A Soft and Hard Mixture-of-Experts Approach for Improved ADR Extraction from Patient-Generated Narratives

Oumayma Elbiach[1], Hanane Grissette[2], El Habib Nfaoui[3]
Faculty of Sciences Dhar El Mahraz-L3IA Laboratory, Sidi Mohamed Ben Abdellah University, Fez, Morocco[1,3]
Faculty of Sciences and Technology-Department of Computer Science, Sidi Mohamed Ben Abdellah University, Fez, Morocco[2]

***Abstract*—Traditional single-architecture neural models, including monolithic transformer-based and sequence-to-sequence architectures, often struggle to extract Adverse Drug Reactions (ADRs) from patient-generated health narratives due to informal language, high linguistic variability, and complex relationships among drugs, diseases, and adverse events. Although Mixture-of-Experts (MoE) architectures have demonstrated strong performance across various Natural Language Processing (NLP) tasks, their effectiveness for ADR extraction from unstructured patient narratives remains largely unexplored. This study investigates the application of MoE architectures, specifically Soft MoE and Hard MoE, for ADR extraction from patient-generated content. The task is formulated as a sequence-to-sequence generation problem and evaluated on the PsyTAR dataset using both strict and relaxed evaluation metrics. Experimental results demonstrate that Soft MoE consistently outperforms Hard MoE, achieving a relaxed F1-score of 80.40% compared to 79.40%. These findings highlight the critical role of expert-routing strategies in capturing linguistic variability in patient narratives and establish MoE architectures as a competitive and reliable approach for automated ADR extraction in biomedical text mining and pharmacovigilance applications.**

***Keywords*—Adverse Drug Reaction; Mixture-of-Experts; Soft and Hard MoE; sequence-to-sequence; patient narratives; biomedical text mining**

I. Introduction

An Adverse Drug Reaction is an unintended and harmful effect that occurs after the use of a drug at its normal therapeutic dose [1]. Adverse Drug Reactions are a significant contributor to morbidity and mortality rates worldwide [2]. Their timely detection and monitoring have immense importance in order to improve the safety of drugs and to support various pharmacovigilance activities. Typically, ADR data is collected through passive spontaneous reporting systems, which usually suffer from under-reporting and delayed reporting [3]. Nevertheless, with the rapid growth in user-generated content on social media and health-related forums, valuable information regarding patient experiences has become increasingly available in real time [4]. Millions of users actively share their medication experiences within online communities, frequently reporting potential side effects almost in real-time [5]. Consequently, it opened a promising avenue for the early ADR detection, potentially overcoming the drawbacks of traditional systems. However, leveraging social media data for ADR detection poses significant challenges because of the informal, unstructured, and noisy nature of text generated by users [6]. Patients often describe their experiences in various ways, using colloquial expressions, abbreviations, or implicit mentions of symptoms, making automatic extraction of ADRs a challenging task [7].

In light of these challenges, more and more researchers have become interested in using Natural Language Processing methods to automatically identify ADRs not explicitly written in structured formats, such as medical forums and social media [8]. Early studies have mainly focused on rule-based systems, machine learning, and deep learning methods. The early rule-based systems were interpretable yet often could not generalize to linguistic variability or unseen expressions [9]. While machine learning methods introduced data-driven learning capabilities, they still depended on extensive manual feature engineering [10]. The emergence of deep learning, especially the transformer-based models [11], has marked a major advancement by enabling models to automatically capture semantic and contextual relationships in text. Despite their substantial success, most of these models generally require a vast amount of annotated biomedical corpora, which limits their applicability in real-world pharmacovigilance scenarios where such data are often limited.

More recently, Large Language Models have shown remarkable potential for biomedical information extraction, due to the rich contextual representation they can learn from massive amounts of text data [12]. LLMs can generalize across a wide range of linguistic patterns, making them particularly apt at dealing with the noisy and unstructured language used on social media and patient narratives. However, monolithic architectures, whether standard transformers or LLMs, rely on a single generalized network to handle all inputs, which can be suboptimal for capturing the highly specific and varied expressions of Adverse Drug Reactions. This limitation motivates the exploration of the Mixture-of-Experts paradigm, which leverages multiple specialized sub-networks to process different types of input patterns dynamically. While MoE models achieved state-of-the-art results in general language understanding and other specialized domains [13], their effectiveness and optimal configuration for the specific domain of ADR extraction have not been largely explored. Thus, the present work systematically investigates the performance of the MoE architectures for ADR extraction from patient narratives. In particular, a comparison is conducted between Soft MoE, where all experts contribute to each input, and Hard MoE, where only a subset of experts is activated, in an

attempt to establish which of the approaches better captures the semantic variability of ADRs. Through this comparison analysis, the present study investigates whether the specialization capabilities of MoE can lead to improvement in extraction performance on this challenging domain, including insights into their applicability in biomedical text mining.

