

TMUNLPG3 AT THE NTCIR-18 RADNLP TASK

Wen-Chao Yeh (speaker), Yan-Chun Hsing, Tzu-Yi Li, Nitisalapa Timsatid, Shih-Chuan Chang, Shih-Hsin Hsiao, Chu-Chun Wang, Pak-Yue Chan, Wen-Lian Hsu and Yung-Chun Chang

2025-06-13

<https://hackmd.io/@wyeh/ntcir18-radnlp>

AGENDA

WHO WE ARE

WHAT DO WE NEED TO DO AND ACHIEVE

HOW DO WE ACCOMPLISH THIS

CONCLUSION

Q&A

WHO ARE TMUNLPG3?

WE ARE AND DOMAIN EXPERTS

Wen-Chao Yeh



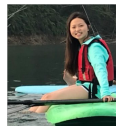
NTHU PhD
Student

Yan-Chun Hsing



TMU Research
Assistant

Tzu-Yi Li



TMU Health Care A.
BS

Nitisalapa Timsatid



TMU Data Science
MS

Shih-Chuan Chang



TMU Data Science
MS

Shih-Hsin Hsiao



TMUH Doctor

Chu-Chun Wang



TMUH Senior PA

Pak-Yue Chan



TMU Medicine BS

Wen-Lian Hsu



NTHU Professor

Yung-Chun Chang



TMU Professor

**WHAT DO WE NEED TO DO AND
ACHIEVE**

RADNLP MAIN TASK

A multi-label document classification to correctly determine **T**, **N**, and **M** categories for each radiology report.

T - the size and/or extension of the primary lesion:

T0, Tis, T1mi, T1a, T1b, T1c, T2a, T2b, T3, T4

N - the extent of lymph node metastasis:

N0, N1, N2, N3

M - the extent of distant metastasis:

M0, M1a, M1b, M1c

RADNLP SUB TASK

a document segmentation (sentence level) to identify up to eight spans related to the following topics

Omittable

Measure

Extension

Atelectasis

Satellite






Lymphadenopathy

Pleural

Distant

WHAT WE'VE ACHIEVED

RadNLP aims to automatically determine the TNM stage of lung cancer from radiology reports.

	English Track	Japanese Track
Main Task TNM Staging	 	
Sub Task Multi-label Sentence Classification		

HOW DO WE ACCOMPLISH THIS?

THREE STEPS TO TACKLE CHALLENGES

Analysis Dataset

Distribution of training set and validation set
Opinions from medical doctor, radiologist and pathologist

Methods

Last year's approach?
LLMs and Pre-trained models (e.g., BERT)

Optimization

Analyze errors
Fine-tune prompts

OPINIONS FROM MEDICAL EXPERTS, CASE 1

T1b / N0 / M0

A 12 mm subpleural nodule in the right lower lobe, believed to be a finding of the known lung cancer. Interstitial pneumonia is suspected in both lungs. No pathological lymph node enlargement. No pleural effusion. Gallstones.

OPINIONS FROM MEDICAL EXPERTS, CASE 2

T2b / N1 / M0

The tumor has spread to the ipsilateral hilar lymph nodes and fused with the tumor.
腫瘤已經擴展到同側的肺門淋巴結，並與腫瘤融合

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

OPINIONS FROM MEDICAL EXPERTS, CASE 3

T2b / N0 / M0

Emphysematous changes.

A 47 mm irregular mass in the left upper lobe, suspicious of lung cancer. Possible invasion into the left pulmonary artery, as far as can be assessed with CT to a limited extent. The mass is also close to the aortic arch.

A small nodule was noted in the right middle lobe. While it could be inflammation or metastasis, it remains uncertain at present. Therefore, in the absence of definitive evidence of metastasis...

指出右中葉有小結節，雖然可能是炎症或轉移，但目前還不確定，因此在缺乏明確轉移證據的情況下

OPINIONS TO GUIDELINES

Medical doctor, radiologist and pathologist

- Annotate training dataset according to their expertise
- Compare with released labels
- Provide analysis of discrepancies between the two

Prompt Writing Guidelines

DATASET QUANTITIES

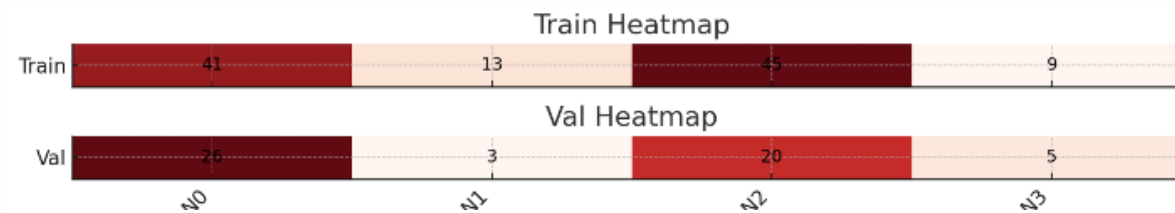
Dataset	English Track	Japanese Track
Main Task (Document Level)		
Train	108	108
Validation	54	54
Test	81	216
Sub Task (Sentence Level)		

