Journal of Theoretical and Applied Information Technology
30th September 2025. Vol.103. No.18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SNOMEDTM: A NOVEL TRANSFORMER-BASED ARCHITECTURE FOR ADVERSE DRUG EVENT EXTRACTION FROM CLINICAL TEXT

SALISU MODI [1], KHAIRUL AZHAR KASMIRAN [2], NURFADHLINA MOHD SHAREF [3], MOHD YUNUS SHARUM [4]

Corresponding author: Khairul Azhar Kasmiran[2]
[1] Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Selangor, Malaysia & Department of Computer Science, Sokoto State University, Sokoto, Nigeria
[2,3,4] Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Selangor, Malaysia.

E-mail: [1]gs63125@student.upm.edu.my , [2]k_azhar@upm.edu.my, [3]nurfadhlina@upm.edu.my, [4]m_yunus@upm.edu.my

ABSTRACT

Extraction of adverse drug event (ADE) mentions and their attributes and relations within electronic health records are crucial for adequate pharmacovigilance studies and drug safety surveillance. Transformer-based large language models (LLMs) have recently shown promising results in this research. However, clinical LLMs are few and have a limited number of parameters. General LLMs are domain-agnostic models developed for varying NLP tasks. However, due to the domain-specific nature of ADE extraction with ambiguous, polysemous and infrequent entities, general LLMs lacking prior medical knowledge have been observed to perform sub-optimally in handling these complex situations within clinical narrative documents. Consequently, researchers further pre-train the general models on domain-related knowledge before finetuning them for the downstream clinical tasks. Nevertheless, this approach is associated with several risks. The model may overfit when the domain-specific data is small. In addition, catastrophic forgetting may occur. To propose a new architecture tailored to extract ADEs called the SNOMED Transformer Model (SNOMEDTM) pre-trained from globally standard medical knowledge bases. The process is in two phases: A new transformer architecture was designed and pre-trained on the medical terminology-based SNOMEDCT and MedDRA. The model is tuned using fine-tuning and soft prompt tuning for multi-task ADE concept and relation extraction tasks. This study experimented with two tuning strategies, frozen and unfrozen model parameters. The model's performance was evaluated using the TAC 2017 and n2c2 2018 clinical challenge datasets. On TAC 2017, the proposed model outperformed the five compared transformer-based models and the top five systems contributing to the TAC 2017 challenge for fine-tuning. On n2c2 2018, the model outperformed GatorTron-base for soft prompting with unfrozen model and JNRF systems. This research demonstrates the potential of incorporating prior medical knowledge into LLMs tailored for clinical research.

Keywords: *Adverse Drug Event, Fine-tuning, Soft Prompt Tuning, Prior Knowledge, Pretraining.*

1. INTRODUCTION

An adverse drug event (ADE) is any unforeseen effect caused by using drugs during patient care. Improving ADE cases identification and extraction from clinical narratives improved the overall patient medication and documentation. In addition, detecting and monitoring drug safety is crucial for pharmacovigilance studies conducted in pre-marketing and post-marketing. Clinical trials with volunteer patients were common during pre-marketing but often lacked complete information due to fewer volunteers and shorter trial durations [1]. The traditional approach of the spontaneous reporting system (SRS) at the post-marketing stage falls short due to the problem of under-reporting by the affected patient or the medical practitioners [2], [3]. Fortunately, electronic health record systems

(EHRs) have been widely used for patient record documentation for a decade. Natural language processing (NLP) approaches have proven to improve the outcome of pharmacovigilance studies in recent years by processing these documents [3]. Several NLP community challenges have been organised to extract ADE and its related information from clinical narrative text such as the Medication, Indication and Adverse Drug Events (MADE 1.0) 2018 challenge [4], the Text Analysis Conference (TAC) 2017 challenge [5] and the n2c2 2018 challenge [6]. Due to the dual nature of ADE extraction, researchers have handled the task using pipeline, joint task, or multi-task learning approaches [7]. Concept extraction involves identifying and extracting the main entities and their attributes mentioned within the dataset. Relation extraction involves identifying and extracting relation types between extracted main entities and their related attributes.

Recent methods for the ADE relation extraction task often achieve a low F1-score, especially for the challenging ADE-reason and drug-ADE relations [8] for several reasons, including the limited number of pair samples within the training dataset, the long distance between relation pair entities in the text, and the ambiguous and polysemous nature of medical terminology[9]. For instance, in the following three examples from the n2c2 2018 dataset extract, the "sedation" concept was annotated differently as reason, drug, and ADE, respectively, as shown below:

Example 1: "Patient had significant delay in recovery of mental status, initially attributed to build up of benzodiazepines used for **sedation**" [B-Reason]

Example 2: "His extubation was initially limited both by agitation requiring **sedation** and by requirements for high PEEP to maintain oxygenation." [B-Drug]

Example 3: "Morphine 15 mg Tablet Extended Release Sig: One (1) Tablet Extended Release PO once a day as needed for pain: hold for **sedation**, RR< 12" [B-ADE]

The traditional embedding models that do not consider word context may generate the same representations for the concept, leading to an incorrect data representation of the input sequence.

Transformer-based models are pre-trained Language Models (PLMs) trained on vast amounts of unlabelled data through self-supervised learning. General language models such as BERT and its variants, Roberta [10], etc., and GPT [11] have been fine-tuned for ADE extraction tasks. However, these models are domain-agnostic, trained on data far from the specific downstream clinical tasks and, as such, may lack some commonsense knowledge of the structures and patterns of language constructs for practical ADE extraction. Recent studies have proposed tackling the domain adaptation problem of LLMs. The most notable is through further pre-training of the general models on domain-specific data. This process risks overfitting and catastrophic forgetting, affecting the generalizability and transfer learning capabilities of the model [12], [13].

Extracting ADEs from clinical documents using domain-agnostic LLMs has shown sub-optimal performance due to a lack of specific clinical knowledge. Pre-training on domain-specific data risks overfitting and catastrophic forgetting. The effectiveness of using globally standardized ADE-related terminologies like SNOMED-CT and MedDRA to develop a task-adaptive LLM for improved ADE extraction has not been investigated.

Prompt tuning approaches have been proposed to bridge the gap between the upstream LLMS' pretraining and downstream task-specific objectives. Prompting is an LLM adaptation technique in which additional tokens control the model for downstream tasks. These can be hard prompts, where non-trainable tokens control the model or soft prompts with learnable embeddings added to the input sequence of the downstream task to control the model. Different strategies are utilised for LLM models, such as keeping the model parameters fixed (frozen) or allowing them to be updated (unfrozen) during training [14].

This paper proposes a new transformer-based architecture trained on globally standard medical terminologies and concepts specifically used to report ADEs. This enables the model tailored for ADE extraction tasks to learn the patterns and construction of how terms are represented in downstream ADE extraction. The aim is to develop a task-adaptive model tailored for ADE extraction with prior medical knowledge of ADE concepts before tuning the model on ADE tasks. Additionally, this research uses the proposed methods [15] of multi-prompt soft prompt tuning with attention-based prompt tokens feature selection to tune the proposed architecture.

In summary, this paper's contributions are as follows.

1. This paper proposes a new transformer architecture called SNOMEDTM that is pre-trained based on masked language modelling objectives on SNOMED-CT and MedDRA general medical terminology and a synthesised dataset from the original knowledge bases as a complement.

2. This paper experimented with the proposed model with fine-tuning and soft prompting on the TAC 2017 and n2c2 2018 datasets. The proposed system outperformed the state-of-the-art models for clinical NLP and the top-ranking system for the challenges of ADE extraction.

The remaining sections of the paper are presented as follows: The following section details the review of the current models for clinical NLP tasks and, more specifically, ADE extraction. The section is followed by the methodology section of the paper, which provides details on how the proposed model was developed, pre-trained, and fine-tuned for multi-task learning problems. The details of the multi-prompt-based learning method is given. In the next section, this paper presents the experiment conducted and the results obtained. The subsequent section is the discussion section. The paper is concluded with a conclusion section.

2. RELATED WORK

This section provides a detailed overview of the literature on transformer-based large language models. It begins by elaborating on the limitations of general models for clinical NLP. Then, various transformation methods proposed to transform general models to domain-specific models, including further pre-training, knowledge distillation, and development of clinical task-adaptive models, are reviewed.

2.1 Transformer-based Clinical Large Language Models

With the current trend of increasing use of LLMs based on the transformer architecture, biomedical literature and unstructured textual documentation are extensively utilised to pre-train models tailored for clinical NLP-related tasks. The goal is to provide a substitute for the general LLMs, which have been shown to perform sub-optimally on biomedical-related downstream tasks and improve clinical healthcare delivery [16], [17]. Recently, there has been a rapid increase in developing multi-domain datasets, for instance, Dai et al. [18] proposed a multi-domain dataset for ADE extraction named dataset-CADECv2 by combining different data sources from clinical NLP, social media and weblogs. The authors experimented with the dataset using GPT-4 and Llama-3. Despite the models performing well across all the datasets, the models still struggle in identifying complex ADE cases and fall short in performance compared to domain-specific models pre-trained from a large collection of in-domain data.

Researchers have proposed many approaches to developing a clinical large language-based model to achieve this goal. One of the most prominent approaches involves creating a new model based on the architecture of a general model such as BERT with biomedical literature — for instance, Lee et al. [19] created BioBERT from PMC full-text articles and PubMed abstracts. Lo et al. [20] created SciBERT by initialising its architecture to that of BERT and pre-trained on full-text articles from semantic scholars. Alsentzer et al. [21] developed ClinicalBERT from clinical text and discharge summaries to generate clinical embeddings, and Liu et al. [22] developed RoBERTa, which eliminated BERT's next sentence prediction objective and changed the static masking of tokens to dynamic.

