ATILF at NTCIR-18 RadNLP 2024 Shared Task: With less radiology reports, comes less performance

Aman Sinha, Ioana Buhnila

NTCIR-18: Proceedings of the 18th NTCIR Conference on Evaluation ofInformation Access Technologies, June 10-13, 2025, Tokyo, Japan
DOI: https://doi.org/10.20736/0002002077

ATILF at NTCIR-18 RadNLP 2024 Shared Task: With less radiology reports, comes less performance

Aman Sinha and Ioana Buhnila
ATILF UMR 7118 (CNRS-Université de Lorraine)
France
{firstname.lastame}@univ-lorraine.fr

ABSTRACT

We present our results on the main task and subtask of the NTCIR-18 RadNLP 2024 shared task on the English language. We tested to what extent Large Language Models (LLMs) and Pretrained Language Models (PLMs) can identify and classify tumor types and subtypes. Our results for the main task showed that LLMs have difficulties in understanding different subtypes of tumors. For the tumor sentence segment classification subtask, we obtained competitive overall score with pretrained language models with an overall score of 0.83 for micro F2.0 metric. Our results showed that in low amount of data setting, we have a better chance with clinical PLMs in comparison to general and domain specific LLMs. Providing additional information such definitions in case clinical staging classification can help LLMs achieve better scores on fine-grained classification.

KEYWORDS

LLM, clinical PLM, radiology report, tumor identification and classification

TEAM NAME

ATILF

SUBTASKS

Multi-label classification subtask (English)

1 INTRODUCTION

Recently, the shift of NLP community towards LLMs has shown how any language related task can be solved with generative language modeling capabilities. Large Language Models (LLMs) have shown effectiveness in various language understanding task, such as summarization, sentiment classification or natural language understanding [13, 18, 21]. Traditional deep learning approaches are data driven and therefore medical tasks such as radiology report classification remains challenging for LLMs. Therefore, domain specific areas such as medical domain NLP tasks require fine-tuning or domain adaptation [7, 17, 22].

Medical domain NLP tasks often suffer from low resource constraints because of the small amount of existing annotated data. Therefore, the translation of existing effective language understanding benchmark becomes difficult. Existing approaches include augmenting the available medical data set via text augmentation techniques [3] or including other publicly available data set for pre-training [20].

split →	TRAIN	DEV	TEST
en	108	54	81
jp	108	54	81

Table 1: RR-TNM Dataset description

In spite of their capabilities, large language models are prone to generate factually incorrect information which can be very problematic in risk based settings such medical domain. The phenomenon of generating incorrect information is called hallucination. Recent approaches such as retrieval augmentation generation (RAG) [10] has found to be effective to avoid hallucinations through self-reflection for medical reasoning [2, 6].

In this work we investigated the performance of general and medical pretrained Large Language Models performance on radiology report identification and classification in English. Further, we study this approach for its reliability in the medical domain. The main contributions of this work are: 1) benchmarking pretrained language models and LLMs for TNM clinical staging classification and sentence segmentation classification task. 2) our proposed system achieves an overall score of 0.76 on the classification subtask.

2 DATASET & TASK DESCRIPTION

The dataset is part of RadNLP challenge [14]. It contains two languages : en and jp, where the en version of the dataset in machine translated version of jp dataset. The dataset statistics is shown in Table 1. The RadNLP challenge involves two tasks, namely, one maintask and one subtask. The **Main task** aims to automatically determine the clinical staging[1] (i.e., the degree of progression) of lung cancer from radiology reports and involves multi class classification for three tumor related questions (T, N, M) where T stands for assessment of the size and/or extension of the primary lesion; N denotes assessment of the extent of lymph node metastasis; and M stands for assessment of the extent of distant metastasis. The individual T,N, and M are subdivided into fine grain classes. T has 10 sub classes, N has 4 sub classes, and M has 4 sub classes. The **Sub task** includes a multi-label sentence binary classification where the aim is to identify up to eight spans related to the following topics: *Omittable* (free of any positive findings or clearly unrelated to lung cancer staging), *Measure* (the existence and diameter of the primary lesion), *Extension* (the range of the primary lesion's extension

