NTCIR-18: Proceedings of the 18th NTCIR Conference on Evaluation ofInformation Access Technologies, June 10-13, 2025, Tokyo, Japan
DOI: https://doi.org/10.20736/0002002070

# NURad at the NTCIR-18 RadNLP Task

Marina Higashi
Nagoya University, Japan
higashi.marina.v9@s.mail.nagoya-u.ac.jp

Rintaro Ito
Nagoya University, Japan
ito.rintaro.k5@f.mail.nagoya-u.ac.jp

Keita Kato
Nagoya University, Japan
kato.keita.m1@s.mail.nagoya-u.ac.jp

Ryota Asai
Nagoya University, Japan
asai.ryota.c8@f.mail.nagoya-u.ac.jp

Shingo Iwano
Nagoya University, Japan
iwano.shingo.u3@f.mail.nagoya-u.ac.jp

Shinji Naganawa
Nagoya University, Japan
naganawa.shinji.m8@f.mail.nagoya-u.ac.jp

## ABSTRACT

Lung cancer is the most common cause of cancer death in Japan. The TNM classification is essential for lung cancer diagnosis and treatment planning, and CT imaging plays a crucial role in its evaluation. However, the number of thoracic radiologists is limited in Japan. The development of a system to automatically extract TNM classification from radiology reports would be beneficial to radiologists and other clinicians. Large language models (LLMs) have recently shown remarkable progress in natural language processing, opening new possibilities for medical applications. The NURad team participated in the NTCIR-18 Natural Language Processing for Radiology (RadNLP) task [1]. This paper describes our approach to the problem and discusses the official results.

We explored different prompts, LLM models (Llama3, Open AI O1pro, Google Gemini 2.0, Google Notebook LM), and data types (Japanese and English). We also investigated fine-tuning with clinical data. The final model, utilizing a short prompt and trained on both Japanese and English datasets using Google Notebook LM, did not incorporate clinical data.

Our final model with Google Notebook LM achieved a TNM (fine) score of 0.93 on the validation dataset. However, the score decreased to 0.54 on the test dataset. This decline was more pronounced for the T classification compared to the N and M classifications.

This study demonstrates the potential of LLMs for automated TNM classification from radiology reports, but also highlights challenges in generalization to unseen data, particularly for T classification. Further research is needed to improve the robustness and accuracy of LLM-based TNM classification systems.

## KEYWORDS

Medical Natural Language Processing, Radiology Reports, Lung Cancer, Cancer Staging

## TEAM NAME

NURad

## SUBTASKS

Main task (Japanese track), No subtask

## 1 INTRODUCTION

Lung cancer is the leading cause of cancer death in Japan. The diagnosis and treatment planning for lung cancer rely on the TNM classification. In imaging evaluation of the TNM classification, CT images play a key role. Accurate evaluation of TNM classification from CT images requires appropriate training. Among radiologists, it is the thoracic radiologists who have received such training. In Japan, the number of radiologists specializing in thoracic imaging is limited. While the Japanese Radiological Society has approximately 10,500 members [2], the Japanese Society of Thoracic Radiology comprises only around 400 members [3]. Consequently, image diagnosis reports based on TNM classification of lung cancer by thoracic radiologists remain scarce.

An AI model which automatically applies TNM classifications from radiology reports would be valuable not only for radiologists but also for many clinicians who rely on these reports. Furthermore, such a system might alert radiologists to missing critical TNM-related information before finalizing the report. Recently, natural language processing using large language models (LLMs) has made great progress, and their applications in the medical field are being actively explored.
In this study, we developed and evaluated a method to automatically apply the TNM classification (8th edition) of lung cancer [4] from radiology reports using LLM.

## 2 MATERIALS AND METHODS

### 2.1 Data

The task organizers provided 164 data consisting of simulated diagnostic imaging reports for CT of lung cancer in both Japanese and English, along with the corresponding TNM classifications. 108 cases were provided as training data, and

56 cases as validation data. For the final evaluation, 216 cases were supplied as test data, without TNM classifications.

