Pertanika J. Sci. & Technol. 34 (3): 1769 - 1800 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Human-in-the-Loop Soft Prompt Tuning for Adverse Drug Event Extraction from Clinical Notes

Salisu Modi[1], Khairul Azhar Kasmiran[2]*, Nurfadhlina Mohd Sharef[2], and Mohd Yunus Sharum[2]

[1]Department of Computer Science, Faculty of Computing, Sokoto State University, 840101 Sokoto, Sokoto State, Nigeria

[2]Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

ABSTRACT

Adverse drug events (ADEs) are significant healthcare issues, increasing costs and medication durations. Extracting ADE information is crucial for enhancing healthcare delivery and drug safety. Transformer-based models' performance has recently improved through soft prompt tuning in the ADE task. Current ADE extraction models require large, annotated datasets and offer limited interpretability for clinical use. In addition, the complexity and black-box nature of transformer-based models limiting incorporating human preference and make interpretation difficult, which are critical in healthcare, hindering its full adoption. To mitigate this problem, this research proposes a human-in-the-loop learning method to tune transformers with domain expert input to complement limited data and incorporate expert preference. The method evolves in two phases: initially soft prompt tuning the model for multi-task learning of dual sequence labelling, ADE extraction with an additional soft prompt that guides the model, followed by iterative prediction, validation, and feedback refinement to optimise the model. Visualisation techniques display model predictions to users, enhancing understanding. The self-attention weights aid in diagnosing and explaining the model using saliency maps and attention flow diagrams. A graphical user interface allows experts to provide corrective labels for misclassified samples, thus refining the model. The result of the final model evaluated on testing sets from TAC 2017 and N2C2 2018 datasets, achieving state-of-the-art performance of 0.9404 and 0.9132 for N2C2 concept and relation extraction and 0.8723 and 0.5506 for TAC 2017 concept and relation extraction. In conclusion, this research

ARTICLE INFO

Article history:
Received: 25 November 2025
Accepted: 07 June 2026
Published: 25 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.21

E-mail addresses:
gs63125@student.upm.edu.my (Salisu Modi)
k_azhar@upm.edu.my (Khairul Azhar Kasmiran)
nurfadhlina@upm.edu.my (Nurfadhlina Mohd Sharef)
m_yunus@upm.edu.my (Mohd Yunus Sharum)
* Corresponding author

e-ISSN: 2231-8526

demonstrates the effectiveness of incorporating human feedback to improve model performance in complex scenarios, demonstrating its effectiveness in improving drug safety surveillance and drug monitoring.

Keywords: Adverse drug event, drug safety surveillance, human-in-the-loop learning, optimisation, soft prompt tuning

INTRODUCTION

Extracting adverse drug events (ADEs) from narrative documents has been a topic of interest in recent years. The goal is to identify the drug entities and their related attributes and then determine the causal relationships between the drug and the adverse effects caused by the drug. This task is significant in improving healthcare delivery and drug safety surveillance. The increasing rate of ADE cases reported worldwide is becoming worrisome as it increases medication costs and causes damage to public health (Xia, 2022). Most ADE cases are preventable when appropriate measures are taken (Ong et al., 2024). The early stage of monitoring and investigating the side effects of a drug is through clinical trials with volunteer patients before the approval and marketing of new drug products. This approach is ineffective due to its shorter duration and lack of volunteer patients, especially with drugs with longer latency (Xia, 2022). Consequently, spontaneous reporting becomes the next option to complement the earlier approach at the post-marketing stage of drug usage. The spontaneous reporting system (SRS) relies on the clinician's assumption during the patient's medication period or the report from the affected patient. This approach is characterised by underreported cases and incompleteness (Kim et al., 2023).

With improved performance, large language models (LLMs) based on transformer neural networks have been widely adopted for various ADE extraction tasks (Gu et al., 2023; Jamil et al., 2026; Ren & Wang, 2023; Sahoo et al., 2024). This can be attributed to their transfer learning capabilities, which enhance downstream-specific performance on downstream tasks. Adapting pre-trained language models (PLMs) through fine-tuning leaves an objective gap between pre-training and downstream task objectives and also provides a memory-efficient approach. The soft prompting mitigates this gap by integrating task-specific trainable prompts to guide the model on the downstream task (Tian et al., 2024). However, this requires a huge amount of annotated data from the downstream task. Healthcare-related data are primarily scarce and restricted due to privacy constraints and a lack of domain expertise, consistent annotations.

Moreover, transformer-based models are black-box and complex. They are composed of complex architectures with billions or trillions of trainable parameters. The models are trained on large volumes of data, making interpreting and diagnosing their decision-making processes challenging. Adverse drug event extraction involves establishing causality

relationships between the drug entities and the adverse effects caused by their usage during patient treatments. This type of connection could sometimes involve complex, ambiguous and polysemous entities that would otherwise be difficult for a machine-learning model to handle accurately. Incorporating human preferences through domain experts into the modelling process could improve model performance and reduce time-consuming upfront data annotations. However, most current approaches treat humans as passive participants (i.e. sending input and receiving output processed by the model) during model development (Gómez-Carmona et al., 2024). On the one hand, an interactive human-in-the-loop learning approach is a form of learning where models are developed with human involvement throughout the process. In this approach, end-users actively contribute to model creation by continuously providing training parameters, reviewing model outputs, and offering feedback on intermediate results to improve performance (Wondimu et al., 2022).

The black-box nature of the LLM models has also been a significant concern for an effective clinical natural language processing (NLP) model for ADE extraction. Incorporating human factors, model interpretation, and explanations in model development is crucial for improving model reliability and capability in handling complex situations (Zitu et al., 2025). Interactive human-in-the-loop learning (IHITLL) enables the incorporation of human judgment and feedback into PLM's model adaptation for ADE extraction, thereby mitigating the risk of opaque decision-making in the model (Wondimu et al., 2022).

This work proposes a novel approach that combines transformer-based model soft prompt tuning with interactive human-in-the-loop interfaces to optimise the model's performance tailored for a multi-task sequence labelling of ADE concept and relation extraction.

This paper introduced the human-in-the-loop process to this task to enable humans to understand, interpret and modify the model that extracts ADE cases from unstructured clinical documents. The work designed interfaces to support model interpretation, diagnosis and refinement in model development. The approach first tunes the model using a soft-prompt tuning approach on the training set. Further, it optimises the model with feedback based on the prediction outcome from the sample validation data. The method leverages visualisation techniques to display the model prediction result and utilises the self-attention weight generated by the model to interpret and diagnose the model decision markings. The proposed approach was experimented on several transformer-based models, such as BERT and SCIBERT, as a baseline. All the experimented models have shown an improvement compared to traditional fine-tuning, indicating the generality of the proposed hybrid approach.

In summary, the objectives of this research are as follows:

1. To propose a novel, interactive human-in-the-loop soft prompt tuning of the transformer model method with three levels of interaction: model understanding, model diagnosis

and model refinement, for improving model interpretation and explanations.

2. To experiment with the proposed approach to multi-task dual sequence labelling of adverse drug event extraction tasks of concepts and relations on two public datasets, N2C2 2018 and TAC 2017.
3. To improve the ADE concept and relation extraction F1-score using human feedback-based soft prompt tuning.

The architecture of PLMs based on transformers is inherently complex. It features multiple nested attention layers and processing blocks that perform numerous linear and non-linear transformations on input sequences to generate output (Vaswani et al., 2017). This intricate structure makes the internal workings of these models challenging to interpret, resulting in opaque decision-making processes (Budd et al., 2021). Integrating human factors into the decision-making process during model development and deployment is crucial, especially for healthcare-related tasks such as ADE extraction.

Interpreting neural network-based models has been a significant concern for effectively developing practical systems for NLP tasks (Accenture, 2019). Enhancing model understanding and mitigating the risks associated with the black-box nature of PLM is an ongoing area of research. Various methods have been developed to improve the explainability of neural network-based models, for instance Kim & Kim, (2022) proposed Shapley Additive exPlanations (SHAP). This facilitates incorporating human factors into the human-in-the-loop process of developing intelligent systems for NLP tasks. Techniques such as Active Learning (AL) (Riesener et al., 2024; Shelmanov et al., 2021), Reinforcement Learning with human-in-the-loop learning (RLHILL) (Bai et al., 2022; Ouyang et al., 2022), Explainable AI (XAI), and IHITLL (Zhao & Liu, 2021) have been applied to various NLP tasks.

Moreover, human feedback can optimise LLM's performance on different downstream tasks. An example is the success of OpenAI's InstructGPT (Ouyang et al., 2022), which incorporated user feedback and intent into its GPT-3.5 model using reinforcement learning from human feedback (RLHF). The initial model had 1.3 billion parameters and was fine-tuned using a feedback corpus to include helpfulness, honesty, and harmlessness. Based on comparisons by human expert evaluators, InstructGPT was preferred over GPT-3.5, which had 175 billion parameters.

The approach has also been explored by Thoppilan et al., (2022), that proposed LaMDA, which fine-tuned LLMs to be interesting, helpful, and factually accurate using supervised learning techniques that combine generative and discriminative approaches. Similarly, (Bai et al., 2022) applied preference modelling and RLHF to fine-tune LLMs, aligning the models to be helpful, harmless, and honest. This approach involves iteratively updating the model with feedback from online crowd workers.

Recently, many parameter-efficient fine-tuning approaches have been promising to make language model adaptation more efficient, and they each have distinct characteristics. Adapter-based fine-tuning (Houlsby et al., 2019) introduces inference latency, as the adapter layers are executed sequentially, unlike the original pre-trained transformer layers executed in parallel. LoRA (Hu et al., 2021) and its variants approximate the pre-trained weights with two matrices, requiring alterations to the structure of the pre-trained model. The soft prompting (Liu et al., 2022; Li & Liang, 2021) It is a lightweight and straightforward approach that requires updating only a few prompt parameters without altering the core parameters of the pre-trained model. The study by Peng et al., (2024) Driven further developments in ADE extraction tasks by adapting a deep prompt-tuning approach to compare four learning strategies: fine-tuning, hard prompting, soft prompting with a frozen model, and soft prompting with an unfrozen model. This system utilised GatorTron clinical LLMs to identify ADE concepts and perform end-to-end RE from clinical documents.

