PolyuCBS at SMM4H 2024: LLM-based Medical Disorder and Adverse Drug Event Detection with Low-rank Adaptation

Yu Zhai[1], **Xiaoyi Bao**[1], **Emmanuele Chersoni**[1], **Beatrice Portelli**[2],
Sophia Yat Mei Lee[1], **Jinghang Gu**[1] **and Chu-Ren Huang**[1]
[1]The Hong Kong Polytechnic University, Hong Kong, China
[2]University of Udine & University of Naples, Italy
{tonyayu.zhai,xiaoyi.bao}@connect.polyu.hk,
portelli.beatrice@spes.uniud.it
{emmanuele.chersoni,ym.lee,jinghang.gu,churen.huang}@polyu.edu.hk

Abstract

This is the demonstration of systems and results of our team's participation in the Social Medical Mining for Health (SMM4H) 2024 Shared Task. Our team participated in two tasks: Task 1 and Task 5. Task 5 requires the detection of tweet sentences that claim children's medical disorders from certain users. Task 1 needs teams to extract and normalize Adverse Drug Event terms in the tweet sentence. The team selected several Pre-trained Language Models and generative Large Language Models to meet the requirements. Strategies to improve the performance include cloze test, prompt engineering, Low Rank Adaptation etc. The test result of our system has an F1 score of 0.935, Precision of 0.954 and Recall of 0.917 in Task 5 and an overall F1 score of 0.08 in Task 1.

1 Introduction

The rise in people using social media for health information has resulted in a significant increase in health-related data, which allows researchers to harness the information, along with NLP and Machine Learning methods, to contribute to public health (Klein et al., 2024). The 9th Social Media Mining for Health Applications (SMM4H) Shared Tasks, aiming to advance methods that utilise social media data for health research, have a special focus on Large Language Models (LLMs) and Generalizability for Social Media NLP.

There are 7 tasks given in the 9th SMM4H workshop (Xu et al., 2024). Our team focused on Task 1 and Task 5. Task 1 has two sub-tasks, which are (1) detecting Adverse Drug Event terms in tweets and (2) normalizing these colloquial mentions to their standard concept to Preferred Term IDs according to MedDRA. Task 5 focuses on classifying tweets reporting children's medical disorders. The challenge of this task is differentiating between tweets from pregnancy users who declared that their child has specific conditions with attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorders (ASD), delayed speech, or asthma, and tweets that just mention the disorder. The main challenges of both tasks are as follows: (1) medical terms conveyed in colloquial language, a common issue in social media data, which might misguide the trained model; (2) How to activate, transfer and utilize the knowledge learnt from the pre-training in Pre-trained Language Models(PLM) and Large Language Models (LLM). To address these issues, we use both PLMs and LLMs as basis and implement parameter-efficient tuning during training. To move a step further, we also conduct prompt engineering along with task decomposition to fully activate the knowledge of the LLM and ensure its understanding of the task, which could bridge the gap between the LLM and the specific downstream task.

2 Methodology

2.1 Task 1

For Task 1, BERT (Devlin et al., 2019) was utilized as the baseline model to conduct domain continual pre-training (Gu et al., 2021; Peng et al., 2021) and fine-tuning on the training dataset, during which BIO tags were utilized. In the extraction task, the models chosen are BERT and LLaMA-2(Touvron et al., 2023). The normalization task was regarded as a classification problem, the models chosen are SapBERT (Liu et al., 2021) and CODER (Yuan et al., 2022).