We collected 876 diagnostic imaging reports for evaluation of lung cancer with clinical TNM classification (8th edition) [4] at Nagoya University Hospital from January 2017 to February 2024. Each report was interpreted and classified by a single radiologist (24 years of experience) specializing in thoracic radiology. The study protocol was approved by the Nagoya University Clinical Research Ethics Committee under the condition that the reports be used exclusively within a secure terminal environment (Approval No.: 2024-0447).

## 2.2 Environment

The experiments were conducted on a Windows 11 computer equipped with an Intel® Core™ i7-14700F CPU, 32GB of memory, and an NVIDIA GeForce RTX 4070 Ti SUPER 16GB GPU. Local LLM were used based on Ollama (ver0.4.7) [5] and Open Web UI (ver0.3.16) [6]. Fine tuning was done with LLaMA-Factory (ver0.9.1) [7], and the trained models were deployed using LM Studio (ver0.3.8) [8].

## 2.3 Model

For this study, the following local LLMs were used:

- Llama-3-ELYZA-JP-8B [5]
- Llama-3.2-90B(Meta) [6]

The following Cloud-based LLMs were used in this study, with analysis conducted from December 2024 to January 2025:

- Gemini 2.0 Flash Thinking (Google) [7]
- GPT-O1 pro (Open AI) [8]
- Google Notebook LM(Google) [9]

Models were selected based on their trainability and performance. Llama-3-ELYZA-JP-8B was chosen because it is specialized for Japanese, and its model-size allows for additional training with our computer. Llama-3.2-90B was selected as the largest model size that could run on our PC. From the cloud-based LLMs, we selected GPT-O1 pro, which was considered to have the best performance at the time this study was conducted. However, GPT-O1 pro has a character limit on the system prompt. Therefore, Gemini 2.0 Flash Thinking, which does not have a character limit, was selected as a comparison target for the verification of the system prompt. Google Notebook LM is superior in Retrieval-Augmented Generation （RAG）, so it was added to the comparison target.

## 2.4 Prompt

We divided the task into the following subtasks:

A. Overall instructions
B. TNM classification of lung cancer
C. TNM classification procedure


**Figure 1-A.** Simple prompt

**Figure 1-B**. Step−by-step prompts


**Figure 1-C.** Detailed prompt

Three types of prompts were prepared to perform this task **(Figure 1.):**
(1) Simple prompt: A short system prompt instructing the model to output only the TNM classification.
(2) Step-by-step prompts: We asked LLMs multiple times and extracted the information separately.
(3) Detailed prompt: An integration of the step-by-step prompts into a system prompt, with instructions to output only the TNM classification.

Overall instructions section and TNM classification of lung cancer section were held constant across the three prompts, while the TNM classification procedure section was varied for each prompt. In Simple prompt, we segmented diagnostic imaging report sentences into some categories (This is a subtask of RadNLP.) and specified the sentence for judgment for each TNM classification. Step-by-step prompts and Detailed prompt were offered a more granular approach to TNM classification determination after segmentation. As an example, for the T classification, tumor size and pleural invasion status were assessed before generating the T classification. We also used three of the Japanese train data to the system prompt as example cases (Few-Shot Prompting).

These prompts were tested using Llama-3-ELYZA-JP-8B, Llama-3.2-90B, and Gemini 2.0 Flash Thinking. For performance comparison, we used 27 Japanese cases extracted from the initial 30 cases (excluding the 3 cases used as examples in the system prompt). 30 train data was given at the beginning of the NTCIR-18 RadNLP Task.
Llama-3-ELYZA-JP-8B and Llama3.2-90B were executed with Open Web UI. Parameters are temperature = 0, Context length = 1001, Max tokens = 1001. Gemini 2.0 Flash Thinking parameters are temperature = 0, Output length = 8192.

