NTCIR-18: Proceedings of the 18th NTCIR Conference on Evaluation ofInformation Access Technologies, June 10-13, 2025, Tokyo, Japan
DOI: https://doi.org/10.20736/0002002065

Hirosaki team at the NTCIR-18 RadNLP2024 Shared Task: Few-Shot Learning and Prompt Engineering for TNM Staging Classification of English Radiology Reports Using Large Language Models.

Ryutaro Mori
Hirosaki University, Japan
d.forest@hirosaki-u.ac.jp

Shota Hosokawa
Hirosaki University, Japan
shosokawa@hirosaki-u.ac.jp

Tsudou Watanabe
Hirosaki University, Japan
h24mh214@hirosaki-u.ac.jp

Koichi Okuda
Hirosaki University, Japan
okuda1@hirosaki-u.ac.jp

Taisei Komoda
Hirosaki University, Japan
h24mh205@hirosaki-u.ac.jp

Yasuyuki Takahashi
Hirosaki University, Japan
ytaka3@hirosaki-u.ac.jp

ABSTRACT

We participated in the NTCIR-18 RadNLP2024 shared task [1] and investigated the automation of TNM classification using large language models (LLMs), specifically GPT-4o-mini, GPT-4o, and o1-mini. Our approach integrates cosine similarity-based retrieval using embedding vectors and few-shot learning to enhance classification accuracy. As a result of the experiment, o1-mini achieved the highest classification accuracy. However, the accuracy on the test data declined by approximately 30% compared to the validation data. In particular, the low classification accuracy of the T factor highlighted challenges in interpreting tumor size and extent of infiltration. In this paper, we analyze these results and report our approach to this task along with official results.

KEYWORDS

Medical Natural Language Processing, Radiology Reports, Lung cancer, Cancer Staging, Large Language Model, Few shot learning, Chain of Thought, Embedding vectors.

TEAM NAME

Hirosaki

SUBTASKS

Main task (English track)

1 INTRODUCTION

1.1 Background and objectives

Since we have been performing high-resolution computed tomography (CT) and magnetic resonance imaging (MRI) scans with thin slice thicknesses, the number of images interpreted by radiologists per examination is increasing. And cancer diagnosis, applying the TNM staging system is essential for determining the treatment plan. However, Manual determination of the TNM staging is time-consuming and increases the burden on radiologists.

Rule and machine learning-based methods have been developed to solve this issue in the past. Several automated TNM staging models developed using these methods demonstrated high accuracy. However, rule-based methods struggle to adapt to new challenges. Therefore, these methods have difficulty improving generalization and collecting data. A large language model (LLM) is based on natural language processing (NLP) that is pre-trained on a large amount of text data. As the sizes of LLMs increase, they have recently acquired higher inferential capabilities. LLMs can accurately execute tasks using the pre-trained knowledge without additional training. Furthermore, LLMs by utilizing prompt engineering and few-shot learning methods have achieved high classification capability with limited additional data. Therefore, LLMs are helpful for determining the TNM classification.

In this study, we investigated methods to predict lung cancer TNM staging from radiology reports as part of the NTCIR-18 RadNLP2024 shared task [1]. Specifically, we used three OpenAI models: GPT-4o-mini, GPT-4o, and o1-mini, and compared the classification accuracy using prompt engineering and few-shot learning. Furthermore, we analyzed factors of these methods contributing to improved accuracy in TNM staging.

1.2 Lung cancer TNM classification

The TNM classification is an indicator used to assess the progression of malignant tumors. The TNM classification by the Union for International Cancer Control (UICC) / the American Joint Committee on Cancer (AJCC) is used as the international standard. Japanese medical society has also established management guidelines for various types of carcinoma. The Japan Lung Cancer Society (JLCS) has established the Classification of Lung Cancer 8th edition [2]. Furthermore, it closely aligns with the 8th edition of the UICC TNM classification [3]. The labeling criteria for lung cancer staging in this task are based on the JLCS's 8th Classification of Lung Cancer [2].

The TNM classification consists of the following three categories.

