OURad at the NTCIR-18 RadNLP Task: Predicting Lung Cancer Clinical Staging from Radiology Reports Using Few-Shot Prompting of Large Language Models

Junya Sato, Kosuke Kita, Daiki Nishigaki, Miyuki Tomiyama, Masatoshi Hori Artificial Intelligence in Diagnostic Radiology, the University of Osaka Graduate School of Medicine

Introduction :

In the RadNLP shared tasks, we compared various LLMs and BERT-based models to evaluate their performance on this task.

Dataset :

Main task :

Radiology reports are paired with TNM stages (8th edition), predicting T, N, and M; accuracy is based on all three being correct. The dataset includes 108 training, 54 validation, and 216 test samples.

Sub task :

The same reports are used, with each sentence labeled into one or more of eight clinical categories. These categories include Omittable, Measure, Extension, Atelectasis, Satellite, Lymphadenopathy, Pleural, and Distant.

Methods :

GPT model :

We employed two GPT models, GPT-4o (gpt-4o-2024-11-20) and GPT-o1 (o1-preview-2024-09-12).

We first applied zero-shot prompting to the training data, then selected only the misclassified cases to use as few-shot examples for subsequent prompts.

The prompt for the main task (English translation)

You are an excellent physician who can always make accurate judgments about lung cancer based on the following text. Please select the appropriate TNM classification from the given text according to the following staging criteria.

[TNM Classification Definition]

Read the text carefully and select the correct TNM classification. Ensure that your output strictly follows this format.

Insert a single space between T and N, and between N and M. Do not output anything else.

T<number>[optional_letter] N<number>[optional_letter] M<number>[optional_letter]

The prompt for the sub task (English translation)

You are an excellent physician who can always make accurate judgments about lung cancer based on the following text. Consider the entire text and determine which of the following labels apply to the specified target sentence.

[Definition of Text Classification] [Full Text to be Considered] [Target Sentence]

For each topic below, output "1" if applicable, or "0" if not. 1. omittable 2. measure 3. extension 4. atelectasis 5. satellite 6. lymphadenopathy 7. pleural 8. distant

Output format: Output a sequence of 0s and 1s separated by spaces.

Example: 1 0 0 0 0 0 0 0

Output:

BERT-based training :

We used two pretrained BERT-based models: UTH-BERT and DeBERTa-v3-large.

UTH-BERT is a domain-specific model developed by the University of Tokyo, pretrained on a large corpus of Japanese medical texts. It is optimized for understanding medical terminology and context, making it well-suited for clinical NLP tasks.

DeBERTa-v3-large, developed by Microsoft, is a general-purpose transformer model with disentangled attention and an improved mask decoder, enhancing contextual comprehension. Although not trained on medical data, it serves as a strong baseline for comparison.

Text preprocessing was model-specific:

UTH-BERT used a MeCab-based WordPiece tokenizer.

DeBERTa used a SentencePiece tokenizer.

To preserve contextual information, stopwords and punctuation were retained, and no additional normalization was applied beyond the tokenizer's default behavior.

Postprocessing :

We applied postprocessing to validate and correct LLM-generated TNM outputs using a regular expression. If the output matches the expected format (e.g., “T1a N0 M0”), it is accepted; otherwise, a valid substring is extracted.

Implementation detail :

All analyses were conducted in Python (v3.11.2). We used LangChain for GPT prompting and Unsloth for fine-tuning open-source LLMs on a single NVIDIA A6000 GPU. Code is available at: https://github.com/ai-radiol-ou/radnlp2024.

Results and Discussions :

This table shows the performance of each model on the validation dataset for the main task.

Model	Setting	Fine-Joint	Fine-T	Fine-N	Fine-M	Coarse-Joint	Coarse-T	Coarse-N	Coarse-M
GPT-o1	Zero-shot	0.907	0.944	0.963	1.000	0.926	0.963	0.963	1.000
	Few-shot	0.926	0.944	0.981	0.981	0.963	0.963	0.981	1.000
GPT-4o	Zero-shot	0.759	0.889	0.963	0.870	0.796	0.925	0.963	0.907
	Few-shot	0.759	0.833	0.926	1.000	0.889	0.963	0.926	1.000
Gemma-2 2b-jpn	Zero-shot	0.074	0.074	0.537	0.537	0.352	0.370	0.537	0.667
	Finetuned	0.315	0.481	0.889	0.852	0.463	0.648	0.889	0.852
Llama-3.1 Swallow-8B	Zero-shot	0.000	0.019	0.481	0.500	0.000	0.019	0.481	0.500
	Finetuned	0.519	0.759	0.926	0.759	0.741	0.926	0.926	0.833
Llama-3.1 Swallow-70B	Zero-shot	0.278	0.389	0.833	0.889	0.537	0.685	0.833	0.981
	Finetuned	0.556	0.704	0.981	0.870	0.852	0.889	0.981	0.981
DeBerta	-	0.204	0.389	0.833	0.722	0.407	0.556	0.833	0.796
UTH-bert	-	0.204	0.426	0.815	0.759	0.241	0.426	0.815	0.815

Based on the validation results for both tasks, we selected the GPT-o1 few-shot prompting model as our final submission model. The main task fine joint accuracy was 0.732, and the sub task overall F2.0 score was 0.688; both rankings placed 3rd on the private leaderboard.

This table shows the performance of each model on the validation dataset for the sub task.

Model	Setting	Overall	Inclusion	Measure	Extension	Atelectasis	Satellite	Lymphadeno pathy	Pleural	Distant
GPT-o1	Zero-shot	0.804	0.904	0.589	0.857	0.890	0.840	0.960	0.878	0.854
	Few-shot	0.850	0.952	0.851	0.868	0.825	0.694	0.865	0.870	0.736
GPT-4o	Zero-shot	0.747	0.840	0.579	0.855	0.860	0.800	0.855	0.859	0.664
	Few-shot	0.735	0.847	0.585	0.718	0.879	0.788	0.926	0.885	0.639
Gemma-2-2b-jpn	Zero-shot	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Finetuned	0.163	0.101	0.062	0.161	0.438	0.078	0.597	0.000	0.000
Llama-3.1 Swallow-8B	Zero-shot	0.270	0.030	0.000	0.000	0.602	0.161	0.806	0.653	0.000
	Finetuned	0.882	0.944	0.970	0.864	0.899	0.777	0.963	0.800	0.761
Llama-3.1 Swallow-70B	Zero-shot	0.685	0.894	0.417	0.778	0.851	0.839	0.884	0.685	0.744
	Finetuned	0.891	0.932	0.966	0.845	0.889	0.960	0.955	0.837	0.708
DeBerta	-	0.803	0.945	0.888	0.885	0.889	0.438	0.928	0.898	0.498
UTH-bert	-	0.757	0.946	0.878	0.841	0.805	0.432	0.903	0.689	0.500

Conclusions :

Our GPT-o1-based few-shot prompting outperformed BERT models and Japanese LLMs in extracting information from radiology reports. These results suggest that LLMs, with prompting or fine-tuning, can understand complex medical texts and may eventually match or surpass human performance.