ORAD at NTCIR-18 RadNLP 2024 Shared Task

Keisuke Hidaka

Osaka Red Cross Hospital Department of Radiology

Introduction

The main and Japanese task of the RadNLP 2024 Shared Task focuses on correctly determining the TNM classification from lung cancer radiology reports using natural language processing. Since radiology reports do not always contain sufficient information to uniquely specify the TNM classification, we created reasoning data to bridge this gap. We investigated whether providing reasoning data to large language models (Gemini 2.0 Flash Experimental / Gemma2-9b-it) could improve the accuracy of TNM classification.

– Methods

Overview of our system

Results

Results on validation data

(Gemini 2.0 Flash Experimental)

Accuracy		Zero-Shot	Many-Shot (non- reasoning)	CoT + Zero-Shot	CoT + Many-Shot (reasoning)
	Joint	0.685	0.667	0.722	0.963
fino	Т	0.833	0.796	0.833	0.981
nne	Ν	0.963	0.889	0.944	1.000
	Μ	0.833	0.981	0.944	0.981
coarse	Joint	0.815	0.815	0.833	0.981
	Т	0.944	0.926	0.889	1.000
	Ν	0.963	0.889	0.944	1.000
	Μ	0.889	1.000	1.000	0.981



✓ Reasoning Data

A board-certified radiologist with 5 years of experience created detailed reasoning processes for TNM classification from reports in several cases; for the remaining cases, initial reasoning was generated by GPT-40 and subsequently manually revised by the radiologist to ensure accuracy. An example is provided below. Since we participated in the Japanese task, all prompts are in Japanese, with corresponding English translations shown alongside.

In total, 108 such examples were created, demonstrating the complete reasoning process from report interpretation to final classification.

Radiology report: 左上葉に最大径 47mm の腫瘤があります。葉間胸膜を越えて、一部、 下葉に浸潤しています。肺門リンパ節は腫瘍と一塊となっています。 縦隔リンパ節腫大は認めません。胸水貯留はありません肝転移、副腎 転移はありません。撮影範囲の腹部臓器に粗大な異常を認めません。	Radiology report: There is a tumor with a maximum diameter of 47 mm in the upper left lobe. It has infiltrated the lower lobe, crossing the interlobar pleura. The hilar lymph nodes are fused with the tumor. No enlargement of the mediastinal lymph nodes is observed. There is no pleural effusion. No liver or adrenal metastasis is observed. No significant abnormalities are observed in the visualized abdominal organs.	
Reasoning: <t分類>まずはT分類を検討します。性状とサイズを考えます。左上葉 に最大径47mmの腫瘤があり、この腫瘤は「腫瘤」と記載があるため 充実成分であると仮定します。径が47mmであるため、40mmより大</t分類>	Reasoning: Classification First, the T classification is considered based on characteristics and size. There is a 47-mm mass located in the left upper lobe. Since it is described explicitly as a "mass," it is assumed to be entirely solid. The diameter of 47 mm places it in the range greater than 40 mm but not exceeding 50.	

Results on test data

(Gemini 2.0 Flash Experimental)

Accuracy		CoT + Many-Shot (reasoning)		
fine	Joint	0.741		
	Т	0.833		
	Ν	0.958		
	М	0.907		
coarse	Joint	0.806		
	Т	0.861		
	N	0.958		
	М	0.944		

Results on validation data

充夫成分であると1仮走します。 全か4/mm であるため、40mmより入				
きく、50mm以下に該当し、T2bに相当します。浸潤や副条件を考えま				
す。腫瘍が葉間胸膜を越えており、下葉に浸潤しているため、この浸				
潤もT2bの範囲内と考えられます。性状・サイズ・浸潤を総合的に評				
価して、T分類はT2bと判断します。分類>				
<n分類>次にN分類を検討します。レポートには「肺門リンパ節は腫</n分類>				
瘍と一塊となっている」と記載されています。これは同側肺門リンパ				
節への転移に該当し、N1に分類されます。縦隔リンパ節腫大が認めら				
れないため、N2以上ではありません。このため、N分類はN1と判断し				
ます。				
<m分類>最後にM分類を検討します。肝転移、副腎転移がなく、撮影</m分類>				
範囲の腹部臓器に粗大な異常も認められないため、遠隔転移を示す所				
見はありません。また、胸水がないため悪性胸水も否定的です。				
これらを踏まえて、M分類はM0と判断します。分類>				
以上より、TNM分類は <answer>T2bN1M0</answer> と考えます。				