Despite the success of these approaches in optimising LLM performance for various NLP tasks, most existing methods are based on generative AI models using decoder-only transformer architectures for Natural Language Generation (NLG) tasks. The Natural Language Understanding (NLU) tasks, such as ADE extraction, have not fully utilised these promising approaches to improve LLM adaptation. In the active learning model, uncertain samples are to be annotated by an oracle or human expert (Quennelle et al., 2025). In RLHF, human feedback shapes reward signals for reinforcement learning, and human judgment guides the model optimisation (Bai et al., 2022). The interactive Machine learning involved a continuous loop of training with user input, while the human iterative refine output (Teso et al., 2023). The proposed IHITLL framework involves iteratively collecting domain expert feedback based on model predictions, then utilising the feedback to refine the model weight on few iterations.

Additionally, while RLHF is the most prominent approach for incorporating human preferences into model fine-tuning, IHITLL has not been extensively explored. Furthermore, current approaches have placed less emphasis on interpreting models to enhance human understanding of their opaque decision-making process. The effectiveness of IHITLL in the domain of ADE extraction is still uncertain. Therefore, it is essential to explore this approach using PLM. Incorporating human feedback into model training can reduce the need for manual labelling and address complex data situations in the ADE concept and relation extraction subtasks through real-time human feedback. To the knowledge of the author, no study explores IHITLL for multi-task ADE extraction tasks, which this study aimed to investigate through soft prompt tuning of PLMs. This paper presents an interactive human-in-the-loop learning method designed to optimise model performance by incorporating human feedback in the model tuning on a token-level sequence modelling task for ADE extraction.

METHODOLOGY

This section details the proposed hybrid method of interactive human-in-the-loop multi-task learning tuning of a transformer-based model to optimise its performance on multi-task sequence labelling ADE extraction tasks. The key novelty of the proposed method is introducing interactive learning with three levels of interaction to tune a transformer-based model. Three levels of user interaction with the model and data are provided to enhance user understanding of model predictions, diagnosis and refinement, with a user interface for model prediction understanding, a user interface for model interpretation, and an interface for interactive model refinement. The overall flow of the procedure is depicted in Figure 1, which comprises two phases. The initialisation phase is where the transformer model is initially tuned using the soft prompt tuning approach. The interactive model optimisation is where the initiated model is tested and iteratively improved from sample data by the domain expert.

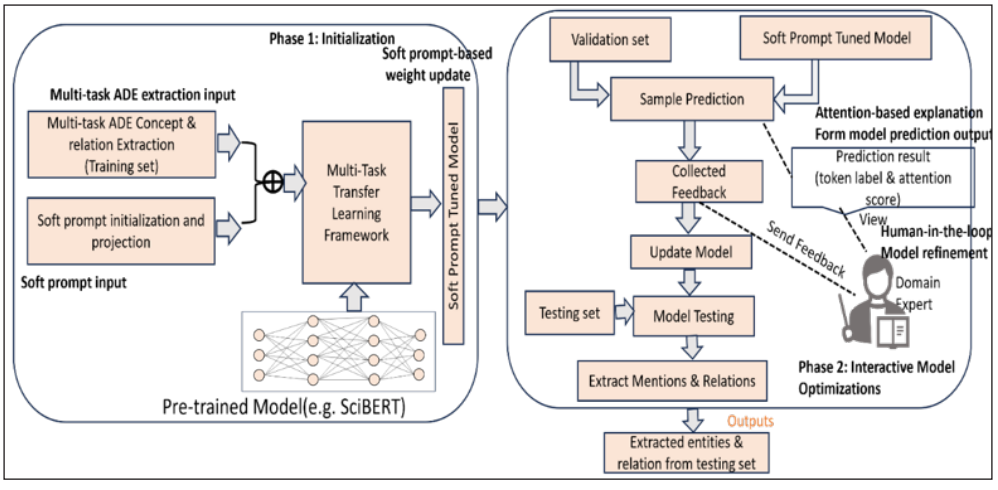

Figure 1. IHITLL-based fine-tuning framework

For more effective design and implementation of interactive human-in-the-loop methods, two human-in-the-loop concerns are as follows:

- Who is the human-in-the-loop, and what role do they play? This research utilises datasets provided by TAC 2017 and N2C2 2018 ADE extraction challenges. The task was handled as sequence labelling, where the model is trained to predict a corresponding label to each token within the sequence. Therefore, the human-in-the-loop can be a clinical expert or data scientist developing and training the model based on the pre-annotated dataset provided by the organisers. The role of the human is to provide corrective feedback based on model prediction according to the gold standard annotation through an interactive interface.

- How to mitigate model and human bias during the interactive model optimisation: The dataset was bifurcated into training and testing sets. For the interactive model tuning, 10% of the training set is used to optimise the model through predictive and refined cycles. These datasets have been carefully annotated based on the annotation guidelines provided by the challenges organisers, detailed in (Demner-Fushman et al., 2018; Henry et al., 2020). During the model optimisation process, corrective feedback is provided only from the gold-standard annotations, thereby avoiding incorporating unverified human factors into the model refinement. The rationale for selecting 10% of the training data for interactive refinement is to allow the model to be initially fine-tuned on a larger portion of the dataset, thereby acquiring domain knowledge, while simultaneously reducing the annotation burden on the human-in-the-loop expert.

Problem Formulation

The task of ADE extraction involves two sub-tasks: concept and relation extraction. The process starts with identifying named entities within a sequence of tokens named concepts, denoted as $S = \{t_1, t_2, \ldots, t_n\}$, where n represents the sequence length. The goal is to extract a set of positive entities $E = \{e_1, e_2, \ldots, e_i\}$ where each entity $e_i \in E$. has one or more relationships with other entities or entities in E, and a corresponding set of relations $R = \{r_1, r_2, \ldots, r_j\}$.

Recently, many parameter-efficient fine-tuning approaches have been promising to make language model adaptation more efficient, and they each have distinct characteristics. Adapter-based fine-tuning (Quentin, 2019) introduces inference latency, as the adapter layers are executed sequentially, unlike the original pre-trained transformer layers executed in parallel, LoRA (Wallis, 2021) and its variants approximate the pre-trained weights with two matrices, requiring alterations to the structure of the pre-trained model. In contrast, soft prompting is a lightweight and straightforward approach that requires updating only a few prompt parameters without altering the core parameters of the pre-trained model (Chen et al., 2024).

The soft prompt tuning method prepends task-specific instructions as a learnable prompt, instructing the pre-trained model on downstream task objectives. The aim is to minimise the multi-task loss together with the additional learnable prompt, as in Equation 1. Here, the model parameters are updated, as shown in Figure 2b.

$$L_{\text{MTL-ADE}}\left(\theta_{\text{pre-trained}}, \theta_{\text{tasks}}, \theta_{\text{prompt}}\right) = \sum_{t=1}^{T} \delta t \left(-\sum_{n=1}^{N_t} \sum_{r=1}^{R_t} y_{nc}^{(t)} \log P_{nc}^{(t)}\left(\theta_{\text{pre-trained}}, \theta_{\text{tasks}}, \theta_{\text{prompt}}\right)\right) \quad [1]$$

For parameter-efficient soft prompt tuning (frozen model), instead of updating the pre-trained parameters $\theta_{\text{pre-trained}}$, only the task-specific parameters θ_{tasks} and the additional θ_{prompt} trainable vectors generated from the embedding of the prompt token of the input sequence

are updated. The overall multi-task loss in Equation 2 is minimised by updating only the θ_{tasks} and θ_{prompt} parameters while leveraging the knowledge of the pre-trained model during the fine-tuning process, as shown in Figure 2a.

$$L_{MTL\text{-}ADE}(\theta_{tasks}, \theta_{prompt}) = \sum_{t=1}^{T} \delta t \left(- \sum_{n=1}^{N_t} \sum_{r=1}^{R_t} y_{nc}^{(t)} \log P_{nc}^{(t)}(\theta_{pre\text{-}trained}, \theta_{tasks}, \theta_{prompt}\right) \quad [2]$$

Where $\theta_{tasks}=[\theta_{ner}; \theta_{re}]$ and $\theta_{prompt} = [\theta_{ner\text{-}prompt}; \theta_{re\text{-}prompt}]$ and N_t and R_t are the numbers of tokens for the ADE-concept and ADE-relation input sequences. $y_{nc}^{(t)}$ is the true label for the n-th token ADE-concept and ADE-relation, and $P_{nc}^{(t)}$ is the predicted probability of the n-th token ADE-concept and ADE-relation. Here, all the parameter sets ($\theta_{pre\text{-}trained}$, θ_{tasks}, θ_{prompt}) are learnable and have been adjusted to minimise the multi-task loss.