## 2.5 Preparing Training Source

The data used for fine-tuning and Google Notebook LLM were divided into the following three categories and formatted in JSON format as shown in **Figure 2.**

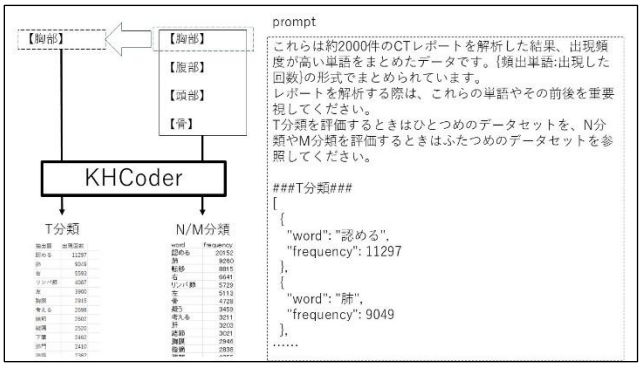
**Figure 2**. Data format


```
[
 {
  "input": "",
  "instruction": "左上葉全体が無気肺になっています。左上葉気管支は閉塞して造影 CT で増強効果の乏しい 74mm の腫瘤があります。肺癌と考えます。左肺門、同側縦隔リンパ節腫大しリンパ節転移と考えます。気管右側にもリンパ節腫大があり、こちらもリンパ節転移を疑います。左下葉気管支も腫瘍により浸潤あり、狭窄しています。胸水貯留は認めません。撮影範囲の腹部臓器に粗大な異常を認めません。",
  "output": "T4N3M0",
  "category": "classification",
  "index": "0"
 },
 {
  "input": "",
  "instruction": "右下葉 S8 に長径 30mm の腫瘤性病変があります。辺縁には spicula を伴っています。肺癌の可能性がありますので、呼吸器科的に精査ください。左下葉には炎症瘢痕様の索状影を認めます。右肺門リンパ節腫大が疑われます。胸水認めません。",
  "output": "T1cN0M0",
  "category": "classification",
  "index": "1"
 },
```


・108 Japanese train cases
・164 cases of English train data and validation data
・876 clinical cases

Next, 876 clinical cases were subjected to text-mining analysis using KHCoder [14] to identify words that appeared frequently in the reports. In addition, we specified the top 50 most frequent words in the chest findings and the top 50 most frequent words in the entire report, then created prompts directing the analysis to focus on these words (**Figure 3**).

**Figure 3**. Text mining




## 2.6 Choose Model

Fine-tuning of Llama-3-ELYZA-JP-8B was performed in an offline environment using the JSON-format Japanese train data and clinical cases prepared in 2.5. The results were then quantized and executed in LM Studio. The training parameters were as follows:

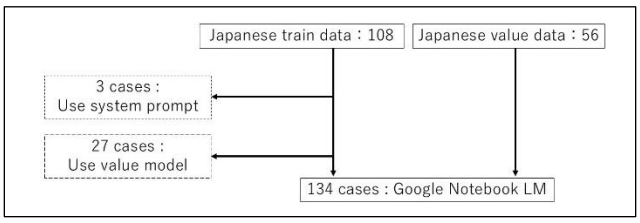
- Fine tuning: learning rate = 2e-4, epoch = 4, maximum gradient norm = 0.3, max samples = 10000, compute type = bf16, cut off length = 1024, batch size = 4, gradient accumulation = 16, valsize = 0, LS scheduler = cosine
- Quantification: max shared size = 5GB, quantization bit = none.
- LM Studio: parameter = 0

We then compared the following three models using 27 training data sets. We used the simple prompts stated in 2.4, further shortened them to fit GPT-O1pro's character limit, and put them into the system prompts of these three models. Due to GPT-O1pro's character limit, specific examples were not given. (Zero-Shot Prompting)

(1) Llama-3-ELYZA-JP-8B with fine-tuning
(2) GPT-O1 pro
(3) Google Notebook LLM + source

We gave Google Notebook LM sources a total of 164 Japanese train and value data, excluding 3 used for prompt creation and 27 used for evaluation, for a total of 134 data (**Figure 4**).

**Figure 4**. Data source for Google Notebook LM




## 2.7 Choose Training Source

We compared which sources could be added to the Google Notebook LM to get the best data. We used 56 Japanese validation data for evaluation. The prompts used were the same as in 2.6.

(1) 108 train data in Japanese
(2) 108 Japanese train data + 164 English data
(3) 108 Japanese train data + text mining
(4) 108 Japanese train data + 164 English data + text mining

## 3 RESULTS

The accuracy of each prompt and model was evaluated using the following evaluation indicators provided by the Task organizer.

