NTCIR-18: Proceedings of the 18th NTCIR Conference on Evaluation ofInformation Access Technologies, June 10-13, 2025, Tokyo, Japan
DOI: https://doi.org/10.20736/0002002062

ASUKAI89 at NTCIR 18 RadNLP Task: Lung Cancer Staging Automatic Classification System Utilizing Large Language Models and Meta-Prompting

Yoshifumi Okura
Kyoto Min-iren Asukai Hospital
Japan
o77164906y@gmail.com

Yuki Kataoka
Kyoto Min-iren Asukai Hospital
Scientific Research Works Peer Support Group
(SRWS-PSG)
Japan
youkiti@gmail.com

ABSTRACT

This study aims to develop and evaluate a system that automatically extracts the TNM classification of lung cancer (T: primary tumor, N: lymph node metastasis, M: distant metastasis) from radiological diagnosis reports.

In the initial experiments, inference was performed using 'gemini-2.0-flash-thinking-exp-1219'. By incorporating explicit TNM classification criteria and unit specifications—features absent in conventional methods—and introducing error analysis and prompt improvements through meta-prompting, an overall accuracy improvement of approximately 15% was achieved after prompt modification.

In the final evaluation, using the 'o1 2024-12-01-preview' model, we achieved approximately 70% joint accuracy (fine), 76% T accuracy, 93% N accuracy, and 95% M accuracy. This paper provides a detailed account of the experimental procedures and the improvement process at each stage.

KEYWORDS

Large Language Models (LLM), Automatic Classification, Meta-Prompting

TEAM NAME

asukai89

SUBTASKS

Main task (Japanese track)

1 INTRODUCTION

Accurate TNM classification is critical for determining the appropriate treatment strategy in lung cancer diagnosis and therapy. Yet, radiological reports often omit explicit staging, forcing clinicians to rely on experience and judgment.

Recently, we have advanced machine learning techniques alongside the emergence of large language models (LLMs). We now automatically extract features from large diagnostic datasets using deep learning and supervised learning, which yields more objective and reproducible classifications.

Traditionally, researchers used strict rule-based methods [6] for pathological TNM classification, but these methods struggle with context-dependent and ambiguous expressions. LLMs now allow us to flexibly extract information through prompt design and tuning, as discussed on Kaggle [4] [5] and shown in recent studies [3] [2].

In this study, we investigate how to improve an automated system that extracts TNM classifications from diagnostic reports [7].

2 RELATED WORK

Researchers have traditionally performed pathological TN classification using strict rule-based methods based on international guidelines. However, diagnostic reports often include ambiguous descriptions and subtle contextual nuances, which conventional methods fail to fully extract. Consequently, clinicians frequently rely on their experience and judgment [6]. Recently, the emergence of large language models (LLMs) has inspired new approaches that overcome these limitations. For example, Yamagishi et al. [9] proposed a zero-shot information extraction and clustering method using an open-source LLM. Their approach, which does not depend on supervised learning, effectively addressed data scarcity issues and leveraged contextual information in Japanese radiological reports, sharing common challenges and results with our approach.

In addition, Chia et al. [2] examined a method that extracts pathological TNM (pTNM) classification from pathology reports using only an open-source clinical LLM, without requiring labeled training data. They demonstrated that proper prompt design enhances performance.

These studies suggest that by improving prompt design and applying meta-prompting, we can fully leverage LLMs' strengths to achieve more accurate TNM classification extraction.

3 METHODS

3.1 Data Acquisition

Data Source: Radiological diagnosis reports stored in .txt format and their corresponding CSV files on Google Drive (Figure 1).

3.2 Inference Process and Prompt Design

For determining the TNM classification, the definition of each candidate was provided as an instruction. The criteria were based on the official guidelines of the internationally adopted "TNM Classification, 8th Edition, 2017" [1] and were organized into JSON format using 'o1 2024-12-01-preview' (o1) model. The output was then manually inspected to ensure there were no errors.

In cases of misclassification, error analysis was performed using the training data and correct labels. Based on the results, a

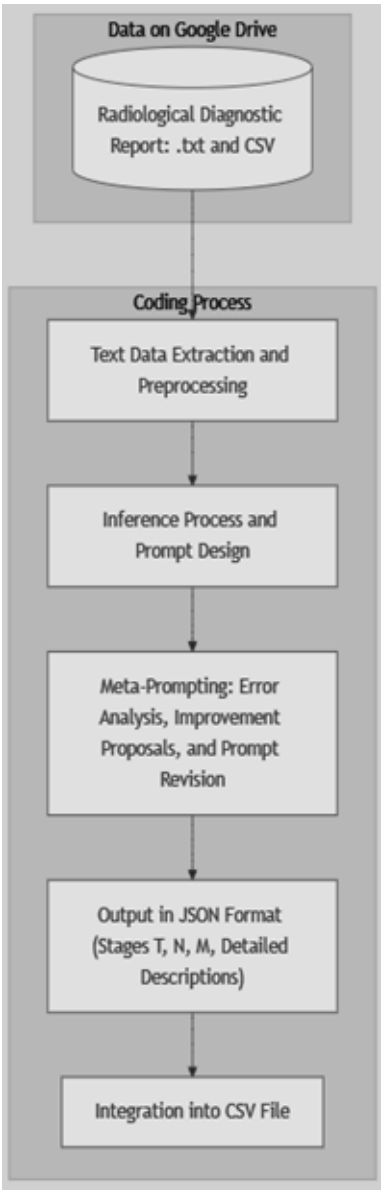

Figure 1: This system consists of the following main pipeline:

meta-prompt was generated to improve the prompt—a process we refer to as "meta-prompting." For meta-prompting, ‘gemini-2.0-flash-thinking-exp-1219‘ (gemini-2.0) was employed. After modifying the prompt according to the improvement proposals, inference was performed again, and further readjustment was carried out using o1 to achieve an overall improvement in accuracy. Details have been published on GitHub.

