NTCIR-18: Proceedings of the 18th NTCIR Conference on Evaluation ofInformation Access Technologies, June 10-13, 2025, Tokyo, Japan
DOI: https://doi.org/10.20736/0002002073

TMUNLPG3 at the NTCIR-18 RadNLP Task

Wen-Chao Yeh
Institute of Information Systems and Applications, National Tsing Hua University
Hsinchu, Taiwan
wyeh@m109.nthu.edu.tw

Yan-Chun Hsing
Graduate Institute of Data Science, Taipei Medical University
Taipei, Taiwan
m946106006@tmu.edu.tw

Tzu-Yi Li
School of Health Care Administration, Taipei Medical University
Taipei, Taiwan
b908108028@tmu.edu.tw

Nitisalapa Timsatid
Graduate Institute of Data Science, Taipei Medical University
Taipei, Taiwan
m946112012@tmu.edu.tw

Shih-Chuan Chang
Graduate Institute of Data Science, Taipei Medical University
Taipei, Taiwan
m946112004@tmu.edu.tw

Shih-Hsin Hsiao
Division of Pulmonary Medicine, Department of Internal Medicine, Taipei Medical University Hospital
Taipei, Taiwan
hsiaomd@gmail.com

Chu-Chun Wang
Division of Pulmonary Medicine, Department of Internal Medicine, Taipei Medical University Hospital
Taipei, Taiwan
judewang1218@gmail.com

Pak-Yue Chan
School of Medicine, Taipei Medical University
Taipei, Taiwan
b101108138@tmu.edu.tw

Wen-Lian Hsu
Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, Taiwan
Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan
hsu@iis.sinica.edu.tw

Yung-Chun Chang*
Graduate Institute of Data Science, Taipei Medical University
Taipei, Taiwan
changyc@tmu.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.

KEYWORDS

Medical Natural Language Processing, Radiology Reports, Lung Cancer, Cancer Staging, LLM, Prompt Engineering

TEAM NAME

TMUNLPG3

SUBTASKS

Main task (Japanese & English tracks)

Sub task (Japanese & English tracks)

1 INTRODUCTION

The TNM Classification of Malignant Tumors represents a globally standardized system for describing and categorizing the anatomical extent of cancer spread. Originally developed in France during the 1940s by Pierre Denoix, this system has evolved to become the cornerstone of cancer staging worldwide [1]. The system's fundamental importance lies in its ability to provide a standardized framework for cancer classification, primarily focusing on solid tumors (though notably excluding leukemia and central nervous system tumors). Its universal acceptance has significantly enhanced communication between healthcare providers and facilitated broader research initiatives across different populations [10].

The NTCIR-18 RadNLP 2024 task [7] represents a significant advancement in automating cancer staging from radiology reports, with particular focus on lung cancer. The task has evolved to incorporate more detailed classification requirements, reflecting real-world clinical needs. The main components of this task involve: **Main Task: Lung Cancer Staging** (including English and Japanese tracks), compare with NTCIR-17 RR-TNM task [12], this year enhanced TNM classification granularity. The system now includes suffix-based classifications (e.g., "T1a") and expanded from 3-label, 2-5 class classification to 3-label, 4-10 class classification.

- T categories: T0, Tis, T1mi, T1a, T1b, T1c, T2a, T2b, T3, T4
- N categories: N0, N1, N2, N3

* Corresponding Author

- M categories: M0, M1a, M1b, M1c

Sub Task: Multi-label Sentence Classification (including English and Japanese tracks) includes a sophisticated sentence-level annotation system with eight distinct categories that require binary classification (0 or 1) for each category, allowing for multiple classifications per sentence.

- Omittable: Sections without positive findings
- Measure: Primary lesion existence and diameter
- Extension: Primary lesion spread beyond lung parenchyma
- Atelectasis: Identification of atelectasis or obstructive pneumonia
- Satellite: Documentation of intrapulmonary metastasis
- Lymphadenopathy: Regional lymph node enlargement
- Pleural: Assessment of pleural/pericardial effusion
- Distant: Identification of distant metastasis

This comprehensive approach to cancer staging automation represents a significant step forward in medical informatics, combining standard TNM classification principles with natural language processing capabilities. The task's structure reflects both the complexity of cancer staging and the need for precise, standardized documentation in clinical settings. Our team, TMUNLPG3, tackled Lung Cancer 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 in the English track, ranking 2nd. All source code of systems in this paper are all accessible through a GitHub Repository[2].