One widely adopted method for developing large language models (LLMs) tailored to clinical and biomedical fields involves initially pre-training general models on domain-specific data, followed by fine-tuning them for specific tasks within the domain. For instance Alrowili and Shankar [23] proposed BioM-ALBERT by further pre-training ALBERT before fine-tuning it on biomedical tasks. Similarly, McMaster et al. [24] created an ADE extraction framework using the DeBERTa model. which included pre-training on unannotated clinical texts and subsequent fine-tuning on labeled discharge summaries to classify documents based on the presence or absence of ADE. Another study [17] introduced a pharmBERT framework, which builds upon the original BERT architecture. This framework undergoes additional pre-training using drug labels obtained from the DailyMade dataset. Subsequently, the pre-trained model is fine-tuned to three NLP tasks: ADR detection, drug-drug interaction extraction, and ADME classification.

Additionally, decoder-based clinical language models have been developed, such as those

of Luo et al. [25]. They developed BioGPT from scratch based on GPT-2-medium architecture from PubMed articles. The model was fine-tuned for text generation and six other biomedical NLP tasks, including end-to-end relation extraction. Wang et al. [26] proposed ClinicalGPT by further pretraining on medical records and multi-round dialogue. Additionally, based on GPT architecture, Zhang et al. [27] proposed BiomedGPT based on masked language modeling and supervised learning pre-training objectives. The model was fine-tuned to multimodal tasks of vision language and image captioning. Furthermore, Yuan et al. [28] proposed BioBART based on BART architecture for text-infilling tasks and fine-tuned biomedical named entity recognition. Similarly, based on LlaMA-3 architecture, Wu et al. [29] proposed PMC-LLaMA by further pre-training the base model on biomedical academic papers and textbooks. The process begins with knowledge injection and instruction, fine-tuning medical conversation and answering medical questions.

However, despite the advantages of further pre-training of general models, as it enables the model to learn more about the distribution of words from the specific domain data, the shift from the initial pre-training parameters can affect the model's generalizability [13]. Additionally, the model can considerably adapt to the domain data, leading to overfitting during fine-tuning [30].

To address this problem, researchers have proposed a novel regularisation during further pre-training through self-distillation, where a student and teacher model was used. Lee et al. [30] introduced a self-distillation model, where a pre-trained model is further pre-trained using masked auto-encoding objective on domain-specific data to serve as a teacher to the student model fine-tuned for downstream tasks. In similar vein, Gu et al. [11], introduced a distillation model aimed at extracting ADEs. This approach utilizes GPT-3.5 as the instructive model to generate labelled sentences from unannotated data through self-supervised learning, which are then used to train the student model, PubMedBERT.

Incorporating domain-specific knowledge into a general domain model through further pre-training or self-distillation perturbs the initial optimal parameters of the general domain model. This can possibly lead to catastrophic forgetting [13]. Researchers have developed new models from scratch from biomedical and clinical tasks to mitigate the challenge. Recently, Yang et al. [31] developed a GatorTron model with about 8.9 billion words from electronic health records documents from UF Health 82 billion, MIMIC III 0.5 billion, PubMed 6 billion, and Wikipedia 2.5 billion. The GatorTron model is significantly improved over other models on most popular biomedical and clinical NLP tasks. However, this model was exceptionally trained from localised generated data from the UF health centre and general knowledge from Wikipedia and PubMed publications.

With the increased availability of globally standard knowledge bases, such as SNOMED-CT and MedDRA terminologies, these important thesauri have not been utilised to develop models tailored for medically related downstream tasks such as ADE named entity recognition and relation extraction, even though these two ontologies are the product of a carefully selected and comprehensive set of terminologies for clinical natural language text and electronic health record systems [32]. To address this problem, this paper proposed a new transformer architecture with about 138 million parameters initiated with pre-training on medical terminology using self-supervised learning with unlabelled data. The model is then fine-tuned for the ADE multi-task learning problem on two public datasets.

3. METHODOLOGY

This section details the proposed SNOMEDTM model, the pre-training and fine-tuning datasets, and the detailed architecture of the model. The model pre-training, fine-tuning, and soft prompt-tuning procedures for ADE extraction are elaborated.

3.1 Pre-training Dataset

3.1.1 SNOMED-CT and MedDRA terminologies

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) was released in 2002 by SNOMED International with 39 member countries. It combines two medical nomenclatures, the SNOMED Reference Term and Clinical Terms Version 3. It is one of the globally accepted comprehensive multilingual medical thesauri with over 350,000 medical concepts and over a million relations between terms. The knowledge base consists of 3 main components: the concepts, concept descriptions and the relationships between concepts [2]. This paper processed the

concepts and descriptions of the ontology component as the text data to pre-train the model. To access the thesaurus, the authors subscribed to nonprofit educational use of the terminologies upon signing a licence agreement from the National Library of Medicine (NLM)[1].

The Medical Dictionary for Regulatory Activities (MedDRA) is a standard medical terminology for drug monitoring and pharmaceutical companies developed by the Internal Conference on Harmonisation (ICH). It is available and translated into various languages for easy access to all nations. It comprises bidirectional hierarchical structures of five tiers for easy search and exploration using the MedDRA browser. At the top are 26 system organ classes (SOCs), followed by 332 high-level group terms (HLGTs), which is then followed by 1688 high-level terms (HLTs). The single medical concept Preferred Terms (PTs) contains over 24,000 terms. Finally, the lowest-level term comprises over 70,000 pharmaceutical and adverse drug events-related terminologies [33]. This study utilised MedDRA version 26.0 to extract the terminologies. The terminologies were accessed upon subscribing as a nonprofit organisation for educational research from the MedDRA organisation[2].

3.1.2 Synthesized data

PLMs work best with large amounts of pretraining data. This gives the model more to learn the vocabulary and construct of a given domain language. Due to the limited amount of data in the SNOMED-CT and MedDRA terminology knowledge bases, synthetic data was created to augment the pre-training dataset. The model consisted of a deep learning architecture based on Long Short-Term Memory (LSTM) to generate text, which has been proven effective in handling sequential data like narrative texts and medical concepts. The synthetic data was generated by training the LSTM model on a large corpus of medical texts from SNOMED-CT and MedDRA terms and their descriptions. The trained model is then used to synthesise new sentences word by word that mimic the style and content of the original data as expressed in Equation 1 below:

$$y_{t+1} = \text{argmax}\ \sigma(W_h h_t + W_x x_t + b) \quad (1)$$

where y_{t+1} is the predicted next word at time t, h_t is the hidden state of the LSTM at time step t, x_t is the current input embeddings, W is the learned weight, and b is the bias. The σ is the softmax function that converts prediction to probability. The synthetic data was combined with the original data to create a larger, more diverse dataset for pretraining the model. The procedure for generating the synthetic data is outlined in Algorithm 1, presented in Figure 1.

[1] https://www.nlm.nih.gov/healthit/snomedct/international.html

[2] https://www.meddra.org/basics

Journal of Theoretical and Applied Information Technology
30th September 2025. Vol.103. No.18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Algorithm 1: Procedure for generating synthetic pre-training data.

```
Input: T: SNOMED-CT, MedDRA terminologies, Model: LSTMmodel
       seed_text, seq_length, tokenizer
Output: Synthesized text, trained_model: Trained LSTM model
text_sequence ← PreprocessText(T)        ► tokenize the text and input sequences
 x_train, y_train  generated_text ← []    ►  initialize lists
for seq ∈ text_sequence do:
    for i in range (seq_length, len(s)) do:
        x_train ← seq[i-seq_length:i]     ► formulate x features, y target.
        y_train ← seq[i]
    end for
end for
trained_model ← TrainModel (Model, x_train, y_train)  ► train the LSTM model shown in Eq.1
 for _ in range(seq_length) do:
        tokenized_seed_text ← tokenizer(seed_text)
        word_g ← zero-initialize word
        predicted ← trained_model(tokenized_seed_text) ►predict the following words from seed text
        for word, index from tokenizer.word_index.items()
            if word == predicted
                word_g ← word      ► sample the next words from the predictions
                seed_text.append (word)
            end if
            generated_text ← seed_text
        end for
        synthesized_text ← generated_text
end for
return synthesized_text, trained_model
end procedure
```

Figure 1: Procedure for generating the synthetic data

The synthetic data was generated by training the LSTM model on a large corpus of medical texts from SNOMED-CT and MedDRA terms and their descriptions, one knowledge base at a time. The synthesised data generated and utilised to complement the original pre-training data constitute 10% of the overall data. During the synthesis, different seed values were utilised to generate new input sequences of certain word lengths. However, during data synthesis, the models were observed to be biased in generating repeated terms. A method was developed to detect and remove consecutive repeating words in the generated sequence to assess and mitigate the LSTM-based model bias in synthesizing the data. To evaluate the effectiveness of the synthesized text, two metrics were employed. The word error rate (WER) and BLUE score was calculated between a selected reference text and generated text. Table 1 below shows sample validation examples.

Table 1: Sample evaluation of the synthesized dataset.

Reference text (real text)	Seed text	Generated text	Evaluation
"Compounding refers to products that are usually made by a pharmacist or physician" "	"Compounding refers to"	"Compounding refers to products that are made by a pharmacist and drug company" "	WER=0.3076 BLEU=0.5783
"Neoplasm of esophagus"	"Neoplasm of"	"Neoplasm of esophagus"	WER=0.0000 BLEU=1.000
"Neoplasm of anterior aspect of epiglottis"	"Neoplasm of anterior"	"Neoplasm of anterior abdominal esophagus of"	WER=0.5000 BLEU=0.2060

"Chronic cold agglutinin disease associated with B-cell neoplasm"	"Chronic cold agglutinin"	"Chromic cold agglutinin disease due another virus infection"	WER=0.6450 BLEU= 0.4464
"Abuse for the purposes of term selection and analysis of MedDRA-coded data, abuse is the intentional, non-therapeutic use by a patient or consumer of a product"	"Abuse for the purposes of term selection and analysis of MedDRA-coded data,"	"Abuse for the purposes of term selection and analysis of MedDRA-coded data, abuse is the intentional, non-therapeutic by use a patient or consumer to of take a drug product"	WER=0.1923 BLEU= 0.6856
"medication error refers to the when a patient is prescribed, dispensed, or administered a drug that is documented in the drug label to cause an expected adverse event with patient's consumed food"	"medication error refers to when a patient is"	"medication error refers to the situation when a patient is prescribed, dispensed, or administered a drug that is documented in the drug label to cause hypersensitivity adverse reaction to in the patient"	WER= 0.2424 BLEU= 0.7298

Note: *A lower value of WER indicates better performance.*

Certain factors such as patient privacy and confidentiality, bias and fairness in data, validity, and reliability, among others, are essential when synthesising clinical data or using real-world clinical data for knowledge incorporation into models. However, the SNOMED-CT and MedDRA terminologies utilized in this research are completely anonymised and de-identified by the organisations in charge of the thesauri terminologies [2], [33]. The knowledge base does not contain any personal information of patients. To this end, the synthesised augmented data does not contain patients' personal or private information, making this research free of privacy concerns. Furthermore, the most observed bias during the data synthesis involves repeated duplication of medical terms. The problem was mitigated by removing consecutive repetition of the words. Synthesized data generated are qualitatively evaluated using BLEU and WER metrics before being incorporated into real-world data and human observation. Nonetheless, it is acknowledged that the synthesised data that constitutes only 10% of the overall pretraining lacks comprehensive representativeness of the original real-world data samples. However, due to the size limitations and its resemblance to actual data, the risks associated with the data may not affect the applicability and generalizability of the SNOMEDTM model to real-world ADE extraction.

