INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Enhanced Adverse Drug Event Extraction Using Prefix-Based Multi-Prompt Tuning in Transformer Models

Salisu Modi [a,b,1], Khairul Azhar Kasmiran [a,2], Nurfadhlina Mohd Sharef [a], Mohd Yunus Sharum [a]

[a] *Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang Selangor Darul Ehsan, Malaysia*
[b] *Department of Computer Science, Sokoto State University, Sokoto, Nigeria*
Corresponding author: [1] gs63125@student.upm.edu.my; [2] k_azhar@upm.edu.my

***Abstract*— Extracting mentions of adverse drug events and relationships between them is crucial for effective pharmacovigilance and drug safety surveillance. Recently, transformer-based models have significantly improved this task through fine-tuning. However, traditional fine-tuning of transformer models, especially those with many parameters, is resource-intensive, memory-inefficient, and often leaves a gap between pre-training and downstream task-specific objectives. Soft prompting is a lightweight approach that updates a trainable prompt to guide task-specific fine-tuning, showing comparable performance to traditional fine-tuning for large language models on simple tasks. However, its effectiveness on complex tasks like token-based sequence labeling requiring multiple predictions for a single input sequence remains underexplored, particularly in multi-task settings. In addition, using holistic prompts in multi-task learning settings may be biased to other subtasks. Additionally, some prompt tokens hurt the model prediction. This study proposes a prefix-based multi-prompt soft tuning method with attention-driven prompt token selection for tuning transformer models on multi-task dual sequence labelling for concept and relation extraction. We experimented with BERT and SciBERT models using frozen and unfrozen parameter strategies. Our approach achieved state-of-the-art performance on the n2c2 2018 and TAC 2017 datasets for adverse drug event extraction, with multi-prompt tuning in unfrozen models surpassing traditional fine-tuning. Moreover, it outperforms the largest clinical natural language processing model, GatorTron, on the n2c2 2018 dataset. This research highlights the potential of soft prompts in efficiently adapting large language models to complex downstream NLP tasks.**

***Keywords*— Adverse drug event; fine-tuning; multi-prompt; multi-task; soft prompt tuning.**

Manuscript received 5 Apr. 2024; revised 10 Aug. 2024; accepted 12 Sep. 2024. Date of publication 30 Nov. 2024.

I. INTRODUCTION

Adverse drug events (ADEs) refer to any harmful or unpleasant reactions that occur due to taking a medication. [1]. Consequently, accurate extraction of adverse drugs is vital for pharmacovigilance studies. In addition, it is significant to the information retrieval research paradigm due to the dual nature of named entity recognition and relation reaction. In the past decade, this task has been handled at different stages of drug usage. The notable stages are the pre-marketing and post-marketing. At the pre-marketing stage, the popular approach was through clinical trials with some volunteer patients, which seriously needed more volunteers. On the one hand, at the post-marketing stage, the spontaneous reporting system (SRS) was the earlier approach to collecting adverse drug event (ADE) cases from affected patients or clinicians, which suffers from underreporting, leading to automated approaches of natural language preprocessing (NLP) [2], [3].

The dual nature of adverse drug event extractions involving named entity recognition [4], [5], [6], and relation extraction [6], [7], [8], makes it a challenging task. The earlier approaches were rules-based [9], [10], machine-learning [11], [13], and deep learning [6], [14], [15] approaches [16]. However, the natural language processing paradigm has recently experienced a rapid increase in performance due to the prevalence of large language models (LLM) [15], [17], [18]. The de facto method of adapting the LLMs was through model fine-tuning. This approach works similarly to the traditional supervised learning approach, which requires much-annotated data to train the model on downstream-specific tasks. The approach is also a top-down through train-test workflow with all model-tuned parameters saved for inference. Thus, it is time-consuming, memory-inefficient and resource-intensive compared to prompting, especially for models with larger parameters [19].

Prompt tuning methods connect LLMs' pre-training objectives with specific downstream task objectives with an additional prompt. Prompting is a technique to adapt LLMs, where additional tokens guide the model for downstream tasks. There are two types of prompts: challenging prompts, which use non-trainable tokens, and soft prompts, which use trainable embeddings added to the input sequence. In prompt-based learning, different strategies are employed, including frozen, where LLM parameters are fixed, and unfrozen, where LLM parameters are updated during training.

Despite its potential, prompting remains in its early stages. A significant performance gap exists compared to fine-tuning, especially for small-size models when frozen, and has yet to be fully leveraged for complex natural language understanding tasks, such as sequence labeling tasks that require multiple predictions per input sequence and in multi-task learning scenarios.

