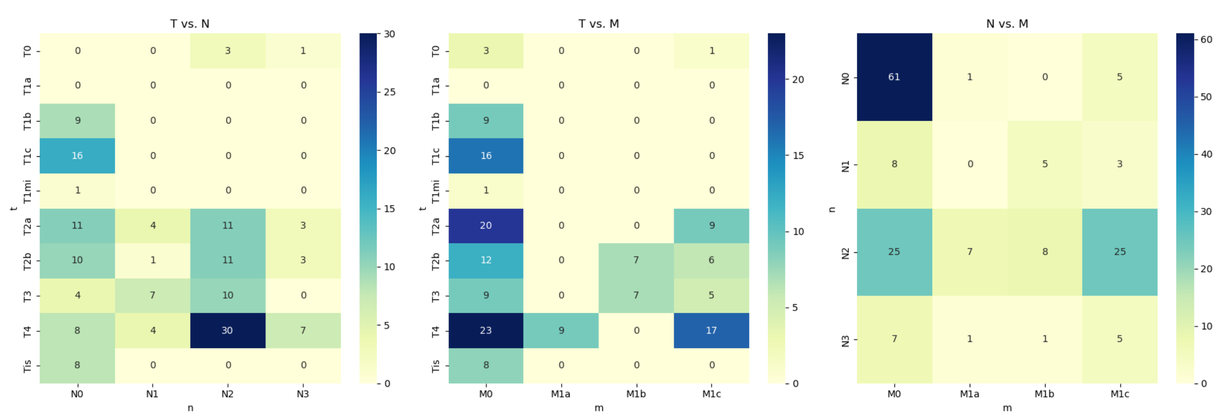
Model	Joint (fine)	T (fine)	N (fine)	M (fine)
Llama3.1-8B-fp16 (mm)	**0.46**	**0.67**	**0.87**	0.85
Llama3.1-8B-fp16 (cm)	0.35	0.52	0.85	**0.89**

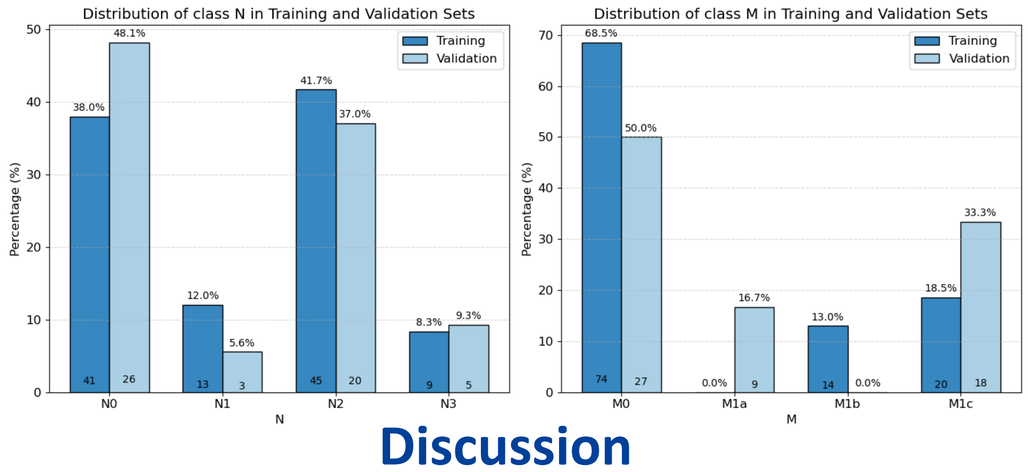
Table 3: Comparison of Different Approaches with Llama-3

Data Distribution and Relationship

The dataset presents an imbalances challenge. It could negatively affect the performance of traditional methods, such as BERT. We also observed a moderate correlation among the T, N, and M staging classes. Experimental results indicated that jointly prompting LLMs to predict all three classes simultaneously yields improved performance.







Result

Model	Joint (fine)	T (fine)	N (fine)	M (fine)	Joint (coarse)	T (coarse)	N (coarse)	M (coarse)
Reasoning Model Ensemble	**0.83**	**0.89**	0.94	**1.00**	**0.93**	**0.98**	**0.94**	**1.00**
DeepSeek-R1	0.81	**0.89**	0.94	0.96	0.89	**0.98**	**0.94**	0.96
OpenAI-o1-mini	0.78	0.81	0.94	0.98	0.87	0.93	**0.94**	0.98
GPT-4o	0.78	0.85	**0.96**	0.93	0.85	0.96	0.96	0.93
GPT-4o-mini (baseline)	0.24	0.56	0.57	0.80	0.41	0.76	0.57	0.85
RoBERTa	0.24	0.39	0.61	0.88	0.27	0.42	0.61	0.91

Table 4: Validation Accuracies for Various Models(mm) (Sorted by Joint (fine)).

Discussion

Report ID: 241752, Misclassify Tis to T1b.
Major Problem: Report don't provide enough information of invasive component, so can't tell is the tumor Tis/T1b.
Report ID: 2318717, Misclassify T1mi to T1b.
Major Problem: Model primarily focusing on its overall size rather than its specific histological/radiological definition.
Report ID: 12646171, Misclassify T4 to T3.
Major Problem: The model fails to follow the staging rules correctly and completely.

Future Work

1. Utilize **Supervised Fine-Tuning (SFT)** to fine-tune LLM.
2. Employ **Reinforcement Learning (RL)** to generate thinking process to enhance model ability.
3. Unlock Few-Shot learning with RAG and BERT fine-tuning can be achieved by **enriching the dataset size**.
4. Integrate **multi-modal data** (e.g., CT images, radiology reports) for richer contextual understanding.