3.2 SNOMEDTM Architecture

The SNOMEDTM is a transformer-based encoder model composed of multi-head self-attention layers and fully connected layers with layer normalisations. The transformer's self-attention makes it powerful in providing the contextual representation of each token in the input sequence [34]. The model's architecture comprises 16 transformer layers; for each layer, there are 16 attention heads, 768 hidden units, and a feed-forward size of 2024. The SNOMEDTM is made up of 138 million parameters. The transformer model takes the embedding vectors as input. Two embedding vectors are generated for each token for this model: the token embedding vectors and the token position embedding vectors. The token embedding vectors are vector representations generated for each token. The transformer includes an additional special classification token known as [CLS] at the beginning of the input sequence and the separator token [SEP] to designate the end of a given sequence. This paper adopts the transformer embedding layer to define the word embedding, as shown in Equation 2. The positional encoding was added to the word embedding to determine the exact position of each word in the input sequence. These enable the self-attention module of the transformer to emphasise each token based on its context for each sequence. The position encoding is defined in Equations 3 and 4 for even and odd positions, respectively. The final vector encoding is obtained by concatenating the two embedding vectors, as shown in Equation 5. Figure 2 shows the overall architecture and the pre-training settings.

$$We = \mathrm{E}(\mathrm{V}_{\mathrm{size}} \,.\, \mathrm{d}_{\mathrm{model}}) \quad (2)$$

Journal of Theoretical and Applied Information Technology
30th September 2025. Vol.103. No.18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

where E is the embedding layer, V_{size} is the vocabulary size, and d_{model} is the dimension of the model [34].

$$Pe(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/dmodel}}\right) \quad (3)$$

$$Pe(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/dmodel}}\right) \quad (4)$$

$$E_{emb} = \text{concat}(We, Pe) \quad (5)$$

The overall transformation of an input sequence (X) by a transformer encoder model can be expressed in the following equations [34]:

$$Q = XW_Q,\ K = XW_K,\ V = XW_V \quad (6)$$

Q, K and V are the query, key and value matrices obtained from the linear transformation of input embeddings using multi-head attention (Eq. 7), and W is the learned weight.

$$MultiHead(Q,K,V) = \text{concat}(H_1, H_2 \ldots H_{16})W_O \quad (7)$$

is the concatenation of linear transformation heads (16 for SNOMEDTM) where:

$$H_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (8)$$

$$Attention(Q_i, K_i, V_i) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (9)$$

where d_k is the hidden size (768 for SNOMEDTM). The residual connection is then performed, which involves adding up the input and the output for a given layer. The concatenated output is then normalised. The normalisation rescales the output to have zero mean and unit variance to stabilise them and speed up training as expressed in Equation 10:

$$\text{LN}(x_i) = \frac{x_i - \mu}{\sqrt{\delta^2 - \varepsilon}} \bullet \gamma + \beta \quad (10)$$

Where μ is the mean, δ is the variance of the input (x_i), and γ & β are learnable parameters and ε is a constant for numerical stability. The output of multi-head attention is normalized in Equation 11:

$$X_t = \text{LN}(X + \text{MultiHead}(Q,K,V)) \quad (11)$$

and the two linear transformations with activation by feed-forward layer are as in Equation 12:

$$FFN(x) = \text{GELU}(XtW_1 + b_1)\ W_2 + b_2 \quad (12)$$

followed by output normalisation in Equation 13:

$$X_{out} = \text{LN}(Xt + FFN(x)) \quad (13)$$

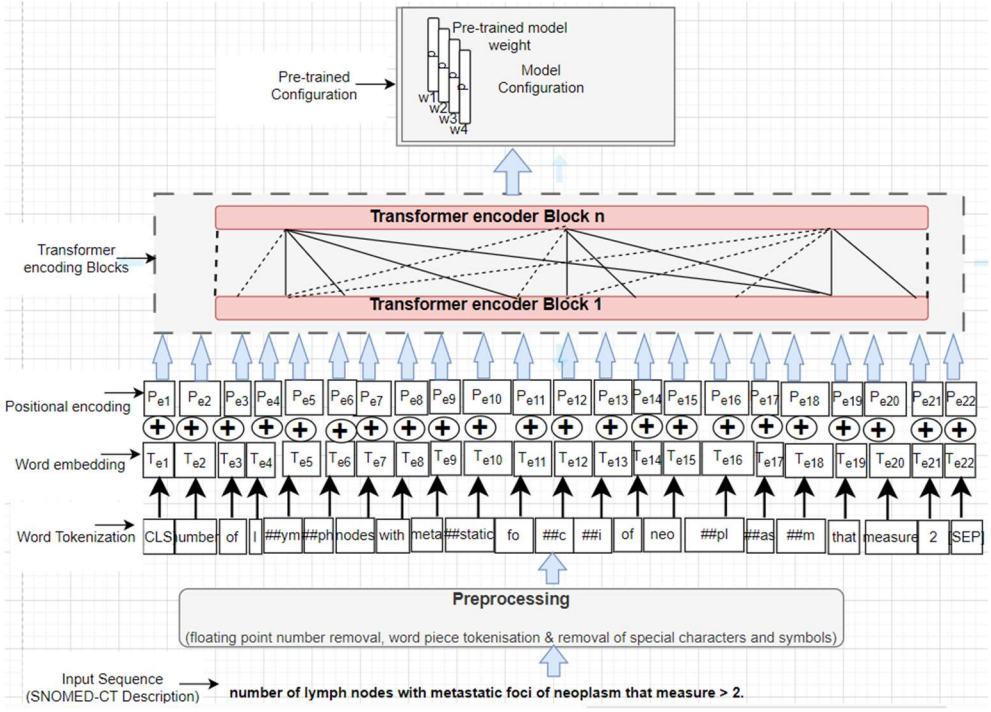

Figure 2: The architecture consists of an input sentence that is tokenised into word pieces and embeddings generated. The model has 16 attention heads, 16 encoder layers and 768 embedding sizes.

3.3 SNOMEDTM Pre-training

The SNOMEDTM was pre-trained from scratch using the masked language modelling objective proposed in BERT [35]. Masked language modelling is a self-supervised learning technique where certain parts of the input tokens are randomly masked, and the model is trained to predict these masked tokens. The 15% of the input tokens were randomly masked and used the model to predict the masked tokens. The model was trained for 200 epochs; since the dataset is small, this will give the model more chance to capture the syntax and the semantics of the terminologies. An early stopping condition was applied during the training. The training stops if there are three consecutive increases in the perplexity of training validation, indicating that the model is overfitting or is performing suboptimally. The larger the perplexity, the less confident the model is in its prediction [36]. The model reaches its optimal stage at about 12 hours. The observed perplexity does not increase or decrease much, with an average of 2.00. The formula for calculating perplexity is shown in Equation 14.

$$\text{Perplexity} = \exp\left(-\frac{1}{n}\sum_{i=0}^{n}\log P(w_i)\right) \qquad (14)$$

where n is the number of words in the test set, and $P(w_i)$ is the probability of the i^{th} word assigned by the model. The overall flowchart of the pretraining is shown in Figure 3a below. Figure 3b below shows the combined curves for the model pretraining accuracy and loss for training and evaluation.

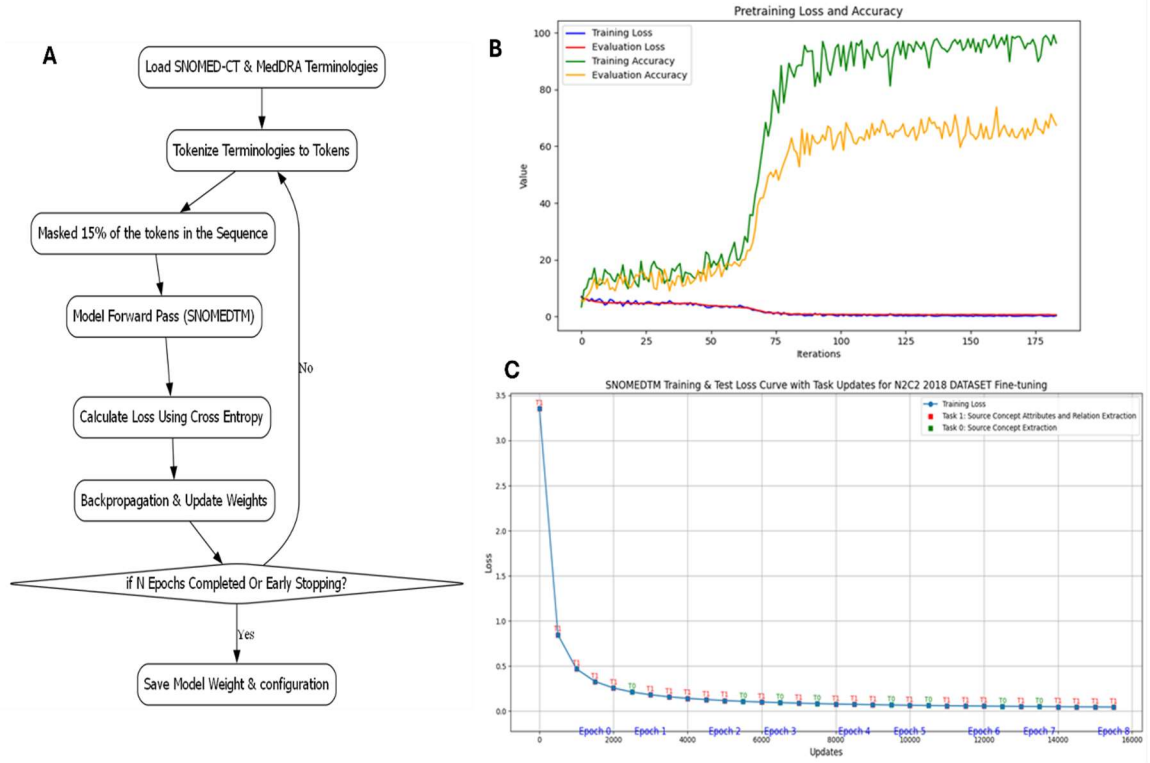

Figure 3:. SNOMEDTM Pre-Training And Fine-Tuning Process. (A) SNOMEDTM Pre-Training Flowchart. (B) Pre-Training Accuracy And Loss Curve. (C) Fine-Tuning Loss Curve For The N2c2 2018 Dataset.