2 RELATED WORK

Prior research in automated cancer staging from clinical texts includes the Radiology Report TNM staging (RR-TNM) subtask, conducted as part of the NTCIR-17 MedNLP-SC shared task in 2023. This initiative focused on developing automated systems for lung cancer staging from radiology reports, utilizing a curated dataset of 243 anonymized Japanese radiology reports. The systems demonstrated promising results, achieving accuracy scores of 67%, 80%, and 93% for T, N, and M categories respectively. This work provided significant insights into leveraging natural language processing for extracting staging information from medical documents. [12]

The NTCIR-17 RR-TNM task showcased diverse approaches to automated cancer staging, with three teams presenting distinct methodological perspectives. Team KRad [6] took an innovative approach by leveraging the capabilities of GPT-3.5-turbo through zero-shot in-context learning. Their strategy involved crafting specialized prompts that positioned the model as a thoracic surgeon, equipped with staging criteria and specific output format requirements. When faced with uncertain predictions, their system defaulted to conservative T0/N0/M0 classifications. Meanwhile, Team kuhp [5] explored a different direction by fine-tuning open-calm-7b4, a Japanese language model. They enhanced their training data through creative augmentation techniques, including character manipulation and sentence restructuring, while also incorporating handcrafted clinical knowledge. Their experimental approach led to three submissions with varying degrees of model fine-tuning. The third competitor, Team NAIST-SOCRR [11], approached the challenge from a traditional document classification perspective, deploying various configurations of bidirectional transformer models. Their submissions ranged from a sophisticated joint inference system using JMedRoBERTa to a pragmatic majority baseline approach.

The evolution of this task from its previous iteration, coupled with the innovative approaches demonstrated by participating teams, has illuminated key strategies for addressing the challenges in automated cancer staging. Success appears to stem from a multi-faceted approach: leveraging both representation and generative pre-trained models, implementing sophisticated prompt engineering techniques, optimizing training methodologies, and incorporating domain expertise through physician-guided correction mechanisms. This comprehensive framework has proven instrumental in achieving enhanced performance in automated TNM classification.

3 METHODS

This section describes our approach to address the challenges from the Lung Cancer TNM Staging main task and the Multi-label Sentence Classification subtask.

An analysis of training and validation data distribution highlights class imbalances across TNM categories. In T classification, T2b and T4 are the most common labels (Train: 20, 31; Val: 9, 18), whereas T1mi and T1a are underrepresented. In N classification, N0 (Train: 41, Val: 26) and N2 (Train: 45, Val: 20) dominate, while N3 cases are sparse. The M classification distribution is also imbalanced, with M0 cases (Train: 74, Val: 27) far outnumbering M1 subcategories, particularly M1a (Train: 0, Val: 9) and M1b (Train: 14, Val: 0). This uneven distribution

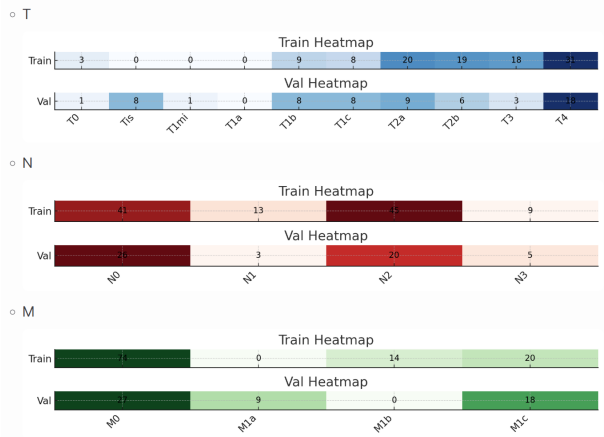

Figure 1: Heatmap Representation of TNM Classification Distribution in Training and Validation Data

[2] https://github.com/nlptmu/NTCIR-18-RadNLP

presents challenges in generalization, particularly for underrepresented categories, as illustrated in Figure 1.

