DOI: https://doi.org/10.20736/0002001299

# Classification of cancer TNM stage from Japanese radiology report using on-premise LLM at NTCIR-17 MedNLP-SC RR-TNM subtask

Koji Fujimoto †
Advanced Imaging in Medical Magnetic Resonance
Kyoto University
Kyoto, Japan
kfb@kuhp.kyoto-u.ac.jp

Mizuho Nishio
Center for Advanced Medical Engineering Research & Development
Kobe University
Kobe, Japan
nishiomizuho@gmail.com

Chikako Tanaka
Real World Data Research and Development
Kyoto University
Kyoto, Japan
tchikako@kuhp.kyoto-u.ac.jp

Morteza Rohanian
Department of Quantitative Biomedicine
University of Zurich
Zurich, Switzerland
m.rohanian@qmul.ac.uk

Farhad Nooralahzadeh
Department of Quantitative Biomedicine
University of Zurich
Zurich, Switzerland
Farhad.Nooralahzadeh@uzh.ch

Michael Krauthammer
Department of Quantitative Biomedicine
University of Zurich
Zurich, Switzerland
michael.krauthammer@uzh.ch

Fabio Rinaldi
Dalle Molle Institute for Artificial Intelligence
Lugano, Switzerland
fabio.rinaldi@idsia.ch

## ABSTRACT

In this manuscript, we describe our submission to the RR-TNM subtask of NTCIR-17 MedNLP-SC shared task. We took an approach to create extensive question-and-answer (Q&A) pairs related to TNM classification as a method of domain-specific augmentation. Compared to the result without data augmentation, improvement in the accuracy especially for the M stage was observed.

## KEYWORDS

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

## TEAM NAME

kuhp

## SUBTASKS

RR-TNM

## 1 INTRODUCTION

### 1.1 Epidemiology of lung cancer

Lung cancer stands as one of the most prevalent and lethal malignancies globally[1]. The incidence of lung cancer varies widely across different regions, with higher rates observed in developed countries, often attributed to factors such as smoking, environmental pollution, and occupational exposure to carcinogens[2]. Mortality rates, on the other hand, have shown a complex trend, with some areas experiencing a decline due to advancements in early detection and treatment, while others continue to face challenges[3]. The multifaceted nature of lung cancer's epidemiology underscores the necessity for a comprehensive understanding of its risk factors, prevention

strategies, and therapeutic interventions, setting the stage for the ongoing battle against this formidable disease.

### 1.2 CT imaging

Computed Tomography (CT) imaging has emerged as an indispensable tool in the diagnosis and staging of lung cancer, offering a non-invasive and highly detailed view of the pulmonary structures. CT imaging provides essential information for treatment planning, including the assessment of tumor size, lymph node involvement, and potential metastasis. The accuracy of CT scans demonstrated sensitivity rates of up to 94% for detecting lung nodules[4].

### 1.3 TNM staging system

The TNM staging system's application in lung cancer care is a vital aspect of diagnosis, treatment planning, and prognosis. In lung cancer, the TNM staging system categorizes the disease based on the size and extent of the tumor (T), the involvement of nearby lymph nodes (N), and the presence of distant metastasis (M)[5]. The stage of lung cancer directs the treatment approach. Early-stage tumors may be suitable for surgical resection, while later stages may require chemotherapy, radiation, targeted therapy, or palliative care. The stage at diagnosis is a key factor in predicting survival rates. Early-stage lung cancer has a significantly better prognosis compared to advanced stages. TNM staging allows for consistent grouping of patients in clinical trials and epidemiological studies, facilitating the comparison of treatment outcomes across different studies and populations.

### 1.4 TNM staging in radiology reports

The inclusion of TNM staging in radiology reports by imaging specialists is a critical aspect of cancer diagnosis and treatment planning, particularly in lung cancer. However, the TNM system requires subjective judgment, especially in assessing lymph node involvement and tumor boundaries, which can lead to inconsistencies. As a result, different radiologists might interpret the same images differently, leading to variations in TNM staging. Other than this inter-reader inconsistency, the TNM stage is not explicitly written in radiology reports. One of the reasons for this is because radiologists are aware of the limitation of the imaging modality and avoid assertive description. Therefore, TNM staging information embedded in narrative radiology reports can be challenging to extract for secondary uses such as research, quality control, or population health management.