3.4 SNOMEDTM Fine-tuning

The proposed model was fine-tuned on multi-task ADE-concept extraction and relation extraction as a dual sequence labelling on popular datasets: TAC 2017 and n2c2 2018.

During dual sequence labelling, the ADR-concept mention identification and ADR-concept attribute relation identification was done. ADR-concept identification classifies input sequences as positive (with relations) or negative (without relations). ADR-concepts attribute relation identification identifies attributes and relationships for positive concepts. An extended BIO tagging scheme

[10] manages discontinuous mentions and sub-words, with additional tags (DB, DI) for discontinuous concept beginning and inside and "X" for tokenizer-generated sub-words.

To implement multi-task transfer learning, the system adopted the MT-DNN framework [37] to simultaneously model the output of dual sequence labelling, promoting parameter sharing between the two related tasks. The framework is made up of three separate layers. The input layer receives the input sequences generated for each task and normalises them to the same length by padding shorter sequences or truncating longer sequences. The shared layer utilizes the pre-trained model to generate the contextual representation of the input sequence. This paper employs the proposed model SNOMEDTM incorporated with prior medical knowledge from standard medical terminology to develop the shared representation for the two related subtasks of ADE-source mention and ADR-source mention attribute relation extraction tasks. The output generated is then passed to the final classification of fully connected layers with SoftMax to obtain the final prediction for each token in the sequence. Figure 4 shows the overview of the fine-tuning architecture of the system.

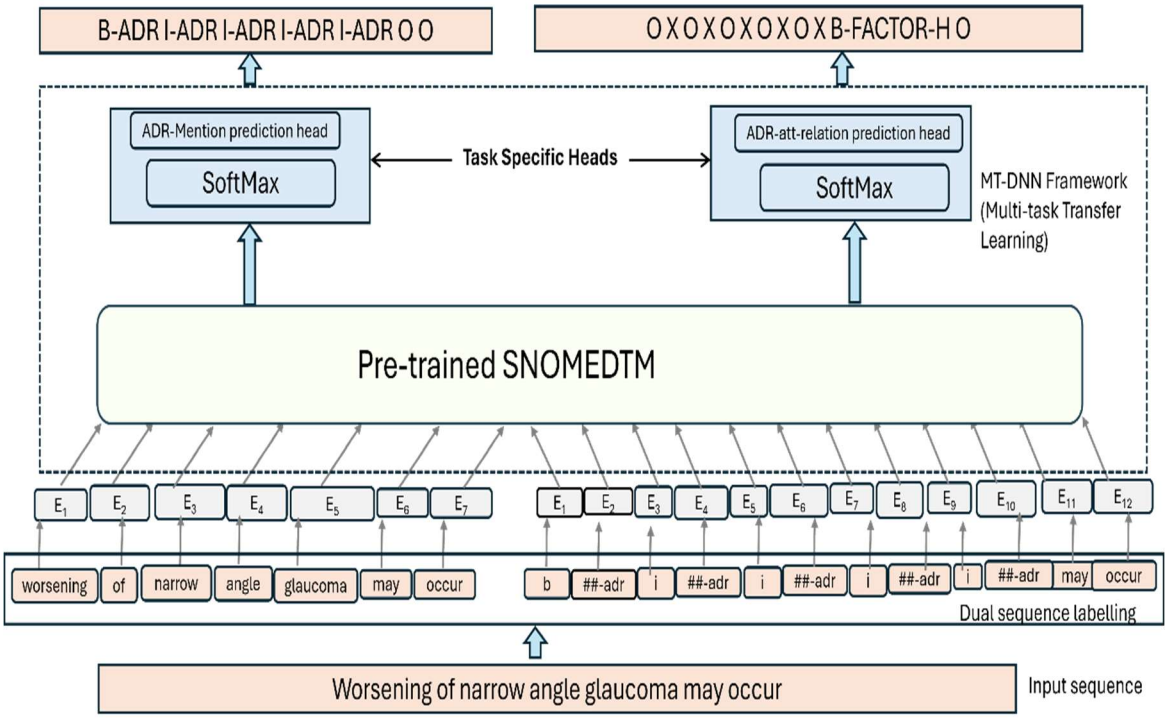

Figure 4: Fine-Tuning Takes An Input Sequence And Generates A Shared Representation For The Two Tasks By SNOMEDTM With A Task-Specific Head For The Final Classification.

3.5 SNOMEDTM Soft Prompt Tuning

To further test the capabilities of the SNOMEDTM model, the model was tuned based on the proposed multi-task soft prompting with prompt feature selection method [15] on the n2c2 2018 dataset. Multi-task learning involves modelling two or more related downstream tasks with different task-specific objectives. Utilizing a single holistic prompt to control the adaption of the LLMs may result in biased treatment of one of the tasks involved in the shared modelling. To deal with the problem, a multi-prompt-based soft prompting method was proposed as shown in Figure 5. Sample example input for the two tasks is shown in Figure 6, presented in a two-sequence labelling task, one for source mention labelling and the other for mention attribute and relation labelling. Two prompt templates were initialized, one for each task. For instance, the task 1 prompt: "Identify drug mentions and label each token in the sequence", and the task 2 prompt: "Identify drug attributes and relation, label each token in the sequence". The textual prompts were then

transformed into trainable embeddings. Firstly, the text was tokenized using a pre-trained tokenizer and then the embedding vector of both the input and the soft prompt was obtained using a pre-trained model embedding layer. Based on task type, the soft prompt was then prepended as a prefix to the input embedding.

However, researchers have argued that some prompt tokens hurt the performance of the LLM's downstream fine-tuning [38]. This paper applied a prompt token feature selection based on feature importance calculated using the transformer's attention mechanism to select the top important prompts to be prepended to the input sequence as trainable parameters to tune the model.

The overall procedure for multi-task ADE extraction using multi-prompt soft tuning is presented in Algorithm 2 in Figure 7.

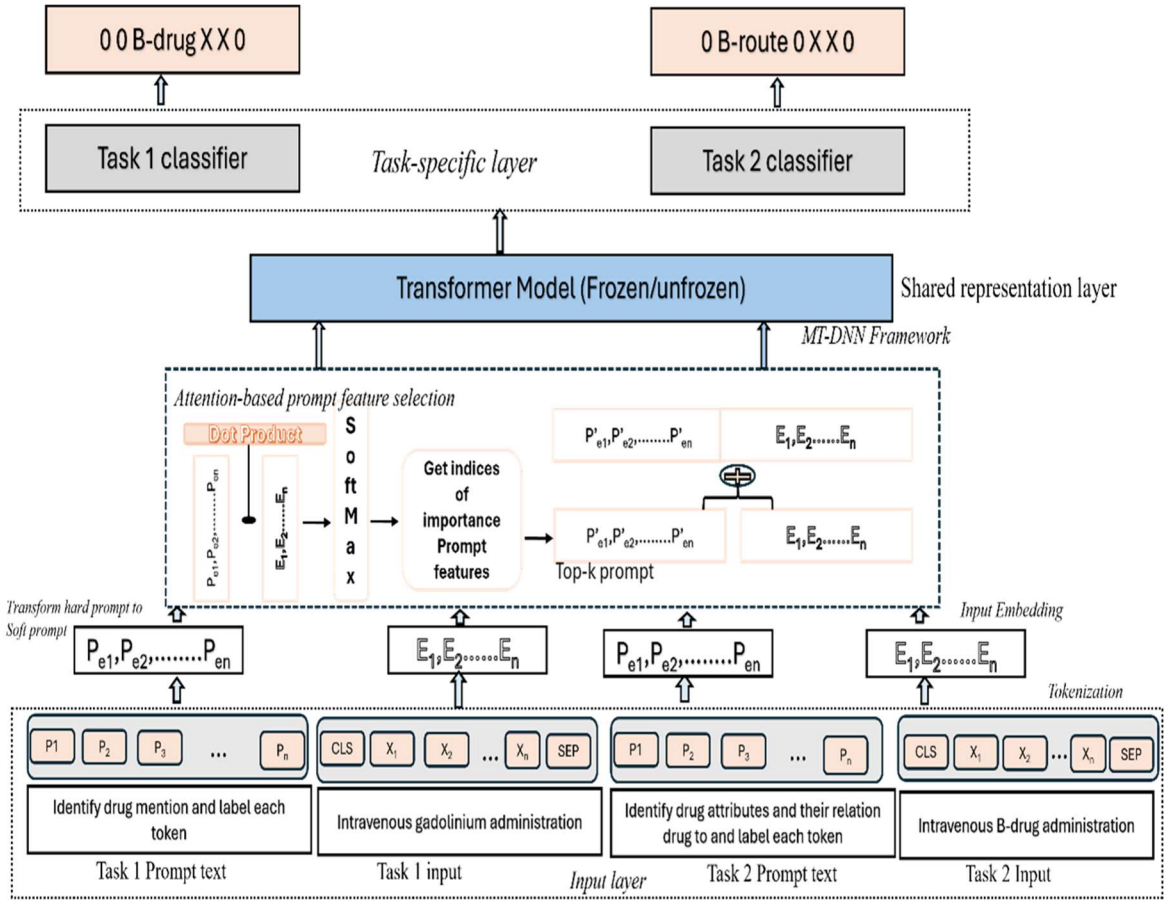

Figure 5: Soft Prompt Tuning Procedure.

