

ORAD at NTCIR-18 RadNLP 2024 Shared Task

Keisuke Hidaka
Department of Radiology
Osaka Red Cross Hospital
Osaka, Japan
hida1211@gmail.com

ABSTRACT

Here, we report our approach to the NTCIR-18 RadNLP2024 Shared Task (Japanese Track, Main Task). In this study, we developed a system to determine the TNM classification from lung cancer using Japanese radiology reports. Specifically, we provided Google DeepMind’s Gemini 2.0 Flash Experimental (gemini-2.0-flash-exp) with a prompt that combines Chain-of-Thought (CoT) and Many-Shot In-Context Learning (ICL), enabling automatic prediction of the T, N, and M factors for each case. Besides accuracy, interpretability is crucial in the medical domain; thus, having the model output the rationale for its TNM classification ensures a degree of transparency. Moreover, by including numerous examples of CoT-based reasoning—written by a radiologist with 5 years of dedicated experience in diagnostic radiology—to explain how the TNM classification is derived, we achieved improved inference accuracy.

Furthermore, to address privacy concerns and the need for local inference without network connectivity in clinical settings, we performed Supervised Fine-Tuning (SFT) using Gemma2-9b-it, a comparatively lightweight open-source model. By providing the model with CoT-based reasoning steps leading to TNM classification as training data, we observed improved inference accuracy.

These findings demonstrate that additional data and prompt strategies to support large language model (LLM)-based inference can be highly effective in automating TNM classification while also indicating the feasibility of realizing interpretability in LLM-based medical applications.

KEYWORDS

Medical Natural Language Processing, Radiology Reports, Lung Cancer, Cancer Staging, TNM Staging, Gemini, Gemma2, In-Context Learning, Many-Shot In-Context Learning, Chain-of-Thought, Supervised Fine-Tuning, Reasoning dataset

TEAM NAME

ORAD

SUBTASKS

RadNLP Main task (Japanese track)

1 Introduction

Because radiology reports are unstructured data written in natural language, direct analysis and automated processing are challenging. In recent years, rapid advancements in large language models (LLMs) have made it feasible to automatically analyze unstructured reports. Consequently, there is growing anticipation for systems that can automatically extract crucial information from clinical sources, such as medical records, radiology reports, and pathology reports, and for their use in clinical practice and research.

1.1 NTCIR-18 RadNLP 2024 Shared Task

In the Japanese Track, Main Task of the NTCIR-18 RadNLP 2024 Shared Task, participants were provided with a dataset of radiology reports related to lung cancer. The challenge was to estimate the TNM classification (tumor size and extent (T), metastasis to nearby lymph nodes (N), and presence or absence of distant metastases (M)) from each report [1]. The RadNLP2024 dataset was constructed based on lung cancer cases obtained from the open-access radiology reference site Radiopaedia (<https://radiopaedia.org/>) with all personal information completely removed.

1.2 TNM staging system and radiology reports

Although the incidence and mortality rates of lung cancer are trending downward—thanks to factors such as declining smoking rates and advancements in immunotherapy—it remains the most common cancer worldwide and continues to have the highest mortality rate [2, 3]. Imaging-based diagnosis is indispensable in lung cancer care, where the stage is typically classified based on TNM, guiding prognostic predictions and treatment planning. The lung cancer staging criteria used in RadNLP 2024 follow the “General Rules for Clinical and Pathological Record of Lung Cancer, 8th edition” by the Japan Lung Cancer Society [4, 5], which is essentially equivalent to the 8th edition of the TNM classification issued by the Union for International Cancer Control (UICC) [6]. The inclusion of the TNM classification in radiology reports is recommended because it can accelerate patient care and help avoid unnecessary diagnostic interventions [7].

However, there is currently no absolute criterion for determining the presence or absence of pleural invasion in CT images. Reports indicate that sensitivity ranges from 46% to 87%, and specificity ranges from 59% to 91%, depending on the radiologist’s experience and knowledge [8]. As a result, it is unclear whether radiologists must consistently perform a detailed TNM classification for lung cancer in everyday practice [9]. Moreover,

unlike in fields such as pathology—where a standardized format for cancer reporting is well established—many radiologists prefer free-text reports. A study in the UK showed that for lung cancer cases, there was substantial variability (25%–100%) among radiologists in how comprehensively their free-text reports included the information required for clinical treatment decisions [10].

Given this lack of standardization and the variability in the completeness of information in such reports, an automated system for extracting staging information, such as TNM classification from imaging reports, could facilitate smoother communication between radiologists and clinicians, enabling faster and more accurate development and implementation of appropriate treatment plans [11].

1.3 Related works

Several prior studies have used LLMs to determine TNM classifications or cancer staging based on medical documents.

Before this shared task, the NTCIR-16 Real-MedNLP [12] and NTCIR-17 MedNLP-SC Radiology Report TNM Classification Subtask (RR-TNM) [13] were conducted. Notably, the NTCIR-17 RR-TNM also aimed to determine TNM classifications from radiology reports, similar to the present study. The labels were aggregated for the T category (T0, T1, T2, T3, or T4), N category (N0, N1, N2, or N3), and M category (M0 or M1). A solution using GPT-3.5 turbo exceeded 80% accuracy for N and M classification but achieved less than 40% accuracy for T classification [14]. Other solutions have also faced the challenge of specifically low accuracy for T classification compared to N and M. One reason for this is that the T classification criteria require consideration of both tumor size and anatomical features, thus demanding more complex reasoning than N or M classification [13]. In the current task, the TNM classification system includes more detailed suffixes (e.g., “T1a” and “T1b”), increasing the task’s difficulty while making it more practical.