To address these challenges, we propose a novel approach that utilizes multiple soft prompts, one for each task, with an attention-driven prompt token selection to optimize the prompt tokens. This multi-prompt soft prompt tuning method selectively highlights the most contributing prompt tokens, enabling more effective model adaptation to downstream tasks.

Extracting ADEs from the vast amount of unstructured clinical notes is highly significant in real-world settings, as it supports drug discovery and pharmacovigilance studies. Various approaches have been employed to improve this task. These include rule-based [9], machine learning [20], and deep learning [18], as well as adopting large language models (LLMs) through fine-tuning [2].

Prompt learning has emerged as a preferred method for adapting LLMs due to the limitations of traditional fine-tuning, which can be cumbersome, memory-intensive, and resource-heavy, particularly for larger models. [21]. Prompt learning encompasses two approaches. One is prompting, which uses discrete tokens to query LLMs, as seen in the success of models like GPT. [22], and prompt tuning, a more efficient method that adds trainable tokens (soft prompts) to the input sequence to guide the model's performance on specific tasks.

Research has explored prompt tuning for both fixed and adaptable models. One approach proposed in [23] involves inserting trainable tokens into various layers of a pre-trained model, including encoder and decoder layers, while keeping the model's parameters frozen. This technique, known as prefix tuning, was later expanded to deep-prompt tuning, demonstrating its versatility across different model sizes and tasks. [24]. As an alternative, P-tuning was introduced, which involves inserting continuous prompts at various points in the input tokens designed by human experts in specific tasks. In this approach, both the prompts and initial model parameters are updated. Building on deep prompt tuning, [19] has developed a system that compares four learning strategies—fine-tuning, hard prompting, soft prompting with a frozen model, and soft prompting with an unfrozen model—using GatorTron clinical LLMs.

Recent research by [25] proposes hierarchical structured prompt pruning based on the lottery ticket hypothesis to identify the winning ticket and eliminate the losing ticket in collecting trained prompt tokens. In this research, the authors designated the positive prompt tokens as the winning tickets and the negative tokens as the losing tickets concerning the lottery hypothesis. The importance of soft prompt tokens is defined as the expected sensitivity of model outputs to the mask variables. A larger score implies a token with a significant contribution, and a lower score implies a negative token with little or no contribution to the model tuning. However, in addition to the fact that the procedure is trial training to obtain the optimal tokens [25], pruning the soft tokens to various levels (token and piece levels) repeatedly to get the optimal soft tokens is resource-intensive, especially for large pre-trained language models (PLM).

While soft prompt-tuning methods have shown promise and match fine-tuning performance, some challenges remain unaddressed. These approaches have mainly been tested on natural language generation tasks and large models using a single, holistic prompt, which may not be suitable for multi-task learning with various objectives. Moreover, the impact of negative prompt tokens is limited, and prefix tuning is constrained by the fixed sequence length of LLMs, resulting in a limited number of trainable parameters. Deep prompt tuning also has limitations, requiring fixed prefix tokens at each layer and needing significant changes to the internal workings of the transformer layers. [25]. In summary, prompt-based learning has yet to be fully explored for multi-task and sequence labeling tasks that require multiple predictions for a single input sequence. [26], highlighting the need for further research in this area.

II. Materials and Method

A. Datasets

The TAC 2017 dataset [27] consists of 200 drug labels in XML format, divided into a training set of 101 labels and a test set of 99. The dataset features five attributes related to Adverse Drug Reactions (ADR): Animal, Drug Class, Factor, Negation, and Severity. Additionally, the dataset includes three types of relationships: Effect (linking severity to ADR), Hypothetical (linking animal, drug class, or factor mentions to ADR), and Negated (linking negation or factor mentions to ADR).

The second dataset, the n2c2 2018 dataset [28], derived from clinical narratives, was used for the adverse drug events extraction challenge. This dataset contains annotations for nine entities (drug, strength, form, dosage, frequency, route, duration, reason, and ADE entities) linked to a drug entity as their source, with eight possible relations between them. Our model was trained and evaluated using the official dataset splits of 303 training records and 202 testing records.