The main contributions of this work are as follows:

- This study investigates the MOE approach for the specialized task of ADR extraction from patient-generated narratives.
- Comparative analysis of the two primary MoE variants (Soft MoE and Hard MoE) to determine their efficacy in capturing the high semantic variability present in user-generated medical text.
- An experimental evaluation is conducted on the PsyTAR dataset using strict and relaxed metrics to assess the effectiveness of MoE in this task.

The rest of the study is organized as follows: Section II reviews related work. Then, Section III describes the materials and methods, covering the dataset, model architecture, and evaluation metrics. The results will be reported and discussed in Section IV. Finally, Section V concludes the study and outlines future work directions.

II. Background and Related Work

Research on Adverse Drug Reaction extraction has evolved rapidly with the advancement of NLP techniques. Early studies focused on applying rule-based and lexicon-driven systems, which relied on handcrafted patterns and domain-specific vocabularies such as MedDRA to identify ADR mentions [14],[15]. Although these approaches offered transparency, they struggled with the linguistic variability and ambiguity present in real-world patient narratives. Subsequent work introduced machine learning-based models to learn statistical patterns from annotated corpora [16]. These methods improved generalization but still depended heavily on manual feature design and lacked the ability to capture deep contextual semantics.

The advent of deep learning marked an important turning point in ADR extraction, with architectures such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) enabling automatic feature learning from raw text [17], a significant improvement over manual feature engineering [18]. Many of these models often struggled with long-range dependencies and contextual subtleties inherent in biomedical language. The breakthrough came with transformer-based architectures, which exploit a self-attention mechanism to capture complex contextual relationships across entire sequences. Models like BERT and its domain-specific variants, such as BioBERT and PubMedBERT, demonstrated unprecedented performance on automatic learning of rich contextual and semantic representations from vast amounts of textual data [11]. This reduced the need to manually perform feature engineering and allowed a more effective handling of the linguistic variations observed in both the formal literature and the informal narratives of patients [19].

Building upon this foundation of specialized transformers, the field has progressed toward Large Language Models, which extend this architecture to much larger and more powerful systems through pre-training on huge and diverse text corpora. This extreme scaling allows LLMs to have superior generalization and powerful few-shot capabilities, which make them very promising in handling the complexity of information extraction tasks even when the use of annotated data is limited [20]. A major innovation within these recent LLMs that has been aimed at handling such a large scale efficiently is the Mixture-of-Experts architecture. In contrast to dense models that activate the full network for every input, MoE relies on a sparse structure where the routing network dynamically routes the tokens to the appropriate specialized experts. The potential of this paradigm for complex biomedical tasks was demonstrated by [21], who developed MOEBQA, an MoE-adapted model for Biomedical Question Answering. Their model used a router to assign different types of professional questions to specialized experts, which significantly improved performance by alleviating parameter competition across diverse question types, achieving state-of-the-art results. In [22], the authors proposed MOELoRA, which combines MoE with parameter-efficient fine-tuning to efficiently learn separate parameters for various tasks while maintaining a small number of trainable parameters, demonstrating superior performance on a multi-task Chinese medical dataset. This approach was later extended to Named Entity Recognition (NER) in work by [23], who proposed BOND-MoE, a model that integrates a MoE layer into a BERT-based framework. In their contribution, experiments verify that assignment of predictions to multiple experts is effective in collectively improving performance as well as in effectively overcoming noisy, distantly supervised labels in various instances of NER tasks across multiple datasets, achieving performance superior to BERT

Though Mixture-of-Experts architecture have demonstrated success in different NLP and biomedical text mining tasks, their application to the Adverse Drug Reaction extraction task remains largely unexplored. Despite substantial progress achieved by deep learning, transformer-based, and large language model approaches, existing ADR extraction methods still largely depend on monolithic architectures that process all inputs using a single shared representation space. This design may impact their ability to effectively capture the high linguistic variability and heterogeneous expression patterns characteristic of patient-generated narratives. Moreover, although MoE-based models have demonstrated effectiveness in related biomedical NLP tasks such as question answering and named entity recognition, prior work has not systematically examined their use for ADR extraction, nor has it investigated how different expert-routing strategies influence performance in this domain. This gap motivates the present study, which evaluates both Soft and Hard MoE variants, assessing their respective strengths and limitations for ADR extraction from patient narratives.

III. Methodology

This study aims to assess the performance of the Mixture-of-Experts mechanism on the task of extracting Adverse Drug Reactions from patient narratives. The primary goal is to test the efficiency of the architecture by performing a comparative performance analysis between the two main variants of this architecture, Soft MoE and Hard MoE, within the same domain. The main hypothesis is that the MoE paradigm, due to

its specialized sub-networks, can more effectively capture the high semantic variability and diverse ADR categories found in the text compared to a monolithic model.