Lee et al. compared the accuracy of ChatGPT with that of physicians in determining the clinical stage from radiology reports of CT and FDG PET/CT for non-small cell lung cancer [11]. GPT-4o achieved an overall staging accuracy of 74.1%, which was significantly higher than those of GPT-4, GPT-3.5, and one radiology resident. However, its accuracy was significantly lower than that of a fellowship-trained radiologist and statistically equivalent to that of a radiology fellow and another resident.

In a task to determine TN classification from breast cancer pathology reports, Chang et al. investigated an ensemble approach called EnsReas [15]. They used an LLM derived from Llama2-70B—fine-tuned with medical knowledge datasets (Med42-70B)—and applied Chain-of-Thought (CoT) prompting to generate multiple inferences and predictions. If those predictions were consistent, the result was used as-is; otherwise, the predictions were grouped, and the LLM was asked to re-predict. They reported that this method achieved an F1-score of 0.800 for T classification and 0.838 for N classification, outperforming both simple CoT prompting and majority-vote strategies for self-consistency.

Kim et al. conducted CoT fine-tuning on the Llama, Mistral, and Orca2 (a fine-tuned version of Llama-2) datasets to generate rationales for TN classification and evaluated their performance [16]. They used exact match ratio (EMR) and semantic match ratio (SMR) as metrics to measure the classification accuracy and contextual consistency of the generated rationale, respectively. The Orca2_13b model achieved the best performance, with an EMR of 0.934 and SMR of 0.864.

2 Methods

2.1 Dataset

Details on how the RadNLP2024 dataset was constructed can be found in the overview paper [1]. A total of 108 documents (12 cases) created by nine radiologists, along with correct TNM classification labels, were provided as the training set, and 54 documents (6 cases) were provided as the validation set. Later, a test set of 216 documents (24 cases), for which the correct labels were not disclosed, was released. The dataset was divided into cases to prevent data leakage.

2.2 Reasoning data

To teach the model how to apply the TNM classification definitions to actual reports, one board-certified radiologist (with 5 years of full-time experience) created detailed reasoning processes (Reasoning data) illustrating how to arrive at the TNM classification from a subset of the training reports. These reasoning data were delineated in the format `<T 分類>...</T 分類>` `> <N 分類>...</N 分類>` `<M 分類>...</M 分類>` `<answer>...</answer>`, allowing the model to learn how each category is reasoned. Next, these reasoning data were provided as Few-Shot examples for the remaining training data to GPT-4o (accessed via a browser) [17] to generate new reasoning processes. The same radiologist manually revised the generated reasoning. Ultimately, we created corresponding reasoning-process data for all 108 documents in the training set (Fig. 1).

2.3 Overview of the system

In this study, we developed a simple system in which each radiology report (on a case-by-case basis) is provided to an LLM to automatically determine the TNM classification for lung cancer. Two main approaches were explored.

2.3.1 Inference using Gemini 2.0 Flash Experimental.

We used gemini-2.0-flash-exp [18], a Gemini 2.0 multimodal model family developed by Google DeepMind, via API from December 2024 to February 2025. The Gemini series can accommodate more context input than competing models and was relatively affordable as of 2024. The Gemini 2.0 series was unveiled in December 2024, with gemini-2.0-flash-exp being the first model made available. It offers fast response times and outperforms previous models in benchmark tests. The following prompt strategies were used for inference with the validation and test data of gemini-2.0-flash-exp:

ORAD at NTCIR-18 RadNLP 2024 Shared Task

NTCIR'18, June, 2025, Tokyo, Japan

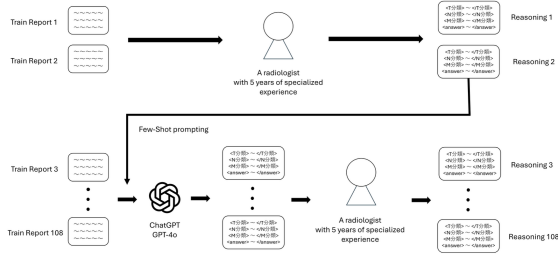


Figure 1. Overview of the reasoning process.

Definition of TNM classification

We included the definition and cautionary notes of the TNM classification 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).

In-Context Learning (ICL)

It is known that LLMs can learn patterns and make appropriate predictions when given few input-output examples within a prompt [19]. Traditionally, Few-Shot prompting has been the most common method, typically involving only a few samples. However, with the recent extensive expansion of context windows, it has become possible to provide several hundred to thousands of samples to the model, drawing attention to Many-Shot ICL [20]. In Many-Shot ICL, studies have shown that increasing the number of examples can improve task performance and inference stability.

Chain-of-Thought (CoT)

CoT is a method for crafting input prompts so that the model outputs its intermediate reasoning steps, which has been shown to improve the accuracy of the results [21]. In this study, instead of using a common instruction like “think step by step” or generating a thought process with <think></think> tags, we employed the same approach as in our creation of the reasoning data. Specifically, we segmented the answer into the format <T 分類>...</T 分類> <N 分類>...</N 分類> <M 分類>...</M 分類> <answer>...</answer>, allowing the model to systematically consider T, N, and M classifications in turn and then produce a final output. Because we instructed the model to output the rationale first, we treated this broadly as a CoT prompt.

By combining these elements, the classification performance of the validation data was compared using the following four methods:

1. **Zero-Shot:** Provides only the target radiology report and TNM classification definition, and the model directly outputs the TNM classification.
2. **Many-Shot (non-reasoning):** In addition to the target report and TNM classification definition, training reports and their correct TNM classifications are included, and the model directly outputs the TNM classification.

3. **CoT prompt + Zero-Shot:** Provides the target report and TNM classification definition, generates a rationale using a CoT prompt, and then the model determines the TNM classification.
4. **CoT prompt + Many-Shot (reasoning):** In addition to the target report and TNM classification definition, the training reports, reasoning processes, and correct TNM classifications are included; a rationale is generated via a CoT prompt; and then the model determines the TNM classification.