- TNM (fine): Exact match accuracy for all T, N, and M factors.
- T(fine): Accuracy for the T factor.
- N(fine): Accuracy for the N factor.
- M(fine): Accuracy for the M factor.
- TNM (coarse): Accuracy allowing for differences among Tis/T1mi/T1a/T1b/T1c, T2a/T2b, and M1a/M1b/M1c.T(coarse): Accuracy for the T classification with permissible differences among Tis/T1mi/T1a/T1b/T1c and T2a/T2b.
- N(coarse): Same as N(fine).
- M(coarse): Accuracy for the M classification allowing for differences among M1a/M1b/M1c.

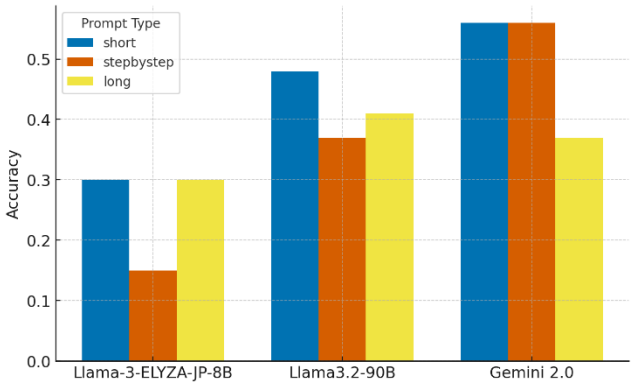
## 3.1 Make Prompt

In Llama-3-ELYZA-JP-8B, Simple prompt and Detailed prompt performed equally well. In Llama3.2-90B, Simple prompt had the best performance. In Gemini 2.0 Flash Thinking, Step-by-step prompts and Simple prompt performed equally well, but Step-by-step prompts required more effort to generate multiple responses **(Table 1, Figure5).**

**Table 1.** Compare prompt

| prompt | Simple | Step-by-step | Detailed | Simple | Step-by-step | Detailed | Simple | Step-by-step | Detailed |
|---|---|---|---|---|---|---|---|---|---|
| model | Llama-3-ELYZA-JP-8B | | | Llama3.2-90B | | | Gemini 2.0 | | |
| TNM(fine) | **0.30** | 0.15 | **0.30** | **0.48** | 0.37 | 0.41 | **0.56** | **0.56** | 0.37 |
| T(fine) | 0.37 | 0.22 | 0.33 | 0.52 | 0.41 | 0.44 | 0.63 | 0.67 | 0.48 |
| N(fine) | 0.78 | 0.67 | 0.78 | 0.96 | 0.89 | 0.96 | 1.00 | 0.96 | 0.96 |
| M(fine) | 0.85 | 0.78 | 0.85 | 0.93 | 0.85 | 0.93 | 0.89 | 0.93 | 0.85 |
| TNM(coarse) | **0.41** | 0.22 | **0.41** | **0.56** | 0.41 | 0.52 | **0.56** | **0.56** | 0.44 |
| T(coarse) | 0.48 | 0.30 | 0.44 | 0.59 | 0.44 | 0.56 | 0.63 | 0.67 | 0.56 |
| N(coarse) | 0.78 | 0.67 | 0.78 | 0.96 | 0.89 | 0.96 | 1.00 | 0.96 | 0.96 |
| M(coarse) | 0.89 | 0.78 | 0.93 | 0.93 | 0.85 | 0.93 | 0.89 | 0.93 | 0.85 |