3.1 System I

System I for TNM Staging Classification leverages LLM-generated inference rationales, Few-Shot Learning, and a multi-step voting mechanism to enhance alignment with expert clinical reasoning. We consulted a thoracic oncology specialist with over 10 years of experience and a pulmonologist with nearly 30 years of clinical practice from our team to review the training dataset. They observed that the annotation logic was generally more conservative than real-world clinical decision-making, leading to discrepancies where assigned TNM categories did not fully align with expert interpretations. Let's discuss following case for reference:

The radiology report *no. 1679413* within training dataset, differences in classification likely stemmed from differing objectives. The annotator took a more conservative approach, whereas the clinician was more proactive and cautionary. Regarding **T classification**, the clinician considered *"Possible invasion into the left pulmonary artery, as far as can be assessed with CT to a limited extent"* as sufficient evidence of vascular invasion, classifying the tumor as *T4*. In contrast, the annotator viewed this invasion as uncertain due to *"limited assessment with CT"* indicating incomplete visualization, and instead assigned *T2b*, relying on tumor size (*"A 47 mm irregular mass in the left upper lobe"*). For **M classification**, the clinician interpreted *"A small nodule is present in the right middle lobe, possibly inflammatory or metastasis"* as metastatic, leading to *M1a*. Meanwhile, the annotator deemed the nodule indeterminate, emphasizing *"possibly inflammatory or metastasis"* and maintained *M0*, pending follow-up (*"re-evaluation recommended in follow-up"*). These observations highlight the subjectivity inherent in TNM staging, particularly in ambiguous cases, underscoring the need for an explicit reasoning process to improve model interpretability and consistency.

3.1.1 System I for Main Task (System-I-MT-En and System-I-MT-Ja)

The GPT-4o [3] was employed to generate structured inference rationales for each annotated training case. Given the pathology report, TNM label, and staging guidelines, the model inferred the underlying annotation logic. These rationales were stored as contextual references for downstream classification tasks, capturing nuanced decision-making processes not explicitly encoded in the original labels. In the validation and testing phase, *Few-shot Learning* was used to align the model with annotation logic. Each test case incorporated randomly selected training samples—including pathology reports, TNM labels, and reasoning—within the prompt, grounding GPT-4o's predictions in real-world annotation patterns and reducing inconsistencies. To further enhance reliability, a *hard-voting mechanism* was implemented. The model performed multiple independent inference runs, each using different randomly sampled Few-shot examples. The predicted *T*, *N*, and *M* categories from these runs were aggregated via majority voting, mitigating stochastic variability and improving classification robustness, as illustrated in Figure 2.

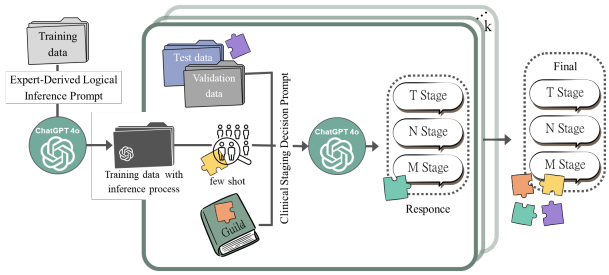

Figure 2: System I Architecture

3.2 System II

3.2.1 System II for Main Task (System-II-MT-En and System-II-MT-Ja)

The system II for main-task implements a robust multi-model approach for TNM staging classification using the DSPy [4] framework, leveraging both GPT-4o and Gemini-2 models in an alternating pattern. The core architecture consists of a DetermineStagingConclude class that simulates a panel of medical experts analyzing radiology reports. For each report, the system generates multiple independent assessments by alternating between the two language models. In the end, the system will collect all independent assessments and make final decision by GPT-4o. These models operate under carefully crafted prompts that embody the role of a skilled radiologist and oncologist, following the IASLC 8th edition lung cancer staging guidelines. The system processes each report through five iterations, collecting diverse opinions which are then synthesized into a final staging decision. To ensure consistency and reproducibility, the implementation incorporates comprehensive error handling, deterministic behavior through seed setting, and standardized data preprocessing. The classification pipeline is optimized using MIPROv2 [9], which fine-tunes the model's performance through bootstrapped demonstrations and labeled examples, ultimately producing TNM classifications with accompanying clinical reasoning.