B. Multi-prompt-based Multi-task Soft Prompting of Large Language Models

Learning multiple related tasks simultaneously can lead to biased results if a single prompt is used to adapt LLMs. To overcome this limitation, we propose a novel approach that uses multiple prompts tailored to each task to guide the adaptation of LLMs and ensure more balanced and effective multi-task learning. Two task-specific prompt templates are generated, one for each task. The text prompts are converted into embedding vectors that can be fine-tuned. This process involves two steps: first, the text is broken down into sub-words using a pre-trained tokenizer, and then, the embedding

layer of a pre-trained model is used to generate vector representations for both the input text and the soft prompt tokens, as in Equation 1.

$$Sp = We(Tp) \text{ and } Ex = We(X) \quad (1)$$

where W_e are the embedding matrix of the model, S_p is the embedding of the soft prompt tokens, and E_x is the embedding of the input sequence. The soft prompt is added to the input embedding as a prefix specific to each task. However, since some prompt tokens can harm LLMs' performance, we use a feature selection method to choose the most important ones based on their attention weights generated by the transformer attention mechanism. This ensures that only beneficial tokens are used to fine-tune the model. The detailed procedure for prompt selection is in the following section.

To allow the model to process the added soft prompt embedding, we extended the model's maximum sequence length to accommodate the combined input sequence. This, in turn, required extending the attention mask, token type IDs, and sequence labels to match the new sequence length.

C. Transformer-based Attention-driven Prompt Token Selection

The self-attention mechanism proposed by [29], is an effective way to determine the contextual relationships between different words within an input sequence regardless of their relative distance. It enables the model to ascertain the importance of each word within the sequence. Because some prompt tokens prepended to the input sequence may negatively impact the model adaptation, we apply an attention-based selection approach to select only the top relevant tokens to the input sequence, thereby reducing the use of negative prompt tokens.

We start by taking the dot product of the prompt input embeddings to the input sequence embeddings, then apply a SoftMax function to obtain the attention weight, as in Equation 2. Finally, compute the weighted sum for each token to get its importance score, as shown in Equation 3. The overall procedure is depicted in Algorithm 1.

$$Attention\_weights = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (2)$$

where Q, K, and V are obtained from the linear transformation of input embeddings.

$$Attention\_output = Attention\_weights \cdot Ptoken \quad (3)$$

Algorithm 1: Procedure for Transformer-based attention-driven prompt tokens selection method

Input:
S:←input-embedding, ***S_{emb}***:←soft-prompt-embed
K:←top-k-features, ***D***:←model-dimension
Output: C_{emb}, `top-selected-prompt`
A; W; I ← [] ▶ Initialize attention score, weight and token importance
for I_{emb} ∈ S do:
  for s_{emb} ∈ S_{emb} do:
    A←***DotProduct***(I_{emb},s_{emb}) ▶ computes the dot product.
    W ← ***Softmax***(*A*) ▶ convert the attention scores to probabilities
  endfor
  I←sum(*W*). s_{emb} ▶ Sum up the attention weight for importance.
  Indices←***GetIndices***(*I*, *K*) ▶ `indices of the top-k soft prompt.`
  $E_{indices}$←***Expand***(Indices, *D*) ▶ expand to the d_{model}
  `top-prompt-tokens`←***GatherSelected***(S_{emb}, $E_{indices}$)
  C_{emb}←***concat***(`top-prompt`, C_{emb}) top prompt toinput embedding.
endfor
return *Cemb* ▶ return the combined input to the model.
end procedure

D. Dual Sequence Labelling for Adverse Drug Event Extraction

Figure 1 illustrates the overall architecture where two tasks—concept identification and attribute relation extraction—are modeled simultaneously. We employed a sequence labeling and multi-task transfer learning approach as proposed by [30].

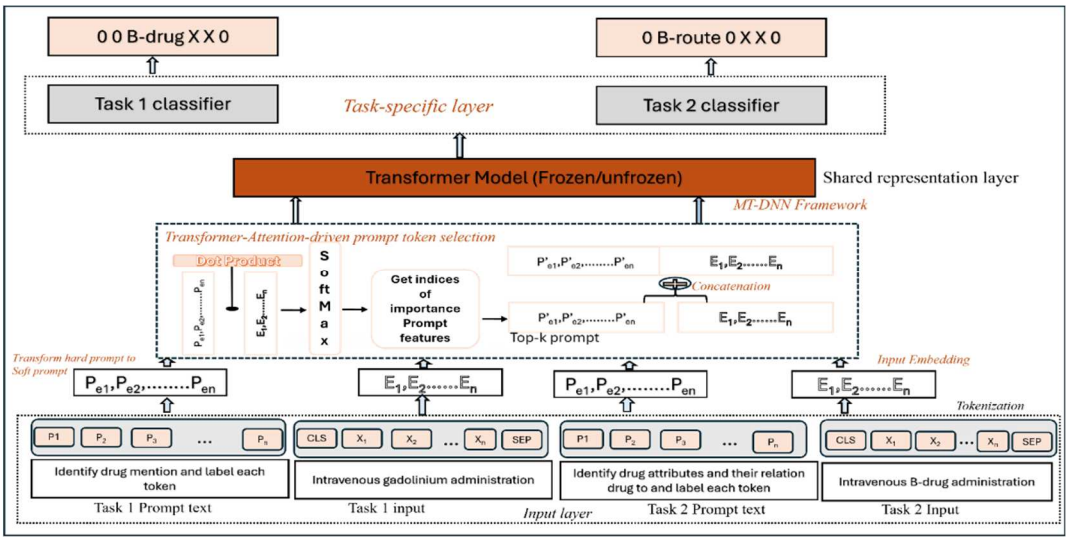

Fig. 1 A multi-prompt model takes two prompt sequences, one for each task.