[1]The staging criteria is aligned with the 8th edition of the TNM Classification of Malignant tumors by the Union for International Cancer Control (UICC)

Model	Joint Acc. (Fine)	T Acc. (Fine)	N Acc. (Fine)	M Acc. (Fine)	Joint Acc. (Coarse)	T Acc. (Coarse)	N Acc. (Coarse)	M Acc. (Coarse)
Llama3.2+P	0.0556	0.1667	0.2407	0.3889	0.0926	0.3519	0.2407	0.5740
BioMedllama+P	0.0556	0.2037	0.3333	0.5000	0.1296	0.2593	0.3333	0.6111
Llama3.2+D	0.0185	0.2592	0.2592	0.4814	0.0555	0.3333	0.2592	0.5556
BioClinicalBERT	0.1481	0.2407	0.3704	0.4444	0.2778	0.3889	0.3704	0.5000
BioBERT	0.0185	0.0741	0.4444	0.4259	0.2037	0.2778	0.4444	0.4815

(a) Maintask; (P: Prompting, D: Definitions)

Model	Overall	Inclusion	Measure	Extension	Atelectasis	Satellite	Lymphadenopathy	Pleural	Distant
BioBERT	0.7618	0.8643	0.8236	0.7409	0.6365	0.2692	0.8345	0.6267	0.6723
BioClinicalBERT	0.7239	0.8511	0.7378	0.6687	0.7425	0.0	0.8301	0.6266	0.6235
ClinicalBigbird	0.7530	0.8605	0.7940	0.7099	0.7426	0.5891	0.7896	0.5768	0.5410

(b) Subtask (Micro F2.0 Scores)

Table 2: Performance on VAL set.

outside the lung parenchyma), *Atelectasis* (atelectasis or obstructive pneumonia), *Satellite* (intrapulmonary metastasis or lymphangiomatosis carcinomatosa), *Lymphadenopathy* (enlarged regional lymph nodes), *Pleural* (pleural/pericardial effusion/dissemination) and *Distant* (distant metastasis outside the lung parenchyma).

3 RELATED WORK

This section briefly discusses the systems submitted in the previous edition of this challenge. Three teams presented 7 systems, from which the top three solutions included a ChatGPT based approach [15], an Open-Calm-7B model based approach [8], and a pre-trained language model based approach [5]. Each of the three teams worked only on Japanese language. Last year's edition only contained **maintask**. Overall, the result showed that category N and M was better modeled by the different methods than category T. Further, we noticed that the joint classification of tumor metastasis still requires better modeling.

4 METHODOLOGY

We prompted different Large Language Models with a small number of parameters to allow easy implementation in medical structures and lower computational cost [16]. For both RadNLP tasks, we tested two recent and efficient small language models (SLM): a general language SLM `llama3.2` (3B parameters) [4], and a clinical/medical LLM `BioMed-LLaMa3` (8B parameters) [19]. We also explored the performance of three pretrained clinical language models (PLM), `BioBERT` [9], `Bio-ClinicalBERT` [1], and `ClinicalBigbird` [11]. We detail our method for each task below.

4.1 Main Task (English Track)

The **main task** comprises of TNM clinical staging classification. We considered the following candidates:

(1) Prompt-based LLM baseline, where we provide as an input : prompt + radiology report to an LLM.
(2) Prompting LLMs with TNM classification tags (i.e. *'T0', 'Tis', 'T1mi'*) and definitions, such as "t_{desc} = "Assessment of the size and/or extension of the primary lesion."
(3) Joint prediction with clinical PLMs using deep learning on 20 epochs.

