Leveraging Large Language Models for Adverse Drug Event Detection: A Comparative Study of Token and Span-Based Named Entity Recognition

Howard Prioleau†, Saurav K. Aryal, and Jeremy Blackstone
EECS, Artificial Intelligence for Positive Change (AI4PC) Lab, Howard University, Washington, DC 20059, USA
†*E-mail: howard.prioleau@bison.howard.edu*
https://ai4pc.howard.edu/

Adverse Drug Events (ADEs) pose a persistent threat to patient safety and public health. This study investigates the use of large language models (LLMs) fine-tuned for both token classification and span-based named entity recognition (NER) to improve ADE detection in clinical text. Using the 2018 n2c2 Track 2 dataset, we evaluate models under both predefined (gold label) and end-to-end settings. RoBERTa Large consistently outperforms other models, particularly in identifying ADEs, which remain more challenging due to their contextual ambiguity. Token-based models generally deliver stronger performance than span-based approaches, and ensemble methods, especially majority voting and XGBoost-based aggregation, further enhance end-to-end relation extraction by mitigating individual model weaknesses. These findings highlight the potential of fine-tuned LLMs, augmented by strategic ensembling, to advance clinical NLP pipelines and support safer, more personalized healthcare.

Keywords: Artificial intelligence, Large language models, Clinical Notes, Adverse Drug Events, n2c2

1. Introduction

Adverse Drug Events (ADEs), defined as "harm caused by the use of a drug",[1] continue to pose a critical threat to patient safety and the healthcare system at large. In 2023, the FDA's Adverse Event Reporting System (FAERS) documented over 2.15 million ADE cases, including nearly 163,000 associated deaths. The economic impact is equally stark, with estimated costs reaching up to $30 billion annually in the United States alone.[2] These figures have doubled over the past decade and are expected to rise, highlighting the urgent need for more effective detection methods. Beyond the financial and human costs, ADEs erode public trust in medical institutions, potentially deterring individuals from adhering to treatments or seeking timely care.[3,4]

While pre-market clinical trials aim to identify potential drug risks, many ADEs surface only after widespread public use, particularly in vulnerable groups such as the elderly.[3] Factors like genetic variability, comorbidities, and even psychosomatic effects like the nocebo phenomenon contribute to the complexity of ADE identification.[5] Developing systems that

can detect ADEs at the individual level is therefore essential, but the fragmented and unstructured nature of clinical documentation challenges such efforts.

This research focuses on improving adverse drug event (ADE) detection using large language models (LLMs) fine-tuned for token classification and span-based named entity recognition (NER). The goal is to accurately identify Drug and ADE mentions in clinical text, and then perform relationship extraction (RE) to determine which drugs are associated with which ADEs. Token classification enables efficient sequence labeling, while span-based methods are better suited for handling overlapping or discontinuous entities common in medical narratives. This dual modeling strategy addresses challenges posed by variable document lengths, complex entity structures, and annotation inconsistencies. To further enhance performance and robustness, we explore ensemble methods that combine diverse model predictions, revealing that aggregating complementary strengths yields more reliable end-to-end relation extraction. Using the n2c2 dataset as a benchmark, our findings demonstrate the potential of LLMs, especially when combined through ensemble strategies, to refine ADE detection and support safer, more personalized clinical decision-making.

2. Relevant Works

This section explores key developments in ADE detection and medical Natural Language Processing (NLP), focusing on Token and sequence classification techniques and relation extraction.

2.1. *ADE Detection Task*

Early research in ADE detection began with the development of the ADE Corpus,[6] which comprises 2,972 curated medical documents containing manual annotations for drugs, adverse effects, dosages, and their interrelations. This corpus resulted in a total of 20,967 sentences, with 4,272 labeled as Positive for containing at least one mention of a drug-related adverse effect, and the remaining 16,695 labeled as Negative, indicating the absence of such mentions. Building upon this foundation, the TAC 2017 shared task[7] introduced an initiative focused on extracting adverse reactions from structured drug labels and mapping them to standardized MedDRA terms. However, the XML format used in this task imposed a rigid document structure, which limited the retention of nuanced contextual information found in natural clinical narratives. This constraint reduced the effectiveness of relation extraction by stripping away the free-form, real-world textual cues essential for accurately identifying ADEs.