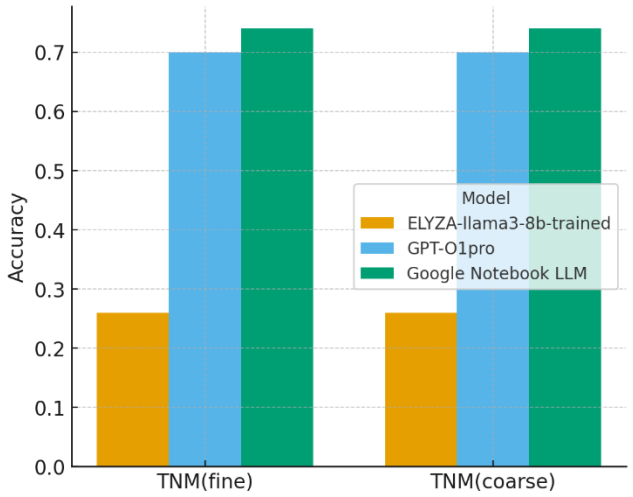
**Figure 5.** Compare prompt




## 3.2 Choose Model

Google Notebook LLM demonstrated the best performance with a TNM (fine) score of 0.74 and a TNM (coarse) score of 0.74. In comparison, GPT-O1pro achieved scores of 0.70 for both metrics, and ELYZA-llama3-8b-trained scored 0.26 for both. **(Figure6.)**

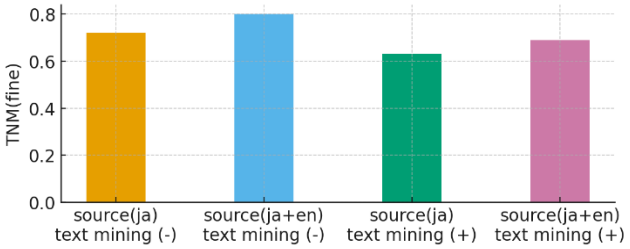
**Figure 6.** Compare Model



## 3.3 Choose Training Source

Using both Japanese and English data ('ja+en') improved the TNM (fine) score compared to using Japanese data only ('ja'), regardless of whether text mining analysis was applied or not. Specifically, the highest TNM (fine) score of 0.80 was achieved using Japanese and English data without text mining analysis. Incorporating text mining seemed to slightly decrease the TNM (fine) score. (**Figure7**)

**Figure 7.** Compare Training Source



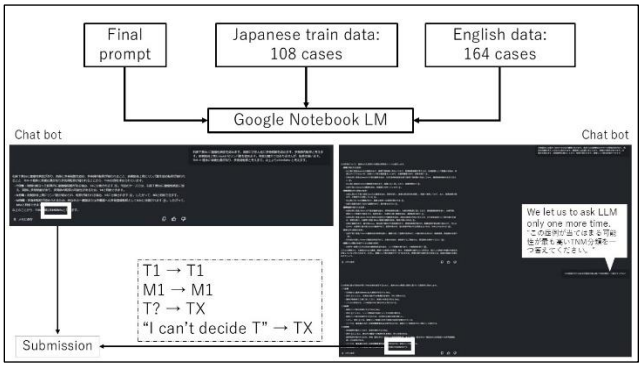

## 3.4 Final Model

From the results of 3.1, we selected a short prompt. From the results of 3.2, we selected Google Notebook LM. From the results of 3.3, we selected to provide Japanese and English data. Finally, the following prompts and data were used to predict the validation data and test data (**Figure 8).**

Incomplete predictions such as T1 and T2 were submitted as responses as they were. Those that were answered as T? or "unpredictable" were answered as TX or MX.

**Figure 9** presents the official results of test sets with our final model. Although the TNM (fine) score was 0.93 for the validation data, it dropped to 0.54 for the test data. The decline was particularly notable for the T classification compared to the N and M classifications.

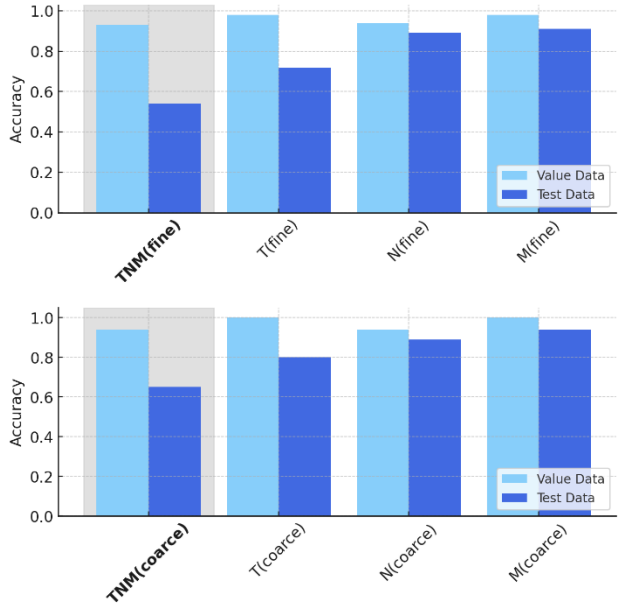
**Figure 8**. Final Our Model.