A. Problem Formulation

In this work, the task of Adverse Drug Reaction extraction from a patient's narrative is formulated as a sequence-to-sequence (Seq2Seq) problem. Given an input text $X = (x_1, x_2, \ldots, x_T)$ that describes a patient's experience with a medication in their own words, the task is to generate an output sequence $Y = (y_1, y_2, \ldots, y_N)$ listing all mentioned adverse reactions separated by commas. Each output token y_i represents a word or a phrase describing an ADR that is mentioned in the input text. This formulation is effective for patient-generated content, which tends to be informal, unstructured, and may contain multiple ADRs using varied expressions. ADR generation as a textual list, rather than performing token-level classification, allows the model to better capture contextual and semantic relationships and to produce concise and accurate ADR outputs directly from the narrative text.

B. Dataset

The PsyTAR (Psychiatric Treatment Adverse Reactions) dataset [24] is a psychiatric adverse reaction dataset that captures how patients describe the perceived effects and adverse reactions associated with psychiatric medications in real-world settings. The dataset consisted of patient-generated reviews, collected from the online health platform Ask a Patient, on four commonly prescribed antidepressants: Zoloft, Lexapro, Cymbalta, and Effexor XR, which belong to two classes: Selective Serotonin Reuptake Inhibitors (SSRIs) and Serotonin–Norepinephrine Reuptake Inhibitors (SNRIs). The dataset was developed in a multi-step process through sentence segmentation and comprehensive manual annotation. 6,009 sentences were extracted from 891 drug reviews and labeled regarding semantic categories such as Adverse Drug Reactions (ADRs) (2,168 sentences), Withdrawal Symptoms (WDs) (438 sentences), Signs and Symptoms of Illness (SSIs) (789 sentences), Drug Indications (DIs) (517 sentences), Drug Effectiveness (EF) (1,087 sentences), and Drug Ineffectiveness (INF) (337 sentences). Within the ADR-labeled subset, 4,813 distinct ADR mentions were identified [28]. For this study, only sentences annotated with ADRs were retained, while non-ADR sentences were excluded. After filtering, the final corpus used for ADR extraction comprised 1,981 sentences, providing a focused and high-quality subset for investigating patient-reported Adverse Drug Reactions in psychiatric treatments.

C. Model Architecture

In this work, the Mixture-of-Experts mechanism is investigated and applied within an encoder-decoder framework for a challenging application: Adverse Drug Reaction extraction from patient narratives. The problem is formulated as a Seq2Seq generation task, and therefore, the model maps a textual patient description to a list of ADR mentions. While the standard Seq2Seq architectures can capture complex linguistic dependencies, they often fail under heterogeneous biomedical language, which may contain ADR expressions of different styles, domains, and granularities. To overcome this limitation, a Mixture-of-Experts mechanism is introduced to enable the model to distribute the learning process across multiple expert networks, each one potentially specializing in different linguistic or semantic aspects of ADR expressions. Fig. 1 illustrates the integration of the Mixture-of-Experts mechanism inside the Transformer encoder-decoder architecture used for ADR extraction. In the standard Transformer block, the Feed-Forward Network (FFN), a two-layer position-wise neural network applied independently to each token, is responsible for transforming intermediate hidden representations after the self-attention operation. In the proposed architecture, this FFN component is replaced by an MoE layer, which routes each token representation to one or more expert networks. Two routing strategies are explored: Soft MoE, where all experts contribute proportionally to their gating probabilities, and Hard MoE, where only the Top-k experts (k=2) are activated for each token. This modification preserves the overall encoder–decoder structure of the model while introducing specialization among experts, enabling more flexible and adaptive representation learning for ADR extraction.

1) Mixture-of-Experts: The Mixture-of-Experts [25] mechanism enhances the model's representational capacity by introducing a set of K expert networks, each designed to learn a specific subspace of the input representation. A gating network determines how much each expert contributes to processing a given input. Formally, let $H = (h_1, h_2, \ldots, h_T)$ denote the encoder's hidden representations of the input text. The gating network G produces a vector of scores $g = (g_1, g_2, \ldots, g_K)$ corresponding to the K experts:

$$g = G(H) \tag{1}$$

where, each g_k denotes the relevance of expert E_k for the current input. The output from the encoder is routed through the MoE layer to generate a refined representation H', which the decoder will take as an input:

$$H' = \sum_{k=1}^{K} \alpha_k \, E_k(H) \tag{2}$$

where, α_k are weights given by the gating mechanism, which are determined differently in the Soft and Hard variants. Then, the decoder subsequently predicts the output sequence of ADR mentions based on H'. This type of structure allows the model to learn different linguistic patterns while still sharing an encoder–decoder architecture.

a) Soft MoE: In the Soft MoE configuration, all experts contribute to the processing of every input, but to a different extent. The gating network uses a softmax function to produce a probability distribution over all experts:

$$\alpha = \text{Softmax}(G(H)) \tag{3}$$

Each expert E_k processes the entire encoder representation H, generating an output $O_k = E_k(H)$. The final representation fed into the decoder is a weighted sum of all the expert outputs:

$$H' = \sum_{k=1}^{K} \alpha_k O_k \tag{4}$$

This allows the model to make use of a mixture of several expert knowledge sources for each input, leading to smooth sharing of information between experts. However, activating