TNM CLASSIFICATION DISTRIBUTION

◦ T



◦ N



SUBTASK LABEL DISTRIBUTION

Train(918)+Val(415) = 1,333

Category	0	1	% of Class 1
omittable	738	595	44.64%
measure	1085	248	18.6%
extension	1175	158	11.85%
atelectasis	1280	53	3.98%
satellite	1253	80	6.0%

METHODS

Total: 5 systems

Three adopted LLMs, two use BERT-related models.

English Track

Main Task
TNM Staging



System I



System II

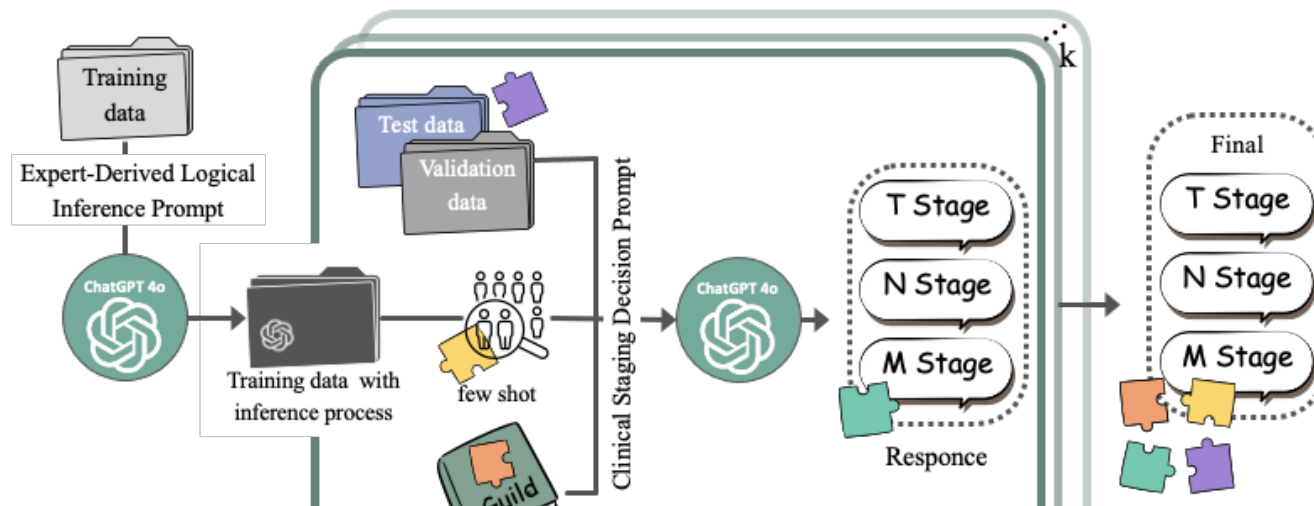
Sub Task
Multi-label Classification



System II

SYSTEM I: ARCHITECTURE

Using LLM and Few-Shots (report+label+reasoning)



SYSTEM I: REASONING

Using ChatGPT-4o to analyze the **rationale** behind TNM annotations in each radiology report within the training dataset.

💡 Reasoning from GPT-4o, Example: 1863157

T: T1b
N: N0
M: M0

Report: A nodule measuring 12 mm in diameter in the right lower lobe S8/9, increasing in size, suspected to be a known lung cancer. Reticular opacities present in the right S6, likely inflammatory changes. Calcified pleural plaques on both sides, suggesting the possibility of asbestos-related disease. No significant enlargement of the mediastinal and hilar lymph nodes or other mediastinal lesions. No pleural effusion. Gallstones.

Reasoning: The nodule is 12mm, which fits T1b (tumor >10mm but ≤20mm). No lymph node enlargement suggests N0 (no regional lymph node metastasis). Absence of distant metastasis or pleural effusion indicates M0 (no distant metastasis). These factors collectively account for the TNM classification T1b/N0/M0.