**Final prompt**

```
###肺癌の TNM 分類についての説明###


==T 分類==
T0：{
原発腫瘍を認めない}
Tis：{
上皮内癌（carcinoma in situ）や AIS：肺野型の場合は，充実成分径 0 cm かつ病変全体径≦3 cm}
T1：[
腫瘍の充実成分径≦3 cm，肺または臓側胸膜に覆われている，葉気管支より中枢への浸潤が気管支鏡上認められない（すなわち主気管支に及んでいない）
さらに、腫瘍の大きさにより以下の 4 つに分類される。
T1mi：{微少浸潤性腺癌：部分充実型を示し，充実成分径≦0.5 cm かつ病変全体径≦3 cm}
T1a：{充実成分径≦1 cm でかつ Tis・T1mi には相当しない}
T1b：{充実成分径＞1 cm でかつ≦2 cm}
T1c：{充実成分径＞2 cm でかつ≦3 cm}
]
T2：[
充実成分径＞3 cm でかつ≦5 cm
または充実成分径≦3 cm でも以下 2 つのいずれかであるもの:{
a)主気管支に及ぶが気管分岐部には及ばない
b)臓側胸膜や葉間胸膜に浸潤、肺門まで連続する部分的または一側全体の無気肺か閉塞性肺炎がある
}
さらに、腫瘍の大きさにより以下の 2 つに分類される。:{
T2a：充実成分径＞3 cm でかつ≦4 cm
T2b：充実成分径＞4 cm でかつ≦5 cm
}
]
T3：[
充実成分径＞5 cm でかつ≦7 cm
または充実成分径≦5 cm でも以下 2 つのいずれかであるもの:{
a)壁側胸膜，胸壁（superior sulcus tumor を含む），横隔神経，心膜のいずれかに直接浸潤。
b)同一葉内の不連続な副腫瘍結節(肺転移、肺内転移)を持つ
}
]
T4：{
充実成分径＞7 cm，または大きさを問わず横隔膜，縦隔，心臓，大血管，気管，反回神経，食道，椎体，気管分岐部への浸潤，あるいは同側の異なった肺葉内に副腫瘍結節がある}


==N 分類==
肺癌の所属リンパ節は以下の通りです。[
鎖骨上窩リンパ節
前斜角筋リンパ節
縦郭リンパ節：{
上部気管傍リンパ節
血管前リンパ節、気管後リンパ節
下部気管傍リンパ節
大動脈下リンパ節
大動脈傍リンパ節
気管分岐下リンパ節
食道傍リンパ節
肺靱帯リンパ節
}
肺門リンパ節：{
主気管支周囲リンパ節
葉気管支間リンパ節
]


これらのリンパ節について、以下のように分類します。
N0：{所属リンパ節転移なし}
N1：{同側の気管支周囲かつ/または同側肺門，肺内リンパ節への転移で原発腫瘍の直接浸潤を含める}
N2：{同側縦隔かつ/または気管分岐下リンパ節への転移}
N3：{対側縦隔，対側肺門，同側あるいは対側の前斜角筋，鎖骨上窩リンパ節への転移}


==M 分類==
M0：{遠隔転移なし}
M1a：{対側肺内の副腫瘍結節，胸膜または心膜の結節，悪性胸水(同側・対側)，悪性心嚢水}
M1b：{肝臓や副腎、骨など肺以外の一臓器への単発遠隔転移}
M1c：{肺以外の一臓器または多臓器への多発遠隔転移}


###解答の形式###
以下は、あなたの回答が従うべき注意点です。{
・"T*N*M*"のように答えをまとめて出力してください。
・最も可能性が高い候補が複数ある場合は、最も低いステージの分類をひとつだけを答えてください。
・Input された文章以外の情報を TNM 分類に使用しないでください。
・ソースで与えられた文章、思考過程、Input された文章を繰り返し出力しないでください。
}
```

**Figure 9**. Official Result




## 4 DISCUSSION

Our final model achieved the following accuracies on the test dataset: TNM (fine): 0.54, T (fine): 0.72, N (fine): 0.81, and M (fine): 0.91. Among these, the T classification showed the lowest accuracy. This lower performance in T classification was due to the more frequent occurrence of both obvious incorrect answers and incomplete responses.