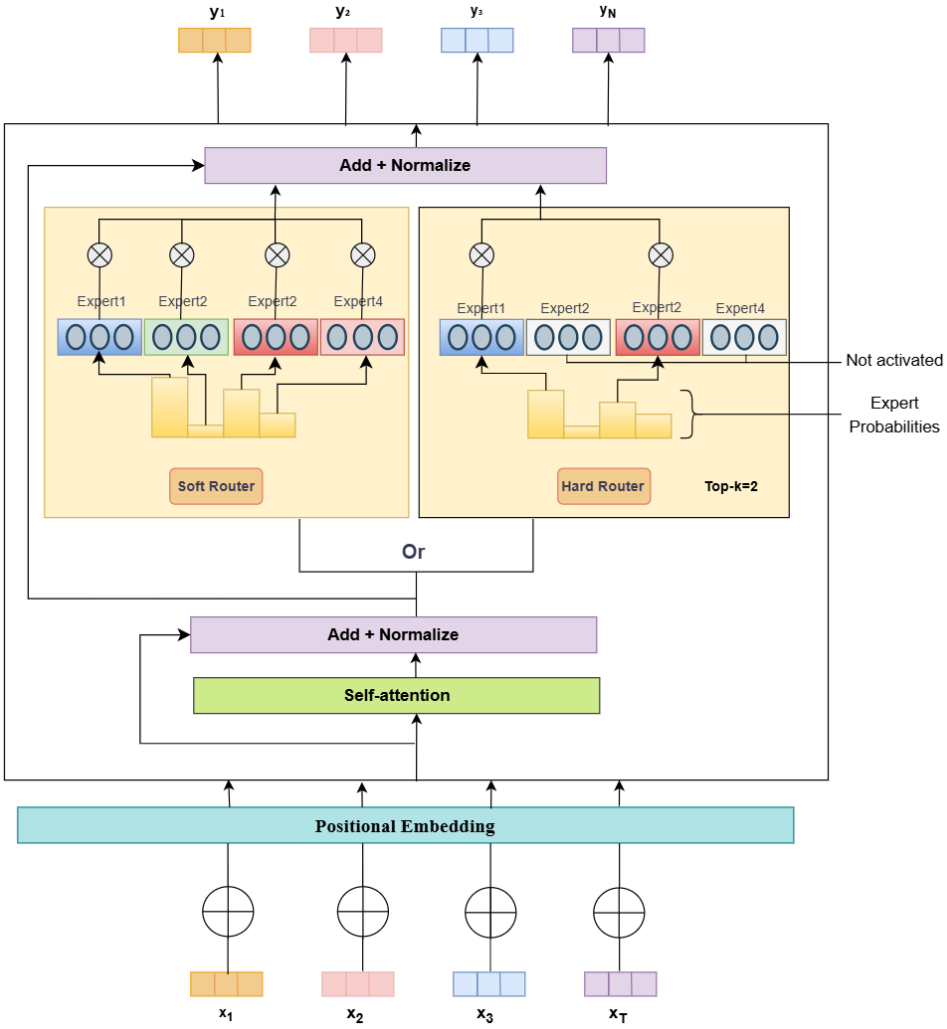

Fig. 1. Overall architecture of the proposed Mixture-of-Experts for ADR extraction.

all experts for every input increases computational cost and may result in redundancy.

b) Hard MoE: The hard version of MoE uses a sparse routing strategy, where only a subset of experts, typically the top-k according to the gating scores, are activated for any given input. The gating network computes scores $g = G(H)$, and only the experts with the highest k scores are selected. The weights α_k for the selected experts are usually set to 1 (a simple sum) or re-normalized via softmax over the top-k scores, while others are set to zero:

$$H' = \sum_{k \in \text{Top-}k(g)} \alpha_k E_k(H) \tag{5}$$

This selective routing greatly reduces computation and allows each expert to specialize in a narrower subset of the input space, thereby promoting more distinct expert behavior. However, without proper intervention, this approach introduces a significant challenge: the gating network may collapse, consistently favoring a small subset of experts while allowing others to stagnate, a phenomenon known as expert collapse. This issue is mitigated, and stable training as well as balanced expert utilization are ensured through an auxiliary load-balancing loss term [26]. This loss penalizes uneven usage of the experts and thus encourages the gating network to distribute the inputs more uniformly across the set of experts. By this regularization, all experts get enough training signals and contribute meaningfully to the overall performance of the model. This prevents a small set of experts from dominating.

Both MoE variants are integrated at the interface between the encoder and decoder, enabling the experts to refine intermediate representations before decoding. This comparative study focuses on how the Soft and Hard routing strategies influence model specialization and performance for the ADR extraction task. A concise comparison of the two variants is presented in Table I.

D. Evaluation Metrics

Since extracting ADRs inherently incorporates ambiguity in defining the entity boundaries of patient narratives, performance evaluation is essential. The standard metrics, such as precision, recall, and F1-score, allow for a basic understanding of how well models can identify correct mentions of ADRs.