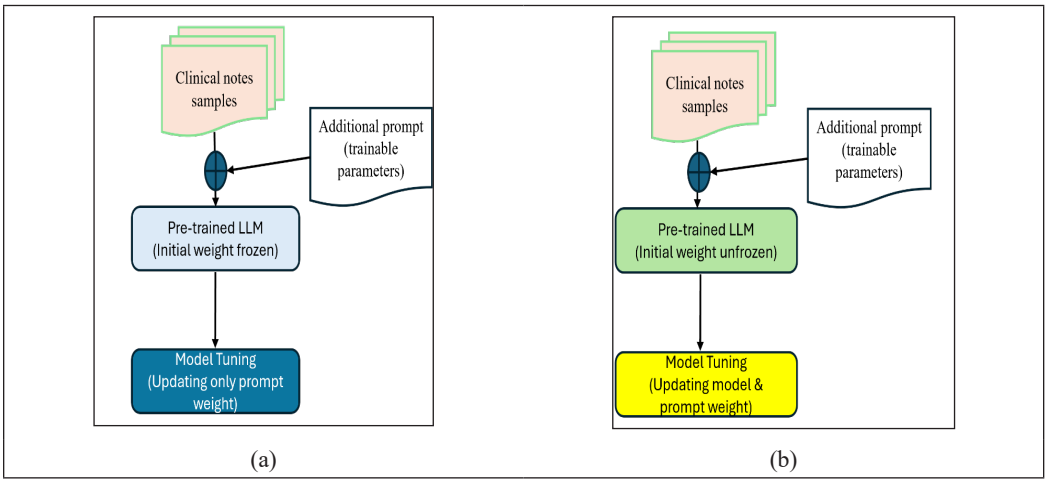

Figure 2. Soft prompt model tuning strategies. (a) Frozen model tuning and (b) unfrozen model tuning

Datasets

The N2C2 2018 dataset (Henry et al., 2020), derived from clinical narratives, was used for the ADE extraction challenge. This dataset contains annotations for nine entities: drug, strength, form, dosage, frequency, route, duration, reason, and ADE entities, all linked to a drug entity as their source. There are eight possible relations between these entities. The proposed model and methods were trained and evaluated using the official dataset splits, which consist of 303 training records and 202 testing records.

The TAC 2017 dataset (Belousov et al., 2019) consists of 200 drug labels in XML format, divided into a training set of 101 labels and a test set of 99. The dataset contains five attributes related to Adverse Drug Reactions (ADR): Animal, Drug Class, Factor,

Negation, and Severity. Moreover, the dataset has 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)

Prompt Initialisation and Projection

To ena ble the adaptation of LLM models on the downstream token-level ADE extraction tasks with task-specific objectives in the tuning process, a learnable prompt is prepended to the input embeddings generated by the embedding layer of the model, as in Equation 3. The procedure initially used a textual prompt and then converted it to a vector as in Equation 4.

$$E=\text{Embed}(X)\in R^{Bz*L_X*d_{model}} \quad [3]$$

Where X is the input sequence, Lx is the length of the input sequence, and Embed is the transformer embedding layer.

$$P \in R^{(1,0)*L_p*d_{model}} \quad [4]$$

Where P is the prompt vector, L_p is the prompt length and d_{model} The model dimension is 768 for the BERT-based models. The prompt is then projected through the linear transformation, as in Equation 5.

$$P= W_p P + b_p \quad [5]$$

Where W_p is the projection weight, and b_p is the bias term. The projected prompt is broadcast to the batch size (B_z) as in Equation 6.

$$P'=\text{broadcast}(P) \in R^{Bz*L_P*d_{model}} \quad [6]$$

One major modification to the original model setting is expanding the maximum sequence length. The experimental PLMs use a fixed sequence length (512 for the BERT-based models). To enable the model to process the broadcast soft prompt embeddings as in Equation 6, the method extended the model's maximum input sequence length to accommodate the combined input sequence as in Equation 7. Consequently, the attention mask in Equation 8, token type IDs in Equation 9, and sequence labels were also padded to match the new sequence length as in Equation 10. The procedure for creating the extended input is outlined in Algorithm 1, presented in Figure 3. Only the prompt projection parameters (Wp, bp) gradient are updated while the model parameters remain frozen.

$$C_{input}=[P';E] \in R^{Bz*(Lp+Lx)} \quad [7]$$

$$M = \left[1^{Bz*Lp}; Mx\right] \in \{0,1\}^{Bz*(Lp+Lx)} \quad [8]$$

$$T = \left[0^{Bz*Lp}; Tx\right] \in \{0,1\}^{Bz*(Lp+Lx)} \quad [9]$$

$$Y = \left[-1^{Bz*Lp}; Yx\right] \in Z^{Bz*(Lp+Lx)} \quad [10]$$

Algorithm 1: Procedure for Soft Prompting Input and Projection

Input: Task Input data, Top-k Prompt Tokens feature select, List of Prompt Template, Pre-trained Model, Number of Task

Steps:

For task in Number of Task **Do:**

1. *Pmt1* ← ***SelectPrompt****(Pmt, c_task_id)* #Get the prompt for the task from the list
2. *Pmt_tokenized* ← ***Tokenizer****(Pmt1)*
3. *Pmt1_emb* ← ***Embedding***(Pmt_*tokenized*) #embed the tokenized prompt using pre-trained model
4. *tokenized_input* ← ***Tokenizer(task_*Input data***) # tokenization
5. *X__input* ←***Embedding***(*tokenized_input*) #Get the embedding of the input data
6. *X_con_input* ←***[Pmt1_emb ;X__input]*** #Concatenate the input and soft prompt
7. *P_input*← ***EmbeddingSpace***(*X_con_input*) #Projection Layer
8. *Mask* ← ***Expand*** *(mask, 1's, top-k*) #Expand the attention mask of batch input to the length of top-k prompt fill with 1's
9. *Pmt_label* ←***fill****(-1, len(top-k))*
10. *Task_ids* ← ***Expand*** *(task_ids, 0's, top-k*) #Expand the tasks ids of batch input to the length of top-k prompt fill with 0's
11. *Input_labels* ← *[X_input[label], pmt_lable]* #Expand the labels of batch input to the length of top-k prompt fill with -1's
12. *Extended_input* ← *[P_input, Mask, Task_ids, Input_label]* *# combine*

end for

Output: Extended_input

Figure 3. Procedure for soft prompt initialisation and projection

Multi-task Model Soft Prompt Tuning on ADE Extraction

The task is presented as a dual sequence labelling problem for the Adverse Drug Event (ADE) concept and relation extraction at the token level. Each token in the input sequence is assigned a corresponding label. An extended BIO tagging scheme, inspired by (El-Allaly et al., 2021) is used to assign labels to tokens in the sequence for each task. In the first task, all Adverse Drug Reactions (ADR) mentioned in drug labels and drug names in the clinical notes are initially identified to create a set of positive entities with at least one relation with mention attributes. In the second task, attributes of the initially identified positive entities are determined, followed by linking the relations between the entities and their attributes. To guide the Language Model (LM) on the downstream task, a soft

prompt tuning approach is employed using multiple prompts, one for each task, as detailed in the previous sections. A multi-task learning framework (Liu et al., 2019) is utilised to model the two tasks simultaneously. The tasks are input at the input layer, and the PLM generates a shared representation, which is then passed to the fully connected layer for final token-level classification. This research investigates the capabilities of small-scale LLMs; as such, the models' parameters are left unfrozen to leverage the transfer learning abilities of the model for improved performance with minimal resources.

Human-in-the-loop Interactive Model Optimisation

During the interactive optimisation process, the initialised best model obtained from soft prompt tuning (unfrozen) is subjected to sample data from the validation set, one task at a time. The model makes an initial prediction on the sample data to predict a label for each token in the sequence. The model output is displayed to the user via an interactive interface using an alignment plot diagram, which shows the token, the label predicted by the model for the token, and the gold standard label of the token side by side for the human-in-the-loop. A typical example is shown in Figure 6. Additionally, the model's attention weight assigned to each feature, which are the tokens in the sequence, indicates the attention given by the transformer self-attention mechanism to each token that influences the model's decision-making. These attention weights are displayed using attention saliency heatmaps and other visualisation methods to improve the human-in-the-loop's understanding of model prediction results and decision-making processes. The human-in-the-loop provides feedback based on model prediction and diagnosis. An interactive interface allows the human-in-the-loop to accept model predictions and proceed to the next instance or provide corrective feedback to the misclassified token via JSON file upload or through manual labelling from the drop-down list of gold standard annotation labels. This process continues iteratively until all validation instances are executed.

Fine-tuning transformer-based models involves updating the initial pre-trained model parameters to adapt the pre-trained knowledge to the new task. This task requires saving the updated parameters for inference. To handle this constraint and avoid continued perturbation of model parameters for each instance, a list is created for each task that stores all the corrective feedback during the iterations. When all the validation instances are exhausted, the model is retrained to incorporate human feedback for a few epochs, saving model checkpoints at each epoch to mitigate catastrophic forgetting by reverting to the previous checkpoint whenever forgetting occurs. Further training for a few epochs prevents the model from overfitting the small sample from human-in-the-loop feedback. In this research, an optimal performance was achieved with 2 epochs. The final refined model is saved for inference.

