TMUNLPG3 at the NTCIR–18 RadNLP Task

10-13, June, 2025

Wen-Chao Yeh[1,7], Yan-Chun Hsing[2,8], Tzu-Yi Li[3,9], Nitisalapa Timsatid[2,10], Shih-Chuan Chang[2,11], Shih-Hsin Hsiao[5,13], Chu-Chun Wang[5,14], Pak-Yue Chan[4,12], Wen-Lian Hsu[6,15], Yung-Chun Chang[2,*]

[1]*Institute of Information Systems and Applications, National Tsing Hua University, Taiwan*

{[2]*Graduate Institute of Data Science,* [3]*School of Health Care Administration,* [4]*School of Medicine*}*, Taipei Medical University, Taiwan*

[5]*Division of Pulmonary Medicine, Department of Internal Medicine, Taipei Medical University Hospital, Taiwan*

[6]*Department of Computer Science and Information Engineering, Asia University, Taiwan.*

[7]wyeh@m109.nthu.edu.tw, {[8]m946106006, [9]b908108028, [10]m946112012, [11]m946112004, [12]b101108138, *changyc}@tmu.edu.tw, {[13]hsiaomd, [14]judewang1218}@gmail.com, [15]hsu@iis.sinica.edu.tw

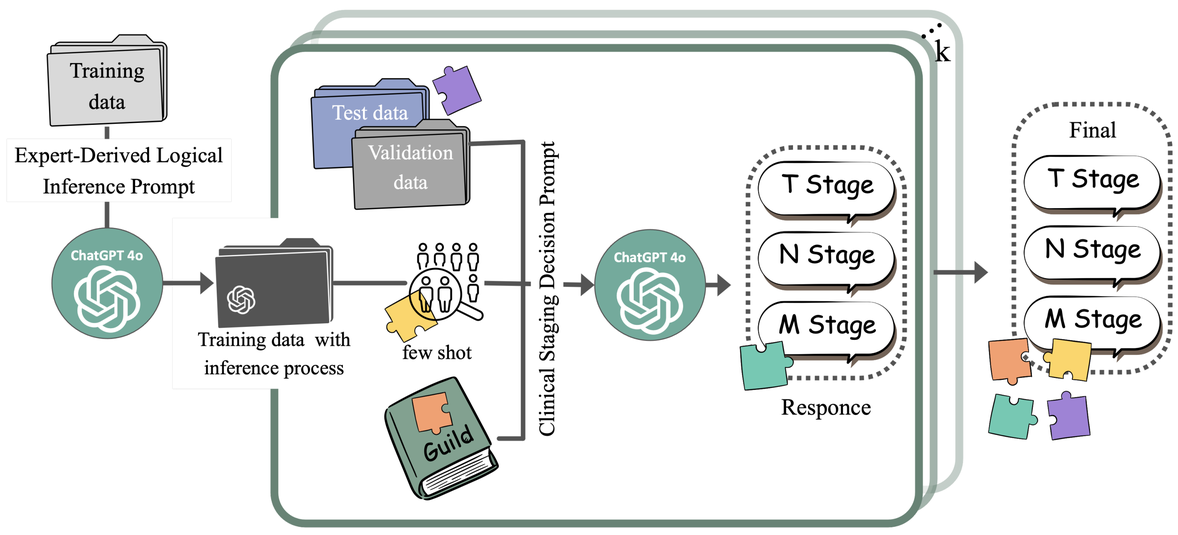
Abstract

The TMUNLPG3 team participated in the Lung Cancer Staging main task and Multi-label Sentence Classification subtask of the NTCIR-18 RadNLP Task. This paper illustrates our approach to address the challenges and discusses the official results. We tackled Lung Cancer TNM Staging maintask to highest among all participants in the English track by adopting LLM and Few-Shot prompt engineering. Our solution also performed excellently in the Multi-label Sentence Classification subtask.

Introduction

The TNM Classification of Malignant Tumors provides a globally standardized system for describing and categorizing the anatomical extent of cancer spread. This system has become a cornerstone for cancer staging worldwide. The NTCIR-18 RadNLP 2024 task aims to automate the staging of cancer from radiological reports. Our team has proposed a novel approach using large language models, combined with few-shot tuning, to significantly enhance the automation of lung cancer TNM staging.