To overcome these limitations, the 2018 National NLP Clinical Challenges Track 2 (n2c2) task[8] offered a more comprehensive and clinically realistic dataset. As the primary data source for this study, the n2c2 dataset significantly broadened the annotation schema to include not only drugs and ADEs but also associated attributes such as strength, form, dosage, frequency, route, duration, and reason. Moreover, it required the identification of relationships among these entities, enhancing the potential for more precise drug-event linkage. Importantly, the data preserved the original structure of electronic health records (EHRs), maintaining the contextual richness and linguistic complexity inherent to clinical documentation. This design makes the n2c2 dataset particularly well-suited for advancing ADE detection models in real-

world healthcare settings.

2.2. *Named Entity Recognition*

NER is a fundamental component of NLP, playing a crucial role in the transformation of unstructured text into structured information. In the medical domain, NER is particularly valuable for identifying key clinical entities such as diseases, medications, procedures, and other relevant concepts. Large language model (LLM) generation-based methods have also been explored for NER,[9,10] but these approaches introduce their own challenges, such as instability and difficulty in evaluation. In this work, we focus on the two more established approaches to NER: token classification and span classification, each offering distinct advantages depending on the complexity and structure of the entities being extracted.

2.2.1. *Token Classification*

Prior to the widespread adoption of large language models (LLMs), state-of-the-art token classification for ADE detection relied heavily on Bidirectional Long Short-Term Memory networks (BiLSTMs). One notable example is the work by,[11] which proposed a cascaded architecture consisting of three sequential stages, each employing BiLSTM-CRF models trained on distinct, non-overlapping entity types. These models incorporated a combination of character-level representations (via CNN or LSTM), word embeddings, and handcrafted linguistic features. An ensemble layer in the final stage aggregated predictions from the earlier stages, resulting in an overall F1 score of 92.87% and an ADE-specific F1 of 38.75%. The shift from BiLSTM-based architectures to LLMs, particularly those with domain-specific pretraining, has significantly enhanced token classification performance, especially for complex and context-dependent entities like ADEs.

A leading example of this evolution is the MC-DRE model,[12] which achieves state-of-the-art performance on the n2c2 dataset. MC-DRE leverages PubMedBERT, a domain-adapted LLM, within a multi-aspect cross-integration framework. This architecture integrates semantic, syntactic, and domain-specific representations using a combination of key-value attention mechanisms, feedforward networks, and cross-learning strategies. The model is further strengthened by auxiliary tasks, such as part-of-speech tagging and general medical NER, along with the use of BIOHD tagging to effectively capture discontinuous entities. As a result, MC-DRE achieves an impressive 98.38% overall F1 and a 61.59% F1 specifically for ADE detection.

2.2.2. *Span Classification*

Span-based classification has gained traction as a powerful alternative to token-level approaches for NER and RE, particularly in addressing overlapping, nested, or discontinuous entities. Unlike traditional sequence tagging, span classification directly predicts entity spans and their corresponding types, allowing for greater flexibility and improved handling of complex clinical text. One prominent model in this area is SpanBERT,[13] which introduced span-level pretraining objectives designed to enhance span representations. Rather than masking indi-

vidual tokens, SpanBERT masks contiguous spans of text and incorporates a span-boundary objective that trains the model to predict the full content of a masked span based solely on its boundary tokens. This approach allows the model to better capture dependencies within and between spans, resulting in notable improvements on span-centric tasks such as question answering and coreference resolution.

Building on this foundation, Packed Levitated Marker (PL-Marker)[14] improves span and span-pair modeling with a marker-based representation method. Instead of processing each span pair separately, PL-Marker groups together spans that share similar boundaries, allowing the model to encode them jointly and improve both boundary precision and contextual understanding. For span-pair classification tasks such as RE, it applies a subject-oriented packing strategy that links one subject span with multiple related object spans. The approach uses levitated markers, which preserve the original token positions while enabling span-specific attention, making it possible to encode many span candidates in parallel. Experiments show that this design yields up to a 4.3% strict F1 improvement over prior state-of-the-art RE models on ACE04 and ACE05 datasets, and also outperforms strong baselines on multiple NER benchmarks.