SYSTEM I: PROMPT DESIGN

Combines guidelines with 7 randomly selected examples from training set (including report, reasoning and answers) to guide the model in generating TNM by following the examples.


example:#1863157

Input:A nodule measuring 12 mm in diameter in the right lower lobe S8/9, increasing

Reasoning: The nodule is 12mm, which fits T1b (tumor >10mm but ≤20mm). No lymph node

Output: {"T":T1b,"N":N0,"M":M0}

SYSTEM I: PREDICTION

- a report  three rounds of predictions
 - Hard vote on the T-stage from the above 3 predictions
 - Hard vote on the N-stage from the above 3 predictions
 - Hard vote on the M-stage from the above 3 predictions
- Determine the conclusive result.

SYSTEM II: ARCHITECTURE



dspy.ai document

```
1 import dspy
2 lm = dspy.LM('openai/gpt-4o-mini', api_key='YOUR_OPENAI_A
3 dspy.configure(lm=lm)
4
5 from typing import Literal
6
7 class Classify(dspy.Signature):
8     """Classify sentiment of a given sentence."""
9
10    sentence: str = dspy.InputField()
11    sentiment: Literal['positive', 'negative', 'neutral']
12    confidence: float = dspy.OutputField()
13
14 classify = dspy.Predict(Classify)
```


SYSTEM II: PROMPT DESIGN

IASLC 8th ed. lung cancer staging system

- Intern doctor 1: gpt-4o uses chain-of-thought to determine TNM staging and opinions.
- Intern doctor 2: gemini-2 uses chain-of-thought to determine TNM staging and opinions.
- Senior doctor: Reviews the staging decisions and reasoning from both intern doctors, then makes the final decision. This output serves as the answer.

SYSTEM II: SPLIT DATASET

- Randomly split 108 training data points into 54 few-shot reference cases and 54 validation cases.
- The original validation set and test set were retained for prediction inference and submission.
- Overfitting prevention.

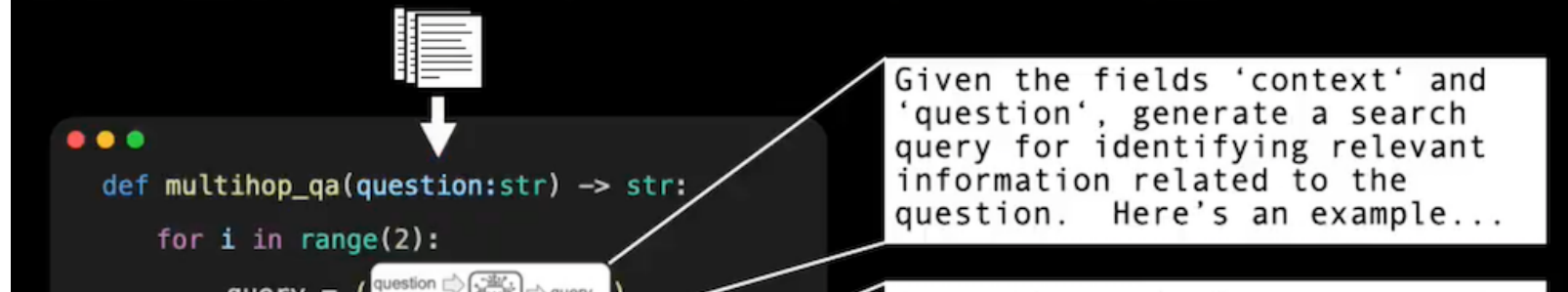
SYSTEM II: OPTIMIZATION

- Used the MIPROv2 method to select 50 few-shot reference cases for auto-prompt instruction fine-tuning design.
- The 54 validation cases were used to test which modifications performed better.
- The prompt was then optimized based on these results.

Krista Opsahl-Ong, Michael J. Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs. arXiv:2406.11695 [cs].