3.2.2 System II for Sub Task (System-II-ST-En and System-II-ST-Ja)

The system II for sub-task implements a sophisticated medical text classification framework utilizing DSPy for model configuration, featuring a specialized prompt design system for Multi-label Sentence Classification. The implementation leverages the Meta Llama-3.3-70B large language model as its core processing engine, demonstrating the application of state-of-the-art language models in medical text analysis. At its core, the system processes eight critical classification categories: omittable, measure, extension, atelectasis, satellite, lymphadenopathy, pleural, and distant. The implementation leverages the MIPROv2 optimizer with customizable parameters for bootstrapped and labeled demonstrations, supporting up to 300 demonstrations for each

category. Performance evaluation is conducted through a robust validation function utilizing F2 score metrics, which prioritizes recall over precision in medical classifications, ensuring high sensitivity in detecting critical medical conditions.

3.3 System III

The system III for the maintask implements a hierarchical pipeline with integrated subtask modules. In this design, the preprocessing stage leverages a BERT-based model for tokenization and semantic enrichment, standardizing raw clinical text according to international TNM guidelines. This stage includes unit conversion, TNM definition standardization, and implicit semantic annotation. Subsequently, an exhaustive data augmentation strategy addresses label imbalances, and subtask-specific text marking directs model attention to key clinical features. Finally, the classification module employs a fine-tuned Bio_ClinicalBERT_freeze_6 model, augmented by an ensemble of Complement Naive Bayes, XGBoost, and SVM for robust T classification. This hierarchical design try to reduces lexical variability and captures essential clinical semantics but also enhances overall performance in TNM staging.

3.3.1 Preprocessing

First, all measurements are converted from centimeters to millimeters to ensure consistency, and variants of *"in situ" (e.g., In situ, insitu, in-situ)* are normalized to the canonical form: *"Carcinoma in situ: abnormal cells are confined to the location where they first formed and have not spread. Irrespective of size, this condition can only be diagnosed after the tumor is resected."* This normalization minimizes lexical variability and ensures consistent interpretation of critical clinical terms. Next, in accordance with the 8th edition TNM guidelines [2], we enforce standardized definitions for Tx, Nx, and Mx by applying a series of regular expression–based replacements—for instance, replacing *"Tis"* with *"Tis, Carcinoma in situ,"* and similarly for another TNM category. This step directly enhances the key TNM information embedded in the text.

To further enrich clinical context, we inject implicit semantic annotations into the text. Based on the AJCC Node category for Regional Lymph Nodes in lung cancer (covering both intrathoracic and extrathoracic nodes), key lymph node terms (e.g., *"pulmonary," "peribronchial," "intrapulmonic," "hilar"*) are augmented with *"(regional lymph node involvement)."* In addition, following the NTCIR 18 subtask definitions, phrases such as *"intrapulmonary metastasis"* (or *"lymphangiomatosis carcinomatosa"*) are annotated with "(satellite lesion)," while additional annotations for pleural and atelectasis-related features are applied as specified. Finally, to disambiguate the term *"extension"*, which clinically indicates that the primary lesion extends beyond the lung, we explicitly augment phrases such as *"extends beyond lung," "involvement of mediastinum,"* and *"infiltration into chest wall"* with *"(extra-pulmonary extension)."* These semantic annotations ensure that the model captures key clinical features for TNM staging.

3.3.2 Data Augmentation

Our analysis revealed significant imbalances in TNM label distributions—certain labels present in the development set are underrepresented or absent in training. To mitigate this, we generate 1,596 additional training samples by systematically permuting t, n, and m definitions using international TNM standards combined with measure-based numeric enumeration.

3.3.3 Subtask-Specific Text Marking

To guide the model's attention to critical clinical features during training, we incorporate explicit task-specific markers into the text:

- T: Sentences identified via the "Measure" metric are enclosed with {T-classification/} and {/T-classification}.
- N: Sentences containing indicators of satellite lesions or lymphadenopathy are wrapped with {N-classification/} and {/N-classification}.
- M: Sentences with descriptors for extension or distant metastasis are marked with {M-classification/} and {/M-classification}.

Highlighting these key sections helps isolate relevant clinical signals for each TNM component, improving model focus during training.

Our multi-label classification framework leverages a fine-tuned Bio_ClinicalBERT_freeze_6 model, with six critical layers frozen to preserve pretrained clinical semantics. The model is tasked with simultaneously predicting T, N, and M labels. Additionally, to enhance T classification—especially in the presence of label imbalance—we apply an ensemble strategy combining Complement Naive Bayes, XGBoost, and SVM. Each classifier generates an independent T prediction, which are then aggregated via a voting mechanism to yield a final decision. Freezing key layers retains essential clinical knowledge, and the ensemble approach mitigates individual model biases, particularly under imbalanced conditions.