Fine-tune: For ADE extraction in Task 1, two fine-tuning modes are being experimented with: the first is by extracting the [CLS] in the last layer of the BERT model and utilizing the data tagged by BIO tag set to train the BERT classification model; the second is by treating the extraction task as a cloze test task. Given a label sets $Y = \{y_1, y_2, \dots y_k\}$, for each tweet sentence $\mathbf{S} = \{t_1, t_2, \dots t_n\}$ where t is the token and n

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 74–78
August 15, 2024 ©2024 Association for Computational Linguistics

is the number of tokens in S. For a token t_q, we create template $\mathbf{P} = \{p_1, p_2, \ldots p_m\}$ where $P = S + t_q + cloze$, eg. *A tweet sentence, a token in this sentence, this is __.* We mask p_i, the ith token in P, into [mask] and construct token-label converter v_y to ensure the label y_1 has a token set $C_{y_1} = \{c_1, c_2 ... c_n\}$ that mapping token c into label y_1. We extract the last layer's representation in the [mask] position to get the fixed-length labels. For a token t_n in a tweet, the possibility of its label equals to y_1 is as follows:

$$\Pr\left(y_1 \mid t_q\right) = \frac{\exp M\left(v_1 \,\middle|\, \frac{P}{\{p_i\}i}\right)}{\sum_{j=1}^{k} \exp\left(v_{y_j} \,\middle|\, \frac{P}{\{p_i\}i}\right)} \tag{1}$$

$\frac{P}{\{p_i\}i}$ means p_i in template P were replaced with [mask]. $M\left(v_1 \,\middle|\, \frac{P}{\{p_i\}i}\right)$ represents the probability that a masked token predicted by the model is mapped to the label y_1.

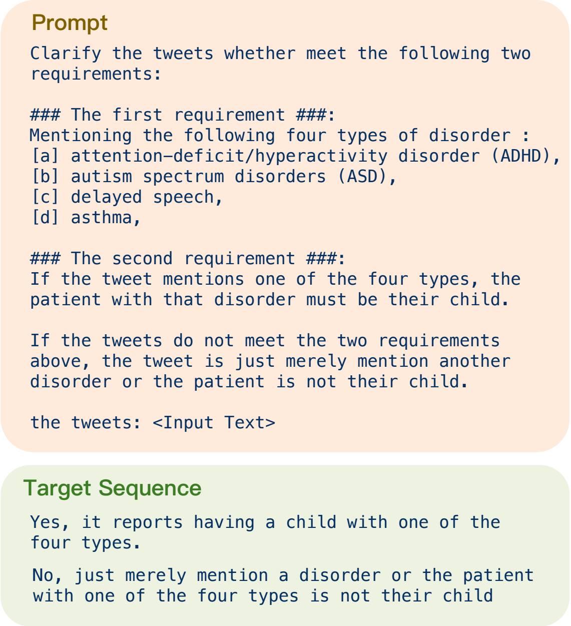
2.2 Task 5

In task 5, we first designed the prompt by several prompt engineering methods, then converted the original data into instructions by the designed template. These instructions were fed into LoRa adapters to finetune the LLaMA-2 model cost-effectively.

Low-Rank Adaptation (LoRA) (Hu et al., 2021) aims to improve the efficiency of fine-tuning large language models by training much smaller low-rank decomposition matrices of certain weights. Consider a weight matrix $\mathbf{W}_0 \in \mathbf{R}^{d \times k}$ from the pre-trained model. During training, the original weight matrix W_0 remains frozen and does not receive gradient updates. The trainable parameters are the matrices $\mathbf{B} \in \mathbf{R}^{d \times r}$ and $\mathbf{A} \in \mathbf{R}^{r \times k}$ ($r \ll \min(d, k)$), which represent the low-rank decomposition. The forward pass with LoRA is as follows:

$$W_0\, x + \Delta W x = W_0\, x + \mathbf{BA} x \tag{2}$$

Instruction Tuning. The importance of prompt templates has been demonstrated in various information extraction studies (Lu et al., 2021; Bao et al., 2022, 2023), particularly in the context of LLM. Task 5 presents a challenge due to its complex requirements. For example, Task 5 involves four types of disorders and requires that the report be from parents regarding their child. If we provide all the requirements in a single pass, LLMs may not accurately capture the semantic information. Based on the analysis above, we propose a decomposed prompt approach. As illustrated in Figure 1, we break down the task into independent units representing each requirement and fine-tune the LLM to address them individually. The first requirement guides the LLM to focus on determining the presence of the four specific diseases, while the second requirement ensures that the patients reported are children and that the reports come from their parents.