2.3. *Relation Extraction*

Relation extraction is often simpler than NER when working with gold entities, as it typically involves determining if a specific relationship exists between given entity pairs, formatted as binary questions or multi-output options. Techniques for RE leverage this reduced complexity to achieve high performance.[15] implemented two random forest classifiers first to detect the existence of relations and then classify their types, achieving a 95% F1 and[11] used SVM classifiers with n-grams, token-level distance, and semantic types as features, considering all drug-entity pairs within one sentence as candidates which achieved a 94% F1. These tasks can also be performed jointly using models like BiLSTM-CRFs, which combine bidirectional long short-term memory networks (BiLSTMs) with conditional random fields (CRFs) to simultaneously handle NER and RE.

3. Methodology

3.1. *Dataset*

This study uses the n2c2 dataset,[8] consisting of 505 discharge summaries from the Medical Information Mart for Intensive Care III (MIMIC-III) database. These medical documents were annotated by seven domain experts, including four physician assistant students and three nurses, for entities such as drugs, adverse drug events (ADEs), strength, form, dosage, frequency, route, duration, and reasons using specific entity tags and attributes.

3.2. *Pre-processing*

Due to the highly structured nature of medical notes, pre-processing was essential to make the text more conducive to entity extraction. Due to the considerable length of the medical notes, we first split each note into sentences using the NLTK project's punkt sentence splitter.[16] This

pre-processing step was necessary because the models struggled to classify tokens effectively when handling very long sequences. We then mapped every token of the sentences to its own entity classification where "ADE" and "Drug" were preserved, while all other entities were consolidated under a generic 'other' label, simplifying the focus on the primary entities of interest.

Our dataset consists of annotated entities across three categories: 'Other', 'Drug', and 'ADE'. The training set contains 96,350 'Other', 18,039 'Drug', and 1,665 'ADE' entities, while the test set includes 74,765 'Other', 11,905 'Drug', and 1,057 'ADE' entities. This distribution highlights a significant class imbalance, with ADE instances being markedly underrepresented. Such imbalance has been shown to negatively affect model performance in prior work, particularly in the reliable detection of ADEs.

3.3. *Modeling*

Table 1: Release year and parameter size of selected models

Model	Release Year	Parameter Size
Qwen2.5-0.5B[17]	2025	500M
RoBERTa-large[18]	2019	355M
ModernBERT-base[19]	2025	149M
BERT-base-cased[20]	2018	110M

This section covers the modeling approaches for each task: Token Classification, Span Classification, and Relationship Extraction. Across these three tasks, we selected four LLMs with different architectures and release years to capture variation in both model design and historical development. The 4 models are showcased here 1, to see if any patterns emerge from model size to when they were trained. The 4 models were: Qwen2.5-0.5B,[17] Roberta-large,[18] ModernBERT-base,[19] and bert-base-cased.[20]

3.3.1. *Token Classification*

For the Token Classification task, sentences from the annotated dataset were tokenized using the tokenizer corresponding to each model. Specifically, sentences longer than 64 tokens were split into overlapping chunks with a maximum token length of 64 and an overlap of 50%. Overall, this approach improved modeling efficiency and ensured more consistent processing times by reducing variability in sentence length and minimizing the need for padding with zeros.

Each token in these sentences was assigned to one of three predefined classes: ADE, Drug, or Other, as illustrated in Figure 1. We used Hugging Face's Trainer[21] API for training and evaluation. The models were fine-tuned for five epochs with a batch size of 32, a learning rate of 2e-5, and mixed-precision training (bf16) enabled for improved computational performance.

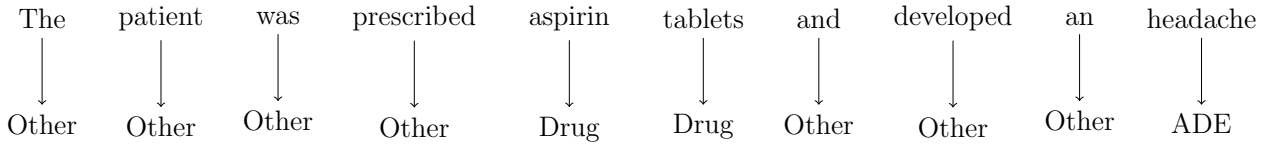


Fig. 1: Token classification for: *"The patient was prescribed aspirin tablets and developed an headache"*. Tokens: **Drug** = "aspirin", "tablets"; **ADE** = "headache"; others = **Other**.