3.4 System IV

The system IV for the English track of Multi-label Sentence Classification Subtask extends the standard fine-tuning process of a pre-trained RadBERT-RoBERTa, which is radiology-specific pre-trained language model, by adding self multi-attention heads after hidden state layer with attention heads which corresponding to number of labels. This approach allows model to attend information from different subspaces at different positions. Additionally, custom attention layer is introduced to help the model to focus on key token that contribute to determining which token belongs to specific category. Since each category has a heavily imbalanced number of instances, sentence augmentation is applied to address this challenge, GPT-4o-mini [8] is utilized to generate synthetic sentences by prompting the model to create additional examples for the minority class. These augmented sentences can enhance model's ability to better understand the context within sentences, improve classification accuracy for underrepresented categories, and reduce the risk of bias toward majority classes.

3.5 System V

System V is developed using the LLM with Zero-Shot approach, leveraging the EnsReas method for TNM staging classification. It follows the Zero-Shot Chain-of-Thought with Self-Consistency (ZS-CoT-SC) framework, which involves two rounds of LLM inference combined with a hard voting mechanism to determine the most appropriate TNM stage for each report.

In the classification process, GPT-4o first references the IASLC 8th edition lung cancer staging guidelines to generate five sets of staging assessments, each accompanied by a concise clinical reasoning explanation. Subsequently, another teacher GPT-4o model reviews the original report along with the five sets of staging results and justifications, ultimately selecting the final TNM stage through hard voting. These models operate based on carefully crafted prompts, enabling them to simulate experienced radiologists and oncologists, while the teacher model further refines classification accuracy by evaluating the outputs of the previous model.

For self-consistency, we set the model temperature to 0.8, reducing randomness in the output and ensuring greater consistency across different inferences. This adjustment is particularly beneficial for TNM staging, a task requiring high stability, ensuring more reliable and consistent results.

4 EXPERIMENTS

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

The NTCIT-18 RadNLP Task was implemented through the Hugging Face Competitions platform, which facilitated automated evaluation of participant submissions through an integrated evaluation script. The competition utilized a dual-leaderboard system: a public leaderboard that displayed validation data performance metrics, and a private leaderboard that tracked test data results. The performance metrics of our various systems in the main task and subtask are presented in Table 1 and Table 2, respectively. In this section, we detail the systematic configuration and parameter optimization implemented throughout the competition process.

Table 3 shows the original dataset from the NTCIT-18 RadNLP task. Each system will mix the training and validation datasets according to development needs to achieve optimal performance.

Table 3: NTCIT-18 RadNLP Task dataset quantities

Dataset	English Track	Japanese Track
Main Task (Document Level)		
Train	108	108
Validation	54	54
Test	81	216
Sub Task (Sentence Level)		
Train	918	1,020
Validation	415	451
Test	568	1,976

4.1 System I

4.1.1 System-I-MT-En

The data processing pipeline utilizes 108 training samples, 54 validation entries, and 81 test records for multi-label document classification, determining T, N, and M categories from English track radiology reports. To ensure synchronization with expert annotation logic, no additional data preprocessing was applied, preserving the raw input structure.

To align the model's decision-making with expert reasoning, training data is used to generate structured inference rationales, which are then incorporated as Few-shot Learning references for TNM classification. Specifically, GPT-4o generates reasoning explanations for each annotated training case using *temperature = 1, max_tokens = 200, and top-p = 1,* allowing for maximum diversity in reasoning while preserving expert annotation logic. These generated rationales are then provided as in-context examples during inference for validation and test cases, ensuring that the model leverages past annotation patterns for more consistent decision-making.

For TNM classification, the model is configured with *temperature = 0.2, top-p = 0.8, and max_tokens = 50*, ensuring controlled variability while maintaining structured output. The *max_tokens = 50* constraint is explicitly set to enforce output format consistency, preventing excessive text generation that may deviate from the expected TNM classification structure. Predictions for T, N, and M categories are obtained through three independent inference iterations, followed by hard voting aggregation, which stabilizes predictions and enhances consistency across different input cases. The introduction of a 7-shot Few-shot Learning setup with 3-vote aggregation was specifically designed to counteract TNM class imbalances by providing more examples for rare categories and reducing the impact of single-instance variability.