Prompt

```
Clarify the tweets whether meet the following two
requirements:

### The first requirement ###:
Mentioning the following four types of disorder :
[a] attention-deficit/hyperactivity disorder (ADHD),
[b] autism spectrum disorders (ASD),
[c] delayed speech,
[d] asthma,

### The second requirement ###:
If the tweet mentions one of the four types, the
patient with that disorder must be their child.

If the tweets do not meet the two requirements
above, the tweet is just merely mention another
disorder or the patient is not their child.

the tweets: <Input Text>
```

Target Sequence

```
Yes, it reports having a child with one of the
four types.

No, just merely mention a disorder or the patient
with one of the four types is not their child
```

Figure 1: Illustration of prompt and target sequence.

As shown in the penultimate paragraph of the prompt in Figure 1, the LLM is asked to make the final decision about whether the tweet fulfills the two requirements at the same time and if not, a No answer should be given. The tag <Input Text> will be replaced by the specific tweets before fed into the model. When it comes to the target sequence, they are proposed under a similar motivation to the prompt, intending to have the LLM aware of the reason for giving the choice. If the tweet cannot fulfill the requirements, then the second line of the target sequence should be given; otherwise, the first one.

3 Experiment Results

3.1 Task 1

In Task 1, tweets in training data were treated as continual pre-training data for the masked language modelling process. For both extraction and normalization tasks with PLMs, we run 30 epochs with a

	Validation(F/P/R)		Test(F/P/R)	
Model	Extraction	Norm-single	Extraction	Norm
BERT-cloze+SapBERT	0.32/0.32/0.32	0.713/0.709/0.718	0.024/0.013/0.136	0.039/0.021/0.224
BERT+SapBERT	0.51/0.49/0.54	0.713/0.709/0.718	0.010/0.005/0.053	0.044/0.024/0.243
LLaMA-2+CODER	**0.659**/0.687/0.632	**0.741**/0.741/0.741	**0.112**/0.062/0.633	**0.080**/0.044/0.450

Table 1: Task 1 overall results on extraction and normalization of the validation and test sets

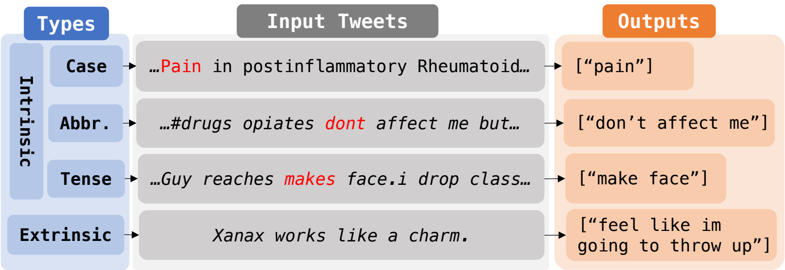

Types
Input Tweets
Outputs
Intrinsic
Case
…Pain in postinflammatory Rheumatoid…
["pain"]
Abbr.
…#drugs opiates dont affect me but…
["don't affect me"]
Tense
…Guy reaches makes face.i drop class…
["make face"]
Extrinsic
Xanax works like a charm.
["feel like im going to throw up"]

Figure 2: Hallucinations from LLaMA-2 in Task 1.

batch size of 4.

In Table 1, Norm task scores on the validation set were calculated based on the assumption that there is no error in the extraction task. For the extraction task, the basic BERT [1] classification method proposed in section 2.1 reached 0.51 in F score, while the cloze test method didn't get the best performance because of the weaker generation ability of encoder-only models. Therefore, LLaMA-2 was introduced.

On the validation set, the F1 score of the extraction saw an around 0.15 rise with LLaMA-2, but the test result still saw a drastic fall for both models (the Task1 test score with LLaMA-2 was uploaded during post-evaluation).

For LLaMA-2 predictions, except for around 1.4% results are empty, 25.9% of the answer in the test set suffers from hallucination problems. This section adopts hallucination categories of "Intrinsic Hallucination" and "Extrinsic Hallucination" (Zhou et al., 2021). Most hallucinations in Task 1 are intrinsic, taking up about 77.2% of the total hallucination errors.