The textual tokens are transformed into trainable embedding vectors and undergo an attention-driven token selection procedure to select top-k prompt tokens (positive tokens). The positive tokens are then prepended to the input embedding of each task to serve as input to the multi-task learning framework to produce shared representation by the transformers model (with frozen or unfrozen parameters). This method converts both tasks into a dual sequence labeling problem, modeled together using a multi-task deep learning framework [31] to generate a shared contextual representation of the input via a transformer-based model. The concatenated input, which includes both the input embeddings and the selected prompt embeddings for the two sub-tasks derived from the proposed multi-prompt tuning procedure detailed above, is fed into the transfer learning framework. The output is then directed to task-specific layers, where the sub-task classification head and SoftMax are applied for the final classification of each token in the sequence.

During the dual sequence labeling stage, the system transforms the tasks into ADR-source mention identification and ADR-mention attribute relation identification. Each dataset contains either ADR or Drug mentions. The ADR-source mention identification task classifies the input sequence into binary classes: positive (source mentions containing one or more relations with mention attributes) and negative (source mentions with no relation with mention attributes). Conversely, the ADR-mention attribute relation identification task involves identifying the attributes and relationships of the positive ADR-mentions identified in the first sub-task.

The system uses an extended beginning inside outside (BIO) tagging scheme to handle discontinuous mentions and sub-words from word piece tokenization during token-based sequence labeling for the two sub-tasks. Additional tags, DB (discontinuous mention beginning) and DI (discontinuous mention inside), are introduced—the "X" tag labels sub-words generated by the tokenizer.

III. . Results and Discussion

A. Large Language Models and Experimental Settings

The BERT model, introduced by [32], is trained on vast text data from English Wikipedia and BooksCorpus. Two pre-trained versions of BERT are available, differing in size: BERT-Base and BERT-Large. We experiment with the base mode for fine-tuning and soft prompting (with frozen and unfrozen model parameters). The SciBERT model [12], this model builds upon the BERT architecture and is pre-trained on a large corpus of 1.14 million full-text papers from Semantic Scholar. There are two available versions of SciBERT: scivocab and base-vocab. We utilize the sci-vocab model.

We configured our model with a maximum sequence length of 512 and a batch size of 8 and 32 for unfrozen and frozen models, respectively. We optimized the learning rate to 2e-5, using the cross-entropy loss function and Adamax optimizer. We applied a weight decay of 0.05 and a dropout rate of 0.1 to prevent overfitting. We trained the model for 10 epochs on the TAC 2017 dataset and 15 on the n2c2 2018 dataset for soft prompt tuning with unfrozen models. Similarly, we trained the model for 200 epochs for soft prompt tuning with frozen models for both datasets.

B. Results

Tables I and II show the reported results of our two experimented models, BERT and SciBERT, on TAC 2017 and n2c2 2018 for concept and end-to-end relation extraction, respectively. On the n2c2 dataset, we can see from Table 1 for concept extraction that the SciBERT model has a better overall performance for both fine-tuning and soft prompt tuning for the unfrozen model. In comparison to BERT, the SciBERT model improved by 3.6%. Similarly, SciBERT outperformed BERT on the TAC 2017 dataset by 1.76%. For the frozen model, SciBERT outperformed BERT by 0.8% and 2.44% for concept extraction.

TABLE I
THE RESULT OF THE TWO EXPERIMENTED MODELS ON THREE TUNING STRATEGIES FOR CLINICAL CONCEPT EXTRACTION.