### 1.5 Purpose of the present study

The purpose of the present study is to automatically extract and classify the TNM staging from radiology reports of lung cancer written in Japanese. To do this, we fine-tuned a Japanese large language model (LLM), openCALM-7B (https://huggingface.co/cyberagent/open-calm-7b), for direct inference of the TNM staging as a Question and Answer task.

## 2 RELATED WORKS

### 2.1 NLP for radiology report

NLP is a set of techniques that involves processing human natural language by computers. NLP has been shown to be able to extract structured information from radiology reports.

In comparison to computer vision in radiology[6,7], NLP has not been popular in the artificial intelligence (AI) community in radiology. However, Casey et al reported in a systematic review published in 2021 publication in 2019 almost tripe those in 2015[8]. Traditionally, rule-based methods have been employed. Recent progress in NLP is accompanied by applications of machine learning and deep learning. In Casey's review, the number of deep learning method increased from five publications in 2017 to twenty-four publications in 2019.

### 2.2 NLP for automatic extraction and classification of TNM staging of lung cancer

As applications of NLP to automatic extraction and classification of the TNM staging, there are several previous studies. Hu et al. aimed to automatically extract staging-related information from CT reports of lung cancer[9]. Their deep-learning model achieved the best performance with a macro-F1 score of 97.13% and a micro-F1 score of 98.37%. Gupta et al. developed a system for automatic classification of the TNM staging using the rule-based method along with a deep learning-based method called long short-term memory (LSTM) [10]. Their empirical experiment showed an accuracy of up to 85%. Aalabdulsalam et al. proposed novel applications of NLP and machine learning to automatically extract and classify TNM stage mentions from records at the Utah Cancer Registry[11]. Their results indicated that the TNM stages can be extracted and classified automatically with high accuracy (extraction sensitivity: 95.5%‐98.4% and classification sensitivity: 83.5%‐87%).

At the previous NTCIR (NTCIR16 Real-MedNLP MexTxt-RR Case Identification task), the task was to find the same radiology report for the same case written by 9 different radiologists. Fujimoto et al. took an approach to first classify each radiology report based on the TNM staging[12]. To do this, they created custom labels, which consists of a pair of a sentence with the corresponding {T or N or M} stage. By using the dataset, they first estimated the probability of {T/N/M} stage for each case. Using this information with the rule-based dedicated algorithm, they implemented a clustering algorithm.

### 2.3 Large language models

LLM has a relatively short but significant history in NLP. A major breakthrough occurred with the advent of deep learning techniques and the development of neural network-based models, such as recurrent neural networks and LSTM networks. The real revolution came with the introduction of transformer-based architectures, such as Bidirectional Encoder Representations from Transformers (BERT)[13] and Generative

Pre-trained Transformers (GPT)[14-16]. These LLMs enabled impressive achievements in NLP tasks, including language translation, sentiment analysis, question-answering, and text generation. They were also adapted for a wide range of applications, from chatbots to content creation, and became a crucial part of AI research and development.

## 2.4 ChatGPT

ChatGPT is based on the GPT architecture developed by OpenAI. The development of ChatGPT began with the original GPT-1 model, released in 2018, which demonstrated the power of transformer-based language models for various language tasks[14]. The subsequent iterations brought significant advancements in terms of model size, performance, and capabilities. GPT-2, introduced in 2019, was a notable breakthrough due to its impressive ability to generate coherent and contextually relevant text[15]. In June 2020, OpenAI released GPT-3, a giant model with 175 billion parameters, marking a new milestone in the field of NLP[16]. GPT-3's capabilities were beyond previous iterations, showcasing remarkable language understanding, question-answering, and even creative text generation skills. Following the success of GPT-3, OpenAI introduced ChatGPT (https://chat.openai.com/), a user-friendly version designed for interactive conversations with users, which became available to the public through the ChatGPT website.