Incorrect answers were more frequent in cases of invasive mucinous adenocarcinoma (IMA). IMA is a subtype of lung adenocarcinoma characterized by abundant mucus production, which often present CT imaging similar to consolidation or pneumonia. These characteristics make it challenging to identify the primary tumor, and consequently, radiologists often find it difficult to evaluate the T stage. Therefore, different prompts from those used for typical lung cancer are needed. Furthermore, training the model with enough IMA cases is crucial.

Incomplete responses were more frequent when diagnosis reports had incomplete TNM staging information. (e.g., ambiguous T1 or T2 descriptors) This is likely because the generated TNM classifications were heavily influenced by the report descriptions. Therefore, further refinement of the prompts is necessary.

In developing our model, we compared offline fine-tuning and multiple cloud-based LLMs and investigated the use of clinical cases. Offline fine-tuning was performed to address privacy concerns associated with utilizing clinical data in cloud-based LLMs. However, this offline fine-tuning resulted in decreased performance compared to the pre-tuned model. Potential contributing factors include the limited number of cases used for fine-tuning and potentially suboptimal fine-tuning parameters. The quality of our data, specifically the incompleteness of staging information, may also contribute to the observed performance decrease. In our institution, a proportion of cases lacked complete staging information for lymph nodes or distant metastasis due to incomplete PET/CT or contrast-enhanced MRI evaluations. These cases were sometimes assigned an 'X' classification, which may have negatively impacted the accuracy of the fine-tuning process.
Prior studies have indicated that when given exemplary responses, Google Notebook LM can outperform GPT-O1 pro [15]. Our findings corroborate this, with Google Notebook LM surpassing the performance of the paid GPT-O1 pro. Since Google Notebook LLM generates responses based on provided sources, carefully selecting and curating the input data may further enhance performance. Although we used text mining to indirectly incorporate clinical cases, this approach did not improve results, suggesting that alternative text mining methods should be explored.

While several studies used LLMs for medical image classification, labeling, and evaluation, few have compared local and commercial LLMs for lung cancer TNM classification using actual clinical cases. Previous work includes a study from NTCIR-17 utilizing fictitious CT radiology reports with GPT 3.5-turbo [16], and another using fictitious reports with Google Notebook LM [15]. Additionally, there is research comparing local and commercial LLMs for chest X-ray report classification [17] and a study using GPT-4 for pancreatic cancer staging in radiology reports [18]. However, to our knowledge, no study has directly compared both local and commercial LLMs for lung cancer TNM classification using real clinical cases. Limitations of this study include the lack of consideration for multifocal lung cancer, the absence of statistical significance testing for model performance differences, and the small number of evaluation cases (27 cases), which may not fully represent lymph nodes or brain metastasis evaluation were sometimes staged with an "X," possibly affecting fine tuning accuracy.

Limitations of this study include the lack of consideration for multifocal lung cancer, the absence of statistical significance testing for model performance differences, and the small number of evaluation cases (27 cases), which may not fully represent clinical accuracy.

In future research, the ambiguity in the interpretation of radiology reports should be addressed. Radiologists may differ in their interpretation of findings suggestive of invasion or metastasis. Some may classify these findings as positive, leading to a higher stage, while others may interpret them as negative, resulting in a lower stage. The cases provided by the

Task organizers in this study seem to be from multiple patients, and the concept of "possibility" appears to be inconsistently applied. This is considered a data set similar to actual clinical practice and is good for evaluation data. However, difficulties arise in terms of training LLM with a small amount of data. In Nagoya University Hospital, The TNM classification is used to stage all cases of lung cancer, according to a standardized format and criteria. We suggest that future studies using our cases could lead to a more consistent model.

This study was mainly performed by a medical researcher with basic programming skills. Developing the LLM environment, processing data, and creating Excel macros were done using online resources and AI tools. The results are preliminary but provide initial insights. The authors thank online experts for sharing their knowledge, which helped in analyzing the data. We also deeply appreciate the support from the team in the Department of Radiology at Nagoya University.