- T： the size of a primary tumor
- N： the presence or absence of lymph node metastasis
- M： the presence or absence of the distant metastasis

The detailed classification of lung cancer is denied as follows: T (T0, Tis, T1mi, T1a, T1b, T1c, T2a, T2b, T3, T4), N (N0, N1, N2, N3), and M (M0, M1a, M1b, M1c). Details of definitions of TNM categories have been published elsewhere in the 8th edition of the UICC-TNM classification [3]. We challenged the English task of the NTCIR-18 RadNLP2024. Thus, we created the English version of the TNM classification based on the JLCS's 8th Classification of Lung Cancer [2] and used it as a criteria for our classification models. The complete text of the classification is as follows.

T: primary lesion

- T0: There is no sign of cancer.
- Tis: An area of cancer cells contained within the inner lining of the lungs.
- T1mi: A staging description for a type of non-small cell lung cancer called adenocarcinoma. It means minimally invasive adenocarcinoma. The cancer is no more than 3 cm at its widest part. It has grown no further than 0.5 cm into deeper lung tissue.
- T1a: The cancer is 1 cm or less at its widest part.
- T1b: The cancer is between 1 cm and 2 cm across.
- T1c: The cancer is between 2 cm and 3 cm across.
- T2a: The cancer is between 3 cm and 4 cm. It involves the main airway (the main bronchus) but is not close to the area where the bronchus divides to go into each lung. It involves the inner lining of the chest cavity (the visceral pleura). Part or all of the lung has collapsed or is blocked due to inflammation.
- T2b: The cancer is between 4 cm and 5 cm. It involves the main airway (the main bronchus) but is not close to the area where the bronchus divides to go into each lung. It involves the inner lining of the chest cavity (the visceral pleura). Part or all of the lung has collapsed or is blocked due to inflammation.
- T3: The cancer is between 5 cm and 7 cm, or there is more than one tumor in the same lobe of the lung, or the cancer has grown into one or more of the following structures: the chest wall (the protective structure around the lungs and other organs in the chest), the outer lining of the chest cavity (the parietal pleura), the nerve close to the lung (phrenic nerve), and the outer covering of the heart (the pericardium).
- T4: The cancer is larger than 7 cm, or it is in more than one lobe of the lung, or it has spread into one or more of the following structures: the muscle below the lungs (the diaphragm), the area between the lungs in the middle of the chest (the mediastinum), the heart, a major blood vessel, the windpipe (trachea), the nerve that controls the voice box, the food pipe (esophagus), a spinal bone, and the area where the main airway divides to go to each lung.

N: nodal involvement

- N0: The lymph nodes do not contain cancer cells.
- N1: There are cancer cells in lymph nodes within the lung or in lymph nodes in the area where the lungs join the airway (the hilum).
- N2: There is cancer in lymph nodes in the center of the chest (mediastinum) on the same side as the affected lung or just under where the windpipe branches off to each lung.
  N3: There is cancer in lymph nodes on the opposite side of the chest from the affected lung, above the collarbone, or at the top of the lung.

M: distant metastasis

- M0: The cancer has not spread to another lobe of the lung or any other part of the body.
- M1a: One or more of the following: there is cancer in both lungs, there are areas of cancer in the lining around the lung or the lining around the heart, and there is fluid around the lung or heart that contains cancer cells—this is called a malignant pleural effusion or a malignant pericardial effusion.
- M1b: There is a single area of cancer outside the chest in an organ (such as the liver or brain) or a lymph node.
- M1c: There is more than one area of cancer in one or several organs.

2 RELATED WORK

2.1 Automatic TNM classification

Several studies have investigated clinical information extraction and automatic TNM classification from electronic medical records (EMR) using NLP. Rule-based methods and machine learning (ML) methods were used in these studies. Handling the diversity of expression, generalizability limitations of rule-based methods, and high computational cost of the BERT model have been reported. Liang Chen et al. developed a system to extract clinical information for hepatocellular carcinoma (HCC) from EMRs [4]. This study used hybrid methods, combining rule-based methods and ML methods. Consequently, the rule-based method demonstrated higher accuracy than the hybrid methods. The limitations of this study were the applicability to other diseases and validation on large datasets. Khushbu Gupta et al. investigated automatic TNM classification of lung cancer using a combination of rule-based methods and deep learning with long short-term memory (LSTM) [5]. The method addresses the limitations of rule-based approaches by incorporating LSTM-based deep learning and achieves up to 85% classification accuracy. Danqing Hu et al. investigated a system for automatically extracting lung cancer TNM staging information from CT reports [6]. Consequently, the BERT-the relation sign constraint (RSC) model achieved the highest accuracy with a Macro-F1 score of 97.13%.