Dataset	**Models**	**Training strategy**		
		Fine-tuning	**Soft prompt Unfrozen**	**Soft prompt Frozen**
		F1-score	**F1-score**	**F1-score**
N2C2 2018	BERT	88.83	88.94	57.25
	SCIBERT	92.38	92.54	58.07
TAC 2017	BERT	83.83	85.32	70.22
	SCIBERT	86.85	87.08	72.66

TABLE II
THE RESULT OF THE TWO EXPERIMENTED MODELS ON THREE TUNING FOR CLINICAL END-TO-END EXTRACTION

Dataset	**Models**	**Training strategy**		
		Fine-tuning	**Soft prompt Unfrozen**	**Soft prompt Frozen**
		F1-score	**F1-score**	**F1-score**
N2C2 2018	BERT	83.83	83.94	32.07
	SCIBERT	89.13	89.25	33.23
TAC 2017	BERT	50.10	51.15	18.10
	SCIBERT	51.26	53.33	19.24

In addition, Table 2 presents the results for end-to-end relation extraction for the experimented models' TAC 2017 and n2c2 2018 datasets. The SciBERT model outperforms BERT by 5.31% on n2c2 and 2.18% on TAC 2017. For the frozen model, it was 1.16% and 1.14%, respectively. These results demonstrate the capability of the SciBERT model over BERT. The observed performances could be attributed to the pre-training data from the scientific document, giving the model more chances to identify concepts and terminologies from the clinical text. Figure 2 depicts the results of the models.

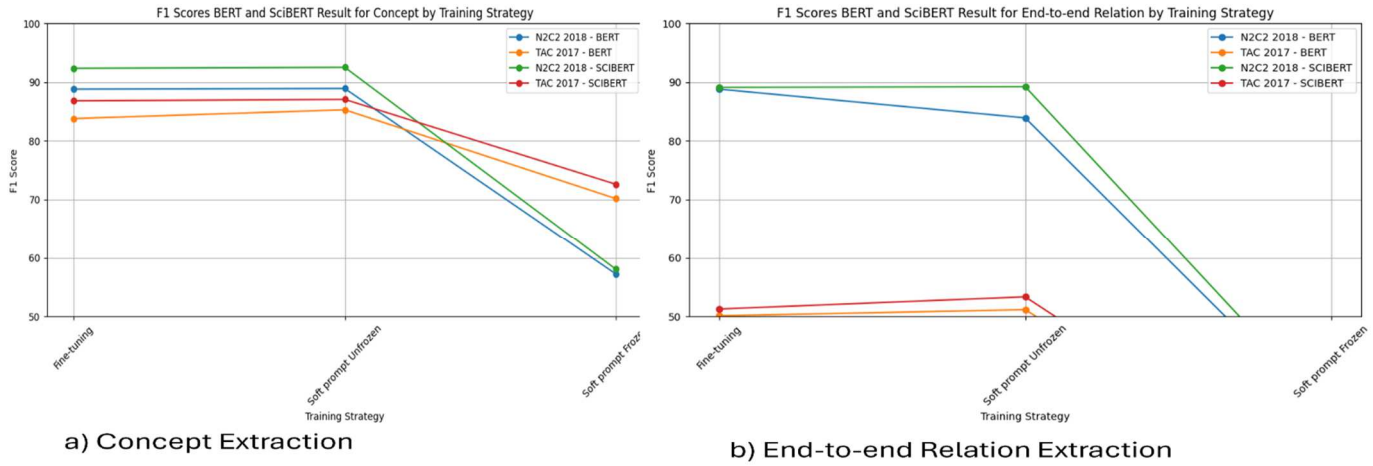

Fig. 2 Summary of the results obtained in the F1 score by the models for concept (a) and relation (b) on both TAC 2017 and n2c2 2018 datasets.

C. Discussion

Prompt-based learning is a lightweight approach to adopting LLM, especially with frozen models. Most existing prompting systems are mainly developed to handle natural language generation problems. The kinds of research developed on natural language processing understanding problems mostly explored large models with many parameters. However, this approach has not fully explored complex NLP problems, such as sequence labeling involving multiple predictions for a single input. In addition, prompt-based learning suffers from the quality of prompt tokens to effectively guide the model on downstream tasks, thereby reducing the gap between pre-training and downstream objectives. This study proposed a multi-prompt-based soft prompting method with a transformer-based attention prompt tokens selection to select the top necessary prompt tokens. We conduct our experiments with two popular small-scale models to investigate the effectiveness of this approach.

To further investigate the potential of our model to state-of-the-art models, we evaluate our models' performance on the n2c2 2018 dataset, comparing them to the GatorTron system [19], a clinical natural language processing model. Notably, GatorTron is the largest clinical model in the literature, pre-trained on an extensive corpus of over 8.9 trillion words from biomedical texts and electronic health records. The model comes in three sizes: GatorTron base (345 million parameters), GatorTron medium (3.6 trillion parameters), and GatorTron large (8.9 trillion parameters) [19]. Tables 3 and 4 show the concept and end-to-end relation extraction comparison results.