4.2 Sub Task (English Track)

The **sub task** comprises of sentence segmentation classification. We consider unsupervised clustering and most frequent labels as our baseline for the validation set. Our submitted system includes a collection of separate binary PLM classifier for each of the 8 categories. For the choice of PLM, we experimented with BioBERT [9], Bio_ClinicalBERT [1], and ClinicalBigbird [11]. To train the models, we use focal loss [12] to account for the class imbalance between 8 categories :

$$L_{\texttt{fl}} = -\alpha_t (1 - \hat{p}_{i,y_i})^{\gamma} \texttt{log}(\hat{p}_{i,y_i}) \tag{1}$$

where $\hat{p}_{i,y_i}$ is the predicted probability for the true class; α_t is the weighting factor for the true class to address class imbalance; γ is the focusing parameter to reduce the contribution of "easy" examples (default set to 2.0).

For the final submission on test set, for the **maintask**, we submitted predictions from `BioClinicalBERT` which was jointly trained for TNM clinical staging prediction simultaneously. For the **subtask**, we submitted an ensemble of three PLMs (BioBERT, BioClinicalBERT, ClinicalBigbird).

5 RESULTS & DISCUSSION

In Table 2, we report the scores for the systems we submitted for public leaderboard and in Table 3 we report the final scores we obtain on test set in private leaderboard. The **maintask** remains challenging for all language models tested. Our best results on the validation set were obtained with the clinical PLM `BioClinicalBERT` were the fine-grained joint accuracy is 0.14 and coarse joint accuracy is 0.27. The tumor size/extension classification class (T) remains the most difficult task for the LLMs. The best results on the T class fine and coarse accuracy were obtained with `llama3.2`+definition on fine grained accuracy (0.2592), and by `BioClinicalBERT` on coarse accuracy (0.3889). For the extent of lymph node metastasis class (N class) we obtained the highest accuracy score of 0.44 with `BioBERT`. The three clinical PLM models

Model	Joint Acc. (Fine)	T Acc. (Fine)	N Acc. (Fine)	M Acc. (Fine)	Joint Acc. (Coarse)	T Acc. (Coarse)	N Acc. (Coarse)	M Acc. (Coarse)
Best	0.6543	0.7037	0.9136	0.8889	0.6914	0.7407	0.9136	0.9136
Mean	0.4105	0.5277	0.7955	0.7646	0.4976	0.5972	0.7955	0.8202
Median	0.5247	0.6172	0.9012	0.8395	0.5617	0.6605	0.9012	0.8827
Ours	0.284	0.4815	0.7531	0.7654	0.4815	0.6296	0.7531	0.9012

(a) Maintask

Model	Overall	Inclusion	Measure	Extension	Atelectasis	Satellite	Lymphadenopathy	Pleural	Distant
Best	0.9433	0.9719	0.844	0.7512	0.8407	0.8631	0.9886	0.9615	0.9122
Mean	0.8545	0.9208	0.7062	0.6677	0.7498	0.5734	0.8824	0.8313	0.7335
Median	0.9005	0.94	0.7748	0.7373	0.8212	0.6886	0.9677	0.9240	0.7990
Ours	0.8353	0.9332	0.6808	0.7005	0.7212	0	0.9387	0.9615	0.7317

(b) Subtask (Micro F2.0 Scores)

Table 3: Performance on TEST set.

were the most efficient on this class, while llama3.2, a general model, achived lower accuracy scores (0.25). The M class (extent of distant metastasis) was the easiest for all LLMs tested in terms of coarse accuracy. The best model was `BioMedllama`+prompting, with a coarse accuracy of 0.61, followed closely by the general language LLM, `llama3.2`+prompting (0.57) and `llama3.2`+definition (0.55). In terms of fine-grained accuracy on the M class, the best performing model remains `BioMedllama`+prompting (0.5). Our addition of classes and subclasses definitions with `llama3.2` helps improve the fine-grained score with +0.10 on the M class (0.48) compared to simple prompting (0.38).

Consequently, we noticed an improvement in the fine-grained scores for the other two classes as well (N and T) when adding definitions in the prompt. We see a +0.09 improvement in the score for T, the most difficult class, and by +0.10 for the M class, when using definitions. This shows the importance of giving textual explanations and definitions of the class tags to the language model, as these labels are very short and coded (*M0, T2b, Tis*). Lexical and semantic matching with the definitions provided helped language models perform better semantic disambiguation.