2.2 Trends in LLM-based methods

The application of LLM has experienced remarkable growth in recent years. LLMs pre-training on a large amount of text data

enables high-precision inference even with limited training data. Additionally, the use of prompt engineering has been reported to improve classification accuracy. Danqing Hu et al. indicated the effectiveness of LLMs in predicting lymph node metastasis of lung cancer by integrating the outputs of ML models with LLMs using the patient clinical data and CT reports [7].

In addition, in fields with many technical terms, such as medicine, pre-training data alone may be insufficient, potentially decreasing classification accuracy. Thus, an approach has been proposed to retrieve information from a created database and supplementary missing information. Specifically, approaches such as k-NN retrieval and embedding-based retrieval exist [8, 9]. Combining these approaches with LLMs enables high-precision information extraction and classification.

3 METHODS

In this study, we worked on automating TNM classification using LLMs and investigated accuracy improvement by combining similarity search with embedding vectors and few-shot learning based on chain of thought (CoT). In all study conditions, we used the aforementioned TNM classification explanation as reference data.

3.1 Datasets

The details of the dataset are described in the overview paper published by the organizers [1]. This dataset consists of radiology reports written in a free-text format by nine radiologists based on lung cancer cases from Radiopeadia. The dataset includes lung cancer cases before initial treatment and does not contain post-treatment cases. The annotations were independently performed by two radiologists. The English corpus was split and distributed as follows.

- Train data: 12 cases (108 reports)
- Validation data: 6 cases (54 reports)
- Test data: 9 cases (81 reports)

Using these datasets, we investigated accuracy improvements in automated TNM classification.

3.2 LLM models

We used the OpenAI API key to compare the performance of three LLM models: GPT-4o-mini, GPT-4o, and o1-mini. We evaluated classification accuracy under the following three conditions for the models and examined the combination of the most accurate approach and the LLM model.

- Baseline: we created this model by referring to the baseline code distributed by the organizers. Predictions were made based on the guideline and a single case.
- Baseline + embedding: high-similarity radiology reports were retrieved using cosine similarity search based on embeddings and incorporated into the prompt.
- Baseline + embedding + few-shot learning: in addition to supplementing information using few-shot learning, CoT was utilized to encourage stepwise reasoning.

By comparing models, we evaluated the impact of LLM size and approach on TNM classification and examined the optimal combination.

3.3 Cosine similarity retrieval using embedding vectors

We aimed to improve classification accuracy by integrating additional medical knowledge into LLMs. In our system, we converted the training and validation data into embedding vectors using a sentence tokenizer and computed their similarity. Biomedical BERT (BioBERT) was used for sentence embedding. BioBERT is a model that has been additionally pre-trained using medical research papers from PubMed [8]. Hence, BioBERT is expected to capture medical characteristics more accurately than a general-purpose pre-trained model. Cosine similarity was used as the similarity search method. Cosine similarity measures the directional alignment between two vectors, indicating a higher similarity between the vectors' representations as the cosine value get closer to 1 [9]. For each validation instance, our approach retrieved the training examples with the highest similarity and incorporated them into the prompt. This approach enables LLMs to utilize previously unlearned knowledge and contextual information, potentially improving TNM classification accuracy.

3.4 Chain of thought (CoT) applied few-shot learning

To improve the inference accuracy of LLMs, we designed prompts to encourage reasoned inference and applied few-shot learning. Specifically, we aimed to improve classification accuracy over batch inference by designing prompts that output the inference process along with the predicted value. We designed the prompt format based on the baseline code provided by the organizers. Additionally, we analyzed the inference results and adjusted the prompt to prioritize incorporated reports from categories with low classification accuracy. The prompt consists of the following three components:

- Input: radiology report description
- Output: correct label
- Explanation: justification for the predicted value

An example of the prompt used is shown below.