The method achieved a joint accuracy of 0.91 on validation data (T: 0.94, N: 1.0, M: 0.96) and 0.65 on test data (T: 0.70, N: 0.9136, M: 0.89), demonstrating strong alignment with expert annotations. These results confirm that model-generated inference rationales, Few-shot Learning, and multi-step voting significantly enhance TNM classification consistency, highlighting the approach's potential for clinical decision-support applications.

4.1.2 System-I-MT-Ja

Like the System-I-MT-En approach, the difference is that the few-shot text is in Japanese, and the inference content is the same as the English track. The best performing combination was 7-shot Few-shot Learning setup with 5-vote.

4.2 System II

The framework utilizes a CUDA-optimized environment with deterministic behavior, employing a fixed random seed across all components for reproducibility.

4.2.1 System-II-MT-En

The experiment was designed to classify lung cancer staging from radiology reports using a multi-model approach. For English track dataset preparation, the training data further divided using a 50% split ratio for validation purposes in optimation step. This is because we want to completely hide the original validation dataset from the model to verify the prediction performance. For model implementation, two advanced language models were employed: OpenAI GPT-4o and Google Gemini-2. The temperature parameter was dynamically set between 0.001 and 0.85 using uniform random distribution. The optimization process utilized the MIPROv2 optimizer with 50 maximum bootstrapped demos and 50 maximum labeled demos.

The best submission predicted by this system ranked second in the English track. Although our program is designed to comprehensively evaluate the classification results and reasoning of 3 GPT-4o and 2 Gemini2 to make the final classification decision, due to cost considerations, we only implemented the ensemble classification results of 1 GPT-4o and 1 Gemini2 in this system.

4.2.2 System-II-MT-Ja

Similar to the approach of System-II-MT-En, the difference lies in that the Few-Shot is using Japanese data, and the Prompt is written in English. One interns' decisions and reasons are inferred, then reviewed by a senior for the final decision, all using GPT-4o.

Our best performance in the Japanese track is seventh place, which is the result of a hard vote using System-II-MT-Ja and System-I-MT-Ja.

4.2.3 System-II-ST-En and System-II-ST-Ja

The data processing pipeline manages a comprehensive dataset comprising 918 training samples, 415 validation entries, and 568 test records. The system architecture leverages the DSPy framework integrated with Meta Llama-3.3-70B as the core language model, utilizing a dynamic temperature range of 0.7 to 0.75 for controlled stochastic prediction. The optimization strategy incorporates the MIPROv2 optimizer with extensive demonstration capabilities, supporting up to 300 bootstrapped and 300 labeled demonstrations. When optimizing the model with 300 bootstrapped and 300 labeled demonstrations is time-consuming, so in the final upload experiment, we used the System-II-ST-En model to predict the System-II-ST-Ja test set, resulting in low performance on the private leaderboard.

4.2.4 Discussion

We consulted with our team's thoracic oncology experts and pulmonologists to review the training dataset and validate the answers and predictions. After that, the thoracic oncology experts and pulmonologists will provide us with professional advice to manually modify the prompts. This is an operation for System I, but we are trying to use the DSPY framework and MIPROv2 mechanism to automatically optimize the prompts and instructions in System II. From the results, although System I still ranks first, the automatic optimization is worthwhile.

The use of LLM and prompt engineering is limited by the context window length, so the main task usually only uses 50 reports. Exceeding this number will reduce prediction performance.

The comparative analysis between System-I and System-II revealed key insights into LLM-based classification effectiveness. A dynamic approach using fewer, carefully selected reference cases proved more effective in guiding LLM classifications compared to using a larger, fixed set of examples. Additionally, the study demonstrated that the initial classification quality and quantity generated by the system's first-phase assessment substantially impacts the final staging decisions. This finding suggests that optimizing the preliminary classification phase is crucial for overall system performance.

4.3 System III

4.3.1 System-III-MT-En and System-III-ST-En

The experiment for system III follows the model architecture described in the methods section. We evaluate our approach on a clinical dataset of radiology reports annotated for TNM staging. To ensure comprehensive label coverage, the training set is augmented with the 1,596 synthetic samples generated by our exhaustive augmentation strategy. All reports are processed through our multi-step preprocessing pipeline, with token lengths capped at 512.