2.3.2 *Supervised Fine-Tuning of Gemma2-9b-it.* From the perspective of real-world clinical implementation, local inference within a hospital system may be preferable for privacy reasons.

We evaluated the performance of the Gemma2-9b-it model, a part of the Gemma2 [22] family from Google DeepMind—which is relatively lightweight and open—by performing supervised fine-tuning (SFT) using the training data.

We adopted Quantized Low-Rank Adaptation (QLoRA) [23], a type of parameter-efficient fine-tuning, as our SFT method.

LoRA and QLoRA

LoRA is a technique that maintains fixed weights of a pretrained model while inserting low-rank approximation matrices into certain layers, thereby reducing the number of trainable parameters. QLoRA further optimizes LoRA by combining 4-bit quantization (NF4) and double quantization to reduce memory usage.

unsloth library

We utilized the “unsloth” library [24], which facilitates faster fine-tuning and inference and reduces memory usage for LLMs. We adopted the 4-bit quantized model (unsloth/gemma-2-9b-it-bnb-4bit) of gemma2-9b-it made available by this library and trained it on Google Colab (mainly using an A100 GPU). The training parameters, such as batch size and learning rate, were as follows:

```
per_device_train_batch_size=1,
gradient_accumulation_steps=8, warmup_steps=5,
num_train_epochs=7, learning_rate=2e-4, fp16=not
torch.cuda.is_bf16_supported(),
bf16=torch.cuda.is_bf16_supported(), logging_steps=10,
optim="adamw_8bit", weight_decay=0.01,
lr_scheduler_type="linear", seed=3407
During inference, we set temperature=0.2 and
repetition_penalty=1.2.
```

We then compared classification performance on the validation data under three methods:

1. **Raw model + CoT prompt:** No fine-tuning; at inference time, provide the target report and TNM classification definition, use a CoT prompt to generate the rationale, and finally determine the TNM classification.
2. **SFT (non-reasoning) + Zero-Shot:** Perform SFT using pairs of training data reports and their correct

NTCIR 18 Conference: Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies,

labels. At inference time, the target report and TNM classification definition are provided, and the model directly outputs the TNM classification.

3. **SFT (reasoning) + CoT prompt:** Perform SFT using pairs of training data reports and their reasoning data. At inference time, provide the target report and TNM classification definition, and then use a CoT prompt to generate the rationale and determine the TNM classification.

2.3.3 Prompt design. In this study, a standardized base prompt was used to provide input to the LLM (Fig. 2). After clarifying the role of the model, we gave it a text that included detailed definitions and cautionary notes on the TNM classification. For cases without a CoT prompt, we instructed the model to produce no output other than `<answer>...</answer>`. If a CoT prompt was included, we instructed the model to output rationale in the format `<T 分類>...</T 分類> <N 分類>...</N 分類> <M 分類>...</M 分類> <answer>...</answer>`.

For inference with Gemini 2.0 Flash Experimental, we either placed the target report directly after this base prompt or under the heading “TNM classification and radiology reports, and the thought process leading to classification” and sequentially appended the training data’s reports and their associated reasoning, followed by the target report. The final prompt for each strategy was structured as shown in Fig. 3.

2.3.4 Pre-processing and Post-processing. The radiology reports were input into the model without any pre-processing, in their original form.

For post-processing, we extracted `<answer></answer>` tags from the model output using regular expressions and separated them into T0/Tis/T1mi/T1a/T1b/T1c/T2a/T2b/T3, N0/N1/N2/N3, M0/M1a/M1b/M1c. No post-processing was performed, such as replacing failed extractions with alternative categories.

2.3.5 Evaluation Metrics. The metrics conform to those employed by the organizer:

Joint accuracy (fine): Proportion of radiology reports with accurate predictions for all T, N, and M factors.

T accuracy (fine): Proportion of radiology reports with accurate predictions for the T factor.

N accuracy (fine): Proportion of radiology reports with accurate predictions for the N factor.

M accuracy (fine): Proportion of radiology reports with accurate predictions for the M factor.

Joint accuracy (coarse): Joint accuracy ignores distinctions between Tis/T1mi/T1a/T1b/T1c, T2a/T2b, and M1a/M1b/M1c.

T accuracy (coarse) T accuracy ignores the distinctions between Tis/T1mi/T1a/T1b/T1c and T2a/T2b.

N accuracy (coarse) Identical to N accuracy (fine).

M accuracy (coarse) M accuracy ignores the distinctions among M1a/M1b/M1c.

3 Results

3.1 Validation Results with Gemini 2.0 Flash Experimental

Tables 1 and 2 show the accuracy of the inference results for the validation and test data, respectively.

	Accuracy	Zero-Shot	Many-Shot (non-reasoning)	CoT prompt + Zero-Shot	CoT prompt + Many-Shot (reasoning)
fine	Joint	0.685	0.667	0.722	0.963
	T	0.833	0.796	0.833	0.981
	N	0.963	0.889	0.944	1.000
	M	0.833	0.981	0.944	0.981
coarse	Joint	0.815	0.815	0.833	0.981
	T	0.944	0.926	0.889	1.000
	N	0.963	0.889	0.944	1.000
	M	0.889	1.000	1.000	0.981

Table 1. Results on validation data using Gemini 2.0 Flash Experimental.

	Accuracy	CoT prompt + Many-Shot (reasoning)
fine	Joint	0.741
	T	0.833
	N	0.958
	M	0.907
coarse	Joint	0.806
	T	0.861
	N	0.958
	M	0.944

Table 2. Results on test data using Gemini 2.0 Flash Experimental.