Input:

{"A nodular infiltrative opacity of 30 mm in diameter is present in the right lower lobe S8, corresponding to known lung cancer. It is in extensive contact with the interlobar pleura, suggesting infiltration.

No findings suspicious of metastatic nodules are observed in both lungs.
Striated opacities are present in both lower lobes, likely inflammatory scarring.
No significant enlargement of the mediastinal or hilar lymph nodes, or other mediastinal organ lesions.
No pleural effusion.
No obvious abnormalities in the visualized upper abdomen."}
Output:
{"T2a"}
Explanation:
{"It is in extensive contact with the interlobar pleura, suggesting infiltration,"
but this infiltration is only within the lung lobes, and there is no mention of a secondary tumor nodule.
Also, since it is 30 mm in diameter and less than 40 mm, it is considered to be "T2a"."}

3.5 Evaluation metrics

For evaluation metrics, category accuracy was calculated for each label, and joint accuracy was calculated for all labels and the correct label. The details are cited from the task overview [1] and shown below.

- Joint accuracy (fine) - the proportion of radiology reports with accurate predictions for all the T, N, and M factors.
- T accuracy (fine) - the proportion of radiology reports with accurate predictions for the T factor.
- N accuracy (fine) - the proportion of radiology reports with accurate predictions for the N factor.
- M accuracy (fine) - the proportion of radiology reports with accurate predictions for the M factor.
- Joint accuracy (coarse) - joint accuracy that ignores distinctions between Tis/T1mi/T1a/T1b/T1c, T2a/T2b, and M1a/M1b/M1c.
- T accuracy (coarse) - T accuracy that ignores distinctions between Tis/T1mi/T1a/T1b/T1c and T2a/T2b.
- N accuracy (coarse) - identical to N accuracy (fine).
- M accuracy (coarse) - M accuracy that ignores distinctions between M1a/M1b/M1c.

Furthermore, in our analysis of joint accuracy, the improvement rate for each method relative to the baseline was calculated using the following formula:

$$Inprovement(\%) = \frac{A - B}{B} \times 100$$

where A represents the accuracy of the proposed methods, and B represents the accuracy of the baseline.

3.6 Experimental conditions

The experiment was conducted on Google Colaboratory using Python 3.11.11 and the OpenAI python package (version 1.60.0). Inference was performed on 54 reports of validation data, and the predicted values were output. Additionally, 108 training reports were used as retrieval information during inference, and classification accuracy was compared based on the presence or absence of each method. Finally, the training data and validation data were integrated, and a total of 162 cases were used as retrieval information. Inference was performed on the test data based on this integrated dataset.

4 RESULTS

We evaluated TNM classification accuracy using three approaches and three LLM models (GPT-4o-mini, GPT-4o, o1-mini). Two evaluation metrics, fine accuracy and coarse accuracy, were used to compute the results. Figures 1 and 2 show the results of three models and approach for accuracy and coarse accuracy, respectively. Figure 3 shows the accuracy improvement over the baseline for joint accuracy in each model.

4.1 Fine accuracy

Fine accuracy represents the detailed accuracy of each factor in TNM classification. Figure 1 shows that T_accuracy_fine improves in all models with the introduction of embedding vectors and further improves with the addition of few-shot learning. In N_accuracy_fine, the effect of embedding vectors was evident, and GPT-4o achieved the highest accuracy. In M_accuracy_fine, the accuracy declined in some models with the addition of embedding vectors, but the addition of few-shot learning resulted in the highest accuracy. Embedding vectors and few-shot learning in o1-mini resulted in the highest joint accuracy. Figure 3 shows that few-shot learning contributed the most to accuracy improvement in all models. GPT-4o showed the highest improvement with few-shot learning.