MIPROV2

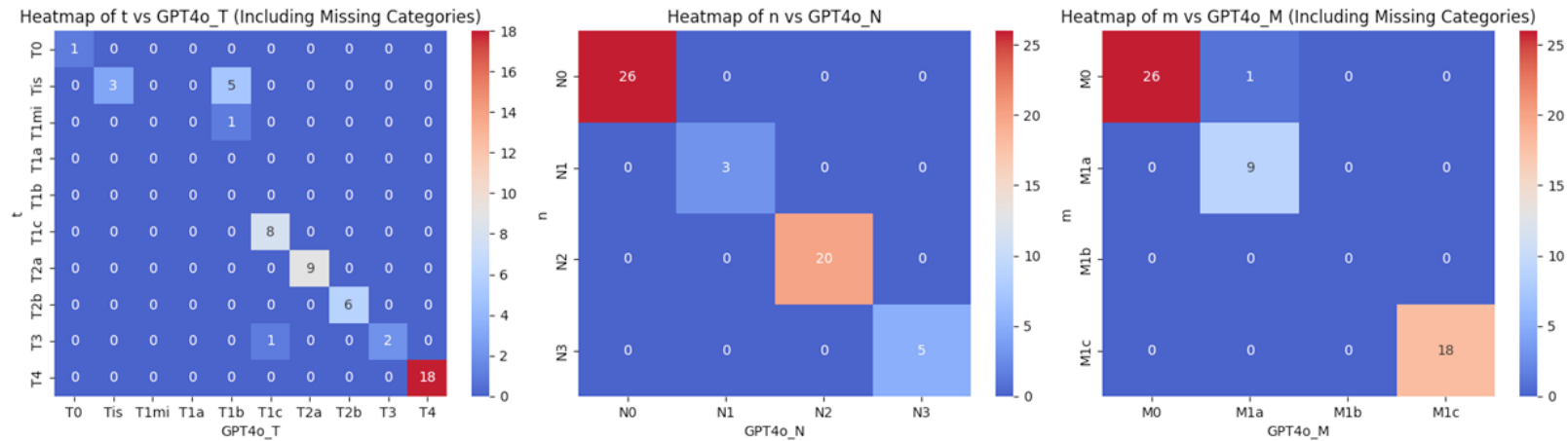
> Given a potentially multistage language program and training inputs with a metric we find optimized instructions and fewshot demonstrations.



SYSTEM II FOR SUBTASK

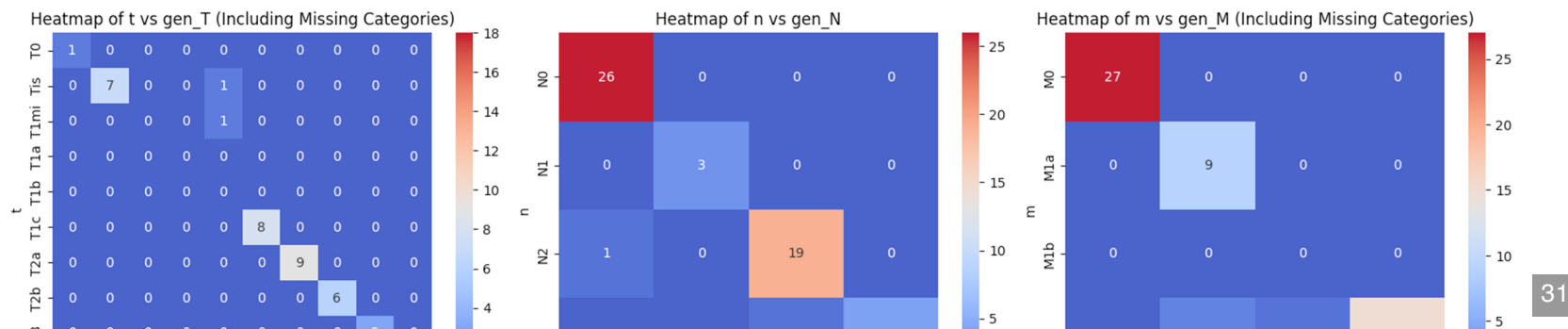
- Same as System II architecture for main task.
- LLM Model: Llama 3.3 70B 16 bits
 - 8 RTX 3090 GPUs (24GB VRAM each)
- 300 shots (LLM model context windows limits)

ANALYZE ERRORS TO REFINE THE PROMPT



ANALYZE ERRORS TO REFINE THE PROMPT

Tumors without evidence of invasion should be prioritized for Tis classification, even if the size approaches criteria for other stages like T1b. Do not classify tumors as T1b based solely on size; clear pathological evidence of invasion is required. If there is uncertainty, always select the most conservative stage.



CONCLUSION

MAIN TASK, ENGLISH TRACK

System-I-MT-En secured first place with impressive metrics

- 65.43% joint fine accuracy
- 69.14% joint coarse accuracy
- The system demonstrated strong individual performance in
 - T (70.37%)
 - N (91.36%)
 - M (88.89%)

SUB TASK, ENGLISH TRACK

System-II-ST-En achieved second place with a notable overall micro F2.0 score of 93.36%.

FINDING

This success is attributed not only to the implementation of **large language models** but also to the application of **few-shot** prompting engineering and structured **reasoning** in TNM classification.

FINDING

A key advantage of our approach is the **integration of expert medical knowledge**, consulting with experienced doctors, radiologist and pathologist to validate and refine the system.

FUTURE WORK

Our efforts validate the potential of artificial intelligence in medical document analysis, establishing a framework for future clinical decision support systems.

QUESTION ?

THANK YOU