## 5 CONCLUSIONS

We successfully developed an AI model using LLMs to automatically apply lung cancer TNM classifications (8th edition) from diagnostic imaging reports, achieving reasonable performance. Further improvements in data volume, fine tuning, and prompt design are needed to enhance accuracy, particularly in the T classification.

## REFERENCES

[1] Yuta Nakamura, Koji Fujimoto, Jonas Kluckert, Michael Krauthammer, Jun Kanzawa, Akira Katayama, Tomohiro Kikuchi, Ryo Kurokawa, Wataru Gonoi, Peitao Han, Kiyoto Hashimoto, Yuki Tashiro, Shouhei Hanaoka, Shuntaro Yada, Eiji Aramaki. 2025. NTCIR-18 RadNLP 2024 Overview: Dataset and Solutions for Automated Lung Cancer Staging. *In Proceedings of the NTCIR-18 Conference*.

[2] Japan Radiological Society https://www.radiology.jp/jrs_about/outline.html Accessed February 25, 2025.

[3] Japanese Society of Thoracic Radiology https://www.jstr.jp/english Accessed February 25, 2025.

[4] The Japan Lung Cancer Society,2017. General Rule for Clinical and Pathological Record of Lung Cancer. Kanehara & Co., Ltd.

[5] Ollama, Ollama. https://ollama.com/ Accessed February 11, 2025.

[6] Open Web UI. Timothy Jaeryang Baek. https://github.com/open-webui/open-webui. Accessed February 11, 2025.

[7] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, Yongqiang Ma. 2024. LLAMAFACTORY: Unified Efficient Fine-Tuning of 100+ Language Models. arXiv:2403.13372

[8] LM Studio Element Labs, Inc. https://lmstudio.ai/ Accessed February 11, 2025.

[9] elyza/Llama-3-ELYZA-JP-8B. Masato Hirakawa and Shintaro Horie and Tomoaki Nakamura and Daisuke Oba and Sam Passaglia and Akira Sasaki. https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B Accessed February 11, 2025.

[10] Llama 3.2 90B. Meta. https://huggingface.co/meta-llama/Llama-3.2-90B-Vision Accessed February 11, 2025.

[11] Gemini 2.0 Flash Thinking. Google DeepMind https://deepmind.google/technologies/gemini/ Accessed February 11, 2025.

[12] GPT-O1 pro. Open AI. https://openai.com/ Accessed February 11, 2025.

[13] Google Notebook LM. Google https://notebooklm.google/. Accessed February 11, 2025.

[14] KH Coder. Kouichi Higuchi. https://khcoder.net/ Accessed February 11, 2025.

[15] Ryota Tozuka, Hisashi Johno, Akitomo Amakawa, Junichi Sato, Mizuki Muto, Shoichiro Seki, Atsushi Komaba, Hiroshi Onishi. 2024. Application of NotebookLM, a large language model with retrieval-augmented generation, for lung cancer staging. *Japanese Journal of Radiology.*

[16] Hidetoshi Matsuo, Mizuho Nishio, Takaaki Matsunaga, Koji Fujimoto and Takamichi Murakami.2024. Exploring Multilingual Large Language Models for Enhanced TNM Classification of Radiology Report in Lung Cancer Staging. *Cancers*

[17] Felix J. Dorfner, Liv Jürgensen, Leonhard Donle, Fares Al Mohamad, Tobias R. Bodenmann, Mason C. Cleveland, Felix Busch, Lisa C. Adams, James Sato, Thomas Schultz, Albert E. Kim, Jameson Merkow, Keno K. Bressem, Christopher P. Bridge. 2024. Comparing Commercial and Open-Source Large Language Models for Labeling Chest Radiograph Reports. *Radiology*.

[18] Kazufumi Suzuki, Hiroki Yamada, Hiroshi Yamazaki, Goro Honda. Shuji Sakai. 2024. Preliminary assessment of TNM classifcation performance for pancreatic cancer in Japanese radiology reports using GPT-4. *Japanese Journal of Radiology*.