# 3 METHODS

In this section, we describe our system to fine-tune a Japanese LLM for extracting and classifying the TNM staging from Japanese radiology reports. For the development of our system, python (version 3.10), a publicly available 7 billion parameter LLM trained with Japanese corpus by a Japanese company (CyberAgent) named OpenCALM (https://huggingface.co/cyberagent/open-calm-7b), and Huggingface's Parameter-Efficient Fine-Tuning (PEFT) package (https://github.com/huggingface/peft) which supports Low-Rank Adaptation of Large Language Models [17] for better memory efficiency was used. One GeForce RTX 3090 with a 24GB memory on a Linux machine (Ubuntu 20.04.6 LTS) was used. The training parameters were: number of epochs=1~1000, learning rate=5e-05, optimizer=adamW, batch size=8.

## 3.1 Dataset of Japanese radiology reports

The details and method to build the RR-TNM dataset are described in the task overview paper of organizers of the RR-TNM subtask[18]. The RR-TNM dataset was labeled under TNM criteria of the 8th edition by the Japan Lung Cancer Society. Radiopaedia (https://radiopaedia.org/), an open-access radiology reference, was used as the source of lung cancer cases, and 27 cases were included as lung cancer cases in the RR-TNM dataset. Japanese nine board-certified radiologists created free-text radiology reports for these cases. Organizer assigned labels of TNM stage to all of these radiology reports. The RR-TNM dataset was split into training, validation, and test sets in the ratio of 12:6:9. The dataset was split on a case basis so that radiology reports diagnosing the same case did not appear in multiple splits. For each report of the training and validation sets, T, N, and M factors are available for participants of the RR-TNM subtask. The radiology reports are available only in Japanese.

## 3.2 Data augmentation

Figure 1 shows the schematic illustration of the data augmentation process. In our system, we used two types of data augmentation as follows:

Augmentation (1)

Radiology reports in a training dataset were processed with random drops (delete randomly 10% of the character), and sentence-wise random shuffle within the same radiology report. At this step, three different degrees of augmentation (i.e. none, 1000x, 500x) were performed.

Augmentation (2)

One of the authors (KF) created extended QAs, which are related to TNM staging. For example,

- What is the T stage in this case? {T0 / T1 / T2 / T3 / T4}
- Does this patient have distant metastasis? {Yes / No}
- Is this tumor larger than 7cm? {Yes / No}

The total number of questions we prepared was approximately 50. A full list of the extended QAs is shown in the appendix. Please note that not all questions have answers. Therefore, by creating variations of QA pairs, the size of the dataset (QA pair) was increased approximately by 35x. By doing this, we expected the model to "understand the concept of the TNM staging system".

## 3.3 Instruction to the LLM

Instruction was given to the LLM based on the following dictionary:

```
PROMPT_DICT = {
    "prompt_input": (
        "以下は、タスクを説明する指示と、文脈のある入力の組み合わせです。"
        "要求を適切に満たす応答を書きなさい。¥n¥n"
        "### 指示:¥n{instruction}¥n¥n### 入力:{input}¥n¥n### 応答:"
    ),
    "prompt_no_input": (
        "以下は、タスクを説明する指示です。"
        "要求を適切に満たす応答を書きなさい。¥n¥n"
        "### 指示:¥n{instruction}¥n¥n### 応答:"
    )
}
```

, where "prompt_input" was used for augmented dataset, and "prompt_no_input" was used for the baseline dataset.