3.3.2. *Span Classification*

For the Span Classification task, we used the Hugging Face implementation of SpanMarker,[22] based on the Packed Levitated Marker (PL-Marker) framework. We initialized the model with bert-base-cased and roberta-large encoders, which were the only ones compatible with SpanMarker due to tokenizer constraints. Attempts to use Qwen2.5-0.5B and ModernBERT-base were unsuccessful, as they do not support the marker-style span encoding required by the architecture. As a result, only encoders with standard Hugging Face tokenizer support were used.

Each candidate span was classified as ADE, Drug, or Other by aggregating token-level predictions into labeled spans, as illustrated in Figure 2. Fine-tuning was performed for five epochs using the Hugging Face Trainer API with a batch size of 32, a learning rate of 5e-5, and mixed-precision training (bf16) to improve efficiency. These settings were kept consistent across encoders to ensure comparability.

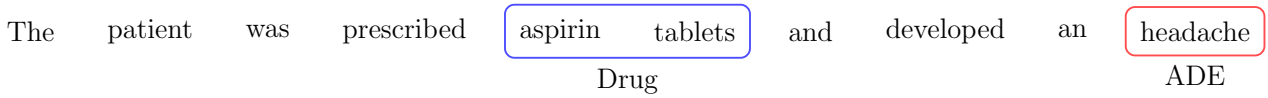


Fig. 2: Span classification for: *"The patient was prescribed aspirin tablets and developed an headache"*. Spans: **Drug** = "aspirin tablets"; **ADE** = "headache".

3.3.3. *Relationship Extraction*

Relationship extraction involves identifying and classifying relationships between drug entities and their corresponding adverse events. The relationship extraction problem was transformed into a binary classification scenario by selecting sentences containing both drug and adverse event mentions, masking these specific entities, and then classifying whether the masked drug-ade pairs were related or unrelated.

This approach was implemented through two distinct experiments: the Golden Labels experiment and the End-to-End experiment. In the Golden Labels experiment, entity spans were predefined (gold standard), and models were tasked with classifying these specified drug-event pairs. In the end-to-end experiment, the models first predicted entities in the test set and then these predicted entities were classified using the models previously trained on gold-standard relation labels.

3.3.4. *Ensembling*

To explore how much additional performance could be gained by leveraging the multiple models we trained, we implemented various token-level ensembling techniques. This effort was motivated by prior work that demonstrated the potential of ensemble approaches in natural language processing tasks,[23] in health-related applications,[24] and more specifically in the ADE extraction domain.[25] We applied voting-based methods, including *majority voting*, where the most frequent label among models was selected, and *minority voting*, which emphasized outlier predictions. Confidence-based techniques were also tested, such as selecting the *most confident prediction* based on the highest individual model probability, and computing the *sum of probabilities* across models to determine the dominant class. Additionally, we trained meta-classifiers including Decision Trees, Random Forests, Support Vector Machines (SVM), LightGBM, XGBoost, Logistic Regression, CatBoost, and Multi-Layer Perceptrons (MLP) on the binary outputs (votes), class probabilities, or a combination of both. This comprehensive framework enabled us to assess the extent to which ensemble methods could enhance token-level ADE classification performance.

3.4. *Evaluation*

Model performance across all tasks was assessed using standard metrics commonly used in NLP research. The evaluation criteria for each task are briefly summarized below.

3.4.1. *Token Classification*

Token classification performance was evaluated at the token level using precision, recall, accuracy, and F1-score metrics.

3.4.2. *Span Classification*

Span classification metrics included precision, recall, and F1-score, which were calculated after mapping span-level predictions back to token-level annotations. A prediction was considered correct only if the predicted span boundaries and associated labels exactly matched the corresponding ground truth tokens for ADE or Drug entities.

3.4.3. *Relationship Extraction*

Relationship extraction was evaluated using precision, recall, and F1-score. Correct predictions required accurately identifying pairs of Drug and ADE entities along with their correct relationship according to the annotated dataset.

In Figure 3, details how performance is evaluated in the End to End setting. Since the evaluation depends on the model's predicted entities and relations, some true Drug ADE pairs cannot be evaluated because the corresponding entities were never identified by the model. In such cases, these missed pairs are counted as false negatives to reflect the model's end to end limitations.