For concept, SciBERT with unfrozen parameters outperformed GatorTron-base, which has the highest score among the GatorTron variants at 1.36%. Similarly, end-to-end relations improved by 5.93%. However, for a frozen model, the performance dropped by 32.86% for concept and 49.76% for end-to-end relation compared to GatorTron-large. This drastic drop is not surprising, as the SciBERT model is 1.23% in parameters compared to GatorTron-large. This parameter difference also indicates the capability of our approach of multi-prompt tuning for multi-task learning settings, achieving the performance of 39% and 40% for concept and relation extraction, respectively, to the highest GatorTron models. Figure 3 displays the comparison of the models.

TABLE III
COMPARISON OF OUR EXPERIMENTED MODELS WITH GATORTRON MODELS ON THE N2C2 2018 DATASET FOR CONCEPT EXTRACTION USING OFFICIAL EVALUATION METRIC (MICRO F1 SCORE).

Dataset	Models	Number of parameters	Training strategy		
			Fine-tuning	Soft prompt Unfrozen	Soft prompt Frozen
			F1 score	F1 score	F1 score
N2C2 2018	BERT	110 million	88.83	88.94	57.25
	SCIBERT	110 million	92.38	92.54	58.07
	GatorTron base	345 million	88.79	91.12	86.56
	GatorTron medium	3.9 billion	88.83	91.18	90.85
	GatorTron large	8.9 billion	88.91	91.15	90.93

TABLE IV
COMPARISON OF OUR EXPERIMENTED MODELS WITH GATORTRON MODELS ON THE N2C2 2018 DATASET FOR END-TO-END RELATION EXTRACTION USING OFFICIAL EVALUATION METRIC (MICRO F1 SCORE).

Dataset	Models	Number of parameters	Training strategy		
			Fine-tuning	Soft prompt Unfrozen	Soft prompt Frozen
			F1 score	F1 score	F1 score
N2C2 2018	BERT	110 million	83.83	83.94	32.07
	SCIBERT	110 million	89.13	89.25	33.23
	GatorTron base	345 million	81.92	83.32	79.21
	GatorTron medium	3.9 billion	82.05	83.21	82.99
	GatorTron large	8.9 billion	82.00	83.30	82.68

This study shows that models with small to medium parameters can perform well while on frozen parameters. However, to attain the performance of traditional fine-tuning, the model's parameters must be scaled up to billions of parameters, as is evident in the GatorTron models. In addition, soft prompt tuning can be successfully applied to complex natural language understanding problems involving multi-tasking with remarkable performance.

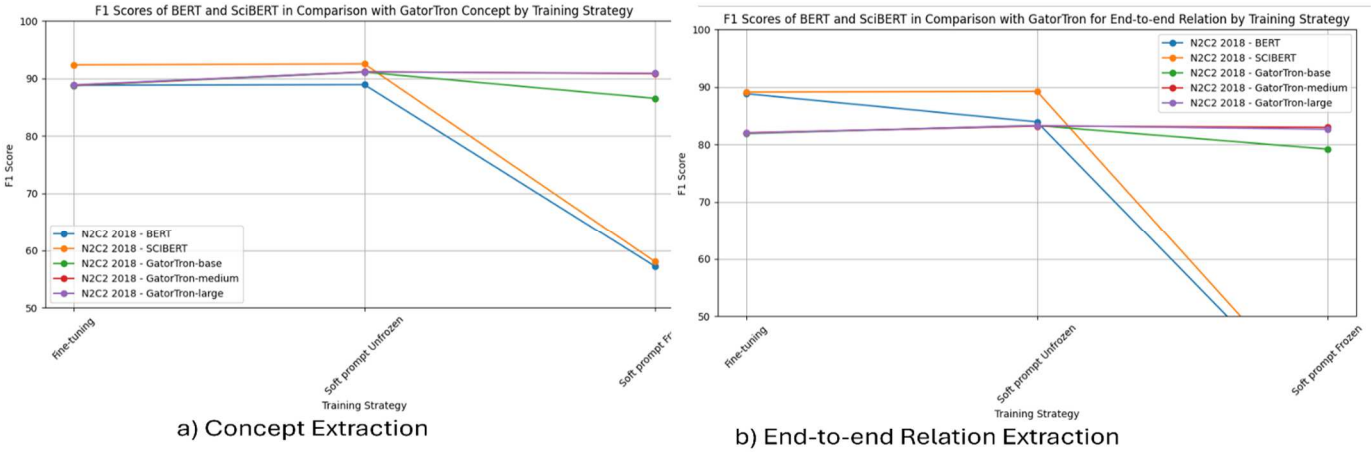

Fig. 3 Summary of the results compared results in the F1 score by the experimented models and GatorTron models for concept (a) and relation (b) on n2c2 2018 datasets

IV. Conclusion

The prefix-based multi-prompt tuning with attention-based prompt token selection proposed in this study has demonstrated the effectiveness of soft prompt tuning in adopting a large language model for natural language understanding problems involving sequence labeling for a multi-task adverse drug extraction. Our approach with unfrozen models outperforms the traditional fine-tuning and GatorTron models for these tasks. In our future work, we plan to investigate our proposed approach with language models of medium to large size and decoder-based models like GPT. In addition, we will explore other NLP tasks like sequence classification.