TABLE I. KEY CHARACTERISTICS COMPARISON BETWEEN SOFT AND HARD MOE MECHANISMS

Features	Soft MoE	Hard MoE
Routing Mechanism	Dense / All Expert Routing	Sparse / Top-k Routing
Expert Selection	Only the top-k experts are selected and activated for a given input	All experts are activated for each input
Gating Function	Use Softmax function to generate a probability distribution for all experts	Top-k discrete routing with gradient approximation
Load Balancing	- KL divergence balance loss - Optional : Naturally balanced due to continious soft gating but can still be benificial	- Explicit balance loss (MSE) - Mandatory : To prevent expert to become overloaded
Training Stability	More stable due to soft combination	Can suffer from expert collapse
Computational Cost	Higher cost because all experts must be computed in full	Lower cost because only k experts are activated and computed
Robustness	Higher: Robust to routing variations since multiple experts contribute jointly, reducing the impact of the individual routing errors	Lower: Highly sensitive to router errors. If an incorrect expert is selected, the performance can significantly degrade

However, the strict evaluation methods, since they require an exact boundary match between the predicted and gold-standard entities, are inflexible and often fail to consider the partially correct predictions that still offer useful ADR information. To handle this limitation, a relaxed evaluation strategy is adopted [27] in which predictions that partially overlap with the gold-standard entities are considered as valid matches. These partial matches are recognized by calculating the cosine similarity between the embeddings of predicted and reference entities by setting a threshold of 0.8, which is sufficient to consider semantic similarity. This dual evaluation framework evaluates not only the model's precision in boundary extraction but also its ability to recognize semantically relevant ADRs when the predicted and true boundaries do not perfectly match.

The predicted entities are grouped into five categories as illustrated in Fig. 2:

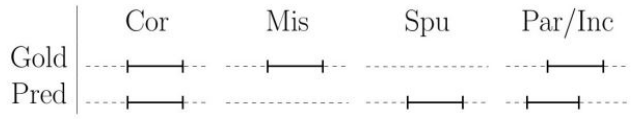

Fig. 2. Visual representation of Correct (Cor), Missed (Mis), Spurious (Spu), and Partial/Incorrect (Par/Inc) entity alignments [27].

- Correct (Cor): Predicted ADRs that match the gold standard exactly in both span and content.
- Partial (Par): Predicted ADRs that partially overlap with a gold-standard entity or are semantically similar, identified using cosine similarity above the threshold.
- Incorrect (Inc): Predicted ADRs overlapping a gold standard entity but not exceeding the similarity threshold.
- Spurious (Spu): Predicted ADRs that do not correspond to any gold-standard entity.
- Missed (Mis): Gold-standard ADRs that were not predicted by the model.

Using these categories, the main evaluation metrics are defined as follows:

Precision:

$$\text{Precision} = \frac{Cor + (Par \times 0.5)}{Cor + Par + Inc + Miss} \tag{6}$$

Recall:

$$\text{Recall} = \frac{Cor + (Par \times 0.5)}{Cor + Par + Inc + Spu} \tag{7}$$

F1-Score:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

For strict evaluation, partial matches are excluded (Par=0), and only exact matches are considered as correct. This provides an exact measure of the model's capability in identifying ADRs correctly. In contrast, the relaxed evaluation includes partial matches, allowing for a more flexible assessment that accounts for semantically relevant extractions even when the predicted span does not align exactly with the gold standard. The proposed framework for this combined evaluation will make sure that both exact and partially correct ADR predictions are suitably assessed to provide a comprehensive view of model performance in extracting Adverse Drug Reactions from patient-generated texts.

IV. EXPERIMENTAL RESULT ANALYSIS AND DISCUSSION

A. Experimental Settings

A set of experiments involving both Soft MoE and Hard MoE architectures has been conducted to estimate the performance of Mixture-of-Experts network architectures. A variety of encoder-decoder models of the T5 family have been investigated in the experiment involving Soft MoE. These variants include T5 base, Flan-T5 base (instruction-tuned), T5 small (efficient small model), BioT5-base (biomedically tuned), and Long-T5-tglobal-base (long-context model). T5-base is used as the tokenizer in this architecture type. For the Hard MoE, experiments were carried out using BART-based architectures, including BART-base, BART-large, mBART-large-50 (multilingual version), and BioBART-base (biomedical variant), with BART-base used as the tokenizer in this experiment as well. A

top-k = 2 routing strategy has been employed to select the two most relevant experts given the input, ensuring both efficiency and expert specialization.

Through various tests of model combinations and hyperparameter values, the final set of values was chosen based on overall performance and the stability of training observed during experimentation. All of these scenarios had been optimized using AdamW optimizer with a learning rate of 3e-5 and trained for 12 epochs with small batch sizes due to computational constraints. A description of the full hyperparameters, including the number of experts, hidden dimensions, optimizer, and batch sizes in both architectures, is provided in Table II.