Journal of Theoretical and Applied Information Technology
30th September 2025. Vol.103. No.18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

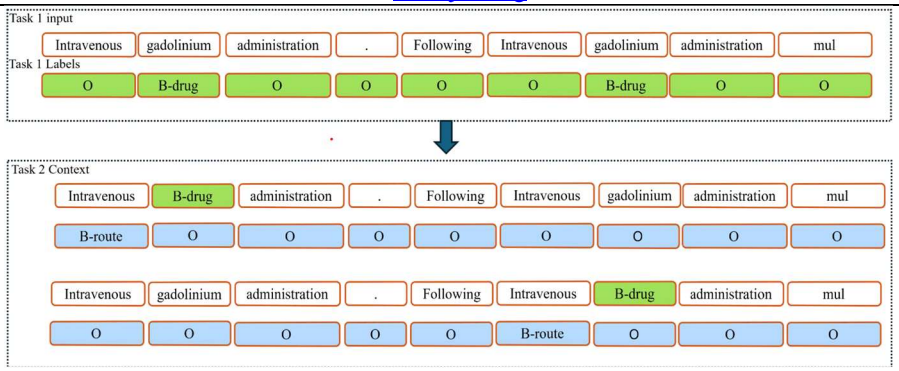

Figure 6: Example Of Task 1 Input And Task 2 Context

Algorithm 2: ADE extraction procedure with multi-prompt soft tuning

Input: ***S***: Clinical text sentences, ***Model***: Pre-trained model, ***P***: Prompt templates
Output: ***M***: a set of tuples of mentions, ***R***: Triplets of relations
M ← [], Initialize list for a set of tuples of mentions.
R ← [], Initialize list for a set of triplets of relations.
for s ∈ S do:
    task_id ← s[task_id] ► gets the id of the current task
    P_1 ← SelectPrompt(P, task_id)
    $P_{tokenized}$ ← tokenize(P_1)
    p_{emb} ← ***GetEmbedding***($p_{tokenized}$)
    I_{emb} ← ***GetEmbedding***(s)
    s_{input} ← ***SelectTopPromptTokens***($I_{emb,}$ p_{emb})
    $s_{expanded}$ ← Expand(s[att-mask], s[token-type-id], s[label], len(s_{input})) ► expand labels
    task1 ←Model($s_{expanded}$) ► feed into the Model
    M_{source} ← ***ExtractAllSourceMentions***(task1)
    for (m, t) ∈ M_{source} do:
        m_tuple ← <m, t> ► m for mention, t for type of mention
        M.append(m_tuple) ► add source mention to set of mentions
    end for
    $M_{p-source}$, C ←***GenerateContextFromPositiveSourceMention***(M_{source}, s)
    for (M_{sp}, t_{sp}) ∈ $M_{p-source}$ & c ∈ C do:
        task_id ← c[task_id] ► get the id of the current task
        p_2 ← ***SelectPrompt***(P, task_id)
        $p_{tokenized}$ ← tokenize(p_2)
        p_{emb} ← ***GetEmbedding***($p_{tokenized}$)
        c_{emb} ← ***GetEmbedding***(c)
        c_{input} ← ***SelectTopPromptTokens***($c_{emb,}$ p_{emb})
        $c_{expanded}$ ← ***Expand***(c[att-mask], c[token-type-id], c[label], len(C_{input}))
        task2 ←Model($c_{expanded}$) ► feed c into the Model to generate a sequence
        $M_{attribute}$, Re ←***ExtractAttributesMentionAndRelation***(task2)
        for (ma, ta) ∈ $M_{attribute}$ and r ∈ Re do:
            m_tuple ← <ma, ta>
            M.append(tuple) ► add source mention to the set of M
            r_triplet ← <msp, r, m>
            R.append(r_triplet) ► add relation triplet to the set of R
        endfor
    end for
end for
return M, R ► Return set of tuples of mentions and triplets of relations

Figure 7: Overall Procedure For Soft Prompt Tuning Of SNOMEDTM On ADE Extraction.

4. RESULTS

This section discusses the experiments conducted and the results obtained. First, the section elaborates on the details of the dataset used to fine-tune the model, followed by evaluation metrics employed to gauge the model performance. Next is the detailed experiment setup and the results for the traditional fine-tuning and soft prompt tuning of the proposed model on the TAC 2017 and n2c2 2018 datasets. The section further present experimental results on the ADE corpus dataset for both sequence classification and relation classification tasks.

4.1 Fine-tuning Datasets

The TAC 2017 [5] dataset consists of 200 drug labels in XML format. Of these, 101 labels are used for training, and 99 are reserved for testing. The primary entity in this dataset is the ADR entity, which has five attributes: Animal, Drug Class, Factor, Negation, and Severity. Additionally, the dataset includes three relation types: Effect (linking severity to ADR), Hypothetical (linking animal, drug class, or factor to ADR), and Negated (linking negation or factor to ADR). The second dataset is from the n2c2 2018 Adverse Drug Events extraction challenge [6]. It includes annotations for eight attributes: strength, form, dosage, frequency, route, duration, cause, and ADE, all linked to a drug entity. The model was trained and evaluated using the official splits: 202 records for testing and 303 for training. Further experiments were done on the ADE corpus dataset [39] to train and evaluate the model. This dataset includes 5,063 drugs, 5,776 adverse effects, and 6,821 relationships between them, all derived from 4,272 unique samples.

4.2 Evaluation Metrics

To measure the performance of SNOMEDTM on pre-training masked language modelling tasks, this research used the perplexity metric, as shown before in Equation 14. Perplexity is an evaluative metric that estimates the effectiveness of a probability model in making predictions for a given sample. It quantifies the level of uncertainty inherent in a model's ability to predict masked text. The system was evaluated using official scripts from the 2017 TAC and n2c2 2018 NLP challenges, with micro-average precision (Equation 15), recall (Equation 16), and F1-score (Equation 17) as the primary metrics for fine-tuning tasks and soft prompt tuning. Additionally, TAC 2017 employs an exact matching score, where a mention is deemed correct if its boundary and type match the gold mention, and a relation is correct if both the relation type and related mentions are accurate.

$$\text{micro-precision} = \frac{\sum TP}{\sum TP + FP} \quad (15)$$

$$\text{micro-recall} = \frac{\sum TP}{\sum TP + FN} \quad (16)$$

$$\text{micro-F1-score} = 2 * \frac{micro-precis \quad * micro-recall}{micro-precision+micro} \quad (17)$$

where TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives regarding prediction of ADE entities and relations.

4.3 Experimental Setup

The implementation of the proposed system was carried out in two phases: pre-training and fine-tuning. A maximum sequence length of 512, batch size 8, the BERT model type, and a learning rate of 3e-5, was used. The masked language modelling probability to mask the tokens was set to 0.15. The experiment was performed on a server with a single GPU Tesla V100 CUDA version 11.7 and a 16-core CPU computer.

Similarly, during the fine-tuning task, a maximum sequence length of 512 and a batch size of 8 were used. The training was run for 20 epochs for the TAC 2017 dataset and 30 epochs for the n2c2 2018 dataset. A 10% of the training set to validate the model and select the best model for inference, utilising the test set for the final evaluation.

A major persistent issue in LLM adaptation is possible catastrophic forgetting. Catastrophic forgetting usually occurs when there is too much perturbation in the initial pre-trained parameters of the model [13]. Approaches that include regularization techniques, early stopping conditions when there is a consecutive increase in the model training loss, and a minimal learning rate value during model adaptation, can be employed to minimise its occurrence. In addition, catastrophic forgetting can be mitigated by saving model updates at various checkpoints. This helps determine the model state that best generalises across tasks and allows the model to be recovered and reverted to its best state whenever catastrophic forgetting occurs. Recently, approaches such as soft prompt tuning [40], [41], adapter [42] and LORA [43]

techniques have been developed for adapting LLM models to downstream tasks where only the additional prompts or modules specific to the task are updated during training, leaving initial pre-trained parameters fixed. To address the issue, this paper employed an early stopping condition during pretraining to stop the model training and save the best model parameters when the validation loss increased consecutively for some experimented number of times. Similarly, at the fine-tuning phase, a weight decay of 0.01, a small learning rate of 5e-5, and dropout of 0.1, as well as saving the model state at intervals of 500 updates to mitigate possible overfitting and catastrophic forgetting as shown in Figure 3c. Furthermore, the study experimented with soft prompt tuning approaches (frozen and unfrozen); the result is shown in Table 4.

4.4 Experimental Results

This research demonstrates the potential of utilising a standard globally accepted medical domain knowledge base to incorporate prior medical knowledge into the transformer model before fine-tuning it on a downstream ADE extraction task. The proposed pre-trained model on SNOMED-CT and MedDRA terminology was used to generate a contextual vector representation for the multi-task ADR named entity exaction and relation extraction tasks. The shared representation is then passed to the individual task-specific classifier head through softmax to predict the final token class. Table 2 and figure 7 shows the results obtained in the TAC 2017 dataset, in comparison with other transformer-based models (BERT, BioBERT, BlueBERT, RoBERTa and SCIBERT) experimented with by [10], and DeepCADRME, proposed by [44]. These results indicate the strength of the task-adaptive SNOMEDTM model over the compared domain-agnostic and domain-specific PLM models in identifying ADR instances. Similarly, Table 3 shows the results obtained in the n2c2 2018 dataset compared with the state-of-the-art JNRF system proposed by [45]. These demonstrate the superiority of the transformer-based model over the foundational model based on the traditional Fourier network.