The saved model is loaded for testing on the test set during inference. The model makes predictions on the test sample. The predicted ADE concepts and relations are then extracted using the extraction module. An official evaluation script provided by the challenge organisers is used to evaluate the performance of the final model. The overall procedure is given in Algorithm 2 in Figure 4**.**

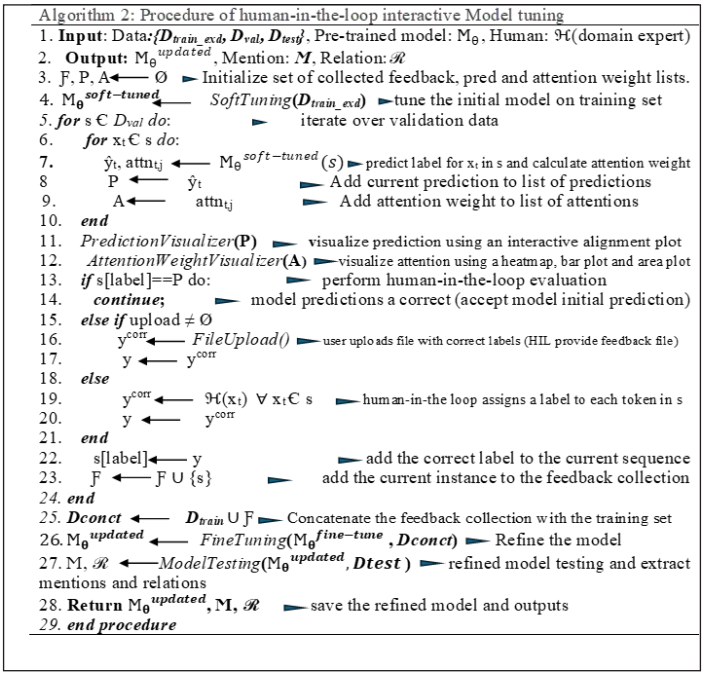

Figure 4. Procedure for human-in-the-loop optimisation

Interactive Interfaces

Interactive machine learning (IML) has recently become a popular approach to achieving human-machine collaboration, a concern of Industry 5.0 (Kumar et al., 2024). However, the two key questions are: How will the human-in-the-loop of model development understand the prediction results provided by the model and how to diagnose the model decision to comprehend the cause of misclassification? To address this concern, the proposed method encompasses two primary avenues for effective interaction between the user and the model development process. On the one hand, the human-in-the-loop component provides continuous feedback with correctional labels to the misclassified samples throughout the model refinement process. On the other hand, the interaction visualisation component allows for human-in-the-loop and model exchange. To this end, three visual interfaces are designed for model explanations and refinement. Firstly, to improve user understanding of the model prediction result, the method utilises an interactive alignment plot to effectively

display each token with its corresponding prediction label, with a hover over the actual label. This enables even non-technical experts to understand the result of the model prediction. Secondly, the method utilises the transformer model self-attention weight that indicates the importance of each token to the overall model decision-making to predict output for interpreting and diagnosing model decision-making. Lastly, the work designed an interface to enable users to interact and provide corrective feedback for model refinement. Details of each interface are given in the subsequent sections.

The diagram in Figure 5 shows a typical interaction for concept extraction subtasks, showcasing the sequence of activities to accomplish model predictions and refinement. Figure 6 shows the main user interface for the system interaction with a human-in-the-loop.

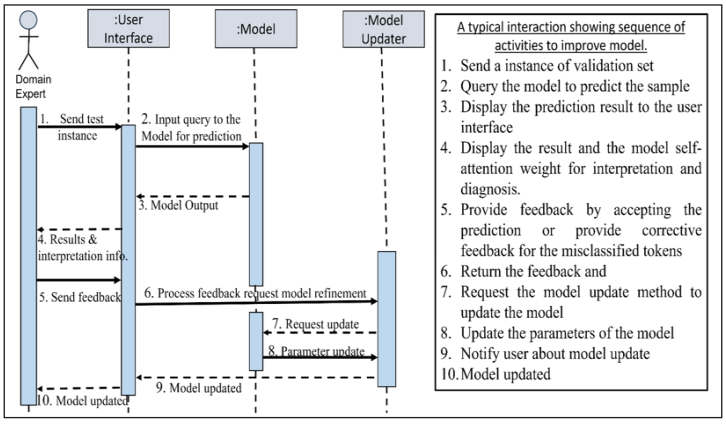

Figure 5. Sample UML interaction diagram showcasing the task of single instance concept extraction

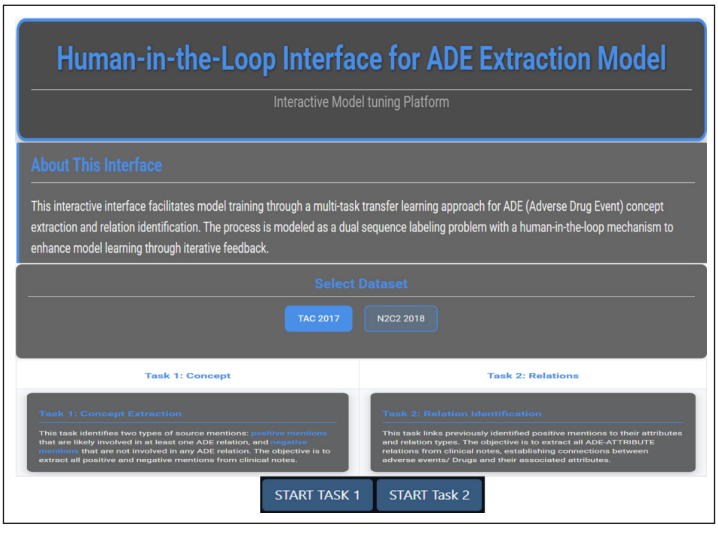

Figure 6. Main page of the interactive system

Model Prediction Understanding Interface

The success of explanations in an IHITLL-based model depends mainly on how well humans can comprehend and interpret their predictions. Inadequate explanations may confuse users and introduce unintended bias during the model refinement (Kumar et al., 2024). Sequence labelling involves training the model to predict a corresponding label for each token within the input sequence, as in Equation 11. Each token is displayed with its corresponding predicted label, as in Equation 12.

$$X= (x_1,x_2……..x_n) \quad [11]$$

Where x_1 is the i^{th} token in the sequence, and n is the number of tokens in the sequence.

$$\hat{y}_n=\text{argmax } (Py_n|x) \quad [12]$$

where $P(y_n|x)$ is the probability distribution over labels for the token x_n given the entire sequence x, using a prediction function defined as $\hat{y}_n=\boldsymbol{f}(x_t;\theta)$ where $\boldsymbol{f}(\cdot;\theta)$ is the model function parameterised by weights θ.

To effectively increase the understanding of model results, especially for non-medical experts, the method uses a Plotly Python library to design an interactive alignment plot diagram sample shown in Figure 7 displaying each token aligned to its corresponding predicted and actual label side-by-side. This gives a granular insight into the model's prediction results, allowing a fine-grained understanding of the prediction. The work employed a colour-coding scheme to differentiate between the correctly predicted tokens (designated by green) and the misclassified tokens (designated by red) as defined in Equation 13.

$$\text{Colour}_t = \begin{cases} \text{green, if } \hat{y}_n = y_n \\ \text{red, if } \hat{y}_n \neq y_n \end{cases} \quad [13]$$

When, $\hat{y}_n$ is the predicted label, and y_n is the true label

Hover text and colour coding present interpretable feedback to the user, enabling even the non-technical expert to comprehend model predictions intuitively. Hovering over predicted and actual labels for each token allows the user to evaluate the model's performance for a given instance quickly. Figure 7 depicts an example outcome for the sequence input “intrathecal methotrexate hists”.

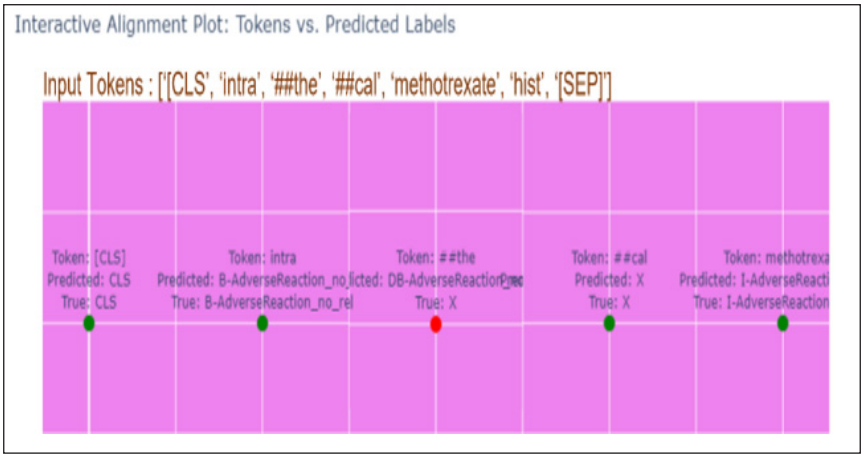

Figure 7. Sample Interactive Alignment Plot diagram for model prediction understanding

Transformer Self-attention Weight-based Model Diagnosis Interface

In sequence labelling tasks, each word or sub-word is a feature for the model's decision-making process. During the contextual representation of the input sequence, the transformer model utilises self-attention mechanisms to determine the importance of each feature within the input sequence. The self-attention mechanisms calculate the score for each token, showing how much attention a particular input should pay to other elements in the given sequence, irrespective of their relative distance. In essence, the transformer self-attention compares all input sequence members with each other and assigns importance, as in Equations 14 and 15.

$$\alpha_{t,i} = \frac{\exp(\text{score}_{t,i})}{\sum_{j=1}^{T} \exp(\text{score}_{t,j})} \quad [14]$$

where $\alpha_{t,i}$ represents the attention weight assigned to token x_i when predicting the label for token x_t, and $\text{score}_{t,i}$ is the attention score, calculated as:

$$\text{Score}_{t,i} = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad [15]$$

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

Expert feedback is considered a feature that the model is backpropagated by the update methods, as in Algorithm 2. The soft prompt weights are appended to the feature representations so that they can be backpropagated jointly. At the classification layer, the