3.3 Post-Processing and Integration of Output

From the JSON outputs obtained for each report, the stage and detailed description for each factor (T, N, M) were extracted and integrated into a CSV file. Logs and intermediate results generated during inference were systematically saved—along with a mechanism for re-execution in the event of errors—to ensure reproducible experimental operations.

4 EXPERIMENTS AND RESULTS

We describe the main experimental procedures and outcomes below.

4.1 Initial Experiment

Model: ‘gemini-2.0-flash-thinking-exp-1219‘ (gemini-2.0)

Approach: A prompt that includes basic TNM classification criteria.

Results:

Metric	Fine	Coarse
Joint accuracy	0.6852	0.7778
T accuracy	0.7778	0.8889
N accuracy	0.9444	0.9444
M accuracy	0.9259	0.9259

Table 1: Initial experiment results using gemini-2.0 model with basic TNM classification criteria

4.2 Prompt Improvement and Meta-Prompting

Model: ‘gemini-2.0-flash-thinking-exp-1219‘ (gemini-2.0)

Approach:

- Explicitly stated the TNM classification criteria at the beginning of the prompt
- Changed unit notation from cm to mm
- Introduced meta-prompting based on error analysis and performed re-adjustment using o1

Results:

Metric	Fine	Coarse
Joint accuracy	0.8333	0.8519
T accuracy	0.9074	0.9259
N accuracy	0.9259	0.9259
M accuracy	0.9815	0.9815

Table 2: Accuracy metrics after prompt improvement and meta-prompting with gemini-2.0 model

4.3 Divided Inference Trial and Discussion

Model: ‘gemini-2.0-flash-thinking-exp-1219‘ (gemini-2.0)

Approach:

- Explored methods to perform inference on each factor (T, N, and M) individually
- Extracted TNM from the entire report via batch extraction
- Conducted specialized inference for N and M through individual extraction
- Determined the final TNM classification by combining all results with information from the original report using ensemble integration

Results:

Metric	Fine	Coarse
Joint accuracy	0.7593	0.7593
T accuracy	0.8889	0.9074
N accuracy	0.8704	0.8704
M accuracy	0.9259	0.9444

Table 3: Results of divided inference approach using individual factor extraction and ensemble integration

4.4 Model Switching and Final Evaluation

Model: ‘o1 2024-12-01-preview‘ (o1)

Approach:

- Used the prompt that achieved good accuracy with gemini-2.0
- Switched to o1 for inference

Final Results (Private Leaderboard):

Metric	Fine	Coarse
Joint accuracy	0.7037	0.7778
T accuracy	0.7685	0.8380
N accuracy	0.9306	0.9306
M accuracy	0.9583	0.9676

Table 4: Final evaluation results on Private Leaderboard using o1 model

The code used in this study is available on GitHub [8].

5 DISCUSSION AND CONCLUSIONS

We evaluated an automated system that extracts lung cancer TNM classifications from radiological diagnosis reports using LLMs. Our results show that leveraging LLMs enables our system to outperform conventional rule-based approaches in interpreting context-dependent and ambiguous expressions.

We found that refining prompt design and applying meta-prompting iteratively played a critical role in boosting overall accuracy. Our o1 model-based inference method achieved approximately 70% joint accuracy, 76% T accuracy, 93% N accuracy, and 95% M accuracy, suggesting that inference models may extract structured data more effectively.

Future work must validate this approach using larger datasets and diverse languages, and develop an integrated method that accounts for factor interdependencies while supporting clinical deployment. These advancements could pave the way for automatic information extraction techniques from radiological diagnosis reports to serve as clinical support systems for accurate lung cancer staging and treatment decision-making.

REFERENCES

[1] Japan Lung Cancer Society. 2017. *Lung Cancer Treatment Guidelines* (8 ed.). Kanehara Shuppan. 4–6 pages.
[2] Can Jin. et al. 2024. Prompt Engineering and Classification with LLMs. *arXiv preprint* (2024). arXiv:2404.01589
[3] Sylvio Barbon Junior. et al. 2024. Recent Advances in Large Language Models and Their Applications. *arXiv preprint* (2024). arXiv:2406.06591
[4] Kaggle. 2023. *Kaggle LLM Science Exam Discussion.* https://www.kaggle.com/competitions/kaggle-llm-science-exam/discussion/446422
[5] Kaggle. 2023. *Kaggle LLM Science Exam Discussion.* https://www.kaggle.com/competitions/kaggle-llm-science-exam/discussion/446248
[6] Sanghwan Kim. et al. 2023. Traditional rule-based approaches to pathologic TN classification. *PubMed Central* (2023).
[7] Y. Nakamura et al. 2024. NTCIR-18 RadNLP 2024 Overview: Dataset and Solutions for Automated Lung Cancer Staging. In *Proceedings of NTCIR-18 RadNLP 2024.*
[8] Yoshifumi Okura. 2024. RadNLP-2024 Repository. https://github.com/okura1406/radnlp-2024.
[9] Yosuke Yamagishi. et al. 2025. Large Language Model Approach for Zero-Shot Information Extraction and Clustering of Japanese Radiology Reports: Algorithm Development and Validation. *JMIR Cancer* 11 (2025), e57275. https://doi.org/10.2196/57275