4.2 Coarse accuracy

Coarse accuracy represents the accuracy of broad classification without considering finer distinctions. Figure 2 shows that T_accuracy_coarse improved with the introduction of few-shot learning, but declined in GPT-4o-mini. O1-mini achieved the highest accuracy. In N_accuracy_coarse, the introduction of embedding vectors improved accuracy, but the effect of few-shot learning varied across models. In M_accuracy_coarse, all models exhibited high accuracy, and the addition of approaches did not lead to considerable changes in accuracy. O1-mini achieved the highest joint_accuracy_coarse, and Figure 3 shows that the introduction of few-shot learning contributed to overall accuracy improvement.

4.3 Result of the private leaderboard

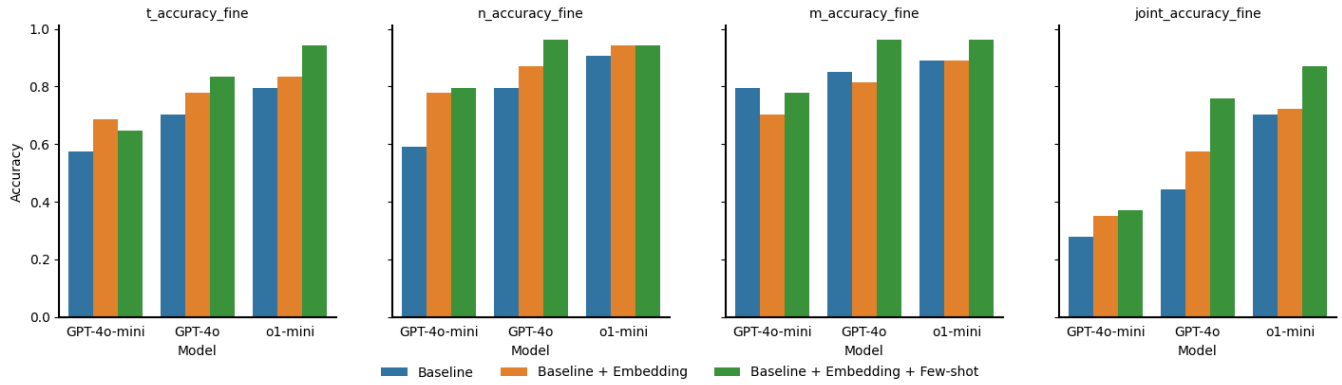

Fig.1 Accuracy (fine) for each factor and overall across the three LLM models.

Figure 1 shows that a total of three methods were evaluated using three models: GPT-4o-mini, GPT-4o, and o1-mini, and their classification accuracy, which was completely consistent across all factors, is presented. The details of the legend are provided below: Baseline: refers only to the guideline during inference. Baseline + embedding: add cosine similarity retrieve using embedding vectors to the baseline. Baseline + embedding + few-shot further incorporates few-shot learning in addition to the baseline.

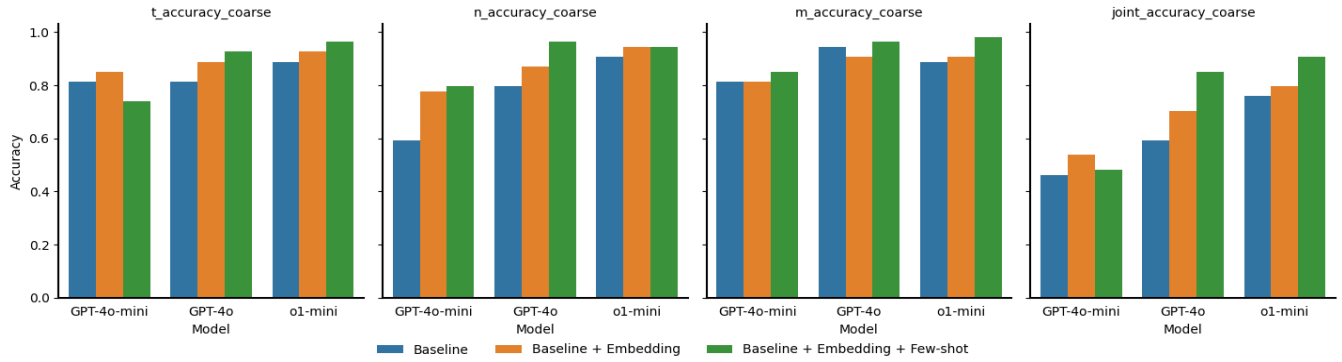

Fig.2 Accuracy (coarse) for each factor and overall across the three LLM models.