Acknowledgments

This research is supported by Universiti Putra Malaysia and the Ministry of Higher Education, Malaysia, under the Fundamental Research Grant Scheme (FRGS/1/2023/ICT02/UPM/02/3).

References

[1] G. Herman Bernardim Andrade *et al.*, "Assessing domain adaptation in adverse drug event extraction on real-world breast cancer records," *Int. J. Med. Inform.*, vol. 191, p. 105539, 2024, doi:10.1016/j.ijmedinf.2024.105539.

[2] E. D. El-Allaly, M. Sarrouti, N. En-Nahnahi, and S. Ouatik El Alaoui, "An attentive joint model with transformer-based weighted graph convolutional network for extracting adverse drug event relation," *J. Biomed. Inform.*, vol. 125, p. 103968, Jan. 2022, doi:10.1016/j.jbi.2021.103968.

[3] Y. B. Gumiel *et al.*, "Temporal Relation Extraction in Clinical Texts," *ACM Comput. Surv.*, vol. 54, no. 7, p. 144, 2022, doi:10.1145/3462475.

[4] B. S. Kaas-Hansen, D. Placido, C. L. Rodríguez, and A. P. Nielsen, "Language-agnostic pharmacovigilant text mining to elicit side effects from clinical notes and hospital medication records," *J. Basic Clin. Pharmacol. Toxicol.*, pp. 282–293, Jul. 2022, doi:10.1111/bcpt.13773.

[5] S. Narayanan, K. Mannam, S. P. Rajan, and P. V. Rangan, "Evaluation of Transfer Learning for Adverse Drug Event (ADE) and Medication Entity Extraction," *Proc. 3rd Clin. Nat. Lang. Process. Work.*, pp. 55–64, 2020, doi: 10.18653/v1/2020.clinicalnlp-1.6.

[6] G. Duan, J. Miao, T. Huang, W. Luo, and D. Hu, "A Relational Adaptive Neural Model for Joint Entity and Relation Extraction," *Front. Neurorobot.*, vol. 15, p. 635492, Mar. 2021, doi:10.3389/fnbot.2021.635492.

[7] Z. Xu, S. Lin, J. Chen, Y. Sheng, and L. Chen, "A Semi-supervised Method for Extracting Multiple Relations of Adverse Drug Events from Biomedical Literature," *IEEE Adv. Inf. Technol. Electron. Autom. Control Conf.*, pp. 934–938, 2021, doi:10.1109/iaeac50856.2021.9390651.

[8] Y. Fan, S. Zhou, Y. Li, and R. Zhang, "Deep learning approaches for extracting adverse events and indications of dietary supplements from clinical text," *J. Am. Med. Informatics Assoc.*, vol. 28, no. 3, pp. 569–577, 2021, doi: 10.1093/jamia/ocaa218.

[9] J. Lamy, "A data science approach to drug safety : Semantic and visual mining of adverse drug events from clinical trials of pain treatments," *Artif. Intell. Med.*, vol. 115, p. 102074, 2021, doi:10.1016/j.artmed.2021.102074.

[10] A. Wasylewicz *et al.*, "Identifying adverse drug reactions from free-text electronic hospital health record notes," *British Journal of Clinical Pharmacology*, vol. 88, no. 3. pp. 1235–1245, 2022. doi:10.1111/bcp.15068.

[11] C. Zhan, E. Roughead, L. Liu, N. Pratt, and J. Li, "Detecting potential signals of adverse drug events from prescription data," *Artif. Intell. Med.*, vol. 104, p. 101839, 2020, doi: 10.1016/j.artmed.2020.101839.

[12] I. Beltagy, K. Lo, and C. Arman, "SciBERT: A Pretrained Language Model for Scientific Text," *Proc. 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process.*, pp. 3615–3620, 2019, doi: 10.18653/v1/D19-1371.

[13] J. Sanyal, D. Rubin, and I. Banerjee, "A weakly supervised model for the automated detection of adverse events using clinical notes," *J. Biomed. Inform.*, vol. 126, p. 103969, 2022, doi:10.1016/j.jbi.2021.103969.