Moreover, Table 4 shows the results of soft tuning of the SNOMEDTM model, this study compared its performance with the system proposed by [14] for both frozen and unfrozen models during training. Despite the larger size of GatorTron-based compared to SNOMEDTM, the ADE-related knowledge incorporated in the SNOMEDTM model improves its performance on ADE extraction over GatorTron-base for the unfrozen model. SNOMEDTM is a transformer-based model; the model's pre-trained weights can be used for further fine-tuning on downstream clinical NLP for transfer learning.

Table 2: TAC 2017 Concept And Relation Extraction Results Compared With Other Transformer-Based Models (Fine-tuning).

Type	Metric (Overall)	BERT	BioBERT	BlueBERT	RoBERTa	SCIBERT	SNOMEDTM
Concept	P	86.14	86.63	86.52	85.35	87.90	88.91
	R	81.64	82.99	81.31	81.55	83.39	84.85
	F1	83.83	84.77	83.83	83.41	85.59	86.83
Relation	P	53.33	56.18	56.68	52.51	58.05	52.36
	R	47.24	48.74	48.55	45.60	49.00	52.92
	F1	50.10	52.19	52.30	48.81	53.15	52.63

Figure 7 depicts the distribution of overall F1 scores compared to TAC 2017 for concept and relation extraction tasks. The SNOMEDTM outperformed all the compared systems for concept extraction and is the second-top model for relation extraction.

Journal of Theoretical and Applied Information Technology
30th September 2025. Vol.103. No.18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

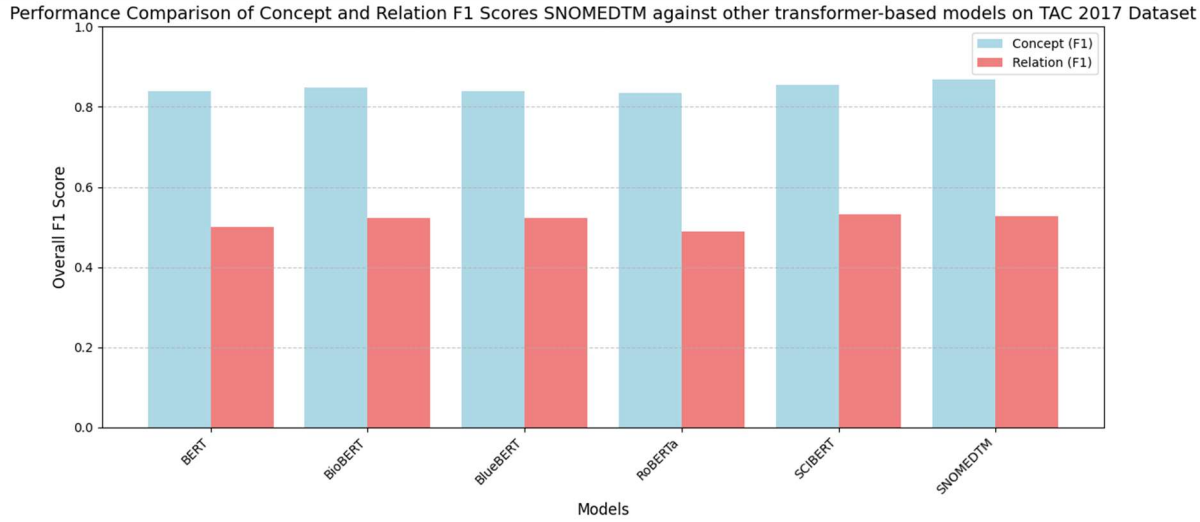

Figure 7: The Distribution Of F1 Scores Of The SNOMEDTM Compared To Other Transformer-Based Models On The TAC 2017 Concept Extraction Relation Extraction Task.

Table 3: Concept And Relation Extraction Results On N2c2 2018 Dataset Compared With The JNRF System (Fine Tuning).

Type	**Metrics (**Overall)	**JNRF**	**SNOMEDTM**
Concept	P	92.95	93.11
	R	84.76	84.81
	F1	88.67	**88.77**
Relation	P	90.97	88.29
	R	72.08	80.29
	F1	80.43	**84.11**

Table 4: Comparison Of SNOMEDTM Soft-Prompting On N2c2 2018 Adverse Drug Events Extraction Dataset For Concept And End-To-End Relation Extraction With A Clinical Transformer-Based Model, And TAC 2017 Results.

System	**Dataset**	**Soft Prompt Unfrozen Model**			**Soft Prompt Frozen Model**		
		P	**R**	**F1**	**P**	**R**	**F1**
Concept							
GatorTron base	n2c2 2018	-	-	91.12	-	-	86.59
SNOMEDTM	n2c2 2018	93.78	90.62	**92.05**	96.58	40.76	57.89
	TAC 2017	89.01	84.93	86.92	79.40	65.07	71.53
Relation							
GatorTron base	n2c2 2018	-	-	83.33	-	-	79.21
SNOMEDTM	n2c2 2018	88.40	81.10	84.59	88.44	19.59	32.08
	TAC 2017	52.73	52.90	52.81	47.61	12.89	20.30

5. DISCUSSION

This paper reports on the proposed new transformer-based model SNOMEDTM, which is pre-trained on standard medical terminologies from SNOMED-CT and MedDRA thesauri. The work fine-tuned the model on multi-task ADE extraction tasks on two publicly available datasets and compared the proposed system with state-of-the-art systems. This section starts by comparing the performance of the proposed model with that of other systems, then analyses the effect of utilising medical

concepts to provide prior medical knowledge to the model tailored for a downstream medical task like ADE extraction on the model's overall performance.

5.1 Performance Benchmarking

This paper compared SNOMEDTM's performance with state-of-the-art systems for ADR extraction on the TAC 2017 challenge (best run only) using the official evaluation script provided by the organisers, as shown in Table 5. From the table, the BUPT-PRIS fifth-ranked system was developed based on a Bi-LSTM-CRF with character and word embedding for entity annotation and an adversarial-based training method for CNN model relation annotation. The system achieved an F1-score of 18.29% for concept extraction and 0.55% for end-to-end relation identification. The fourth-ranked is the MC-UC3M system; the authors utilised some medical knowledge base as a look-up dictionary to extract mentions with SVM for relation classification. The system achieved an F1-score of 60.01% for concept extraction and 10.67% for end-to-end relation identification. The system, PRNA-SUNY, ranked third, was developed based on a conditional random field (CRF) for concept extraction and a rule-based approach based on MetaMap for the end-to-end tasks. The IBM-Research system's second-ranked system was on Bi-LSTM and Attention-Bi-LSTM for extracting concepts and relations while handling disjoint mentions. The UTH-CCB is the top-ranked system at TAC 2017 ADR extraction challenges for concept and end-to-end relation identification. The system is based on combined rules-based techniques to extract mentions, and Bi-LSTM-CRF was used as two cascade sequence labellers for end-to-end ADR relation extraction tasks. The system achieved an F1-score of 82.41% for the concept extraction task, which is about 64.12% different from the fifth-ranked system. Similarly, the system obtained 49.00% for the end-to-end relation extraction task, a difference of 48.45% compared to the fifth-ranked system.

This paper further compared the performance of the system with other state-of-the-art systems that have shown to outperform all the top-ranking systems on the TAC 2017 dataset. These systems are based on an improved deep learning architecture-based transformer model. The DeepCADRME systems proposed an N-level sequence modelling to handle complex ADR mentions such as discontinuous, nested, and overlapping ones. The system adopted the biomedical-based transformer model BioBERT to generate the contextual representation used at various system levels. The system achieved 85.35% for concept extraction, 2.94% higher than the challenge's top-ranking system. El-allay et al. [10] proposed the MTTLADE system, a multi-task transfer learning-based dual sequence modelling method based on large language models. The system fine-tuned five models: SCIBERT, BERT, BioBERT, BluBERT, and RoBERTa for ADR concept and end-to-end relation extraction tasks. The system achieved 85.59% for concept extraction, 0.24% higher than the DeepCADRME system. Similarly, the system achieved 53.15% for end-to-end relation extraction, which is 4.15% higher than the top-ranking system in the challenge. The study further compares the performance of SNOMEDTM with the NeuroADR method proposed by [46], the result is shown in Table 5.

The proposed system in this study, which utilises a pre-trained model based on globally accepted medical terminologies to generate a shared representation of the input sequence, demonstrated its superiority with an F1 score of 86.83%, 1.24% higher than the MTTLADE system for concept extraction. It also achieves an F1 score of 52.36%, 3.36% higher than the top-performing systems for the TAC 2017 challenge and 0.79% less than the MTTLADE for the relation extraction task. The consistent improvement shown by the model across the TAC 2017 dataset indicated the effectiveness of the medical knowledge to the model.

Table 5: Comparison Of SNOMEDTM Fine-Tuning On TAC 2017 With Other Systems For ADR Concept And End-To-End Relation Extraction Tasks.

System	Concept			Relation		
	P	R	F1	P	R	F1
SNOMEDTM (ours)	88.91	84.85	**86.83**	52.36	52.36	**52.36**
NeuroADR [46]	82.45	80.63	81.53	42.05	35.64	38.58
DeepCADRME [44]	85.45	85.24	85.35	-	-	-

UTH-CCB [47]	82.54	82.42	82.48	50.24	47.82	49.00
IBM-Research [48]	80.90	75.30	78.00	48.13	32.54	38.83
PRNA-SUNY [49]	77.71	63.90	70.13	50.48	22.36	30.99
MC-UC3M [50]	54.79	66.33	60.01	10.41	10.95	10.67
BUPT-PRIS [51]	40.47	11.81	18.29	0.97	0.38	0.55

This paper further experimented with SNOMEDTM on the ADE-corpus dataset [39], and benchmarked with two state-of-the-art methods, the result is presented in Table 6. For SMAN [52], this span-based method built a multi-model attention network to capture the interactions between spans and to model information like tokens and labels. It simultaneously extracted context and span position information. TpT-ADE [53] is a two-phase approach that fine-tunes the BERT model for ADE extraction. Firstly, the system identified and normalized the concepts to a standard UML knowledge base then the second phase utilizes BERT to process the text, and extract mentions then classify relations between them.

Table 6: Comparison of SNOMEDTM against SMAN and TpT-ADE on ADE-corpus dataset.