Figure 2 shows that a total of three methods were evaluated using three models: GPT-4o-mini, GPT-4o, and o1-mini, and their classification accuracy, where distinctions between factors were ignored, is presented. The legend is the same as in Figure 1.

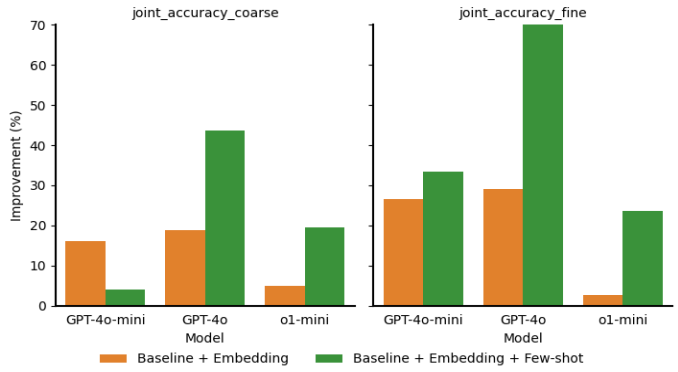

Fig.3 improvement rates of joint and coarse accuracy for each method relative to the baseline accuracy.

Figure 3 shows how the accuracy changed with each approach relative to the baseline accuracy.

In our approach, we ultimately integrated the training and validation data. A total of 162 cases were used as retrieve information, and inference was performed on the test data. The results of evaluation metrics on the private leaderboard are as shown below:

- Joint Accuracy (fine): 0.5185
- T Accuracy (fine): 0.6543

- N Accuracy (fine): 0.9259
- M Accuracy (fine): 0.8395
- Joint Accuracy (coarse): 0.5432
- T Accuracy (coarse): 0.6667
- N Accuracy (coarse): 0.9259
- M Accuracy (coarse): 0.8642

Joint accuracy was approximately 50% for both fine and coarse. Compared to joint_accuracy_fine on the validation data, the results showed a decline of over 30%. Among the categories, T class had the lowest classification accuracy, showing the same trend as in the validation data.

5 DISCUSSIONS

In this paper, we demonstrated that retrieving similarity cases using embedding vectors and applying few-shot learning contributed to improving accuracy in automated TNM classification. O1-mini achieved the highest accuracy in the three models but had the smallest accuracy improvement by embedding vectors and few-shot learning. Additionally, GPT-4o showed the highest accuracy improvement rate across all methods. This result may be attributed to the high inference ability of o1-mini [10]. However, since the actual number of parameters and the amount of training data were not disclosed, it is difficult to determine to what extent model size and training scale contribute to classification accuracy.

The evaluation result on the test data showed a substantial decline compared to the validation data. In particular, the lower accuracy in T-classification is related to the main factor contributing to the overall accuracy decline. This is because the T-factor requires the most detailed categorization within the TNM classification. To classify the T factor, LLMs must accurately interpret detailed information such as tumor size, involved organs, and extent of involvement. Thus, classification may be difficult based solely on information from radiology reports. For example, we observed cases where LLMs made unnecessary inferences and reached incorrect predictions despite the tumor size being explicitly stated in the reports. Additionally, in cases where the description of tumor invasion was ambiguous, LLMs failed to process the information accurately [11]. In this prompt design, we applied a unified format to each TNM factor. However, optimizing the prompt design specifically for the T factor is necessary to address these challenges. Specifically, one possible approach is to explicitly describe additional information regarding the extent of involvement based on guidelines. In addition, incorporating not only correct cases but also incorrect cases and their reasons into few-shot learning may help prevent LLMs from overinterpreting [11].

Our approach achieved high accuracy on validation data. On the other hand, the test data showed a substantial decline in accuracy, highlighting issues with the models' generalization performance. This is likely due to the limited retrieval dataset of 162 cases and the fact that the cases included in the few-shot learning prompt were manually selected. This tendency may also be attributed to the effect of LLMs overfitting to specific data [12]. Utilizing more diverse data may be an effective approach to addressing this issue. For example, an ensemble method that combines the outputs of multiple LLMs [12] or adding pseudo-reports with different writing styles as retrieval information could be considered.