While Gemini 2.0 Flash Experimental demonstrated high Zero-Shot performance for N classification (N accuracy (fine): 0.963), its performance for T and M classification was relatively lower (M accuracy (fine): 0.833). Under Many-Shot ICL (Many-Shot (non-reasoning)), where we directly included the training data in the prompt, the accuracy for M classification improved (M accuracy (fine): 0.833 \rightarrow 0.981), but T and N classification accuracy showed a slight decline (T accuracy (fine): 0.833 \rightarrow 0.796, N accuracy (fine): 0.963 \rightarrow 0.889). Introducing CoT prompts led to a modest accuracy increase, even with Zero-Shot (Joint accuracy (fine): 0.685 \rightarrow 0.722). Moreover, providing multiple inference examples through Many-Shot (reasoning) resulted in near-perfect classification accuracy, including T classification (Joint accuracy (fine): 0.963).

On the test data, the accuracy decreased (Joint accuracy (fine): 0.741), with a particular drop in T classification (T accuracy (fine): 0.833). Nevertheless, this was the second-highest joint accuracy among the participants in the Japanese main task.

Below, the misclassifications that occurred in the validation data when using the CoT prompt/Many-Shot (reasoning) method are highlighted.

In the validation data, there was one case where T classification was incorrect.

ORAD at NTCIR-18 RadNLP 2024 Shared Task

NTCIR'18, June, 2025, Tokyo, Japan

「左肺門部に長径 37mm 大の腫瘤を認め、スピクラと内部空洞がみられます。肺癌を疑います。腫瘤は上葉に位置していますが、葉間胸膜をまたいで、下葉にも浸潤しています。また、肺門部リンパ節転移と一塊となっている可能性があります。両肺に数 mm 大の結節が散見されます。肉芽腫や肺内リンパ節を疑いますが、転移の除外目的に経過フォローください。有意な縦隔リンパ節腫大や腹水なし。」
The correct answer for this report was “T2aN1M0.” The output was

「<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。左肺門部に長径 37mm の腫瘤があり、スピクラと内部空洞を伴うと記載されています。この腫瘤は「腫瘤」と記載されているため、充実成分であると仮定します。径が 37mm であり、30mm より大きく 40mm 以下であるため、T2a に該当します。浸潤や副条件を考えます。腫瘤は上葉に位置していますが、葉間胸膜をまたいで下葉にも浸潤していると記載されています。これは T2b の条件を満たします。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。<T 分類> <N 分類>次に N 分類を検討します。レポートには「肺門部リンパ節転移と一塊となっている可能性がある」と記載されています。これは同側肺門リンパ節への転移を示唆しており、N1 に該当します。縦隔リンパ節腫大は認められないため、N2 以上には該当しません。N 分類は N1 と判断します。<N 分類> <M 分類>最後に M 分類を検討します。レポートには「両肺に数 mm 大の結節が散見される」と記載されていますが、「肉芽腫や肺内リンパ節を疑う」とあり、転移の可能性は低いと判断できます。また、有意な縦隔リンパ節腫大や腹水は認められず、遠隔転移を示唆する所見はありません。したがって、M 分類は M0 と判断します。<M 分類> <answer>T2bN1M0</answer>」

The model identified infiltration crossing the interlobar pleura as T2b. Despite correctly extracting the relevant description from the report, its conclusion did not align with the TNM classification criteria. This is an inference error.

There was also one case where M classification was incorrect.

「左肺門部に長径 103mm の腫瘤を認め、原発性肺癌を疑います。左主気管支は腫瘍浸潤による閉塞を認めます。左肺野は虚脱しています。左肺動脈にも腫瘍が浸潤、狭小化しています。縦隔腫瘍浸潤を疑います。左肺門部、左縦隔リンパ節と腫瘍は一塊となっています。右側縦隔リンパ節も腫大、転移を疑います。右肺野には副腫瘍結節を疑う結節が散見されます。左胸水貯留を軽度認めます。撮像範囲の上腹部臓器に明らかな異常は認めません。」

The correct answer for this report was “T4N3M1a.” The output was

「<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。左肺門部に長径 103mm の腫瘤を認めます。この腫瘤は「腫瘤」と記載されているため、充実成分であると仮定します。径が 103mm であり、70mm を超えるため、T4 に

該当します。浸潤や副条件を考えます。左主気管支は腫瘍浸潤により閉塞しており、左肺野は虚脱しています。また、左肺動脈にも腫瘍が浸潤、狭小化しており、縦隔腫瘍浸潤も疑われます。これらの所見は T4 副条件を満たします。性状・サイズ・浸潤を総合的に評価して、T 分類は T4 と判断します。<T 分類> <N 分類>次に N 分類を検討します。左肺門部、左縦隔リンパ節と腫瘍は一塊となっており、同側肺門および縦隔リンパ節転移を示唆します。また、右側縦隔リンパ節も腫大しており、転移が疑われるため、対側縦隔リンパ節転移を示唆し、N3 に該当します。N 分類は N3 と判断します。<N 分類> <M 分類>最後に M 分類を検討します。右肺野に副腫瘍結節を疑う結節が散見されると記載されていますが、対側肺転移とは断定できません。左胸水貯留を軽度認めますが、悪性胸水との記載はありません。撮像範囲の上腹部臓器に明らかな異常は認められないため、遠隔転移を示す所見はありません。M 分類は M0 と判断します。<M 分類> <answer>T4N3M0</answer>」

Despite recognizing the presence of suspected metastatic nodules in the contralateral lung field, the model failed to identify this as contralateral lung metastasis. This can also be considered an inference error.

3.2 Results of Gemma2-9b-it Fine-tuning

The results for the validation data are shown in Table 3.

Without fine-tuning, using a CoT prompt yielded only limited accuracy (Joint accuracy (fine): 0.241). In particular, the exact match accuracy for T classification was low (T accuracy (fine): 0.426), following the same trend reported in NTCIR-17. Even without providing a reasoning process, simply using the reports and correct TNM classifications as training data for SFT improved the Zero-Shot accuracy for N classification and also improved T classification accuracy (T accuracy (fine): 0.426 → 0.463, N accuracy (fine): 0.778 → 0.963).