For Token and Span Classification the macro average will be reported due to the imbalanced nature of the dataset we want to treat all classes as equal when it comes to gauge

Actual\Predicted	**Related**	**Not Related**
Related	**True Positive (TP)** Drug-ADE pair is related and correctly predicted as related	**False Negative (FN)** Drug-ADE pair is related but predicted as not related (or entity in pair was not predicted)
Not Related	**False Positive (FP)** Drug-ADE pair is not related but predicted as related	**True Negative (TN)** Drug-ADE pair is not related and correctly predicted as not related

Fig. 3: Confusion matrix for Drug–ADE relationship extraction showing classification categories with criteria for Relation

performance.

4. Results

4.1. *Token & Span Classification*

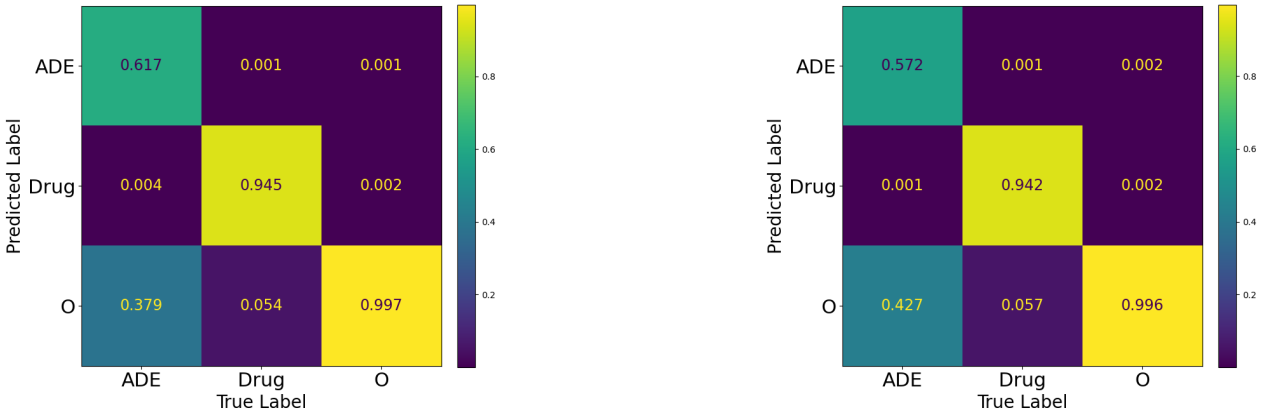
model name	Exp. Type	ADE-f1	ADE-pr	ADE-re	Drug-f1	Drug-pr	Drug-re	macro-f1
ModernBERT	token	0.49258	0.66083	0.39262	0.94026	0.95026	0.93047	0.80990
Roberta large	token	**0.54813**	0.61730	**0.49290**	**0.94612**	0.94549	**0.94676**	**0.83043**
bert_base_cased	token	0.46102	0.56612	0.38884	0.93781	0.93979	0.93584	0.79849
Qwen2.5	token	0.21021	0.48299	0.13434	0.79887	0.85807	0.74731	0.66714
bert_base_cased	span	0.42832	0.56135	0.34626	0.93247	0.94275	0.92241	0.78575
Roberta large	span	0.47740	0.57199	0.40965	0.94052	0.94219	0.93886	0.80488
majority_vote	ensemble	0.50898	0.73656	0.38884	0.94579	0.95860	0.93332	0.81730
minority_vote	ensemble	0.32744	0.32205	0.33302	0.76664	0.81949	0.72019	0.69496
highest_proba_vote	ensemble	0.33554	0.75497	0.21570	0.93311	0.95839	0.90914	0.75509
sum_proba_vote	ensemble	0.42004	0.76500	0.28950	0.94584	0.96084	0.93131	0.78765
meta_vote_XGB	ensemble	0.50234	0.65899	0.40587	0.94013	0.96146	0.91972	0.81312
meta_proba_LGBM	ensemble	0.44869	0.68750	0.33302	0.94599	0.96069	0.93173	0.79724
meta_vote_and_proba_MLP	ensemble	0.37536	**0.77286**	0.24787	0.94166	**0.96616**	0.91837	0.77131

Table 2: Test set performance of token, span, and ensemble models for ADE and Drug classification using per-class and macro-averaged metrics. Only the best macro-F1 per meta-ensemble strategy is shown.