The entire preprocessing pipeline is implemented using regular expressions. Data augmentation is performed by modifying the TNM definition text to generate diverse instances. For example, *"9 mm tumor surrounded by lung or visceral pleura, Ipsilateral peribronchial and/or hilar nodes detected, Distant metastasis found in contralateral lung."* is transformed into a new instance with labels T1a, N1, and M1a by substituting the numeral "9" with other clinically relevant integers within defined ranges. In this process, numeric values are converted into corresponding textual representations that yield different TNM labels. This strategy ensures every possible TNM label is represented in the training set, reducing bias and improving model generalization.

Finally, System-III-MT-En leverages a fine-tuned Bio_ClinicalBERT_freeze_6 model for robust TNM classification, further augmented by an ensemble of machine learning classifiers (Complement Naive Bayes, XGBoost, and SVM) to address label imbalances, particularly for T classification. Crucially, our hierarchical pipeline integrates a dedicated subtask text marking module—implemented as System-III-ST-En—which uses the same Bio_ClinicalBERT_freeze_6 model to embed task-specific markers into the training text. These markers explicitly delineate critical clinical features for T, N, and M classifications, effectively guiding the model's attention during fine-tuning. The refined outputs from the subtask marking module are subsequently incorporated into System-III-MT-En, yielding more precise and consistent final staging decisions.

4.3.2 Discussion

The performance of System III modestly demonstrates the potential benefits of our hierarchical pipeline. Specifically, the joint accuracy of the main task module (System-III-MT-En) was 0.5556, while the subtask module (System-III-ST-En) achieved a joint accuracy of 0.9155. These findings suggest that integrating task-specific text marking can notably enhance model attention and yield more consistent TNM classification. Moreover, our results indicate that data augmentation—even when applied to non-generative models—can contribute to improved performance by mitigating label imbalance and reducing lexical variability.

4.4 System IV

4.4.1 System-IV-ST-En

The experiment for system IV follows the model architechture described in method section. The process begins with sentence tokenization, followed by handling class imbalance, model training, and evaluation, using RadBERT-RoBERTa, truncated/padded to 128 tokens for consistency. Class weights were calculated based on inverse class frequency to prioritize minority classes. The model is trained for 10 epochs using AdamW optimizer (learning rate=2e-5) with weighted Binary Cross-Entropy (BCE). This model is evaluated based on micro F2-score, which prioritizes recall. These evaluation metrics follow the organizer's official guidelines. To address class imbalance, GPT-4o-mini model is applied to generate synthetic sentences for minority class. For each original sentence, one of synthetic sentence is generated, effectively doubling dataset size. The model is prompted with an original radiology sentence along with its corresponding labels. The model assumes the role of a radiologist. Temperature (1.2) and top-p (0.9) are used to enhance diversity while maintaining coherence, improving classification performance and dataset balance.

4.4.2 Discussion

The results demonstrate potential customized BERT model by adding muti attention head and multi attention layer to let model better understand context within sentence make correct prediction.

The effectiveness of this approach is validated by achieving an overall micro F2-score of 91.50

To address class imbalance within dataset, GPT-4o-mini was used to generate synthetic sentences from underrepresented class. Incoorperating both original and synthetic sentences enhanced classification performance, achieving a micro F_2-score of 91.50 overall and achieving 69 of micro F_2-score for inclusion category, which is highest score, indicating that systhetic sentence help model distinguish lung cancer-related sentences more effectively. However, model struggled with the Extension, Atelectasis, and Satellite classes, which contain complex contextual patterns. these classes contain complex contextual patterns. This suggest that additional fine-tuning or domain-specific augmentation may be needed to further improve classification in these categories.

4.5 System V

In TNM classification, the model is configured with a temperature of 0.7 and max_tokens set to 250 to ensure that it can produce complete results. We have set a maximum output length of 50 and obtain predictions for the T, N, and M categories through two phases and five independent inferences, followed by hard voting to determine the best answer. This method stabilizes the predictions and enhances consistency across different input cases. In the model's prompt, we also adjusted the second-phase teacher model. Since in the test data, our system model often misclassifies T1b as Tis, we tried to focus the teacher model more on the Tis category; and in the M category, the first-phase model often considers the situation more severe than it is. We also adjusted the prompt for the categories with classification errors.