When SFT (reasoning) was performed and a CoT prompt was used, T classification accuracy increased markedly (T accuracy (fine): 0.426 → 0.648), and overall accuracy improved (Joint accuracy (fine): 0.241 → 0.463). However, the N and M classification accuracies were slightly higher under SFT (non-reasoning) + Zero-Shot (N accuracy (fine): 0.926 vs. 0.963, M accuracy (fine): 0.796 vs. 0.833).

Accuracy	Raw model + CoT prompt	SFT (non-reasoning)	SFT (reasoning)
		+ Zero-Shot	+ CoT prompt
fine	Joint	0.241	0.463
	T	0.426	0.648
	N	0.778	0.926
	M	0.759	0.796
coarse	Joint	0.519	0.778
	T	0.722	0.852
	N	0.778	0.926
	M	0.944	0.944

Table 3. Results on validation data using Gemma2-9b-it.

NTCIR 18 Conference: Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies,

4 Discussion

In this study, we attempted to automatically extract TNM classifications from radiology reports using Gemini 2.0 Flash and a relatively lightweight open model, Gemma2.

Compared to NTCIR-17 MedNLP-SC RR-TNM, LLMs have progressed significantly, and Gemini 2.0 Flash demonstrated relatively good Zero-Shot accuracy (Joint accuracy (fine): 0.685). Because T classification was anticipated to be particularly challenging, a dataset showing the reasoning process up to the determination of the TNM classification from the training data was created before building our system. Consequently, using Many-Shot ICL with CoT and reasoning data suggests that we can improve the accuracy of T classification, which requires complex reasoning regarding tumor size, internal characteristics, and additional conditions. In contrast, for classifications of N and M, where relatively simple anatomical knowledge or the presence/absence of distant metastases are the main determinants, overly verbose reasoning in the model output sometimes leads to misclassifications, underscoring the intriguing differences in model behavior. It may be necessary to adjust the amount of detail or verbosity in the model’s reasoning according to the complexity of each classification task.

Furthermore, examining examples where the model failed in TNM classification revealed that it generally extracted the relevant portions of the report correctly but likely experienced hallucinations during inference. However, because CoT prompts allow humans to inspect the intermediate reasoning output by the model, there is potential to correct or supplement erroneous logic as needed [15]. In the medical domain, safety is paramount, and not only accuracy but also interpretability and the possibility of human intervention are highly valued. Therefore, methods that make the rationale visible are considered extremely valuable [25, 26].

5 Limitations

This study has the following limitations:

Data Privacy Challenges: While the system using Gemini 2.0 Flash Experimental performed best in this study, employing the Gemini API requires network transmission. Although future solutions, such as anonymizing input data and connections, may emerge, sending actual reports to external APIs is currently difficult from a privacy perspective.

Lack of Evaluation of the Latest Version: At the time of writing, the official version of Gemini 2.0 Flash [27] had been released; however, our study only evaluated the Experimental version. Additionally, models such as Gemini 2.0 Flash Thinking, which have been trained to perform inference even without CoT prompts, are gaining attention [28]. Therefore, it may also be necessary to evaluate other models.

Insufficient Fine-Tuning Optimization: We did not adequately tune the hyperparameters in our SFT of Gemma2-9b-it, and the training data were limited. Thus, there is likely room for further performance improvements.

Quantity and Quality of Reasoning Data: The reasoning data in the training set were created by a single specialist, and their objective quality was not assessed. Additionally, because the radiology reports and reasoning processes were in Japanese, we did not evaluate the performance in English or other languages. We also did not investigate how varying the number of examples included in the prompt affects the performance in Many-Shot ICL. Although thousands of examples are reported to be effective [20], there are practical limitations to manually creating reasoning data, making synthetic data a promising option. Because the context window in Gemini 2.0 Flash can be as large as 100M [27], there is ample opportunity for data augmentation.

6 Conclusion

In this study, we developed a method for automatically determining lung cancer TNM classifications from Japanese radiology reports and validated it using the NTCIR-18 RadNLP 2024 Shared Task. By providing the reasoning process created by a radiologist to an LLM through Many-Shot ICL or fine-tuning, and thereby emulating the reasoning process using CoT, we demonstrated the potential to improve overall accuracy, including T classification. This is particularly relevant in the medical field, where the interpretability of the system’s output and opportunities for human intervention are crucial. CoT-based reasoning data appear to be highly beneficial in this regard. Future work will require evaluating generalizability across more extensive and diverse clinical datasets and performing more precise parameter tuning for SFT.