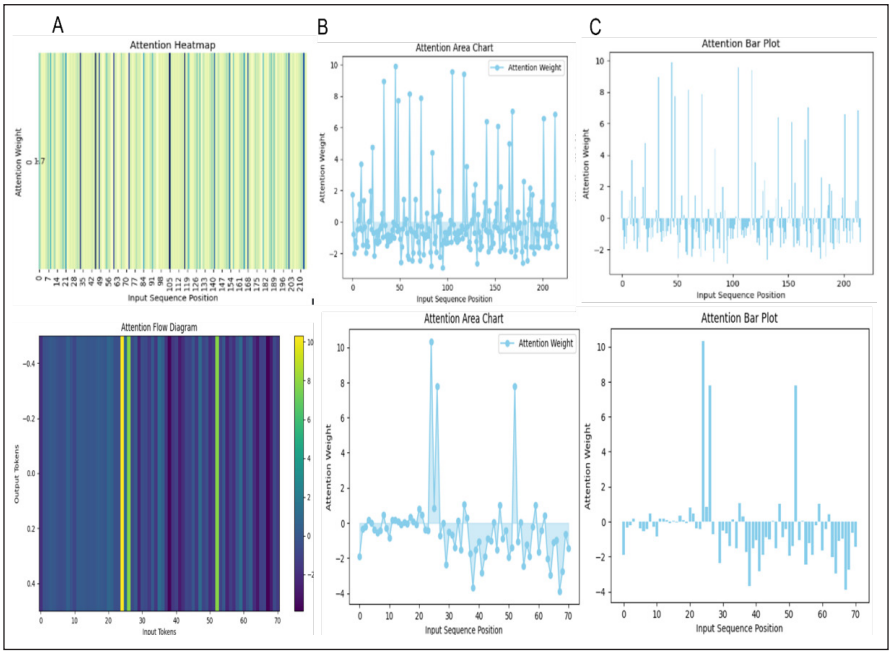

Figure 8. Sample attention interpretation for some input sequence. (a) Attention heatmap/flow (b) Attention area plot (c) Attention bar plot

soft prompt tokens are excluded from the softmax operation, as they are assigned negative label values. The weight update strategy modifies the trained model parameters for a limited number of iterations to incorporate human feedback. To bring the model transparency and interpretation to the human-in-the-loop, the work leverages the attention weight assigned to each token by the transformer self-attention mechanism to visualise its feature importance, thereby explaining the model diagnosis. Visualising each token's importance score helps the user understand which part of the input contributes significantly to the final model prediction, thereby enabling the user to ascertain whether a correct emphasis is given to each token or requires further adjustment of the model decision-making.

Although SHAP provides feature importance scores, it is computationally expensive to combine with transformer-based models and is less intuitive for textual data. In contrast, LIME relies on random perturbations, which may lead to unstable or unreliable explanations (Qian et al., 2023). Both methods are more suitable for post-hoc model interpretation, whereas attention-based approaches offer real-time explanations that can be directly leveraged by human experts.

To this end, the method utilises various alternative visualisation tools, including an attention saliency heatmap, attention flow heatmap, attention bar plot and attention area plot to display the attention weight for the input sequence -- typical examples are shown in Figure 8. The attention heatmap (a) uses a colour-coding scheme with a score to indicate

the relevant token: the darker the colour, the higher the importance score, and the relevance assigned to the input position, the lighter the colour. The bar plot (c) and area plot (b) showcase the relevant percentage of each input position in the sequence.

Visualisation of model judgements increases the chances of a practical model refinement process and the reliability of the end users.

Model Refinement Through Annotation Feedback

The interactive alignment plot and colour coding techniques to increase model prediction understanding allow for immediate visibility of misclassification and areas that require indispensable improvement. This visualisation presents an actionable insight into areas where the model should be refined, suggesting further model tuning. In addition, highlighting the prediction of the results can help spot misclassification, thereby facilitating error analysis of the model. This gives room for understanding why particular predictions go wrong and the pattern of how the model works. In addition, visualising model diagnosis gives clear insight into the model interpretation and explains its internal operation and decision-making to the human-in-the-loop.

To incorporate human feedback into the model's further optimisation process, the work developed an interactive interface to collect user input, as shown in Figure 8. When the model predicts a given instance, the human-in-the-loop can accept if all tokens are correctly labelled (Figure 9a). Otherwise, an expert provides corrective feedback for the misclassified tokens. Considering the longer sequences in clinical documents, the method developed an interface that allows users to upload a JSON file containing the corresponding labels for each token in the sequence (Figure 9b). This reduces the tedious burden of manually labelling each sequence, especially the ones with many misclassified tokens. Alternatively, iterative

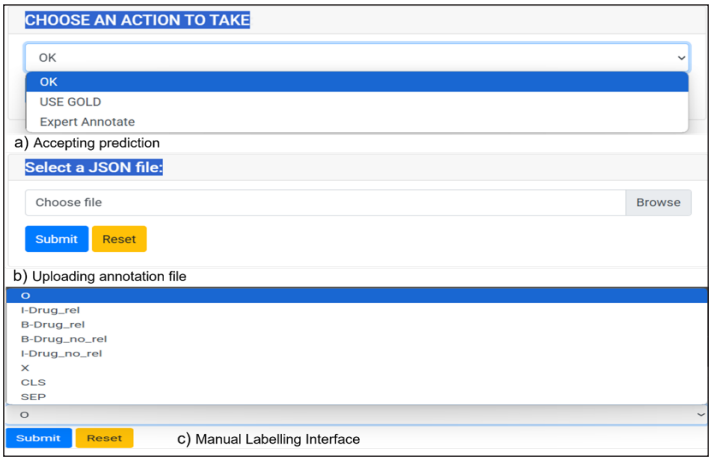

Figure 9. Designed interfaces. (a) Accepting model prediction (b) Uploading feedback (c) Manual labelling interface

labelling for each token in the sequence interface was equally provided to increase the system's usability, flexibility, and robustness for the end users (Figure 9c). A gold standard annotation was used as the reference lookup for the human-in-the-loop to make corrections to ensure unbiased feedback.

To incorporate human feedback, corrected labels are replaced with the model prediction for the misclassified tokens, and the model parameters are updated with new concatenated data during retraining, as shown in Equation 16.

$$\theta updated = \theta - \eta \quad _{\theta} L(D_{new}; \theta) \qquad [16]$$

Where θ are the model parameters, $\mathscr{L}$(Dnew;θ) is the loss function over the updated dataset, and η is the learning rate.

RESULTS AND DISCUSSION

This section summarises the experimentation conducted with the chosen model to optimise the soft-tuned model performance. This paper showcases how sample validation sets are utilised to further optimise the model based on the prediction result provided by the model and the corrective feedback provided by the human-in-the-loop model development. Table 1 shows sample experiments of model predictions on some data instances with user feedback.

Base Models

In this research, transformer-based models are utilised that have a stronger capability of contextually handling longer and more diverse sequences (Elbiach et al., 2025). Based on the soft prompting tuning performance of the base models on the training set, utilising 10% for validation, the work selected the best-performing model among the base models for the interactive human-in-the-loop optimisation experiment. The BERT model (Devlin et al., 2019) It is a PLM based on Masked Language Modelling for NLU tasks; the research utilised BERT-base. The SciBERT model (Beltagy et al., 2019) 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 Sci-BERT: sci-vocab and base-vocab. The work utilised the sci-vocab model. The models are tuned using the soft prompt tuning unfrozen (STU) strategy on multi-task ADE extraction.

Experimental Setup

The experiment was conducted on a single Tesla V100 GPU server running CUDA version 11.7 and a 16-core CPU. A maximum sequence length of 512 for fine-tuning and a batch size of 2 were used. The learning rate was set to 5e-5, with the cross-entropy loss function and the Adamax optimiser. A weight decay of 0.05 and a dropout rate of 0.1 to prevent overfitting

Table 1
Sample experiment of human-in-the-loop interaction for model tuning (TAC 2017 & N2C2 2018 datasets)

Input Sequence	Dataset	Task	Correctly Classified Tokens	Misclassified Tokens	Expert Feedback
"Heparin-induced thrombocytopenia. Secondary 1. End"	TAC 2017	ADE-concept	10	5	Provide a correct label to the misclassified tokens
"Seven-day course of ciprofloxacin and remained asymptomatic. 13. HTN Pt was normotensive with an episode of hypotension as described above. BP medications were held, and BP was monitored throughout his stay. It is recommended that his medications be started as an outpatient after monitoring his BP for hypotension and reevaluating his hypertension. 14. H o alcohol abuse Pt was monitored on CIWA, with prn Ativan. 15. FEN cardiac diabetic diet. 16. Prophylaxis Initially, SC heparin, pneumoboots when off heparin. Ambulatory towards the end of his stay. Bowel regimen, PPI. 17. Access PICC"	N2C2 2018	ADE-concept	111	38	Provide a correct label to the misclassified tokens
"Sedation with Versed 3.5, Fentanyl 75, and Phenergen 25 pt was noted to be apneic with an O2 sat of 77%. Bag ventilation was initiated with an increase in his Sats to 100%. He was given Narcan 400 mcg IM and Flumazenil 200 mg IV."	N2C2 2018	ADE-relation	50	12	Provide a correct label to the misclassified tokens
"Intrathecal methotrexate Hist"	TAC 2017	ADE-concept	6	1	Provide a correct label to the misclassified tokens
ASA 325, Amlodipine 10, Carvedilol 25 Hospital 1, Clobetasol q AM, Furosemide 40, Glipizide 5, Moexipril 30, Rosuvastatin 5, OxyContin 20. Discharge Medication"	N2C2 2018	ADE-relation	30	32	Provide a correct label to the misclassified tokens
"Contrast material opacifying"	N2C2 2018	ADE-concept	3	0	Accept model predictions
"Other serious infections: increased risk of bacterial, viral, fungal, and protozoal infections, including opportunistic infections and tuberculosis."	TAC 2017	ADE-relation	24	24	Accept model predictions
"Potential worsening of infections (e.g., existing tuberculosis; fungal, bacterial, viral, or parasitic infections; ocular herpes simplex)."	TAC 2017	ADE-relation	20	5	Provide a correct label to the misclassified tokens

were applied. The models were retrained for two epochs on the TAC 2017 dataset and ten on the N2C2 2018 dataset. Additionally, 10% of the training data was used for validation to optimise the model using an IHITLL process, and the test set was used for the final evaluation.

Experimental Results

This section presents the result of an interactive human-in-the-loop model fine-tuning of the transformer-based model for multi-task ADE extraction (IHITLFT-MADE) for concept and relation from N2C2 2018 and TAC 2017 datasets.