Table 2 presents performance metrics for all evaluated models across token-based, span-based, and ensemble approaches for ADE and Drug entity recognition. Among the individual models, Roberta large (token) achieves the best overall performance, with the highest macro-F1 score and leading F1 for both ADE and Drug classes. It balances precision and recall particularly well, demonstrating strong generalization to both frequent and less frequent entities. In contrast, ModernBERT yields the highest precision for both classes, especially ADE, but this comes at the cost of reduced recall, resulting in a slightly lower macro-F1. BERT base cased, across both token and span variants, maintains consistent performance, though it underperforms relative to Roberta in ADE recognition. Qwen2.5 shows the weakest performance

overall, struggling particularly with ADE recall, which substantially limits its macro-F1 despite acceptable Drug classification. This suggests that while high Drug precision is relatively easy to achieve, due to consistent lexical cues and training frequency. ADE extraction remains a more discriminative benchmark of model capability.

When comparing token-based to span-based approaches, token models generally outperform span models across both entity types, especially in ADE F1 and macro-F1. This is likely due to finer token-level supervision and broader compatibility with standard pre-trained architectures. For example, span-based RoBERTa lags behind its token counterpart in ADE and macro-F1, despite comparable Drug performance. Still, span models remain competitive for Drug entities and may be useful where precise boundary detection is critical. Ensemble methods provide a strong alternative, with strategies such as majority_vote and meta_vote_XGB surpassing most single models in macro-F1 and narrowing the precision–recall gap for ADEs. Notably, some ensembles (e.g., meta_vote_and_proba_MLP) achieve very high precision at the cost of recall, while minority_vote sometimes identifies difficult ADEs missed by the majority, though with low consensus and reliability. These results suggest that model diversity can uncover challenging cases that uniform agreement overlooks, highlighting the importance of complementary perspectives in ensemble design.



(a) Token Classification Roberta Large

(b) Span Classification Roberta Large

Fig. 4: Confusion matrices for Token Classification models on the test set.

Figures 4a and 4b present a comparison of Roberta large's performance in token-based versus span-based classification, using column-normalized confusion matrices to show per-class precision. In the token classification setting, Roberta large demonstrates strong performance, particularly for the Drug class, with minimal confusion and high precision, while still exhibiting some misclassification of ADE instances as Other. In the span-based setting, this challenge becomes more pronounced: the model's precision for ADE decreases noticeably, with a larger proportion of predictions incorrectly assigned to the Other class, despite maintaining reliable precision for Drug entities. Notably, ADE predictions are rarely confused with Drug, suggesting

that while the model retains a semantic distinction between the two, it is less confident in identifying ADE spans. This performance gap mirrors the quantitative trends observed in Table 2, where the span-based Roberta model underperforms its token counterpart in ADE F1 and macro-F1. These results indicate that token-level classification offers improved precision for rare and syntactically diverse classes like ADE, likely due to denser contextual supervision, whereas span-level approaches may introduce greater uncertainty for minority labels, despite their utility in precise boundary modeling.

4.2. *Relationship Extraction*

There will be 2 sections from the Golden Label RE and End to End RE

4.2.1. *Golden Label*

Model	Accuracy	Macro-pr	Macro-re	Macro-F1
ModernBERT	0.7607	0.7185	0.6750	0.6875
Qwen2.5	0.7451	0.6981	0.7005	0.6992
RoBERTa Large	0.7671	0.7249	0.6930	0.7040
BERT Base Cased	**0.7814**	**0.7405**	**0.7410**	**0.7407**

Table 3: Relation classification performance on the test set using golden labels. Metrics include macro-averaged precision, recall, and F1.

As shown in Table 3, BERT base cased achieves the highest performance across all metrics for relation classification using golden entity labels, outperforming RoBERTa Large, ModernBERT, and Qwen2.5. This result suggests that strong performance in relation extraction does not necessarily correlate with superior entity recognition, as the best-performing model here differs from those leading in token or span classification tasks, highlighting the distinct challenges and model strengths required for each subtask.