ORAD at NTCIR-18 RadNLP 2024 Shared Task

NTCIR'18, June, 2025, Tokyo, Japan

Base prompt: with CoT

```
<role> あなたは与えられた放射線科レポートから、肺癌のTNM分類を正確に判断し、出力する呼吸器外科専門のassistantです。</role>

<task> TNM分類の定義を厳密に理解し、#注意事項の内容を遵守しながらレポート内の情報のみを用いて分類を行ってください。医学用語の区別には細心の注意を払ってください。</task>

# how to submission
<f分類>[f分類]についての詳細な検討</f分類>
<n分類>[n分類]についての詳細な検討</n分類>
<m分類>[m分類]についての詳細な検討</m分類>
<answer>[最終的なTNM分類]</answer>

# 注意事項
## T分類に関する情報の検討時の注意
・まず腫瘍の位置、大きさ、性状を確認し、<TNM分類の定義>と照らし合わせてどのT分類に相当するかを判断します。
・次に、浸潤や副条件を考え、<TNM分類の定義>と照らし合わせてどのT分類に相当するかを判断します。
・最後に、総合的にどのT分類に相当するかを判断します。
・腫瘍が存在する場合、全体径、充実部径、すりガラス部径を明確に区別してください。
・腫瘍の内部性状は明確に記載されていないこともあります。すりガラスであることが明記されていない場合は充実腫瘍として構いませんが、すりガラスの医学的な言い換え (ground glass opacityなど) に注意してください。
・腫瘍径・充実部径・すりガラス部径が全て未記載の場合は、(T2a副条件),(T3副条件),(T4副条件)を参照し、満たす場合はそれぞれのカテゴリと判断します。
・腫瘍径・充実部径・すりガラス部径が全て未記載かついずれの副条件も満たさない場合は「T0」と判断されます。これは、本来のTNM分類ではTXと表記されるべきものですが、データセットのラベル付けルールに準拠しています。
・腫瘍の大きさはすべてmmに変換し、<TNM分類の定義>と照らし合わせます。「以上」と「以下」はその値を含み、「より大きい」と「より小さい」はその値を含まないことに注意して下さい。
・浸潤と転移は明確に区別してください。

## N分類に関する情報の検討時の注意
・リンパ節転移の有無と、位置を確認します。
・所属リンパ節に該当するかどうか、腫瘍と左右が一致するかどうかに注意してください。

## M分類に関する情報の検討時の注意
・<TNM分類の定義>のN分類で検討する必要がある以外の場所のリンパ節転移は、遠隔転移として扱います(attentionを参照)。
・遠隔転移がある場合には、M1a,M1b,M1cの区別を慎重に行ってください。

## 全体に関する注意
・最も可能性の高い分類を採用してください。判断に迷う場合には【TNM分類と放射線科レポート、TNM分類を決定するまでの思考過程】を参考にしてください。"最終結果に分類不能はありません"。
・各分類は、T0/T1s/T1mi/T1a/T1b/T1c/T2a/T2b/T3, N0/N1/N2/N3, M0/M1a/M1b/M1cの中から選ぶ必要があります(T1,T2だけといったような、大分類では不適です。また、「今回の検討にはTXは含まれません」)。
・レポートには曖昧な表現が多数含まれています。レポートの文章のみから確定できない事項はTNM分類に反映させないことが原則です。ただし、「～を疑います」という表現は、放射線科レポートでは一般的に確定的な事項として扱われます。前後の文脈を加味して都度判断してください。
```

Base prompt: without CoT

```
<role> あなたは与えられた放射線科レポートから、肺癌のTNM分類を正確に判断し、出力する呼吸器外科専門のassistantです。</role>

<task> TNM分類の定義を厳密に理解し、#注意事項の内容を遵守しながらレポート内の情報のみを用いて分類を行ってください。医学用語の区別には細心の注意を払ってください。</task>

# how to submission: 回答は最終的なTNM分類を<answer>タグに囲んで出力してください。それ以外には何も出力しないでください。
answerの例1: <answer>T1bN2M1a</answer>
answerの例2: <answer>T4N0M1c</answer>

# 注意事項
## T分類に関する情報の検討時の注意
・まず腫瘍の位置、大きさ、性状を確認し、<TNM分類の定義>と照らし合わせてどのT分類に相当するかを判断します。
・次に、浸潤や副条件を考え、<TNM分類の定義>と照らし合わせてどのT分類に相当するかを判断します。
・最後に、総合的にどのT分類に相当するかを判断します。
・腫瘍が存在する場合、全体径、充実部径、すりガラス部径を明確に区別してください。
・腫瘍の内部性状は明確に記載されていないこともあります。すりガラスであることが明記されていない場合は充実腫瘍として構いませんが、すりガラスの医学的な言い換え (ground glass opacityなど) に注意してください。
・腫瘍径・充実部径・すりガラス部径が全て未記載の場合は、(T2a副条件),(T3副条件),(T4副条件)を参照し、満たす場合はそれぞれのカテゴリと判断します。
・腫瘍径・充実部径・すりガラス部径が全て未記載かついずれの副条件も満たさない場合は「T0」と判断されます。これは、本来のTNM分類ではTXと表記されるべきものですが、データセットのラベル付けルールに準拠しています。
・腫瘍の大きさはすべてmmに変換し、<TNM分類の定義>と照らし合わせます。「以上」と「以下」はその値を含み、「より大きい」と「より小さい」はその値を含まないことに注意して下さい。
・浸潤と転移は明確に区別してください。

## N分類に関する情報の検討時の注意
・リンパ節転移の有無と、位置を確認します。
・所属リンパ節に該当するかどうか、腫瘍と左右が一致するかどうかに注意してください。

## M分類に関する情報の検討時の注意
・<TNM分類の定義>のN分類で検討する必要がある以外の場所のリンパ節転移は、遠隔転移として扱います(attentionを参照)。
・遠隔転移がある場合には、M1a,M1b,M1cの区別を慎重に行ってください。

## 全体に関する注意
・最も可能性の高い分類を採用してください。判断に迷う場合には【TNM分類と放射線科レポート、TNM分類を決定するまでの思考過程】を参考にしてください。"最終結果に分類不能はありません"。
・各分類は、T0/T1s/T1mi/T1a/T1b/T1c/T2a/T2b/T3, N0/N1/N2/N3, M0/M1a/M1b/M1cの中から選ぶ必要があります(T1,T2だけといったような、大分類では不適です。また、「今回の検討にはTXは含まれません」)。
・レポートには曖昧な表現が多数含まれています。レポートの文章のみから確定できない事項はTNM分類に反映させないことが原則です。ただし、「～を疑います」という表現は、放射線科レポートでは一般的に確定的な事項として扱われます。前後の文脈を加味して都度判断してください。
```

Figure 2. The base prompt with and without CoT prompt.