In Figure 2, we demonstrated three common types of intrinsic hallucinations: Case, Abbreviation, and Tense. Some of the intrinsic hallucination happened due to the colloquial nature of the tweet sentence. There are 22.8% extrinsic hallucinations generated without clear grounding in the input. As shown in Figure 2, the model generated *"feel like im going to throw up"* which is irrelevant to the original input tweet. Hallucination issues might cause severe consequences when facing health-sensitive data. We used heuristic rules to correct part of the intrinsic hallucinations in this task and got a 0.015 improvement on F1. Another solution could be Retrieval Augmented Generation (Chen et al., 2024) which we will implement in future work.

3.2 Task 5

For our LLM, we employ LLaMA-2-7B[2] and LoRA fine-tune the adapter parameters. We tune the parameters of our models by grid searching on the validation dataset. We fine-tune the model with 20 epochs and save the model parameters for inference. The LoRA alpha is set to 128 and the LoRA rank is set to 64.

Model	Validation (F/P/R)	Test(F/P/R)
BERTweet	0.85/0.84/0.85	-
GPT-2	0.81/0.79/0.83	-
LLaMA-2	**0.932**/0.947/0.919	**0.935**/0.954/0.917
SMM4H Mean	-	0.822/0.818/0.838

Table 2: Task 5 results on validation and test sets

The model parameters are optimized by Adam (Kingma and Ba, 2015), and the learning rate of fine-tuning is 5e-5. The batch size is set to 4 with a cut-off length of 1024. The LoRA adapter would be merged with the original LLaMA-2-7B parameters and frozen during the inference process. During inference, we do the greedy search. Our experiments are carried out with an Nvidia RTX 4090 GPU. The model finally got an F1 score of 0.935 on the test set.

[1] Bert-base-uncased,https://huggingface.co/google-bert/bert-base-uncased

[2] LLaMA-2-7B-Chat,https://huggingface.co/meta-LLaMA/LLaMA-2-7b-chat-hf

4 Conclusion

In conclusion, this paper presents PLM and LLM-based methods for Task 1 and Task 5 in SMM4H 2024. Our team balanced efficiency and efficacy by employing strategies like cloze test shifting, instruction tuning, and low-rank adaptation. Through empirical evaluations, this paper proved that the LLM did surpass certain PLMs in the two tasks, though it suffered from the hallucination issue. More methods need to be explored to ensure solid and reliable results from LLM in tasks of the public health domain.

Acknowledgments

This research work is supported by a General Research Fund (GRF) project sponsored by the Research Grants Council, Hong Kong (Project No. 15611021).

References

Xiaoyi Bao, Zhongqing Wang, Xiaotong Jiang, Rong Xiao, and Shoushan Li. 2022. Aspect-based sentiment analysis with opinion tree generation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4044–4050. ijcai.org.

Xiaoyi Bao, Zhongqing Wang, and Guodong Zhou. 2023. Exploring graph pre-training for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3623–3634, Singapore. Association for Computational Linguistics.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ari Z Klein, Juan M Banda, Yuting Guo, Ana Lucia Schmidt, Dongfang Xu, Ivan Flores Amaro, Raul Rodriguez-Esteban, Abeed Sarker, and Graciela Gonzalez-Hernandez. 2024. Overview of the 8th social media mining for health applications (smm4h) shared tasks at the amia 2023 annual symposium. *Journal of the American Medical Informatics Association*, 31(4):991–996.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Is domain adaptation worth your investment? comparing BERT and FinBERT on financial tasks. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 37–44, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Sai Tharuni Samineni, and Graciela Gonzalez-Hernandez. 2024. Overview of the 9th social media mining for health applications (#SMM4H) shared tasks at ACL 2024. In *Proceedings of the 9th Social Media Mining for Health Applications Workshop and Shared Task*, Bangkok, Thailand. Association for Computational Linguistics.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of Biomedical Informatics*, 126:103983.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.