4.2.2. *End to End*

Model / Method	Exp. Type	Accuracy	Macro pr	Macro re	Macro F1
ModernBERT	token	0.2470	0.2304	0.2253	0.2277
Roberta large	token	0.2876	0.2715	0.2439	0.2542
Roberta large	span	0.2751	0.2556	0.2413	0.2474
BERT base cased	token	**0.3213**	**0.2766**	**0.2762**	**0.2764**
BERT base cased	span	0.3005	0.2564	0.2718	0.2630
Qwen2.5	token	0.1601	0.1309	0.2119	0.1469
Majority Vote	Ensemble	**0.3262**	**0.2808**	**0.2901**	**0.2850**
XGB Vote	Ensemble	0.3199	0.2750	0.2868	0.2802
MLP Vote and Proba	Ensemble	0.1175	0.0865	0.1339	0.1051

Table 4: End-to-End Relation Classification Metrics on the Test Set

Table 4 presents performance metrics for the classification of end-to-end relationships, where the models must identify both entities and their relationships without access to gold spans. All models show a drop in performance compared to the golden label setting, reflecting the compounded difficulty of the task, since a correct relation prediction requires both the Drug and ADE entities to be correctly identified, and errors in entity detection directly suppress overall performance. Although BERT base cased achieves the strongest individual performance, ensemble methods such as majority vote and XGB voting further improve results across all metrics, highlighting the benefit of leveraging diverse model outputs to mitigate individual weaknesses. These ensembles demonstrate greater stability and robustness, especially in handling harder ADE cases, where entity-level errors otherwise propagate into relation misclassifications. Conversely, weaker ensemble strategies like MLP Vote and Proba show poor generalization, suggesting that not all aggregation methods are equally effective under noisy conditions. In general, these results underscore the critical role of accurate entity recognition in end-to-end settings and the potential of ensemble learning to improve reliability in the extraction of complex biomedical relationships.

5. Discussion

Across all tasks, models consistently struggled more with identifying ADE entities than Drug entities, with ADE recall notably lower despite strong Drug performance. This highlights the linguistic complexity and context dependence of ADE mentions. Model performance did not align with size or recency: for example, Qwen2.5 underperformed while the smaller BERT-base-cased remained competitive, suggesting that architecture and task alignment matter more than scale. RoBERTa Large emerged as the strongest overall model, particularly for ADE detection, likely due to its pretraining strategy with dynamic masking on a diverse corpus.

Token-based models generally outperformed span-based ones for ADEs, while span models remained competitive for Drug recognition where boundary precision is important. Ensemble methods, especially majority voting and XGB-based approaches, improved stability in end-to-end relation extraction by mitigating error propagation from entity recognition. The sharp performance drop in the end-to-end setting underscored the dependence of relation classification on accurate entity detection. Notably, minority-vote ensembles occasionally identified difficult ADE cases missed by the majority, emphasizing the value of model diversity. Overall, these findings show that robust entity recognition, supported by effective ensemble strategies, is key to reliable biomedical NLP systems.

6. Conclusion

This study evaluated the effectiveness of fine-tuned LLMs for ADE detection through entity recognition and relation extraction in clinical text. Using both token and span classification approaches on the n2c2 dataset, we found that models like RoBERTa Large consistently outperformed others, particularly in handling the contextual complexity of ADEs. While Drug entities were identified reliably across models, ADE detection remained more challenging, highlighting the importance of contextual modeling. Notably, our results show that entity

recognition and relation extraction success do not always align across models, underscoring the distinct demands of each task. Ensemble methods, especially majority voting and XGB-based strategies, proved to be effective in boosting end-to-end relation extraction by leveraging complementary model predictions and mitigating individual weaknesses. These findings reinforce the value of ensembling and task-specific modeling in building robust biomedical NLP systems and point toward practical strategies for improving clinical information extraction pipelines.

7. Limitations and Future Work

While RoBERTa Large showed strong performance and the study produced valuable insights, several limitations warrant consideration. Models consistently struggled to identify ADE entities, likely due to class imbalance and their linguistic variability, which was reflected in lower recall scores. In addition, architectural constraints prevented a uniform comparison across models, as some could not be evaluated in both token and span settings. Finally, the Relationship Extraction model could be optimized to leverage the best performer and not the same model making the initial entity predictions.