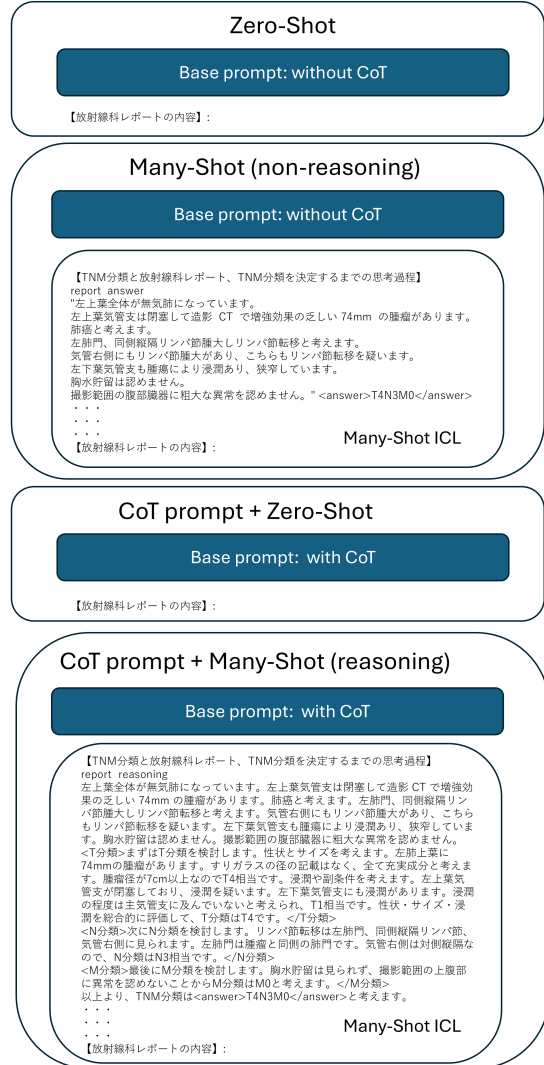


Figure 3. Prompts for each strategy when using Gemini 2.0 Flash Experimental.

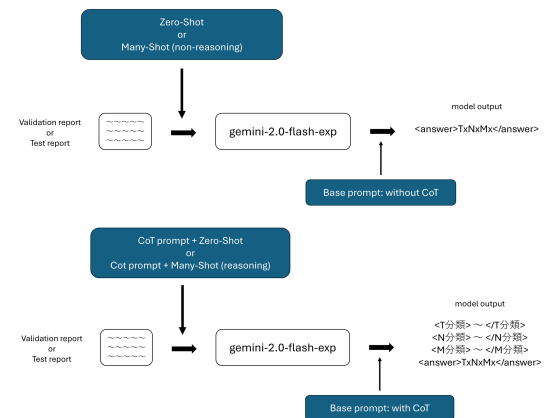


Figure 4a. Overview of the proposed system with Gemini 2.0 Flash Experimental.

NTCIR 18 Conference: Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies,

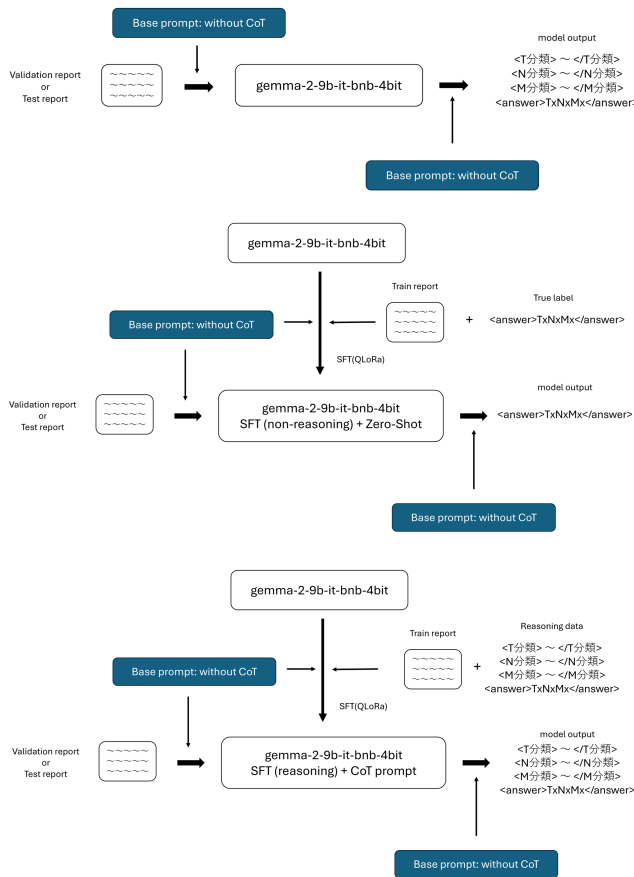


Figure 4b. Overview of the proposed system with Gemma2-9b-it.

ACKNOWLEDGMENTS

REFERENCES

[1]. Y. Nakamura, K. Fujimoto, J. Kluckert, M. Krauthammer, J. Kanazawa, A. Katayama, T. Kikuchi, R. Kurokawa, W. Gono, P. Han, K. Hashimoto, Y. Tashiro, S. Hanaoka, S. Yada, and E. Aramaki. 2025. NTCIR-18 RadNLP 2024 Overview: Dataset and Solutions for Automated Lung Cancer Staging. In Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies.

[2]. F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal. 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 74(3):229–263.

[3]. R. L. Siegel, T. B. Kratzer, A. N. Giaquinto, H. Sung, and A. Jemal. 2025. Cancer statistics, 2025. *CA Cancer J Clin* 75(1):10–45.

[4]. Japan Lung Cancer Society (ed.). 2017. General Rules for Clinical and Pathological Record of Lung Cancer. 8th ed. Tokyo: Kanehara & Co., Ltd.