For the baseline dataset (without data augmentation), "input" was blank and the radiology report was put in the instruction section along with a question about the TNM stage. Representative examples of "input" and "instruction" data for the baseline dataset are shown below.

```
{
    "index": "56344",
    "category": "rad_rpt",
    "input": "",
    "instruction": "次の症例のTNM stageは何ですか？ ：「 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"
},
```

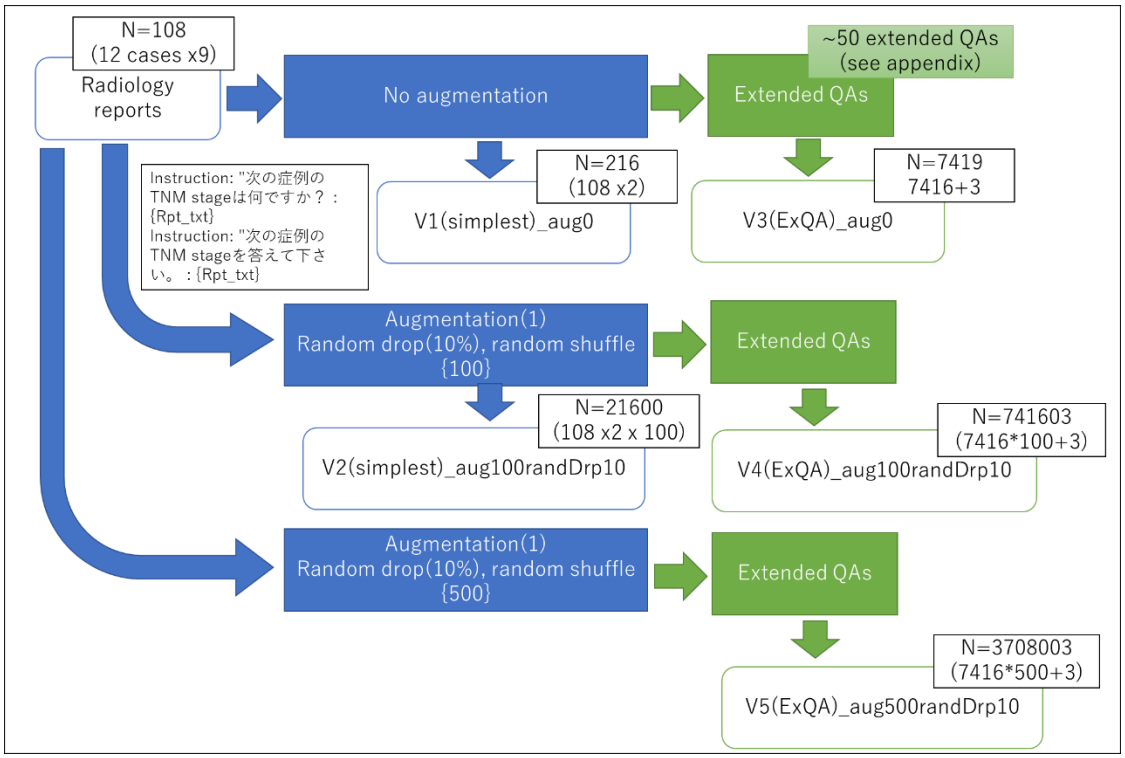



Figure 1. Overview of the data augmentation process. Two similar instructions to ask for the TNM stage were used even without the augmentation. Extended QAs consited of approximately 50 extended QAs. (Please note that not all questions have correspoinding answers.) Three basic questions about the definition of T, N, M staging were additionally included in the extended QA corpus.

For the extended QA corpus, the radiology report was used as an “input” and the “instruction” contained only a question.


```
{
    "index": "56344",
    "category": "rad_rpt_QA",
    "input": "「 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相大な異常を認めません。 」
    ",
    "instruction": "腫瘍の大きさは7cm以上ですか？ : ",
    "output": "はい"
},
```


### 3.4 Evaluation metrics

The organizers used four types of accuracies for evaluating systems of the TNM staging based on the label of the test set (i.e. accuracy for T, N, M stage and joint accuracy). We also used this metric to compare our results in the ablation study.

### 3.5 Chat GPT

Other than the models presented above, the results of zero-shot inference by the chat GPT (version 3.5 and 4.0) by using the following prompt were manually collected and summarized in the spreadsheet in the Microsoft excel.