The **subtask** involving sentence segment classification was effectively handled by pre-trained language models. Inspite of class imbalance across the 8 categories with less than 12% positive examples for 'extension', 'atelectasis', 'satellite', 'lymphadenopathy', 'pleural', 'distant'. The models were able to obtain more than 60% with an exception of 'satellite'. This was further consistent in test set where our submitted ensemble obtained 0 score for 'satellite' category.

6 CONCLUSIONS

We present in this work, our attempt to investigate the effectiveness of LLMs and PLMs for TNM clinical staging and sentence segment classification. Overall, we noticed that given low amount of data setting, domain specific PLMs are better choice for modeling radiology clinical staging and sentence segmentation in comparison to LLMs despite of provided additional information such definitions in case clinical staging classification.

REFERENCES

[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* (2019).

[2] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511* (2023).

[3] Jaime Collado-Montañez, María-Teresa Martín-Valdivia, and Eugenio Martínez-Cámara. 2025. Data augmentation based on large language models for radiological report classification. *Knowledge-Based Systems* 308 (2025), 112745.

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[5] Takuya Fukushima, Yuka Otsuki, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2023. NAISTSOCRR at the NTCIR-17 MedNLP-SC Radiology Report Subtask. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17. National Institute of Informatics (NII).* https://api.semanticscholar.org/CorpusID:266493244

[6] Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving Medical Reasoning through Retrieval and Self-Reflection with Retrieval-Augmented Large Language Models. *arXiv preprint arXiv:2401.15269* (2024).

[7] Jun Kanzawa, Koichiro Yasaka, Nana Fujita, Shin Fujiwara, and Osamu Abe. 2024. Automated classification of brain MRI reports using fine-tuned large language models. *Neuroradiology* (2024), 1–7.

[8] † KojiFujimoto, Morteza Rohanian, Fabio Rinaldi, Mizuho Nishio, Farhad Nooralahzadeh, Chikako Tanaka, and Michael Krauthammer. 2023. Classification of cancer TNM stage from Japanese radiology report using on-premise LLM at NTCIR-17 MedNLP-SC RR-TNM subtask. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17. National Institute of Informatics (NII).* https://api.semanticscholar.org/CorpusID:266378384

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[10] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[11] Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838* (2022).

[12] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision.* 2980–2988.

[13] Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Richard Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024. On Learning to Summarize with Large Language Models as References. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).* 8639–8656.

[INST]
Given this radiology report : [rr_text],

classify it to [var]-clinical staging based on [desc], provided your options are as follows: option_list. **GIVE ONLY THE CORRECT ANSWER** (Pick only one option).
/[/INST]

Figure 1: Default prompt for TNM clinical staging

[14] Yuta Nakamura, Koji Fujimoto, Jonas Kluckert, Michael Krauthammer, Jun Kanzawa, Akira Katayama, Tomohiro Kikuchi, Ryo Kurokawa, Wataru Gonoi, Yuki Tashiro, Shouhei Hanaoka, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2024. NTCIR-18 RadNLP 2024 Overview: Dataset and Solutions for Automated Lung Cancer Staging. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-18. National Institute of Informatics (NII).*

[15] Mizuho Nishio, Hidetoshi Matsuo, Takaaki Matsunaga, Koji Fujimoto, Morteza Rohanian, Farhad Nooralahzadeh, Fabio Rinaldi, and Michael Krauthammer. 2023. Zero-shot classification of TNM staging for Japanese radiology report using ChatGPT at RR-TNM subtask of NTCIR-17 MedNLP-SC. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR*, Vol. 17.

[16] Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118* (2020).

[17] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine* 30, 4 (2024), 1134–1142.

[18] Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 6095–6104.

[19] Nour Eddine Zekaoui, Mounia Mikram, Maryem Rhanoui, and Siham Yousfi. 2025. BioMed-LLaMa-3: Instruction-Efficient Fine-Tuning of Large Language Models for Improved Biomedical Language Understanding. In *Multi-disciplinary Trends in Artificial Intelligence*, Chattrakul Sombattheera, Paul Weng, and Jun Pang (Eds.). Springer Nature Singapore, Singapore, 399–410.