Method	NER			RE		
	P	R	F1	P	R	F1
SMAN	-	-	90.95	-	-	82.25
TpT-ADE	89.24	93.2	91.17	81.91	85.83	**83.82**
SNOMEDTM	93.68	93.63	**93.68**	82.37	83.32	82.84

5.2 Ablation Study

To further identify the contribution of different components of the proposed architecture, this study conducted two ablation studies. In the first instance this paper pretrained the SNOMEDTM complete architecture on only SNOMED-CT terminologies. This is because the terminology is larger (amounting to 65% of the total pretraining data) than the MedDRA and most of the MedDRA terminologies has corresponding mapping terminologies within the SNOMED-CT. SNOMED CT is the most extensive biomedical ontology, covering a diverse array of biomedical and clinical concepts, such as signs, symptoms, diseases, procedures, and social contexts [54]. In the second experiment, this paper created a base model called SNOMEDTM-base, this model consists of 12 encoder layers and 12 self-attention layers. The base model was also pretrained on the complete pretraining data. The experimental results shown in Table 7 on TAC 2017 and n2c2 2018 for concept and relation extraction tasks, shows a drastic drop in performance for the two models compared to SNOMEDTM. Nonetheless, the model on SNOMED-CT does outperform the SNOMEDTM-base model revealing the impact of the medical terminology to the overall performance of the model for ADE extraction tasks.

Table 7: Ablation experiment for different components of the SNOMEDTM architecture.

Model	**Dataset**	**Concept**			**Relation**		
		P	**R**	**F1**	**P**	**R**	**F1**
SNOMEDTM-SNOMED-CT-only	n2c2 2018	87.84	82.18	84.92	76.20	69.21	72.54
	TAC 2017	80.57	75.14	77.76	43.85	31.78	36.85
SNOMEDTM-base	n2c2 2018	89.36	69.85	78.41	74.84	58.04	65.38
	TAC 2017	67.26	63.82	65.49	34.20	29.85	31.88
SNOMEDTM	n2c2 2018	93.11	84.81	88.77	88.29	80.29	84.11
	TAC 2017	88.91	84.85	86.83	52.36	52.92	52.63

5.2 Effect of Prior Medical Knowledge on the Model's Performance

The SNOMEDTM model is comprised of 16 self-attention heads and 16 encoder layers. Contextual representation generated by the model makes the proposed system more powerful in identifying complex situations within the dataset despite the limited pre-training data compared to large general models with trillions of training samples. The proposed model competed with large models and produced comparable results for the two tasks. This

capability shown by the model could be attributed to the influence of medical concepts utilised to pre-train the model despite the smaller data size. Compared to other clinical-based large language models, most of the models are trained from locally generated domain data, but for SNOMEDTM, it uses globally standard medical terminology and specific terminology related to ADEs, which gave the model the ability to recognise tokens related to ADEs.

It is acknowledged that clinical LLMs are limited; however, this research direction is recently gaining momentum. Researchers are developing more domain-specific models, for instance, BioGPT [25] from PubMed abstracts, [28], and task-specific models like PharmBERT [17] from drug labels and GatorTron [31] from University of Florida health data, an LLM for clinical NLP.

Nonetheless, SNOMEDTM pre-training data is from globally accepted medical terminology related to drug safety and adverse drug event cases. This innovative use of ADE-related terminologies to develop the SNOMEDTM model tailored for ADE case extraction demonstrated the potential for mitigating the sparsity of clinical LLMs. The improved performance shown by the model on diverse and multiple clinical datasets of TAC 2017 (drug labels) and n2c2 2018 (discharge summaries) and on various tasks of named entity recognition and relation extraction over the baselines and benchmark models indicated its robustness and applicability to various clinical tasks. This paper plans to incorporate more clinical data sources to pre-train the model in future work. Additionally, this paper plans to transform the SNOMEDTM model into a multi-lingual and multi-modal model to address a broader range of clinical NLP tasks from multiple languages.

This study focuses on improving ADE extraction from clinical textual documentation using LLM-based techniques. The research holds promise for supporting natural language processing and practical healthcare applications. The proposed task-adaptive SNOMEDTM model was pre-trained based on globally standardised ADE-related thesauri. Utilising SNOMED-CT and MedDRA knowledge bases to develop clinical-based LLM language exemplified an innovative strategy for fully utilising this knowledge to develop a state-of-the-art foundational model for improving clinical outcomes. SNOMEDTM has shown its potential to advance clinical NLP significantly. Accurate extraction of ADE cases from unstructured clinical text is critical for improving patient care, enhancing clinical documentation and supporting drug safety surveillance. As a trained transformer-based model, SNOMEDTM can be fine-tuned on various clinical NLP tasks and different tuning strategies in transfer learning settings, as demonstrated in this research. Thus, it can contribute to overall healthcare benefiting both patients and practitioners.

The proposed model can serve as a foundation in addition to its immediate practical applications. It can be further scaled up using additional data sources or transformed into a multilingual or multimodal model for handling various clinical applications, thereby improving its generalizability in clinical domain-specific tasks.

6. CONCLUSION AND FUTURE WORK

This paper introduced a novel transformer architecture tailored for biomedical information extraction. The model comprises 138 million parameters pre-trained with around 15 million tokens. The pre-training was on standard global medical terminologies, the SNOMED-CT and MedDRA, to incorporate prior domain-related medical knowledge into the architecture. The pre-trained model is then fine-tuned to the multi-task adverse drug event extraction of mentions and relation extraction problems on two publicly available datasets provided by the TAC 2017 and n2c2 2018 challenges. The experimental results show that the proposed model showed promising results; despite the small number of pre-training datasets and model parameters compared to larger models like GatorTron, the proposed model outperformed many state-of-the-art models like BERT and GatorTron-base on ADE extraction. In future research, this paper intends to extend model capability by utilising more medical-related data from electronic health record systems, synthesising data, and exploring other pre-training objectives. SNOMED-CT and MedDRA terminologies are available in various languages and international standards. Pre-training the model on multilingual or multimodal objectives and on multiple data formats will improve its generalizability and applicability for various ADE extraction tasks in multiple languages. Additionally, this paper plans to employ a human-in-the-loop learning approach with domain experts to fine-tune the model.

It is important to acknowledge the potential limitations of the proposed techniques and areas where this research can be improved. The SNOMEDTM model was pre-trained on less data than other clinical LLMS like GatorTron. This limited data can affect the model's generalisation and acceptance for clinical NLP tasks. Moreover, LSTM-based model biases in data synthesis may influence the synthesised dataset and lack clinical representativeness. Additionally, the ethical implications of using synthesised data have not been adequately addressed. However, the research aimed to investigate the impact of two globally standard ADE-related terminologies in improving ADE extraction. The use of other data sources for data augmentation will be considered in future research.

Knowledge bases such as Side Effect Resource (SIDER) exist that collect, normalize, and encode ADE-related terminologies [33]. Other agencies, such as DrugMAP [55], DrugBank, and VARIDT (Variability of Drug Transporter Database) [56], provide valuable information about drugs, drug research, and safety monitoring. In future research this paper plans to obtain more terminologies from these knowledge bases to expand the pretraining data to integrate more knowledge to SNOMEDTM.

ACKNOWLEGEMENTS

This research was funded by the Ministry of Higher Education Malaysia, under the Fundamental Research Grant Scheme (FRGS/1/2023/ICT02/UPM/02/3).

REFERENCES

[1]. Henriksson A, Kvist M, Dalianis H, Duneld M. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. J Biomed Inform [Internet]. 2015;57:333–49. Available from: http://dx.doi.org/10.1016/j.jbi.2015.08.013

[2]. Gonzalez-Hernandez G, Krallinger M, Muñoz M, Rodriguez-Esteban R, Uzuner Ö, Hirschman L. Challenges and opportunities for mining adverse drug reactions: perspectives from pharma, regulatory agencies, healthcare providers and consumers. Database (Oxford) [Internet]. 2022;2022(00):1–10. Available from: https://doi.org/https://doi.org/10.1093/database/baac071

[3]. Fan B, Fan W, Smith C, Garner H "Skip." Adverse drug event detection and extraction from open data: A deep learning approach. Inf Process Manag [Internet]. 2020;57(1):102131. Available from: https://doi.org/10.1016/j.ipm.2019.102131

[4]. Jagannatha A, Liu F, Liu W, Yu H. Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0). Drug Saf [Internet]. 2019;42(1):99–111. Available from: https://doi.org/10.1007/s40264-018-0762-z

[5]. Roberts K, Demner-Fushman D, Tonning JM. Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track. Text Anal Conf Proc [Internet]. 2017;1–13. Available from: https://dailymed.nlm.nih.gov/

[6]. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 N2C2 Shared Task on Adverse Drug Events and Medication Extraction in Electronic Health Records. J Am Med Informatics Assoc. 2020;27(1):3–12.

[7]. Modi S, Kasmiran KA, Mohd Sharef N, Sharum MY. Extracting adverse drug events from clinical Notes: A systematic review of approaches used. J Biomed Inform [Internet]. 2024;151(1):104603. Available from: https://doi.org/10.1016/j.jbi.2024.104603

[8]. Yang X, Bian J, Fang R, Bjarnadottir RI, Hogan WR, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. J Am Med Informatics Assoc. 2020;27(1):65–72.

[9]. Florez E, Precioso F, Pighetti R, Riveill M. Deep learning for identification of adverse drug reaction relations. ACM Int Conf Proceeding Ser. 2019;149–53.