テキスト解析ソフトの性能検証をしたいです。次の文章は私がつくったプログラムが生成したテキストです。次の症例の画像所見に基づく TNM stage を「TxNxMx」の形で簡潔に答えて下さい。今回は推論の理由の呈示は不要です。「{radiology report}」

## 4 RESULTS

### 4.1 Main results (submitted to the organizer)

The four types of accuracies calculated by the organizers for the test set are shown in Figure 2. We submitted results of inference by models with extended QAs with 100x augmentation (i.e. v4). Model weights at two different checkpoints (kuhp_1 for 32K~=0.35 epoch, kuhp_2 and kuhp_3 for 64K~=0.69 epoch) were submitted. The best scores in our submission for T, N, and M accuracies were 0.4815, 0.6049, and 0.7407, respectively. The best joint accuracy of our submission was 0.2346.

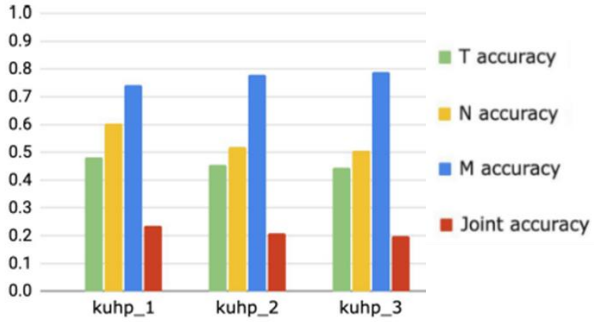
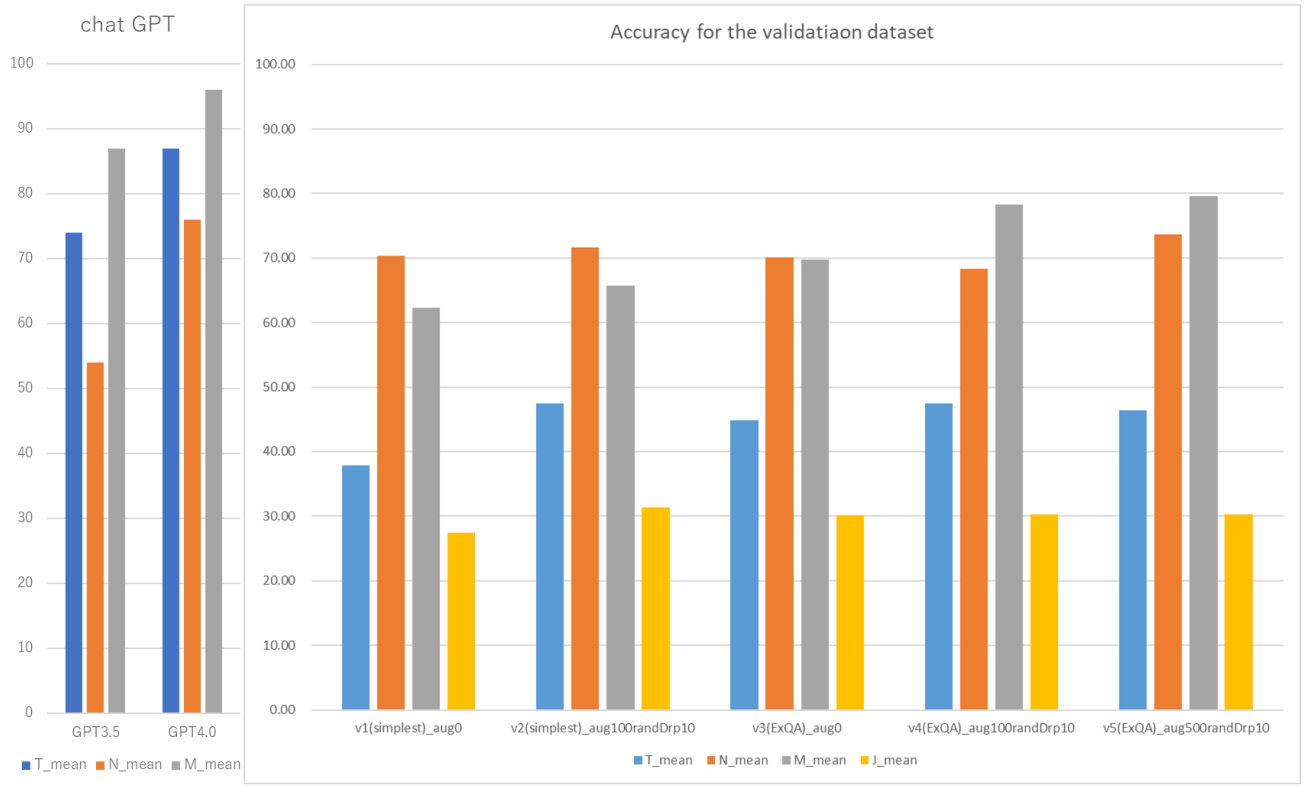