Methods & Results



System I Utilizes GPT-4o with Few-shot prompt engineering

- Extracts reasoning patterns from training data
- Stores contextual references for classification tasks
- hard-voting mechanism with multiple inference runs
- Aggregates predictions via majority voting

Table 1: The result of Main Task in Private Leaderboard

System	Rank	Joint	T	N	M	Joint	T	N	M
		Fine Accuracy				Coarse Accuracy			
English Track									
System-I-MT-En	1	65.43	70.37	91.36	88.89	69.14	74.07	91.36	91.36
System-II-MT-En	2	62.96	72.84	93.83	83.95	66.67	74.07	93.83	88.89
System-III-MT-En	6	55.56	64.20	88.89	83.95	58.02	65.43	88.89	88.89
System-V-MT-En	7	53.09	65.43	91.36	85.19	58.02	66.67	91.36	92.59
Japanese Track									
Vote(System-I-MT-Ja, System-II-MT-Ja)	7	69.44	79.17	91.67	91.20	77.31	84.72	91.67	94.44
System-I-MT-Ja	8	68.52	77.78	92.13	92.13	78.24	87.04	92.13	94.44

Table 2: The result of Sub Task in Private Leaderboard

System	Rank	Overall	Inclusion	Measure	Extension	Atlectasis	Satellite	Lymphade nopathy	Pleurak	Distant
		Micro F2.0								
English Track										
System-II-ST-En	2	93.36	92.97	82.07	75.22	86.96	78.31	97.70	96.15	91.22
System-III-ST-En	5	91.55	95.08	79.40	72.73	84.07	68.45	98.08	96.15	82.76
System-IV-ST-En	6	91.50	96.69	81.91	73.39	71.43	75.30	96.15	88.24	83.62
Japanese Track										
System-II-ST-Ja	4	16.53	20.34	13.57	10.90	12.43	07.10	07.88	07.17	12.93

System II Implements a robust multi-model approach for TNM staging classification using DSPy framework

- Leverages GPT-4o and Gemini-2 models in an alternating pattern. Generates multiple independent assessments by alternating between language models.
  - Uses prompts embodying skilled radiologist and oncologist roles
  - Follows IASLC 8th edition lung cancer staging guidelines
- Core architecture simulates a panel of medical experts analyzing radiology reports and make final decision by GPT-4o
- Optimized using MIPROv2 for fine-tuned performance

For Multi-label Sentence Classification:

- Features specialized prompt design for Multi-label Sentence Classification
- Leverages Meta Llama-3.3-70B as core processing engine
- Uses MIPROv2 optimizer with customizable parameters
- Supports up to 300 few-shot

System III employs a hierarchical pipeline featuring BERT-based preprocessing for semantic enrichment, data augmentation to address label imbalances, and subtask-specific text marking. Its classification module combines fine-tuned Bio_ClinicalBERT with frozen layers and an ensemble approach (Naive Bayes, XGBoost, SVM) to enhance TNM staging accuracy.

System IV enhances RadBERT-RoBERTa with multi-attention heads and custom attention layers, using GPT-4o-mini for sentence augmentation to address imbalanced classification categories.

System V employs GPT-4o in a Zero-Shot Chain-of-Thought framework, generating multiple assessments with clinical reasoning before teacher-model validation for TNM staging.

Conclusion

We excelled in the NTCIR-18 RadNLP lung cancer staging task. System-I-MT-En ranked first in the English track with 65.43% joint fine accuracy and 69.14% joint coarse accuracy, showing strong performance across T (70.37%), N (91.36%), and M (88.89%) classifications. System-II-ST-En placed second in multi-label sentence classification with a 93.36% micro F2.0 score. The implementing LLM with few-shot prompting, structured reasoning, and expert medical knowledge integration, demonstrating AI's potential in medical document analysis.

This work was supported by the National Science and Technology Council of Taiwan under grants NSTC 112-2622-E-038-001, NSTC 113-2221-E-038-019, NSTC 113-2627-M-A49-002, NSTC 113-2321-B-038-012, and NSTC 113-2321-B-038-006. This research was partially supported by the National Science and Technology Council of Taiwan, under the program of AI Thematic Research Program to Cope with National Grand Challenges, project NSTC 113-2634-F-A49-004, in collaboration with the Pervasive Artificial Intelligence Research Labs of the National Yang Ming Chiao Tung University.