Result on N2C2 2018 Dataset

Figure 10 shows the experimental result for concept extraction. For the overall F1 score, the IHITLFT-MADE obtained 0.9404, which is ahead of BERT-STU with 0.9244 by 1.6% and SciBERT-STU with 0.9254 by 1.5%. Similarly, for RE, for the overall F1 score, the IHITLFT-MADE obtained 0.9132, which is ahead of the BERT-STU with 0.8916 by 2.16% and SciBERT-STU with 0.8925 by 2.07%.

A paired Student's T-test was employed to evaluate the performance of IHITLFT-MADE against the baseline model's performance and determine its significance. Table 2 shows that the results are statistically significant (p-values of 0.0010 and 0.0100) for concept extraction and (p-values of 0.0050 and 0.0010) for relation extraction.

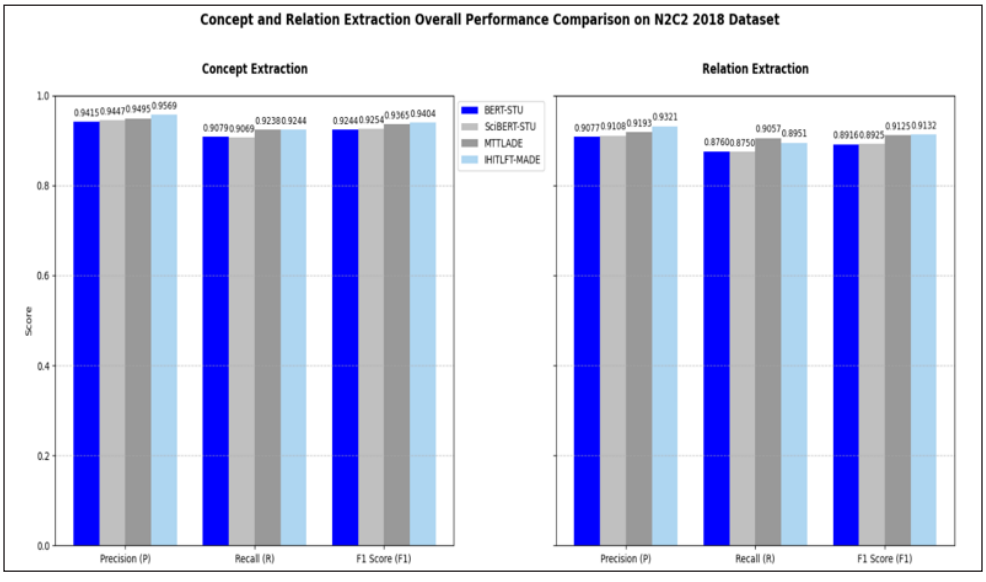

Figure 10. IHITLF-MADE result compared with baseline models: BERT-STU and SciBERT-STU for concepts and relations extraction on the N2C2 2018 dataset

Table 2
Significance test analysis of IHITLFT-MADE against baseline models on N2C2 2018

Measurement	System1	System 2	Statistics	Conclusion
		Concept		
Precision, Recall and F1-score (N=3)	IHITLFT-MADE Mean = 0.9405 SD =0.0162	BERT-STU Mean = 0.9246 SD = 0.0168	t = 50.2130 p = 0.0010	Significant
		SciBERT-STU Mean = 0.9256 SD = 0.0189	t = 9.7340 p = 0.0100	Significant
		Relation		
Precision, Recall and F1-score (N=3)	IHITLFT-MADE Mean = 0.9134 SD = 0.0185	BERT-STU Mean = 0.8917 SD = 0.0158	t = 14.1760 p = 0.0050	Significant
		SciBERT-STU Mean = 0.8927 SD = 0.0179	t = 59.7560 p = 0.0010	Significant

Note. Mean is the mean of precision, recall and F1 score, SD is the standard deviation of the score, N is the number of measurements, t is the statistical t-score, and p is the statistical p-value. $P < 0.05\%$ (95% confidence interval) indicates the performance is significant between the two systems. STU: soft prompt tuning unfrozen

Result on TAC 2017 Dataset

Figure 11 shows the experimental result for concept extraction; for the overall F1-score, the IHITLFT-MADE obtained 0.8723, which is ahead of the BERT-STU with 0.8532 by 1.91% and SciBERT-STU with 0.8708 by 0.15%. Similarly, for the relation extraction, for the overall F1-score, the IHITLFT-MADE obtained 0.5506, which is ahead of the BERT-STU with 0.5115 by 3.91% and SciBERT-STU with 0.5333 by 1.73%.

Table 3 shows that IHITLFT-MADE is statistically significant (p-values of 0.0050) over BERT-STU for concept extraction and (p-values of 0.0140) over SciBERT-STU for relation extraction.

Ablation Study

To evaluate the contribution of each component within the proposed framework, this section presents the results of experiments conducted with three configurations: baseline model fine-tuning, soft prompt only, and the human feedback mechanism. The three interaction interfaces. The first two interaction components, understanding and diagnosis, primarily utilise the model's output to enhance interpretability. In contrast, the refinement component directly contributes to model performance by incorporating human feedback. The outcomes of the ablation study are summarised in Table 4.

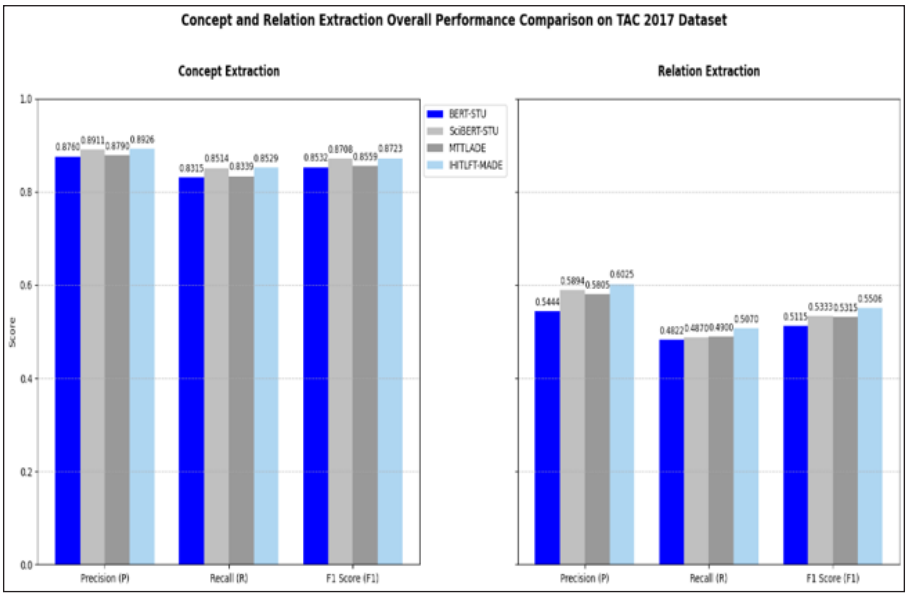

Figure 11. IHITLF-MADE result compared with baseline models: BERT and SciBERT soft prompt tuning (unfrozen) for concepts and relations extraction on the TAC 2017 dataset

Table 3
Significance test analysis of IHITLFT-MADE against baseline models on TAC 2017

Measurement	System 1	System 2	Statistics	Conclusion
		Concept		
Precision, Recall and F1-score (N=3)	IHITLFT-MADE Mean =0.8726 SD =0.0198	BERT-STU Mean = 0.8535 SD = 0.0222	t = 13.7320 p = 0.005	Significant
		SciBERT-STU Mean = 0.8711 SD = 0.0198	t = - p = - SRD = 0	Not Significant
		Relation		
Precision, Recall and F1-score (N=3)	IHITLFT-MADE Mean =0.5533 SD =0.0478	BERT-STU Mean = 0.5127 SD = 0.0311	t = 4.2160 p = 0.052	Not significant
		SciBERT-STU Mean = 0.5365 SD = 0.0512	t = 8.3690 p = 0.0140	Significant

Note. Mean is the mean of precision, recall and F1 score, SD is the standard deviation of the score, N is the number of measurements, t is the statistical t-score and p is the statistical p-value. $P < 0.05\%$ (95% confidence interval) indicates the performance is significant between the two systems. STU: soft prompt tuning unfrozen

Table 4
Ablation study

Metric	Baseline Model Fine-tuning				Baseline Model Soft Tuning				IHITLF-MADE	
	N2C2 2018 Dataset									
	BERT		SciBERT		BERT		SciBERT			
	concept	relation	concept	relation	concept	relation	concept	relation	concept	relation
P	0.9383	0.8885	0.9433	0.9121	0.9415	0.9077	0.9447	0.9108	**0.9569**	**0.9321**
R	0.9026	0.8257	0.9051	0.8715	0.9079	0.8760	0.9069	0.8750	**0.9244**	**0.8951**
F1	0.9201	0.8560	**0.9238**	**0.8913**	0.9244	0.8916	**0.9254**	**0.8925**	**0.9404**	**0.9132**
	TAC 2017 Dataset									
P	0.8614	0.5333	0.8919	0.5805	0.8760	0.5444	0.8911	0.5894	0.8926	0.6025
R	0.8164	0.4724	0.8464	0.4900	0.8315	0.4822	0.8514	0.4870	0.8529	0.5070
F1	0.8383	0.5010	**0.8685**	**0.5315**	0.8532	0.5115	**0.8708**	**0.5333**	0.8723	0.5506

ADE Extraction Outcomes

This section presents sample outcomes of the ADE extraction for concept and relation tasks. Tables 5 and 6 show the extracted concepts and relations for the N2C2 2018 and TAC 2017 datasets, respectively.

The models' predictions for each task extract the entities mentioned within the clinical text. An extraction module was created to extract all the mentioned segments. For instance, each mention has a segment, text, section start, and length for the TAC drug label. Based on the extracted positive mentions, their related attributes and relation pairs are extracted and written to the annotation files for comparison with the gold annotation files.