TABLE II. EXPERIMENTAL HYPERPARAMETER SETTINGS FOR SOFT AND HARD MOE

Model Hyperparameters	Soft MoE	Hard MoE
num_expert	5	4
hidden_dim	512	512
num_beans	4	4
optimizer	AdamW	AdamW
learning rate	3e-5	3e-5
train batch_size	4	6
test batch_size	4	4
epochs	12	7

B. Result Analysis

Table III presents the performance comparison between the Soft and Hard MoE models under both strict and relaxed evaluation settings. The findings show a consistent and notable performance difference that can be directly related to their underlying architectural principles. As the table demonstrates, the Soft MoE model performs better overall, attaining a strict F1-score of 72.90% compared to the 71.60% achieved by the Hard MoE. This performance gap of 1.3 F1-score under the most strict parameters is not negligible and is sustained under relaxed parameters as well, where Soft MoE outperforms by 1% F1-score (80.40% vs. 79.40%). To further contextualize these results, the analysis is extended by including benchmark results from the MultiADE research study [29], which evaluated span-based, generative, and large language model approaches on the PsyTAR dataset. As shown in Table III, the proposed Soft MoE achieves a strict F1-score of 72.90%, outperforming both the best-performing span-based (68.50%) and generative models (68.20%) models, GPT-4 and Llama-3 models in a 5-shot setting, which achieved 58.80% and 57.70%, respectively. This comparison underscores that the proposed Soft MoE model not only surpasses its Hard MoE counterpart but also achieves competitive performance among existing approaches for ADR extraction from the PsyTAR dataset.

The performance gain by Soft MoE might be due to its adaptive and collaborative gating mechanism, which allows multiple experts to jointly contribute towards processing each input. The flexible routing enables smooth information sharing between the experts, enabling the model to capture fine-grained contextual cues and semantic subtleties that frequently occur in user-generated medical texts. Adaptability to such heterogeneity is particularly helpful for ADR extraction from social media, where the expressions of adverse effects are often informal, fragmented, or context-dependent. By leveraging the combined expertise of multiple sub-networks, the Soft MoE demonstrates a more nuanced understanding of these heterogeneous textual patterns. In contrast, Hard MoE activates only a limited set of experts for each token or sequence. Hence, there are more discrete decision boundaries. This can be beneficial because it is computationally efficient, but might suppress contributions by relevant experts for more complex levels of linguistic and ambiguous representations. This may be the most reasonable explanation for the observed gap: the Hard MoE fails to fully leverage its capacity when dealing with noisy and contextually rich data.

Overall, it is evident that the results presented above have established strong empirical support for the routing strategy playing a crucial role in influencing the performance of MoE-based architectures in ADR extraction tasks. The capability of the Soft MoE to leverage distributed knowledge makes it well-suited for the context of real-world digital pharmacovigilance applications.

C. Discussion

The experimental results demontrate effectiveness of the MoE paradigm in the extraction of ADRs from patient-generated narratives. Both Soft and Hard MoE variants leverage specialized sub-networks, enabling the models to handle the semantic variability and informal language commonly used by patients, an aspect where conventional monolithic models often struggle. However, the Soft MoE model shows superior performance due to its ability to aggregate knowledge from various experts simultaneously. This collaborative routing technique enables the model to better understand nuanced expressions, implicit mentions, and context-dependent ADRs, which are frequent in patient narratives. In contrast, Hard MoE, while computationally efficient, may miss some complex interactions between experts, leading to slightly lower performance. This architectural difference also affects response time and deployability. The sparse expert routing of Hard MoE enables faster inference, making it more suitable for latency-sensitive or high-throughput pharmacovigilance systems. By comparison, the denser expert activation in Soft MoE incurs higher computational overhead per prediction in exchange for enhanced extraction accuracy.

The MoE architecture has several practical advantages. Its modularity can be scaled with minimal increase in training cost, and the routing mechanism promotes balanced expert utilization, avoiding overfitting to specific patterns. Compared to standard transformer models, MoE achieves competitive performance without significantly increasing model size or resource requirements, making it a cost-effective and flexible approach for real-world applications.

Further, this architecture naturally supports the integration of many expert knowledge sources, which can be extended to domain-specific biomedical experts or external knowledge bases in future studies. This adaptability makes MoE a robust solution for pharmacovigilance tasks, capable of handling linguistic diversity, improving semantic precision, and ultimately improving automated monitoring of drug safety.

Despite these advantages, the Mixture-of-Experts architecture presents inherent limitations related to routing behavior

TABLE III. PERFORMANCE COMPARISON OF SOFT AND HARD MOE MODELS WITH BENCHMARK RESULTS ON THE PSYTAR DATASET

Model		Strict			Relaxed		
		P	**R**	**F1**	**P**	**R**	**F1**
Benchmark Models (MultiADE Study [29])	Llama-3 (5-shot)	–	–	57.70%	–	–	–
	GPT-4 (5-shot)	–	–	58.80%	–	–	–
	Fine-tuned (BART Generative Model)	–	–	68.20%	–	–	–
	Fine-tuned (Span-based)	–	–	68.50%	–	–	–
Our Models	**Hard MoE**	72.40%	70.70%	71.60%	80.40%	78.50%	79.40%
	Soft MoE	73.20%	72.60%	**72.90%**	80.80%	80.10%	**80.40%**

and computational complexity. Model performance remains sensitive to the quality of the gating mechanism, particularly in the Hard MoE setting, where suboptimal expert selection may lead to information loss for linguistically ambiguous ADR expressions. In addition, the Soft MoE configuration requires activating all experts for each input, which increases computational cost and may limit scalability in resource-constrained pharmacovigilance applications. Furthermore, expert specialization emerges implicitly during training and is not explicitly enforced, which may result in overlapping expert representations rather than clearly distinct expertise.

