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[1], Koichi OKUDA[1], Shota HOSOKAWA[1], Taisei KOMODA[1], Tsudou WATANABE[1], Yasuyuki TAKAHASHI[1]

[1] Dept. of Radiation Science Hirosaki University, Graduate School of Health Sciences

Introduction

- Radiologists face an increasing workload due to the large volume of medical images that must be interpreted.
- The application of the TNM staging system is essential in cancer diagnosis and treatment planning,
- However, manual determination of TNM staging is time-consuming and places a significant burden on radiologists.

This study investigates methods to automatically predict TNM stage of lung cancer from radiology reports. (As part of the NTCIR-18 RadNLP2024 shared task).

Material and Method

1. Dataset

The dataset details are described in the overview paper published by the organizers [1].

The dataset consists of free-text **radiology reports** written by nine radiologists, derived from lung cancer cases available on Radiopaedia.

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

2. Approaches

We compared the accuracy of TNM classification following three approaches:

1. **Model Comparison:** We evaluated three OpenAI models: GPT-4o-mini, GPT-4o, and o1-mini.
2. **Retrieval-augmented prompting:** Similar reports were retrieved from the training data using cosine similarity search [2] and added to the prompt.
3. **Few-shot prompting:** Several examples from the training data were added to the prompt using the following format:
   - **Input**: Radiology report description
   - **Output**: Correct label
   - **Explanation**: Justification for the predicted value

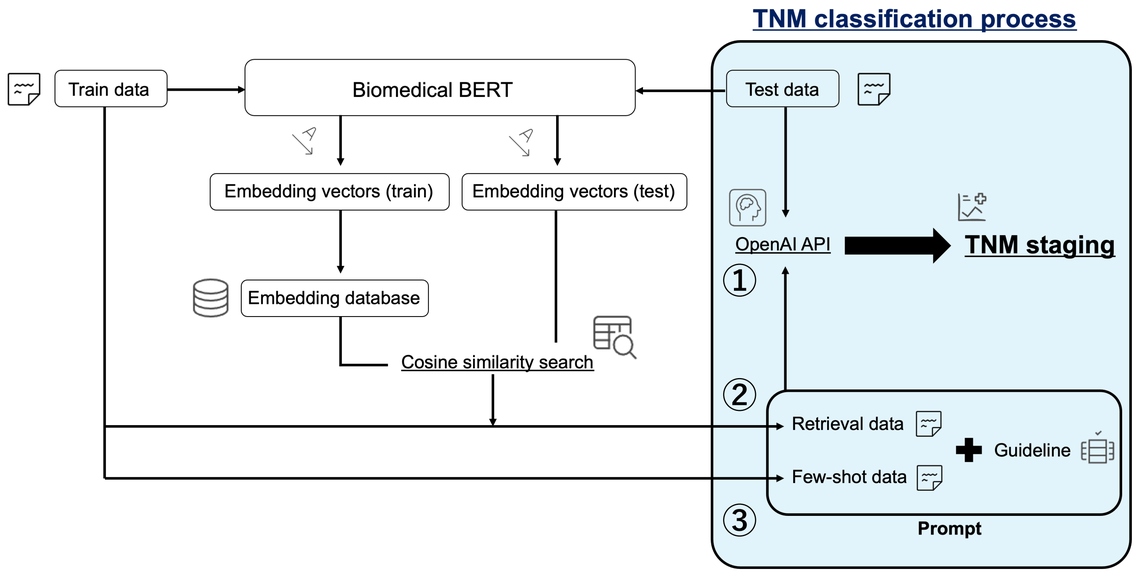

Fig.1 Workflow and component diagram for LLM-based TNM classification.

3. Evaluation metrics

- **Accuracy Metrics:**

  We evaluated performance using joint and individual factor accuracies (for T, N, and M) in **fine** and **coarse** settings.
  - **Fine**: Accurate prediction of all T, N, and M labels is required.
  - **Coarse**: Grouping of similar labels by ignoring distinctions such as Tis/T1mi/T1a/T1b/T1c, T2a/T2b, and M1a/M1b/M1c is permitted.

- **Improvement Rate:**

  The percentage of accuracy improvement over the baseline is calculated.