Test data results show that using EnRease, compared to using hard voting alone, improves performance by about 4%. If the prompt for the M category is adjusted in the second phase, although the overall performance remains the same, the accuracy for M increases from 0.87 to 0.92, but there is a slight decline in performance for T and N categories; when adjusting the prompt for the T category, the overall score increases by about 4%, but considering the potential for overfitting, we ultimately decided to use the original model configuration.

5 CONCLUSIONS

The NTCIR-18 RadNLP 2024 task marks a significant advancement in the automation of lung cancer staging from radiology reports. The TMUNLPG3 team achieved remarkable success in this challenge, particularly in the English track, where our System-I-MT-En secured first place with impressive metrics: 65.43% joint fine accuracy and 69.14% joint coarse accuracy. The system demonstrated strong individual performance in T (70.37%), N (91.36%), and M (88.89%) classifications. Additionally, in the multi-label sentence classification subtask, System-II-ST-En achieved second place with a notable overall micro F2.0 score of 93.36%. 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. A key advantage of our approach is the integration of expert medical knowledge, consulting with experienced thoracic oncologists and pulmonologists to validate and refine the system. Our efforts validate the potential of artificial intelligence in medical document analysis, establishing a framework for future clinical decision support systems. This work highlights the achievements and areas for improvement in automated cancer staging, contributing valuable insights to the advancement of medical natural language processing.

ACKNOWLEDGEMENTS

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.

REFERENCES

[1] Brierley, J. 2006. The evolving TNM cancer staging system: an essential component of cancer care. Canadian Medical Association Journal. 174, 2 (Jan. 2006), 155–156. DOI:https://doi.org/10.1503/cmaj.045113.

[2] Detterbeck, F.C., Boffa, D.J., Kim, A.W. and Tanoue, L.T. 2017. The eighth edition lung cancer stage classification. Chest. 151, 1 (2017), 193–203.

[3] Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., and others 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276. (2024).

[4] Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., A, S.V., Haq, S., Sharma, A., Joshi, T.T., Moazam, H., Miller, H., Zaharia, M. and Potts, C. 2024. DSPy: Compiling Declarative Language Model Calls into State-of-the-Art Pipelines. The Twelfth International Conference on Learning Representations (2024).

[5] Koji Fujimoto, Mizuho Nishio, Chikako Tanaka, Morteza Rohanian, Farhad Nooralahzadeh, Michael Krauthammer and Fabio Rinaldi 2023. Classification of cancer TNM stage from Japanese radiology report using on-premise LLM at NTCIR-17 MedNLP-SC RR-TNM subtask. NII Institutional Repository.

[6] Mizuho Nishio, Hidetoshi Matsuo, Takaaki Matsunaga, Koji Fujimoto, Morteza Rohanian, Farhad Nooralahzadeh, Fabio Rinaldi and Michael Krauthammer 2023. Zero-shot classification of TNM staging for Japanese radiology report using ChatGPT at RR-TNM subtask of NTCIR-17 MedNLP-SC. NII Institutional Repository.

[7] Nakamura, Y., Fujimoto, K. and Kluckert, J. NTCIR-18 RadNLP 2024 Overview: Dataset and Solutions for Automated Lung Cancer Staging. NII Institutional Repository.

[8] OpenAI, G. 2024. 4o mini: Advancing cost-efficient intelligence, 2024. URL: https://openai. com/index/gpt-4o-mini-advancing-cost-efficient-intelligence. (2024).

[9] Opsahl-Ong, K., Ryan, M.J., Purtell, J., Broman, D., Potts, C., Zaharia, M. and Khattab, O. 2024. Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs. arXiv.

[10] Rosen, R.D. and Sapra, A. 2023. TNM classification. StatPearls [Internet]. StatPearls Publishing.

[11] Takuya Fukushima, Yuka Otsuki, Shuntaro Yada, Shoko Wakamiya and Eiji Aramaki 2023. NAISTSOCRR at the NTCIR-17 MedNLP-SC Radiology Report Subtask. NII Institutional Repository.

[12] Yuta Nakamura, Shouhei Hanaoka, Shuntaro Yada, Shoko Wakamiya and Eiji Aramaki 2023. NTCIR-17 MedNLP-SC Radiology Report Subtask Overview: Dataset and Solutions for Automated Lung Cancer Staging. NII Institutional Repository.