In summary, the study confirms that MoE architectures not only deliver strong results in ADR extraction but also provide a modular, efficient, and adaptable framework that can be further optimized for biomedical text mining

V. CONCLUSION

In this study, the use of a Mixture-of-Experts architecture was investigated for the task of Adverse Drug Reaction extraction from patient-generated narratives, where the task is challenging due to the informal, diverse, and context-dependent nature of such texts. The performance of Soft MoE and Hard MoE architectures was compared to highlight the impact of routing strategies on model performance. Experimental results demonstrate that the Soft MoE outperforms the Hard MoE in terms of precision, recall, and F1-measure in both strict and relaxed evaluations. This improvement can be attributed to the Soft MoE's ability to integrate contributions from multiple experts simultaneously, enabling a richer contextual understanding and better handling of informal and subtle language typical for social media and patient reviews.

As a direction for future work, it would be promising to enhance the MoE framework through the integration of external biomedical knowledge bases, such as medical ontologies, pharmacological databases, or drug–disease interaction graphs, to better reinforce the model's domain awareness and semantic precision. Additionally, incorporating mechanisms to capture complex drug–drug interactions and contextual dependencies could further refine ADR extraction and contribute to a more comprehensive view of patient-reported pharmacovigilance data. Ultimately, these advancements could bring MoE-based systems closer to real-world applications in automated drug safety monitoring.

REFERENCES

[1] Y. Zhang, S. Cui, and H. Gao, "Adverse drug reaction detection on social media with deep linguistic features", *Journal of Biomedical Informatics*, 2020, doi: 10.1016/j.jbi.2020.103437.

[2] O. Oyebode and R. Orji, "Identifying adverse drug reactions from patient reviews on social media using natural language processing", *Health Informatics Journal*, 2023, doi: 10.1177/14604582221136712.

[3] P. García-Abeijon, C. Costa, M. Taracido, M. Teresa Herdeiro, C. Torre, A. Figueiras, "Factors Associated with Underreporting of Adverse Drug Reactions by Health Care Professionals: A Systematic Review Update", *Journal of Drug Safety*, 2023, doi: 10.1007/s40264-023-01302-7.

[4] S. Golder, K. O'Connor, Y. Wang, A. Klein, and G. Gonzalez Hernandez, "The value of social media analysis for adverse events detection and pharmacovigilance: A scoping review", *Journal of Medical Internet Research*, 2024, doi: 10.2196/59167.

[5] A. Dirkson, S. Verberne, W. Kraaij, G. van Oortmerssen, and H. Gelderblom, "Automated gathering of real-world data from online patient forums can complement pharmacovigilance for rare cancers", *Scientific Reports*, 2022, doi: 10.1038/s41598-022-13894-8.

[6] R. Deepalakshmi, P. Manikandan, V. Manikandan, "Detecting Adverse Drug Reactions from Twitter Data Using Natural Language Processing and Deep Learning", *Journal of Information Systems Engineering and Management*, 2025, doi: 10.52783/jisem.v10i30s.4896.

[7] C. Wu, F. Wu, J. Liu, S. Wu, Y. Huang, and X. Xie, "Detecting Tweets Mentioning Drug Name and Adverse Drug Reaction with Hierarchical Tweet Representation and Multi-Head Self-Attention", in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 34–37, doi: 10.18653/v1/W18-5909.

[8] J.Y. Huang, W.P. Lee 1, K.D. Lee, "Predicting Adverse Drug Reactions from Social Media Posts: Data Balance, Feature Selection and Deep Learning", *Healthcare*, 2022, doi: 10.3390/healthcare10040618.

[9] M. Tiftikci, A. Özgür, Y. He, and J. Hur, "Machine learning-based identification and rule-based normalization of adverse drug reactions in drug labels", *BMC Bioinformatics*, 2019, doi: 10.1186/s12859-019-3195-5.

[10] O. Elbiach, H. Grissette, and E. H. Nfaoui, "Adverse Drug Reactions Detection from Social Media: an Empirical Evaluation of Machine Learning Techniques", in *Conference: 2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA)*, 2023, doi: 10.1109/SITA60746.2023.10373604.

[11] O. Elbiach, H. Grissette, and E. H. Nfaoui, "Leveraging Transformer Models for Enhanced Pharmacovigilance: A Comparative Analysis of ADR Extraction from Biomedical and Social Media Texts", *Artificial Intelligence (AI)*, 2025, doi: 10.3390/ai6020031.