[14] Y. Mohammadi, F. Ghasemian, J. Varshosaz, and M. Sattari, "Informatics in Medicine Unlocked Classifying referring / non-referring ADR in biomedical text using deep learning," *Informatics Med. Unlocked*, vol. 39, p. 101246, 2023, doi:10.1016/j.imu.2023.101246.

[15] T. ValizadehAslani *et al.*, "PharmBERT: a domain-specific BERT model for drug labels," *Brief. Bioinform.*, vol. 24, p. bbad226, Jul. 2023, doi: 10.1093/bib/bbad226.

[16] S. Modi, K. A. Kasmiran, N. Mohd Sharef, and M. Y. Sharum, "Extracting adverse drug events from clinical Notes: A systematic review of approaches used," *J. Biomed. Inform.*, vol. 151, p. 104603, 2024, doi: 10.1016/j.jbi.2024.104603.

[17] W. Liu, L. Zhou, D. Zeng, and H. Qu, "Document-Level Relation Extraction with Structure Enhanced Transformer Encoder," *2022 Int. Jt. Conf. Neural Networks*, pp. 1–8, 2022, doi:10.1109/ijcnn55064.2022.9892647.

[18] C. Mcmaster *et al.*, "Developing a deep learning natural language processing algorithm for automated reporting of adverse drug reactions," *J. Biomed. Inform.*, vol. 137, p. 104265, 2023, doi:10.1016/j.jbi.2022.104265.

[19] C. Peng *et al.*, "Model tuning or prompt Tuning? a study of large language models for clinical concept and relation extraction," *J. Biomed. Inform.*, vol. 153, p. 104630, 2024, doi:10.1016/j.jbi.2024.104630.

[20] J. Li, X. Ji, and L. Hua, "Improving the Prediction of Adverse Drug Events Using Feature Fusion-Based Predictive Network Models," *IEEE Access*, vol. 8, pp. 48812–48821, 2020, doi:10.1109/access.2020.2979452.

[21] G. Qin and J. Eisner, “Learning How to Ask: Querying LMs with Mixtures of Soft Prompts,” *NAACL-HLT 2021 - 2021 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, pp. 5203–5212, 2021, doi: 10.18653/v1/2021.naacl-main.410.

[22] X. Liu *et al.*, “GPT understands, too,” *AI Open*, 2023, doi:10.1016/j.aiopen.2023.08.012.

[23] X. L. Li and P. Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” *ACL-IJCNLP 2021 - 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 4582–4597, 2021, doi: 10.18653/v1/2021.acl-long.353.

[24] X. Liu *et al.*, “P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks,” *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 2, pp. 61–68, 2022, doi: 10.18653/v1/2022.acl-short.8.

[25] F. Ma *et al.*, “XPROMPT: Exploring the Extreme of Prompt Tuning,” *Proc. 2022 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2022*, pp. 11033–11047, 2022, doi: 10.18653/v1/2022.emnlp-main.758.

[26] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing,” *ACM Comput. Surv.*, vol. 55, no. 9, p. 195, 2023, doi: 10.1145/3560815.

[27] M. Belousov, N. Milosevic, and W. Dixon, “Extracting adverse drug reactions and their context using sequence labelling ensembles in TAC2017,” *TAC2017 Conf.*, pp. 1–11, 2019, doi:10.48550/arXiv.1905.11716.

[28] S. Henry, K. Buchan, M. Filannino, A. Stubbs, and O. Uzuner, “2018 N2C2 Shared Task on Adverse Drug Events and Medication Extraction in Electronic Health Records,” *J. Am. Med. Informatics Assoc.*, vol. 27, no. 1, pp. 3–12, 2020, doi: 10.1093/jamia/ocz166.

[29] A. Vaswani *et al.*, “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, pp. 6000–6010, 2017, doi: 10.5555/3295222.3295349.

[30] E. D. El-Allaly, M. Sarrouti, N. En-Nahnahi, and S. Ouatik El Alaoui, “MTTLADE: A multi-task transfer learning-based method for adverse drug events extraction,” *Inf. Process. Manag.*, vol. 58, no. 3, p. 102473, 2021, doi: 10.1016/j.ipm.2020.102473.

[31] X. Liu, P. He, W. Chen, and J. Gao, “Multi-Task Deep Neural Networks for Natural Language Understanding,” *Proc. 57th Annu. Meet. Assoc. Comput. Linguist.*, pp. 4487–4496, 2019, doi:10.18653/v1/P19-1441.

[32] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, 2019, doi: 10.18653/v1/N19-1423.