[10]. El-Allaly ED, Sarrouti M, En-Nahnahi N, Ouatik El Alaoui S. MTTLADE: A multi-task transfer learning-based method for adverse drug events extraction. Inf Process Manag [Internet]. 2021;58(3):102473. Available from: https://doi.org/10.1016/j.ipm.2020.102473

[11]. Gu Y, Zhang S, Usuyama N, Woldesenbet Y, Wong C, Sanapathi P, et al. Distilling Large Language Models for Biomedical Knowledge Extraction: A Case Study on Adverse Drug Events. ArXiv [Internet]. 2023;1–15. Available from: http://arxiv.org/abs/2307.06439

[12]. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural

Language Processing. ACM Comput Surv. 2023;55(9):195.
[13]. Gupta K, Th B, Mats I, Quentin LR. Continual Pre-Training of Large Language Models : How to (re) warm your. In: 40th International Conference on Machine Learning. 2023.
[14]. Peng C, Yang X, Smith KE, Yu Z, Chen A, Bian J, et al. Model tuning or prompt Tuning? a study of large language models for clinical concept and relation extraction. J Biomed Inform [Internet]. 2024;153:104630. Available from: https://doi.org/10.1016/j.jbi.2024.104630
[15]. Modi S, Kasmiran KA, Mohd Sharef N, Sharum MY. Enhanced Adverse Drug Event Extraction Using Prefix-Based Multi-Prompt Tuning in Transformer Models. IINTERNATIONAL J INFORMATICS Vis. 2024;8(3–2):1713–9.
[16]. Lamproudis A, Henriksson A, Dalianis H. Developing a Clinical Language Model for Swedish: Continued Pretraining of Generic BERT with In-Domain Data. Int Conf Recent Adv Nat Lang Process RANLP [Internet]. 2021;790–7. Available from: https://aclanthology.org/2021.ranlp-1.90.pdf
[17]. Valizadeh Aslani T, Shi Y, Ren P, Wang J, Zhang Y, Hu M, et al. PharmBERT: a domain-specific BERT model for drug labels. Brief Bioinform. 2023;24(4):bbad226.
[18]. Dai X, Karimi S, Sarker A, Hachey B, Paris C. MultiADE: A Multi-domain Benchmark for Adverse Drug Event Extraction. J Biomed Inform [Internet]. 2024;160:104744. Available from: https://data.csiro.au/collection/csiro:62387
[19]. Lee J, Kim S, Yoon W, Kim S, So CH, Kang J, et al. BioBERT : a pre-trained biomedical language representation model for biomedical text mining. 2020;36(September 2019):1234–40.
[20]. Beltagy I, Lo K, Arman C. SciBERT: A Pretrained Language Model for Scientific Text. Proc 2019 Conf Empir Methods Nat Lang Process 9th Int Jt Conf Nat Lang Process [Internet]. 2019;3615–3620. Available from: https://aclanthology.org/D19-1371
[21]. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. Proc 2nd Clin Nat Lang Process Work [Internet]. 2019;72–8. Available from: http://arxiv.org/abs/1904.03323
[22]. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv [Internet]. 2019;(1). Available from: http://arxiv.org/abs/1907.11692
[23]. Alrowili S, Vijay-Shanker K. BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA. Proc 20th Work Biomed Lang Process BioNLP 2021. 2021;221–7.
[24]. Mcmaster C, Chan J, Liew DFL, Su E, Frauman AG, Chapman WW, et al. Developing a deep learning natural language processing algorithm for automated reporting of adverse drug reactions. J Biomed Inform [Internet]. 2023;137:104265. Available from: https://doi.org/10.1016/j.jbi.2022.104265
[25]. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022;23(6):1–12.
[26]. Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation. ArXiv [Internet]. 2023; Available from: http://arxiv.org/abs/2306.09968
[27]. Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J. BiomedGPT : A generalist vision – language foundation model for diverse biomedical tasks Abstract Introduction. ArXiv. 2024;1–57.
[28]. Yuan H, Yuan Z, Gan R, Zhang J, Xie Y, Yu S. BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model. Proc Annu Meet Assoc Comput Linguist. 2022;97–109.
[29]. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: Toward building open-source language models for medicine. J Am Med Informatics Assoc. 2024;31(9):1833–43.
[30]. Lee S, Kang M, Lee J, Hwang SJ, Kawaguchi K. Self-Distillation for Further Pre-training of Transformers. Int Conf Learn Represent 2023 [Internet]. 2022;1–22. Available from: http://arxiv.org/abs/2210.02871
[31]. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. npj Digit Med. 2022;5(1):1–9.
[32]. Gaudet-Blavignac C, Foufi V, Bjelogrlic M, Lovis C. Use of the systematized nomenclature of medicine clinical terms (snomed ct) for processing free text in health care: Systematic

scoping review. J Med Internet Res. 2021;23(1):1–18.
[33]. Große-Michaelis I, Proestel S, Rao RM, Dillman BS, Bader-Weder S, Macdonald L, et al. MedDRA Labeling Groupings to Improve Safety Communication in Product Labels. Ther Innov Regul Sci [Internet]. 2023;57(1):1–6. Available from: https://doi.org/10.1007/s43441-022-00393-1
[34]. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst [Internet]. 2017;6000–10. Available from: https://dl.acm.org/doi/10.5555/3295222.3295349
[35]. Devlin J, Chang MW, Lee K, Toutanova K. 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. 2019;1:4171–86.
[36]. Zhang NL, Xie W, Lin Z, Dong G, Li XH, Cao CC, et al. Example Perplexity. ArXiv [Internet]. 2022;1–16. Available from: http://arxiv.org/abs/2203.08813
[37]. Liu X, He P, Chen W, Gao J. Multi-Task Deep Neural Networks for Natural Language Understanding. Proc 57th Annu Meet Assoc Comput Linguist [Internet]. 2019;4487–4496. Available from: https://aclanthology.org/P19-1441
[38]. Ma F, Zhang C, Ren L, Wang J, Wang Q, Wu W, et al. XPROMPT: Exploring the Extreme of Prompt Tuning. Proc 2022 Conf Empir Methods Nat Lang Process EMNLP 2022. 2022;11033–47.
[39]. Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J Biomed Inform. 2012;45(5):885–92.
[40]. Liu X, Ji K, Fu Y, Tam WL, Du Z, Yang Z, et al. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. Proc Annu Meet Assoc Comput Linguist. 2022;2:61–8.
[41]. Liu X, Zheng Y, Du Z, Ding M, Qian Y, Yang Z, et al. GPT understands, too. AI Open. 2023;
[42]. Quentin M. Parameter-Efficient Transfer Learning for NLP. ArXiv. 2019;
[43]. Wallis P. LORA: LOW-RANK ADAPTATION OF LARGE LAN- GUAGE MODELS. ArXiv. 2021;1–26.
[44]. El-allaly E drissiya, Sarrouti M, En-nahnahi N. DeepCADRME : A deep neural model for complex adverse drug reaction mentions extraction. Pattern Recognit Lett J [Internet]. 2021;143:27–35. Available from: https://pdf.sciencedirectassets.com/271524/1-s2.0-S0167865521X00025/1-s2.0-S016786552030444X/main.pdf?
[45]. Yazdani A, Proios D, Rouhizadeh H, Teodoro D. Efficient Joint Learning for Clinical Named Entity Recognition and Relation Extraction Using Fourier Networks: A Use Case in Adverse Drug Events. Proc 19th Int Conf Nat Lang Process [Internet]. 2023;212–223. Available from: http://arxiv.org/abs/2302.04185
[46]. Liu F, Zheng X, Yu H, Tjia J. Neural Multi-Task Learning for Adverse Drug Reaction Extraction. In: AMIA . Annual Symposium proceedings AMIA Symposium [Internet]. AMIA Annu Symp Proc.; 2020. p. 756–62. Available from: https://pubmed.ncbi.nlm.nih.gov/33936450/
[47]. Xu J, Lee HJ, Ji Z, Wang J, Wei Q, Xu H. UTH_CCB System for Adverse Drug Reaction Extraction from Drug Labels at TAC-ADR 2017. In: Proceedings of Text Analysis Conference [Internet]. 2017. Available from: https://tac.nist.gov/publications/2017/participant.papers/TAC2017.UTH_CCB.proceedings.pdf
[48]. Dandala B, Diwakar M, Murthy D. IBM Research System at TAC 2017: Adverse Drug Reactions Extraction from Drug Labels. In: Text Analysis Conference (TAC2017) [Internet]. 2017. p. 2–9. Available from: https://tac.nist.gov/publications/2017/participant.papers/TAC2017.IBM_Research.proceedings.pdf
[49]. Tao C. Extracting and Normalizing Adverse Drug Reactions from Drug Labels. In: A Text Analysis Conference (TAC) 2017 [Internet]. 2017. p. 1–9. Available from: https://www.semanticscholar.org/paper/Extracting-and-Normalizing-Adverse-Drug-Reactions-Tao-Lee/4fe1095a50731d74dee6ef93c699c81de744496f
[50]. Martinez JL, Martinez P, Segura-Bedmar I, Carruana A, Naderi A, Polo C. MC-UC3M Participation at TAC 2017 Adverse Drug Reaction Extraction from Drug Labels. Text Anal Conf 2017. 2017;1–13.
[51]. Gu X, Ding C, Li S, Xu W. BUPT-PRIS System for TAC 2017 Event Nugget Detection, Event Argument Linking and ADR Tracks. In: TAC2017 Conference [Internet]. 2017. p. 1–9.

Available from: https://tac.nist.gov/publications/2017/papers.html

[52]. Wan Q, Wei L, Zhao S, Liu J. A Span-based Multi-Modal Attention Network for joint entity-relation extraction. Knowledge-Based Syst [Internet]. 2023;262:110228. Available from: https://doi.org/10.1016/j.knosys.2022.110228

[53]. Kuchibhotla S, Singh M. TpT-ADE: Transformer Based Two-Phase ADE Extraction. In: CoNLL 2024 - 28th Conference on Computational Natural Language Learning, Proceedings of the Conference. 2024. p. 209–18.

[54]. Chang E, Mostafa J. The use of SNOMED CT, 2013-2020: A literature review. J Am Med Informatics Assoc. 2021;28(9):2017–26.

[55]. Li F, Yin J, Lu M, Mou M, Li Z, Zeng Z, et al. DrugMAP: molecular atlas and pharma-information of all drugs. Nucleic Acids Res. 2023;51(D1):D1288–99.

[56]. Yin J, Chen Z, You N, Li F, Zhang H, Zhao Q, et al. VARIDT 3 . 0 : the phenotypic and regulatory v ar iability of drug tr anspor t er. Nucleic Acids Res [Internet]. 2024;52(October 2023):1490–502. Available from: https://watermark.silverchair.com/gkad818.