[12] O. Elbiach, H. Grissette, and E. H. Nfaoui, "Benchmarking large language models for adverse drug reaction extraction in social media and clinical texts", *Results in Engineering*, 2025, doi: 10.1016/j.rineng.2025.107362.

[13] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A Survey on Mixture of Experts in Large Language Models", *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, pp. 3896-3915, 2025, doi: 10.1109/TKDE.2025.3554028.

[14] S. Lee, H. Woo, C. Chun Lee, G. Kim, J.Y Kim, and S. Lee, "Drug_SNSMiner: standard pharmacovigilance pipeline for detection of adverse drug reaction using SNS data", *Scientific Reports*, vol. 20, 2023, doi: 10.1038/s41598-023-28912-6.

[15] S. Modi, K. A. Kasmiran, N. M. Sharef, and M. Y. Sharum, "Extracting adverse drug events from clinical Notes: A systematic review of approaches used", *Journal of Biomedical Informatics*, vol. 151, 2024, doi: 10.1016/j.jbi.2024.104603.

[16] P. Joshi, M. Vedhanayagam, and R. Ramesh, "An Ensembled SVM-Based Approach for Predicting Adverse Drug Reactions", *Current Bioinformatics*, vol. 16, pp. 422-432, 2019, doi: 10.2174/1574893615999200707141420.

[17] d. Fan,S. Zhou1,Y. Li, R. Zhang, "Deep learning approaches for extracting adverse events and indications of dietary supplements from clinical text", *Journal of the American Medical Informatics Association*, vol. 16, pp. 422-432, 2021, doi: 10.1093/jamia/ocaa218.

[18] B. Khemani, S. Malave, S. Shinde, M. Shukla, R. Shikalgar, and H. Talwar, "AI-driven pharmacovigilance: Enhancing adverse drug reaction detection with deep learning and NLP", *MethodsX*, 2025, doi: 10.1016/j.mex.2025.103460.

[19] F. Dong, W. Guo, J. Liu, T. A. Patterson, and H. Hong, "BERT-based language model for accurate drug adverse event extraction from social media: implementation, evaluation, and contributions to pharmacovigilance practices", *Frontiers in Public Health*, 2024, doi: 10.3389/fpubh.2024.1392180.

[20] M. M. Zitu, D. Owen, A. Manne, P. Wei, and L. Li, "Large Language Models for Adverse Drug Events: A Clinical Perspective", *Journal of Clinical Medicine*, 2025, doi: 10.3390/jcm14155490.

[21] D. Dai, W. Jiang, J. Zhang, W. Peng, Y. Lyu, Z. Sui et al., "Mixture of Experts for Biomedical Question Answering", *arXiv*, 2022, doi: 10.48550/arXiv.2204.07469.

[22] Q. Liu, X. Wu, X. Zhao, Y. Zhu, D. Xu, F. Tian et al., "MOELoRA: An MOE-based Parameter Efficient Fine-Tuning Method for Multi-task Medical Applications", *arXiv*, 2024, doi: 10.48550/arXiv.2310.18339.

[23] X. Chen, K. Li, T. Song, and J. Guo, "'Mix of Experts Language Model for Named Entity Recognition", *arXiv*, 2024, doi: 10.48550/arXiv.2404.19192.

[24] M. Zolnoori, K. W. Fung, T. B. Patrick, P. Fontelo, H. Kharrazi, A. Faiola et al., "The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications", *Data in Brief*, vol. 24, 2019, doi: 10.1016/j.dib.2019.103838.

[25] T. Shen, M. Ott, M. Auli, and M. Ranzato, "Mixture Models for Diverse Machine Translation: Tricks of the Trade", *arXiv*, 2019, doi: 10.48550/arXiv.1902.07816.

[26] N. Omi, S. Sen, and A. Farhadi, "Load Balancing Mixture of Experts with Similarity Preserving Routers", *arXiv*, 2025, doi: 10.48550/arXiv.2506.14038.

[27] S. Scaboro, B. Portelli, E. Chersoni, E. Santus, and G. Serra, "Extensive Evaluation of Transformer-based Architectures for Adverse Drug Events Extraction", *Knowledge-Based Systems*, vol. 275, 2023, doi: 10.48550/arXiv.2306.05276.

[28] M. Zolnoori, K. W. Fung, T. B. Patrick, P. Fontelo, H. Kharrazi, A. Faiola et al., "A systematic approach for developing a corpus of patient-reported adverse drug events: A case study for SSRI and SNRI medications", *Journal of Biomedical Informatics*, vol. 90, 2019, doi: 10.1007/s001090000086.

[29] X. Dai, S. Karimi, A. Sarker, B. Hachey, and C. Paris, "MultiADE: A Multi-domain benchmark for Adverse Drug Event extraction", *Journal of Biomedical Informatics*, 2024, doi: 10.1016/j.jbi.2024.104744.