[5]. The Japan Lung Cancer Society (ed.). 2022. Lung Cancer Clinical Practice Guidelines 2022 Edition. Tokyo: Kanehara & Co., Ltd.

[6]. J. D. Brierley, M. K. Gospodarowicz, and C. Wittekind. 2016. TNM Classification of Malignant Tumours (8th ed.). Union for International Cancer Control, 1–272.

[7]. J. P. Lichtenberger 3rd. 2022. Reporting of Clinical Stage for Lung Cancer: Point-Yes, Report an Overall TNM Stage. *AJR Am J Roentgenol* 218(6):954–955.

[8]. S. H. Feng and S. T. Yang. 2019. The new 8th TNM staging system of lung cancer and its potential imaging interpretation pitfalls and limitations with CT image demonstrations. *Diagn Interv Radiol* 25(4):270–279.

[9]. R. Perez-Johnston and S. McDermott. 2022. Reporting of Clinical Stage for Lung Cancer: Counterpoint—We Are Not There Yet! *AJR Am J Roentgenol* 218(6):956–957.

[10]. A. Patel, A. Rockall, A. Guthrie, F. Gleeson, S. Worthy, S. Grubnic, D. Burling, C. Allen, A. Padhani, B. Carey, P. Cavanagh, M. D. Peake, and G. Brown. 2018. Can the completeness of radiological cancer staging reports be improved using proforma reporting? A prospective multicentre non-blinded interventional study across 21 centres in the UK. *BMJ Open* 8(10):e018499.

[11]. J. E. Lee, K. S. Park, Y. H. Kim, H. C. Song, B. Park, and Y. J. Jeong. 2024. Lung Cancer Staging Using Chest CT and FDG PET/CT Free-Text Reports: Comparison Among Three ChatGPT Large Language Models and Six Human Readers of Varying Experience. *AJR Am J Roentgenol* 223(6):e2431696.

[12]. S. Yada, Y. Nakamura, S. Wakamiya, and E. Aramaki. 2022. Real-MedNLP: Overview of REAL Document-Based MEDical Natural Language Processing Task. In Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.

[13]. Y. Nakamura, S. Hanaoka, S. Yada, S. Wakamiya, and E. Aramaki. 2023. NTCIR-17 MedNLP-SC Radiology Report Subtask Overview: Dataset and Solutions for Automated Lung Cancer Staging. In Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies.

[14]. M. Nishio, H. Matsuo, T. Matsunaga, K. Fujimoto, M. Rohanian, F. Nooralahzadeh, F. Rinaldi, and M. Krauthammer. 2023. Zero-Shot Classification of TNM Staging for Japanese Radiology Report Using ChatGPT at RR-TNM Subtask of NTCIR-17 MedNLP-SC. In Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, 155.

[15]. C. H. Chang, M. M. Lucas, Y. Lee, C. C. Yang, and G. Lu-Yao. 2024. Beyond Self-Consistency: Ensemble Reasoning Boosts Consistency and Accuracy of LLMs in Cancer Staging. In Proceedings of Artificial Intelligence in Medicine (AIME 2024), 224–228.

[16]. S. Kim, S. Jang, B. Kim, L. Sunwoo, S. Kim, J. H. Chung, S. Nam, H. Cho, D. Lee, K. Lee, and S. Yoo. 2024. Automated Pathologic TN Classification Prediction and Rationale Generation From Lung Cancer Surgical Pathology Reports Using a Large Language Model Fine-Tuned With Chain-of-Thought: Algorithm Development and Validation Study. *JMIR Med Inform* 12:e67056.

[17]. OpenAI. 2024. ChatGPT (GPT-4o version) [Large language model]. Retrieved December 2024 – January 2025, from <https://chat.openai.com/>.

[18]. Google DeepMind. 2024. Gemini 2.0 flash experimental. Retrieved January 2025, from <https://deepmind.google/technologies/gemini/flash/>.

[19]. K. He, R. Mao, Q. Lin, Y. Ruan, X. Lan, M. Feng, and E. Cambria. 2025. A Survey of Large Language Models for Healthcare: From Data, Technology, and Applications to Accountability and Ethics. *Information Fusion* 118:102963.

[20]. R. Agarwal, A. Singh, L. M. Zhang, B. Bohnet, L. Rosias, S. Chan, B. Zhang, A. Anand, Z. Abbas, A. Nova, J. D. Co-Reyes, E. Chu, F. Behbahani, A. Faust, and H. Larochelle. 2024. Many-Shot In-Context Learning. In Proceedings of Neural Information Processing Systems (NeurIPS) 2024 (Spotlight Presentation).

[21]. J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

[22]. Gemma Team. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv preprint, arXiv:2408.00118.

[23]. T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv preprint, arXiv:2305.14314.

[24]. D. Han, M. Han, and Unsloth team. 2023. Unsloth [Software]. Retrieved January 2025, from <https://github.com/unslothai/unsloth>.

[25]. Z. Sadeghi, R. Alizadehsani, M. A. CIFCI, S. Kausar, R. Rehman, P. Mahanta, P. K. Bora, A. Almasri, R. S. Alkhalwaldeh, S. Hussain, B. Alatas, A. Shoeibi, H. Moosaei, M. Hladik, S. Nahavandi, and P. M. Pardalos. 2024. A review of Explainable Artificial Intelligence in healthcare. *Comput Elect Eng* 118:109370.

[26]. S. Sim and T. Chen. 2024. Critique of Impure Reason: Unveiling the reasoning behaviour of medical Large Language Models. arXiv preprint, December 20, 2024.

[27]. Google DeepMind. 2024. Gemini 2.0 flash. Retrieved February 2025, from <https://deepmind.google/technologies/gemini/flash/>.

[28]. Google DeepMind. 2024. Gemini 2.0 flash Thinking. Retrieved February 2025, from <https://deepmind.google/technologies/gemini/flash/>.