Figure 2. Accuracies of T, N, and M factors and Joint accuracy of our system for the test set. kuhp_1-3 were the results of the same model with different number of training epochs.

## 4.2 Accuracy during the training

The accuracy for the validation dataset and the training dataset during the training is shown in figure 3. Except for the training with simple dataset without augmentation (v1), the accuracy reached to almost 100% in the training dataset, which suggests overfitting.


Figure 3. Results of the accuracy for the validation dataset (top) and the training dataset (bottom). The accuracy was shown in blue for T, in orange for N, in green for M. The joint accuracy is shown in red.




| #QA pairs | 216 | 21600 | 7419 | 741603 | 3708003 |
|---|---|---|---|---|---|
| Augmentation | 0 | 100 | 0 | 100 | 500 |
| Random drop | 0 | 10 | 0 | 10 | 10 |
| Random shuffle | - | + | - | + | + |
| ExQA | - | - | + | + | + |

Figure 4. Results of ablation study and the results of chat GPT. The accuracy for the validation dataset were shown for different training dataset with different degree of data augmentation were shown. The numbers in a blue background shows the number of QA pairs after augmentation.

## 4.2 Ablation study

The results of the ablation study are summarized in Figure 4. Other than the submitted results, the results of the zero-shot inference by the chat GPT (version 3.5 and 4.0) were also shown as a reference.

Augmentation (1)

Compared with no augmentation (random drop and random shuffle), the T and the M scores were improved for both simple dataset (v1 vs. v2) and for extended QA dataset (v3 vs. v4-5). For extended QA dataset, augmentation of 500x (v5) improved the accuracy of N staging compared to 100x (v4).

Augmentation (2)

Accuracy for the T and the M stage was higher for extended QA dataset without augmentation (v3) compared with simple dataset without augmentation (v1). Accuracy for the M stage was higher for extended QA dataset with augmentation (v4) compared with simple dataset with augmentation (v2).

## 5 DISCUSSION

In this research, we created individual question-and-answer (Q&A) pairs related to TNM classification as a method of domain-specific augmentation. By focusing on specific Q&A pairs related to TNM classification, this method ensures that the augmentation is directly relevant to the task at hand, potentially increasing the model's performance.

Q&A pairs can help the model to understand the relationships between different aspects of TNM classification, such as how tumor size relates to the T stage, or how lymph node involvement affects the N stage. For creating or validating these Q&A pairs, medical experts were engaged in this research, ensuring that they are clinically accurate and relevant.

The gain in the accuracy for the T stage by doing data augmentation was small. On the other hand, the score for the M stage was improved by the data augmentation. The meaning of medical terms can be highly context-dependent. Altering a word or phrase might inadvertently change the clinical meaning,

leading to incorrect or misleading information. The T stage contains several types of criteria other than the tumor size, and therefore more complex compared with the M stage. Therefore, our approach of data augmentation (random deletion and sentence-wise random shuffle) might not be helpful for the estimation of the T stage.

Regarding the model fitting to the training dataset, the model trained with the dataset without augmentation was quickly over-fitted. By doing data augmentation, we were able to slow this down. However, the model still reached the state of overfitting, showing a discrepancy between the accuracy in the training dataset and the validation dataset. To avoid overfitting, we would require a lot more datasets, ensuring that the dataset captures the diversity of real-world scenarios.