[20] Deshiwei Zhang, Xiaojuan Xue, Peng Gao, Zhijuan Jin, Menghan Hu, Yue Wu, and Xiayang Ying. 2024. A survey of datasets in medicine for large language models. *Intelligence & Robotics* 4, 4 (2024), 457–478.

[21] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment Analysis in the Era of Large Language Models: A Reality Check. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 3881–3906.

[22] Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112* (2023).

APPENDIX

A PROMPT FOR DIFFERENT LLMS

Default prompt. The default prompt (denoted by P, in table 2a) we use as below (Fig 1), where, option_list can be $t_{options}$ = ['T0', 'Tis', 'T1mi', 'T1a', 'T1b', 'T1c', 'T2a', 'T2b', 'T3', 'T4'] or $n_{options}$ = ['N0', 'N1', 'N2', 'N3'] or $m_{options}$ = ['M0', 'M1a', 'M1b', 'M1c']. The desc variable can be one of the following:

- t_{desc} = "Assessment of the size and/or extension of the primary lesion. "
- n_{desc} = "Assessment of the extent of lymph node metastasis."
- m_{desc} = "Assessment of the extent of distant metastasis."

Definition prompt. The definition prompt (denoted by D, in table 2a) we used with definition is shows in Figure 2.

The TNM staging system is a method for classifying the extent of cancer spread. It stands for Tumor, Node, and Metastasis, and is used to describe the size of the primary tumor (T), the extent of lymph node involvement (N), and the presence of distant metastasis (M).

Tumor (T): Describes the size and extent of the primary tumor. The classes are:

['T0|No primary tumor', 'Tis|Ground-glass nodule without solid component with the total diameter ≤3 cm', 'T1mi|Ground-glass nodule with solid component ≤0.5 cm and the total diameter ≤3 cm', 'T1a|Solid component diameter ≤1 cm', 'T1b|Solid component diameter >1 cm and ≤2 cm', 'T1c|Solid component diameter >2 cm and ≤3 cm', 'T2a|Solid component diameter >3 cm and ≤4 cm. Otherwise, extension to main bronchus or visceral pleura, or atelectasis or obstructive pneumonia extending to hilum," with the solid component diameter <3 cm or unknown', 'T2b|Solid component diameter >4 cm and ≤5 cm', 'T3|Solid component diameter >5 cm and ≤7 cm. Otherwise, solid component diameter ≤5 cm and either condition holds: direct invasion of parietal pleura, chest wall (including superior sulcus tumor), mediastinal nerve, or pericardium; separate tumor nodule(s) in the same lobe', 'T4|Solid component diameter >7 cm. Otherwise, either condition holds: invasion of diaphragm, mediastinum, heart, great vessels, trachea, recurrent laryngeal nerve, esophagus, spine, or carina; tumor nodule(s) in a different ipsilateral lobe']

Node (N): Describes the degree of spread to regional lymph nodes. The classes are:

['N0|No regional lymph node metastasis', 'N1|Metastasis to ipsilateral peribronchial, hilar, or pulmonary lymph nodes, including direct invasion of the primary tumor', 'N2|Metastasis to ipsilateral mediastinal or subcarinal lymph nodes', 'N3|Metastasis to contralateral mediastinal, hilar, anterior scalene, or supraclavicular lymph nodes']

Metastasis (M): Indicates whether the cancer has spread to distant parts of the body. The classes are:

['M0|No distant metastasis', 'M1a|Contralateral tumor nodule(s), pleural or pericardial nodule(s), malignant pleural effusion, or malignant pericardial effusion', 'M1b|Single extrathoracic metastasis', 'M1c|Multiple extrathoracic metastases']

I want you to read the radiology report and assess it for clinical staging. Give me an answer for T, one for N, and one for M only using the classes mentioned above. Give only the answer, no additional text. Use this json format to answer 'T':" , 'N':" , 'M':" ; radiology report = ###

Figure 2: Prompt for TNM clinical staging with definitions