Table 5
Sample extraction outcome from testing set in N2C2 2018

Document	Concept Extraction	Relation Extraction
Clinical notes-100130	Drugs: Decadron, Oxacillin, Sulphate, Lipitor, Vanco ADE: eye discharge Dosage: 2 Form: drops Reason: seizure Prophylaxis, ophthalmic Involvement Duration: one week, 3 days Frequency: q.i.d Route: Intravenous Strength: 10 mg	Duration-Drug: Arg1: Decadron Arg2: one week Frequency-Drug: Arg1: Decadron, Arg2: q.i.d Reason-Drug: Arg1: ophthalmic Involvement, Arg2: Oxacillin Route-Drug: Arg1: Intravenous, Arg2: Oxacillin ADE-Drug: Arg1: eye discharge, Arg2: Dilantin Form-Drug: Arg1: drops, Arg2: Sulphate Strength-Drug: Arg1: 10 mg, Arg2: Lipitor Duration-Drug: Arg1: for 3 days Arg2: Vanco Dosage-Drug: Arg1: Lipitor, Arg2: 2

Table 5 (*continued*)

Document	Concept Extraction	Relation Extraction
Clinical-10059	Drugs: Vancomycin, Chemotherapy, Fluid, Metoprolol, Furosemide, Carbamazepine, Chlorhexidine Gluconate ADE: Tumour lysis Syndrome, beta-blocker toxicity. Dosage: 15 ml, 1 Duration: for 8 days Frequency: ONCE Form: Oral Rinse Route: PO Reason: VAP	Duration-Drug: Arg1: for 8 days, Arg2: Vancomycin Frequency-Drug: Arg1: ONCE, Arg2: Furosemide Reason-Drug: Arg1: VAP, Arg2: Vancomycin Route-Drug: Arg1: PO, Arg2: Carbamazepine ADE-Drug: Arg1: Tumour lysis Syndrome, Arg2: Chemotherapy Form-Drug: Arg1: Oral Rinse, Arg2: Chlorhexidine Gluconate Strength-Drug: Arg1: 400 mg, Arg2: Carbamazepine Dosage-Drug: Arg1: boluses, Arg2: Fluid Dosage-Drug: Arg1: 15 ml, Arg2: Chlorhexidine Gluconate

Table 6
Sample extraction outcome from testing set in TAC 2017

Drug Label	Concept	Relation
ACTEMRA	ADRs: infection, decrease in neutrophil counts, decrease in platelet count, Stevens-Johnson Syndrome, viral reactivation Severity: serious, below 1000 per mm 3 Negation: without, no Factor: Risk, may Drug class: immunosuppressive biologic therapies	ADR-Severity (Effect): Arg1: Decrease in neutrophil counts Arg2: below 1000 per mm 3 ADR-Severity (effect): Arg1: Infection Arg2: Serious ADR-Negation (Negated): Arg1: Decrease in neutrophil counts Arg2: without ADR-Factor (Hypothetical): Arg1: Stevens-Johnson Syndrome Arg2: Risk ADR-Drug class (Hypothetical): Arg1: Viral Reactivation, Arg2: immunosuppressive biologic therapies
ASCLERA	ADRs: Decrease visual acuity, Increase in bilirubin, ALT elevation, embryotoxic Severity: Grade 3-4, Grade 1-2 Negation: No Factor: Can, Intestinal perforation Animal: Rats	ADR-Factor (Hypothetical): Arg1: Decrease visual acuity Arg2: Can ADR-Negation (Negated): Arg1: Increase in bilirubin Arg2: No ADR-Severity (Effect): Arg1: ALT elevation Arg2: Grade 3-4 ADR-Animal (Hypothetical): Arg1: embryotoxic, Arg2: Rats

Complexity Analysis

Although the performance gain and improved extraction outcome, the model still struggles with certain complex entities and relations. The common one is the case of ambiguous entities within the dataset. For example, in the TAC 2017 dataset, the text input "no clear relationship between decreases in neutrophils". The term "no" here is ambiguous, as it can stand as a normal term or as a negation term. Similarly, for N2C2 2018 dataset, for inputs "patient had significant delay in recovery of mental status, initially attributed to build up of benzodiazepines used for sedation" and "His extubating was initially limited both by agitation requiring sedation and by requirement for high PEEP to maintain oxygenation.". the sedation here is ambiguous standing as reason for drug in the first input and as a drug in the second input.

Another form of complexity experienced in this implementation is the expansion of the standard sequence length of the BERT model to accommodate the appended soft prompt. This expansion could increase the computational resource usage. However, as a trade-off, it led to improved model performance.

DISCUSSION

The results presented for the proposed IHILTLFT-MADE for N2C2 2018 (Figure 10) and TAC 2017 (Figure 11) for concept and relation, respectively. The proposed method employed an interactive human-in-the-loop to optimise the adaptation of transformer-based models with human feedback. As shown in the tables, the results of the best baseline models before the incorporation of the human feedback and benchmarks are included for comparison. The IHITLTFT-MADE consistently outperforms all the interim models across all tasks and datasets. For instance, the challenging entities in N2C2 are ADE and reason entities due to their polysemous nature. For example, the dataset labelled the gold annotation "headache" as ADE and the reason. Similarly, relation types connecting the entities to drugs have been challenging types. The proposed method achieved 0.6717 and 0.7879 for ADE and reason entities and 0.6454 and 0.7385 for ADE-drug and reason-drug relations, ahead of the best baseline model SciBERT-STU with 0.5966 and 0.7330 for the entities and 0.5536 and 0.6747 for the relations. The consistent improvement of the proposed optimisation method over the interim methods demonstrated the effectiveness of the proposed hybrid framework for ADE extraction. This signifies the impact of incorporating human-in-the-loop into fine-tuning to refine the model prediction with correct labels for the misclassified instances. This study focuses on improving ADE extraction from clinical textual documentation using LLM-based techniques.

The statistical analysis test confirmed the significant difference between the IHITLF-MADE and all the baselines across all tasks and experimental datasets. This implies that the final model has a significantly greater prediction capability in detecting ADE cases than the interim models.

Erroneous extraction of clinical concepts such as drug names, adverse events, or indications for drug administration can introduce misinformation into patient records, thereby compromising treatment decisions and patient safety. Understanding the characteristics of low-confidence entities and the complex relations that models fail to capture is essential for designing effective post-processing checks and for incorporating real-world pharmacovigilance experts as human-in-the-loop participants during model redevelopment. The performance gains achieved by the proposed methods provide evidence that integrating human-in-the-loop learning with pretrained language model adaptation can bridge the performance gap left by traditional approaches. This research has demonstrated the potential to improve ADE extraction, drug safety surveillance, and clinical research in general. The proposed model is a foundation in addition to its immediate practical applications. The model achieves state-of-the-art performance on benchmark datasets, demonstrating its feasibility in supporting large-scale pharmacovigilance and clinical text mining. The graphical user interface and interaction interfaces provided enable direct involvement by domain experts during system development and practical use.

Comparison with Benchmark

Figures 10 and 11 summarise the comparison of the results between IHITLFT-MADE and MTTLADE for both N2C2 2018 and TAC 2017. The MTTLADE system has been a state-of-the-art study that has recently been shown to outperform all the systems that participated in the challenge organised by the two benchmark datasets (El-Allaly et al., 2021). For the N2C2 dataset, IHITLFT-MADE achieved an overall F1 score of 0.9404, ahead of MTTLADE with 0.9365 by 0.39% for the concept extraction task, and IHITLFT-MADE achieved an F1 score of 0.9132, ahead of MTTLADE with 0.9125 by 0.07% for the RE task. Similarly, for the TAC dataset, IHITLFT-MADE achieved an overall F1 score of 0.8723, ahead of MTTLADE with 0.8559 by 1.34% for the concept extraction task, and IHITLFT-MADE achieved an F1 score of 0.5506, ahead of MTTLADE with 0.5315 by 1.91% for the RE task.

The statistically significant analysis shown by Table 7 between IHILFT-MADE and the benchmark MTTLADE using the paired student test for TAC (p-value of 0.0110 for concept and 0.0060 for relation) and approximate randomisation test with randomisation value of 10,000 (p-value of 1.9996×10^{-5}) using the official script provided by the N2C2 challenge organiser shows that IHILFT-MADE's performance in predicting ADE cases is statistically significant compared to MTTLADE's performance.

Error Analysis

Although the proposed hybrid system has demonstrated state-of-the-art performance, the system cannot handle some complex entities and relations within the experimented datasets.

Table 7
IHTLFT-MADE against benchmark MTTLADE statistical significance analysis on TAC 2017 result

Measurement	Task	IHITLFT-MADE	MTTLADE	Statistics	Conclusion
Precision, Recall and F1-score (N=3)	Concept	Mean=0.8726 SD = 0.0198	Mean=0.8552 SD = 0.0226	t = 9.2680 p= 0.0110	Significant
	Relation	Mean= 0.5533 SD = 0.0478	Mean=0.5340 SD = 0.0453	t=13.3160 p = 0.0060	Significant

Note. Mean is the mean of precision, recall and F1 score, SD is the standard deviation of the score, N is the number of measurements, t is the statistical t-score, and p is the statistical p-value. $P < 0.05\%$ (95% confidence interval) indicates the performance is significant between the two systems

The most notable error cases in ADE concept extraction are on concepts with a small number of instances and abbreviations in the training set. For instance, in the TAC 2017 dataset, the text input "TMA was reported in 3 of 2258 patients", the ADR "TMA" only appeared a few times in the training set, so the system failed to extract a similar entity in the test set. Similarly, on N2C2 2018, rare instances of like "QID" drug frequency and "p.o." route of taking a drug are rarely mentioned within the training set. These rare instances with abbreviations are some of the specific cases in which the model failed to extract correctly. Other cases in concept extraction include ambiguous and polysemous entities. For the relation task, failures are related to multiple relations in a single sentence. For example, in the TAC 2017 dataset "The incidence of Grade 3 and 4 depressive disorders", the "depressive disorders" here is related to both "Grade 3" and "Grade 4" severity of ADR with an effect type of relation. The model identifies only one relationship. Similarly, in the N2C2 dataset, identifying the relation between the challenging entities of ADE and Reason entities to the drug was among the failure cases observed in the system.