  $$Improvement\ (\%) = ((Proposed - Baseline) / Baseline) \times 100$$

Result

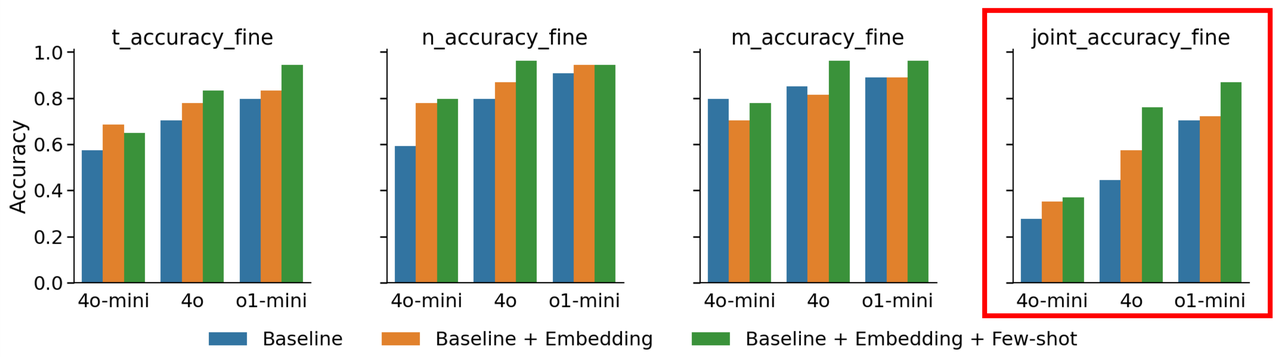

Fig.2 Accuracy (fine) for each factor and overall across the three LLM models.

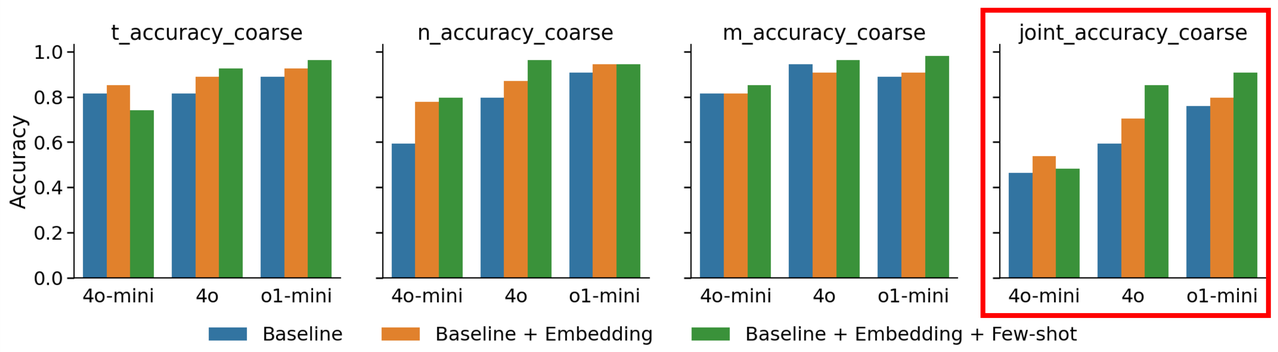

Fig.3 Accuracy (coarse) for each factor and overall across the three LLM models.

- **Baseline**: refers only to the guideline during inference.
- **Baseline + embedding**: add cosine similarity-based retrieve to the baseline.
- **Baseline + embedding + Few-shot**: further incorporates few-shot learning in addition to the baseline.

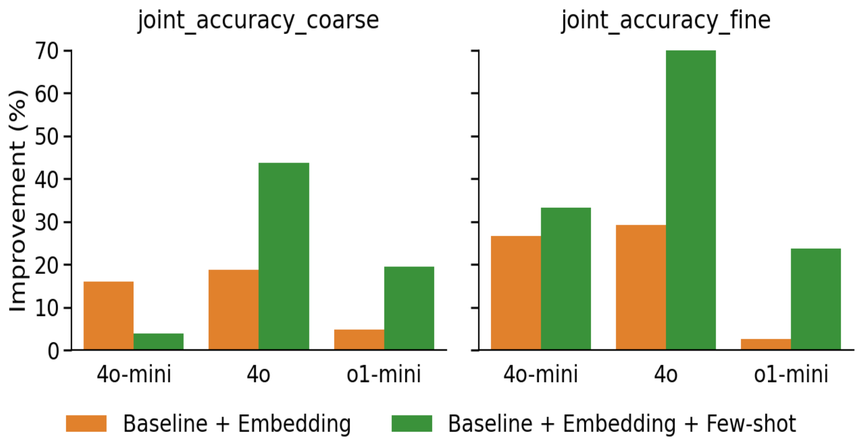

Fig.4 Improvement of joint and coarse accuracy for each method relative to the baseline accuracy.

The highest joint accuracy was achieved by the o1-mini model combined with embedding and few-shot method.

On the private leaderboard, the final joint accuracy was 51.9%.

- Consistent with the validation set, **the T factor remained the most challenging** to classify.
- The significant decrease in performance from the validation to the test datasets highlights an important issue for future research

Conclusion

- We demonstrated that LLMs with embedding-based similar report retrieval and few-shot learning improve the accuracy of automated TNM classification.
- Although the o1-mini model achieved the highest accuracy, issues regarding API cost and inference time remain.
- Low classification accuracy for the T factor was the primary issue, causing a significant performance drop on the test data.

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