LLMs Used and Strategies

1. Gemini 2.0 Flash Experimental

Leveraging the extensive input context, we included both the radiology reports and the prepared reasoning data within the prompts. We anticipated that the model would learn how to align the diversity of the provided radiology reports with the TNM classification definitions through the reasoning data.

Prompt Engineering

diameter of 47 mm places it in the range greater than 40 mm but not exceeding 50 mm, corresponding to T2b. Regarding infiltration or additional conditions, the tumor crosses the interlobar pleura and infiltrates into the lower lobe; however, this infiltration remains within the scope of T2b. Taking into account its size, characteristics, and infiltration collectively, the final T classification is determined as T2b.

<N classification>Next, the N classification is assessed. The report describes the hilar lymph nodes as contiguous with the tumor. This represents metastasis to ipsilateral hilar lymph nodes, thus classified as N1. Since no mediastinal lymph node enlargement is identified, the condition does not meet criteria for N2 or higher. Therefore, the N classification is determined to be N1.</N classification> </M classification>Finally, the M classification is evaluated. No hepatic or adrenal metastasis is present, nor any notable abnormalities in abdominal organs within the imaging field, indicating absence of distant metastasis. Additionally, pleural effusion is not observed, ruling out malignant pleural effusion. Based on these findings, the M classification is considered to be: Mo.</m classification> In summary, the TNM classification is considered to be: <a href="#relations.com No hepatic or adrenal">

2. Gemma2-9b-it

We also evaluated Gemma2, which is lightweight and suitable for local environments. Supervised fine-tuning (SFT) was performed using radiology reports and reasoning data as training data.

X SFT strategies

•4-bit quantization (NF4) and double quantization (unsloth model)
•Learning rate=2e-4, epochs=7, Lora_rank=256

(Gemma2-9b-it)

Accuracy		Raw model + CoT	SFT (non-reasoning) + Zero-Shot	SFT (reasoning) + CoT			
fine	Joint	0.241	0.370	0.463			
	Т	0.426	0.463	0.648			
	Ν	0.778	0.963	0.926			
	Μ	0.759	0.833	0.796			
coarse	Joint	0.519	0.741	0.778			
	Т	0.722	0.796	0.852			
	Ν	0.778	0.963	0.926			
	Μ	0.944	0.963	0.944			

- Discussion

Gemini 2.0 Flash Experimental achieved strong accuracy even in zero-shot settings, and significantly improved Tclassification accuracy when combined with CoT prompting and many-shot reasoning data.

For Gemma2-9b-it, SFT with reasoning data improved Tclassification inference compared to using TNM labels alone.

Definition of TNM classification

We include the definition and cautionary notes of the TNM classification of lung cancer in the prompt so that the LLM would strictly follow the established criteria and fully understand the rules specific to this task(e.g., "TX" does not exist).

Chain-of-Thought prompt (CoT)

To guide the LLM through a sequential examination of T, N, and M classifications, we adopted a type of chain-of-thought (CoT) prompt that utilizes XML-like tags to elicit a step-by-step TNM classification output.

In-Context Learning (ICL)

Considering that conventional few-shot prompting would be insufficient for learning diverse reasoning patterns, we adopted a Many-Shot In-Context Learning approach, incorporating all available training data into the prompts.

The CoT approach, by explicitly outputting the reasoning process, can also be expected to enhance transparency and interpretability—important factors for medical applications.

- Conclusion

This study demonstrated that providing reasoning processes to LLMs through Many-Shot ICL or fine-tuning can significantly improve TNM classification accuracy, particularly in T-staging while maintaining interpretability.