Furthermore, o1-mini achieved the highest accuracy in our approach, but API costs and inference time remain challenges when considering practical deployment. O1-mini's API cost is 20 times higher, and its inference time is 10 times longer compared to GPT-4o-mini. Therefore, in implementation environments that process large amounts of data, cost performance may become a challenge.

To address this challenge, it is necessary to explore methods that use lightweight LLMs, such as GPT-4o-mini while keeping costs down [13]. Specifically, lightweight models can be used for the N and M factors, which have fewer classification categories. Meanwhile, high-accuracy inference can be performed using o1-mini only for the T factor, leading to overall cost reduction.

6 CONCLUSIONS

In our approach, we attempted to automate TNM classification using LLMs and demonstrated that cosine similarity retrieval using embedding vectors and few-shot learning contributed to improving accuracy. Specially, o1-mini achieved highest accuracy; however, API cost and inference time remain challenges. Additionally, the test data results showed a decline of over 30%, confirming that the classification accuracy of the T factor was a major issue. In future studies, it will be necessary to optimize prompt designs to improve T factor classification accuracy and ensure data diversity to enhance generalization performance. Furthermore, to reduce the burden of implementation, we should explore the use of lightweight LLMs and optimize model selection based on classification difficulty.

REFERENCES

[1] Nakamura, Y., Fujimoto, K., Kluckert, J., Krauthammer, M., Uszch, et al. NTCIR-18 RadNLP 2024 Overview: Dataset and Solutions for Automated Lung Cancer Staging.

[2] The Japan Lung Cancer Society. General Rule for Clinical and Pathological Record of Lung Cancer (8th ed.). 2021.

[3] The Union for International Cancer Control. The TNM Classification of Malignant Tumours (8th ed.). 2016.

[4] Chen, L., Song, L., Shao, Y., Li, D., Ding, K. Using natural language processing to extract clinically useful information from Chinese electronic medical records. International Journal of Medical Informatics. 2019;124:6 - 12. doi:10.1016/j.ijmedinf.2019.01.004

[5] Gupta, K., Thammasudjarit, R., Thakkinstian, A. NLP Automation to Read Radiological Reports to Detect the Stage of Cancer Among Lung Cancer Patients. Proceedings of the 2019 Workshop on Widening NLP. 2019;138-141.

[6] 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 Medical Informatics. 2021;9(7). doi:10.2196/27955

[7] Hu, D., Liu, B., Zhu, X., Wu, N. The Power of Combining Data and Knowledge: GPT-4o is an Effective Interpreter of Machine Learning Models in Predicting Lymph Node Metastasis of Lung Cancer. arXiv. 2024. http://arxiv.org/abs/2407.17900

[8] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234‑1240. doi:10.1093/bioinformatics/btz682

[9] Yamagiwa, H., Oyama, M., Shimodaira, H. Revisiting Cosine Similarity via Normalized ICA-transformed Embeddings. arXiv. 2024. http://arxiv.org/abs/2406.10984

[10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., et al. Language Models are Few-Shot Learners. arXiv. 2020. http://arxiv.org/abs/2005.14165

[11] Chang, C.-H., Lucas, M. M., Lu-Yao, G., Yang, C. C. Classifying Cancer Stage with Open-Source Clinical Large Language Models. arXiv. 2024. https://doi.org/10.1109/ICHI61247.2024.00018.

[12] Zhu, M., Zhu, Z., Chen, S., Chen, C., Wu, B. Enhanced Few-Shot Class-Incremental Learning via Ensemble Models. arXiv. 2024. http://arxiv.org/abs/2401.07208

[13] Wei, Q., Cui, Y., Ding, M., Wang, Y., Xiang, L., et al. Performance Evaluation of Lightweight Open-source Large Language Models in Pediatric Consultations: A Comparative Analysis. arXiv. 2024. https://arxiv.org/abs/2407.15862