CONCLUSION

This work presented a proposed interactive human-in-the-loop-based method to further optimise the model tuning through human feedback during the fine-tuning process. Current ADE extraction models require large, annotated datasets and offer limited interpretability for clinical use. The proposed method adopts the best-performing baseline and applies interactive learning techniques to optimise its performance on multi-task ADE extraction. Visualisation techniques are employed to improve human-in-the-loop understanding of model prediction and diagnosis. A transformer-based attention score assigned to features is utilised to interpret model decision-making. To refine the model, interactive interfaces are provided to incorporate human feedback with corrective labels for the misclassified samples.

The proposed method outperformed all the baseline models and the benchmark system in the two experimental datasets, demonstrating its capabilities in extracting more intricate entities and relations from clinical documents. The statistically significant analysis shows that the performance of the proposed method is statistically significant compared to the baseline and benchmark methods. The final evaluation of the proposed framework using an official evaluation script demonstrates its capability, showing a comparable extraction performance to the gold standard annotation provided by the challenge organisers. This implies that the proposed system can be employed for real-time, accurate and reliable extraction of ADE cases from narrative documents.

In future research, the work plan is to employ real-time clinical data and engage a real-time pharmacovigilance expert to interact with the system for identifying ADE instances within the narrative documents and to assess its robustness in handling imperfect or conflicting annotations. In addition, large language models with larger parameters will be investigated using a frozen model tuning strategy for parameter-efficient model adaptation and resource limitations.

ACKNOWLEDGEMENT

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).

SUPPLEMENTAL MATERIAL/DATA

All the datasets utilised for this research were obtained based on data usage agreements. Part of the agreement is that the datasets should not be shared with any third party without the organisation's prior consent. Consent and access to data can be obtained from https://N2C2.dbmi.hms.harvard.edu/data-sets. Codes are available at https://github.com/salisushagari/IHILT-MADE.

REFERENCES

Accenture. (2019). *Explainable AI: The next stage of human–machine collaboration*. https://www.accenture.com/gb-en/insights/technology/explainable-ai-human-machine

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., . . . Kaplan, J. (2022). *Training a helpful and harmless assistant with reinforcement learning from human feedback* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2204.05862

Belousov, M., Milosevic, N., & Dixon, W. (2019). *Extracting adverse drug reactions and their context using sequence labelling ensembles in TAC2017* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1905.11716

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3615-3620). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1371

Budd, S., Robinson, E. C., & Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis, 71*, Article 102062. https://doi.org/10.1016/j.media.2021.102062

Chen, K., Pang, Y., & Yang, Z. (2024). *Parameter-efficient fine-tuning with adapters* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2405.05493

Demner-Fushman, D., Fung, K. W., Do, P., Boyce, R. D., & Goodwin, T. R. (2018). Overview of the TAC 2018 drug–drug interaction extraction from drug labels track. In *Proceedings of the 2018 Text Analysis Conference (TAC 2018)* (pp. 1-10). National Institute of Standards and Technology. https://tac.nist.gov/publications/2018/additional.papers/TAC2018.DDI.overview.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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* (Vol. 1, pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

El-Allaly, E. D., Sarrouti, M., En-Nahnahi, N., & Ouatik El Alaoui, S. (2021). MTTLADE: A multi-task transfer learning-based method for adverse drug events extraction. *Information Processing & Management, 58*(3), Article 102473. https://doi.org/10.1016/j.ipm.2020.102473

Elbiach, O., Grissette, H., & Habib, E. (2025). Benchmarking large language models for adverse drug reaction extraction in social media and clinical texts. *Results in Engineering, 28*, Article 107362. https://doi.org/10.1016/j.rineng.2025.107362

Gómez-Carmona, O., Casado-Mansilla, D., López-de-Ipiña, D., & García-Zubia, J. (2024). Human-in-the-loop machine learning: Reconceptualising the role of the user in interactive approaches. *Internet of Things, 25*, Article 101048. https://doi.org/10.1016/j.iot.2023.101048

Gu, Y., Zhang, S., Usuyama, N., Woldesenbet, Y., Wong, C., Sanapathi, P., Wei, M., Valluri, N., Strandberg, E., Naumann, T., & Poon, H. (2023). *Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.06439

Henry, S., Buchan, K., Filannino, M., Stubbs, A., & Uzuner, O. (2020). 2018 N2C2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association, 27*(1), 3-12. https://doi.org/10.1093/jamia/ocz166

Jamil, S., Saha, S., & Misra, R. (2026). Enhancing adverse drug event extraction and summarisation for cancer drugs through large language models. *Journal of Biomedical Informatics, 178*, Article 105009. https://doi.org/10.1016/j.jbi.2026.105009

Kim, S., Kang, T., Chung, T. K., Choi, Y., Hong, Y. S., Jung, K., & Lee, H. (2023). Automatic extraction of comprehensive drug safety information from adverse drug event narratives in the Korea Adverse Event Reporting System using natural language processing techniques. *Drug Safety, 46*(8), 781-795. https://

doi.org/10.1007/s40264-023-01323-2

Kim, Y., & Kim, Y. (2022). Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models. *Sustainable Cities and Society, 79*, Article 103677. https://doi.org/10.1016/j.scs.2022.103677

Kumar, S., Datta, S., Singh, V., Datta, D., Singh, S. K., & Sharma, R. (2024). Applications, challenges, and future directions of human-in-the-loop learning. *IEEE Access, 12*, 75735-75760. https://doi.org/10.1109/ACCESS.2024.3401547

Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimising continuous prompts for generation. 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)* (pp. 4582-4597). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.353

Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4487-4496). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1441

Liu, X., Ji, K., Fu, Y., Tam, W. L., Du, Z., Yang, Z., & Tang, J. (2022). P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 61-68). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.8

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730-27744.

Peng, C., Yang, X., Smith, K. E., Yu, Z., Chen, A., Bian, J., & Wu, Y. (2024). Model tuning or prompt tuning? A study of large language models for clinical concept and relation extraction. *Journal of Biomedical Informatics, 153*, Article 104630. https://doi.org/10.1016/j.jbi.2024.104630

Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys, 55*(9), Article 194, 1-33. https://doi.org/10.1145/3561048

Quennelle, S., Malekzadeh-Milani, S., Garcelon, N., Faour, H., Burgun, A., Faviez, C., Tsopra, R., Bonnet, D., & Neuraz, A. (2025). Active learning for extracting rare adverse events from electronic health records: A study in paediatric cardiology. *International Journal of Medical Informatics, 195*, Article 105761. https://doi.org/10.1016/j.ijmedinf.2024.105761

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). *Parameter-efficient transfer learning for NLP* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1902.00751

Ren, Y., & Wang, Z. (2023). A tree-structured neural network model for joint extraction of adverse drug events. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 4938-4941). IEEE. https://doi.org/10.1109/BIBM58861.2023.10386041

Riesener, M., Kuhn, M., Schümmelfeder, S., Xiao, D., Norheim, J., Rebentisch, E., & Schuh, G. (2024). Active learning with pre-trained language models for named entity recognition in requirements engineering. *Procedia CIRP, 128*, 339-344. https://doi.org/10.1016/j.procir.2024.06.027

Sahoo, P., Singh, A., Saha, S., Chadha, A., & Mondal, S. (2024). Enhancing adverse drug event detection with multimodal dataset: Corpus creation and model development. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 11214-11226). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.667

Shelmanov, A., Puzyrev, D., Kupriyanova, L., Belyakov, D., Larionov, D., Khromov, N., Kozlova, O., Artemova, E., Dylov, D. V., & Panchenko, A. (2021). Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 1698-1712). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.145

Teso, S., Alkan, Ö., Stammer, W., & Daly, E. (2023). Leveraging explanations in interactive machine learning: An overview. *Frontiers in Artificial Intelligence, 6*, Article 1066049. https://doi.org/10.3389/frai.2023.1066049

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., . . . Le, Q. V. (2022). *LaMDA: Language models for dialogue applications* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2201.08239

Tian, Z., Zhang, H., & Wang, Y. (2024). Personalised soft prompt tuning in pre-trained language models: Bridging multitask transfer learning and crowdsourcing learning. *Knowledge-Based Systems, 305*, Article 112646. https://doi.org/10.1016/j.knosys.2024.112646

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998-6008). Curran Associates, Inc.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-rank adaptation of large language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2106.09685

Ong, J. C. L., Chen, M., Ng, N., Elangovan, K., Tan, N. Y. T., Jin, L., Xie, Q., Ting, D. S. W., Rodriguez-Monguio, R., Bates, D. W., & Liu, N. (2024). *Generative AI and large language models in reducing medication-related harm and adverse drug events: A scoping review* [Preprint]. medRxiv. https://doi.org/10.1101/2024.09.13.24313606

Wondimu, N. A., Buche, C., & Visser, U. (2022). *Interactive machine learning: A state-of-the-art review* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2207.06196

Xia, L. (2022). Historical profile will tell? A deep learning-based multi-level embedding framework for adverse drug event detection and extraction. *Decision Support Systems, 160*, Article 113832. https://doi.org/10.1016/j.dss.2022.113832

Zhao, Y., & Liu, J. (2021). Human-in-the-loop based named entity recognition. In *2021 International Conference on Big Data Engineering and Education (BDEE)* (pp. 170-176). IEEE. https://doi.org/10.1109/BDEE52938.2021.00037

Zitu, M. M., Owen, D., Manne, A., Wei, P., & Li, L. (2025). Large language models for adverse drug events: A clinical perspective. *Journal of Clinical Medicine, 14*(15), Article 5490. https://doi.org/10.3390/jcm14155490