Compared to the previous NTCIR16 RealMed-NLP task (named entity recognition, NER), the approach we used does not require special annotation in the radiology report. For the task of NER, one of the common approaches was to use BERT, along with the token-wise labels for each entity. Since the reference standard annotation labels were given by the organizer, we did not have to put effort into dataset preparation. In this NTCIR17 MedNLP-SC, RR-TNM task, such kind of annotation was not required. The only required label was the TNM stage, which is relatively easy to check in the health record system. Considering that the NER is not a “purpose” of the text processing, but a “method” to extract information (such as TNM stage) from the medical text (such as radiology report), recent progress in the method to train an LLM enabled to reduce the burden for the dataset preparation.

There are several limitations in this research.

- Comparison of the performance with a simple document classification was not performed.
- The size of the model we used was relatively small (7B parameters).
- As shown in figure 3, our model resulted in overfitting. For data augmentation, there are other ways, including word swapping, synonym replacement, sentence rephrasing, etc., which are not evaluated here. However, medical texts often contain highly specialized terminology and complex structures. Standard text augmentation techniques like synonym replacement may not be applicable, as medical terms often don't have simple synonyms that can be swapped without changing the meaning.
- We did not perform hyperparameter tuning.
- There was a limit on the size of the max tokens, which in our case was 1024. As a result, some of the long instructions and the responses were truncated.
- The model we used was trained on the general text. Utilizing models pre-trained on medical text corpora can leverage knowledge learned from related tasks.
- TNM stage is determined using multiple modalities, including ultrasound, bronchoscopy, etc. Therefore, it must be noted that the TNM stage obtained from the medical record sometimes does not match that in the radiology report
-

## 6 Future perspectives

The integration of radiologists' interpretation reports with LLM offers an exciting frontier in the automation and enhancement of TNM staging for lung cancer. in addition, this approach is not specific to lung cancer but could be adapted to different types of cancer or different staging systems, making it a versatile method of augmentation.

As we have done in this project, LLM can be trained on a dataset of radiology reports and corresponding TNM stages, learning to predict the TNM stage based on the text of the report. If the prediction is accurate enough, the predicted TNM stage can be integrated into the patient's EHR, which could be usable for treatment planning, prognosis assessment, and research. As a potential benefit, automation of TNM staging can reduce the time required for this task, allowing radiologists to focus on more complex interpretive tasks, although, at the current performance of the LLM, careful human supervision should be required. For secondary use mainly focusing on research, automated extraction of TNM staging information may facilitate population health management. Once integrated into the EHR, the system can continuously learn from new data. Being able to adapt to changes in the TNM staging system may also be important for the continuous usage of the LLM integrated into the EHR.

## 7 CONCLUSION

In this research, we used an approach to create extended question-and-answer (Q&A) pairs related to TNM classification as a method of domain-specific augmentation. This targeted strategy has the potential to enhance the model's understanding of TNM classification and offers a pathway for expert involvement.

## ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI (Grant Number: 22K07665, and 23K07154).

### Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT-4 in order to prepare the draft of this paper. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

### Contribution

KF performed most of the experiments described in this paper. KF and NM performed most of the manuscript preparation and editing. All authors participated in several discussions and exchanged ideas. All authors read the paper and approved it.

## REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68: 394‑424.

2. Thun M, Peto R, Boreham J, Lopez AD. Stages of the cigarette epidemic on entering its second century. Tob Control. 2012;21: 96‐101.

3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019;69: 7‐34.

4. McWilliams A, Tammemagi MC, Mayo JR, Roberts H, Liu G, Soghrati K, et al. Probability of cancer in pulmonary nodules detected on first screening CT. N Engl J Med. 2013;369: 910‐919.

5. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WEE, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. J Thorac Oncol. 2016;11: 39‐51.

6. Matsuo H, Nishio M, Nogami M, Zeng F, Kurimoto T, Kaushik S, et al. Unsupervised-learning-based method for chest MRI‐CT transformation using structure constrained unsupervised generative attention networks. Sci Rep. 2022;12: 1‐15.

7. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018;9: 611‐629.

8. Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak. 2021;21: 179.

9. Hu D, Zhang H, Li S, Wang Y, Wu N, Lu X. Automatic Extraction of Lung Cancer Staging Information From Computed Tomography Reports: Deep Learning Approach. JMIR Med Inform. 2021;9: e27955.

10. Gupta EK, Thamma- R, Thakkin- A. NLP automation to read radiological reports to detect the stage of cancer among lung cancer patients. [cited 31 Aug 2023]. Available: http://www.winlp.org/wp-content/uploads/2019/final_papers/189_Paper.pdf

11. AAlAbdulsalam AK, Garvin JH, Redd A, Carter ME, Sweeny C, Meystre SM. Automated Extraction and Classification of Cancer Stage Mentions fromUnstructured Text Fields in a Central Cancer Registry. AMIA Jt Summits Transl Sci Proc. 2018;2017: 16‐25.

12. Fujimoto K, Nishio M, Sugiyama O, Ichikawa K, Cornelius J, Lithgow-Serrano O, et al. Approach for Named Entity Recognition and Case Identification Implemented by ZuKyo-JA Sub-team at the NTCIR-16 Real-MedNLP Task. 2022. pp. 322‐329.

13. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019. pp. 4171‐4186.

14. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. [cited 31 Aug 2023]. Available: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

15. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. [cited 31 Aug 2023]. Available: https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf

16. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. arXiv [cs.CL]. 2020. Available: http://arxiv.org/abs/2005.14165

17. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of large language models. arXiv [cs.CL]. 2021. Available: http://arxiv.org/abs/2106.09685

18. Nakamura Y, Hanaoka S, Yada S, Wakamiya S, Aramaki E. 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, NTCIR-17. National Institute of Informatics (NII).

Appendix

List of extended QAs

この症例の T(腫瘍)ステージは T" + {Tstage} + "ですか？
この症例の N(腫瘍)ステージは N" + {Tstage} + "ですか？
この症例の M(腫瘍)ステージは M" + {Tstage} + "ですか？
次の肺癌症例の T(腫瘍）のステージは何ですか？鍵となる文章を抽出して説明してください。
次の肺癌症例の N(腫瘍）のステージは何ですか？鍵となる文章を抽出して説明してください。
次の肺癌症例の M(腫瘍）のステージは何ですか？鍵となる文章を抽出して説明してください。
この症例では遠隔転移はありますか？
原発巣のサイズは？
N stage が NX である理由は？
ンパ節腫大はある？
ンパ節転移の部位は？
同側の気管支周囲リンパ節への転移は？
同側の肺門リンパ節への転移は？
同側のリンパ節への直接浸潤は？
同側の縦隔リンパ節への転移は？
同側の気管分岐下リンパ節への転移は？
同側の鎖骨上窩リンパ節への転移は？
同側の腋窩リンパ節への転移は？
対側の縦隔リンパ節への転移は？
対側の肺門リンパ節への転移は？
対側の鎖骨上窩リンパ節への転移は？
M stage が M1 である理由は？
遠隔転移はある？
対側肺への肺内転移は？
胸膜播種は？
悪性胸水は？
肝臓に転移は？
副腎に転移は？
骨に転移は？
T stage が TX である理由は？
縦隔への浸潤は？
心臓への浸潤は？
気管への浸潤は？
椎体への浸潤は？
気管分岐部への浸潤は？
同側別肺葉への浸潤は？
主気管支への浸潤は？
臓側胸膜への浸潤は？
肺門に連続する無気肺や閉塞性肺炎は？
壁側胸膜への浸潤は？
胸壁への浸潤は？
同側肺葉の副腫瘍結節は？
原発巣（肺癌）は右、左どちらか？
原発腫瘍を認める？
腫瘍の大きさは 3cm 以下ですか？
腫瘍の大きさは 3cm 以上、5cm 以下ですか？
腫瘍の大きさは 5cm 以上、7cm 以下ですか？
腫瘍の大きさは 7cm 以上ですか？'