Future work should prioritize improving ADE recognition through methods that address class imbalance and context sensitivity. Multi-task learning and modular pipeline designs may help reduce the impact of upstream errors by coupling entity and relation prediction. Incorporating external biomedical resources for pretraining could improve model performance. Expanding the usable corpus of ADE datasets that represent real-world data would also be beneficial. In addition, adopting a two-step approach similar to strategies used in multilingual sentiment classification[26] could help cut down on over-prediction by first determining whether a sentence contains an ADE before applying fine-grained extraction.

8. Acknowledgments

This research was supported by the Office of Naval Research, Department of the Navy (FAIN N00014-22-1-2714), the Office of the Director, National Institutes of Health (NIH) Common Fund under award number 1OT2OD032581-01, and the Amazon Research Award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views or official policies of the Office of Naval Research, the U.S. Government, AIM-AHEAD, the National Institutes of Health, or Amazon.

References

1. J. R. Nebeker, P. Barach and M. H. Samore, Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting, *Annals of internal medicine* **140**, 795 (2004).
2. J. Sultana, P. Cutroneo and G. Trifirò, Clinical and economic burden of adverse drug reactions, *Journal of Pharmacology and Pharmacotherapeutics* **4**, S73 (2013).
3. K. M. Cresswell, B. Fernando, B. McKinstry and A. Sheikh, Adverse drug events in the elderly, *British medical bulletin* **83**, 259 (2007).
4. D. Balog-Way, D. Evensen, R. Löfstedt and F. Bouder, Effects of public trust on behavioural intentions in the pharmaceutical sector: data from six european countries, *Journal of Risk Research* **24**, 645 (2021).

5. K. Faasse and K. J. Petrie, The nocebo effect: patient expectations and medication side effects, *Postgraduate medical journal* **89**, 540 (2013).
6. H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius and L. Toldo, Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports, *Journal of Biomedical Informatics* **45**, 885 (2012), Text Mining and Natural Language Processing in Pharmacogenomics.
7. K. Roberts, D. Demner-Fushman and J. M. Tonning, Overview of the tac 2017 adverse reaction extraction from drug labels track. (2017).
8. S. Henry, K. Buchan, M. Filannino, A. Stubbs and O. Uzuner, 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records, *Journal of the American Medical Informatics Association* **27**, 3 (2020).
9. H. Prioleau and S. Aryal, Entity only vs. inline approaches: Evaluating llms for adverse drug event detection in clinical text (student abstract), in *Proceedings of the AAAI Conference on Artificial Intelligence*, (28)2025.
10. H. Prioleau, S. K. Aryal and L. Burge, Evaluating llama-3.1 for adverse drug event entity and relationship extraction across prompting techniques, in *International Conference on Advances in Computing Research*, 2025.
11. H.-J. Dai, C.-H. Su and C.-S. Wu, Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings, *Journal of the American Medical Informatics Association* **27**, 47 (2020).
12. J. Yang, S. C. Han, S. Long, J. Poon and G. Nenadic, Mc-dre: Multi-aspect cross integration for drug event/entity extraction, 4385 (2023).
13. M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer and O. Levy, Spanbert: Improving pre-training by representing and predicting spans, *Transactions of the association for computational linguistics* **8**, 64 (2020).
14. D. Ye, Y. Lin, P. Li and M. Sun, Packed levitated marker for entity and relation extraction, *arXiv preprint arXiv:2109.06067* (2021).
15. A. B. Chapman, K. S. Peterson, P. R. Alba, S. L. DuVall and O. V. Patterson, Hybrid system for adverse drug event detection, in *Proceedings of the 1st International Workshop on Medication and Adverse Drug Event Detection*, eds. F. Liu, A. Jagannatha and H. Yu, Proceedings of Machine Learning Research, Vol. 90 (PMLR, 04 May 2018).
16. S. Bird, E. Klein and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit* (" O'Reilly Media, Inc.", 2009).
17. Q. Team, Qwen2.5: A party of foundation models (September 2024).
18. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, *CoRR* **abs/1907.11692** (2019).
19. B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard and I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference (2024).
20. J. Devlin, M. Chang, K. Lee and K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, *CoRR* **abs/1810.04805** (2018).
21. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, Huggingface's transformers: State-of-the-art natural language processing, *arXiv preprint arXiv:1910.03771* (2019).
22. T. Aarsen, F. M. del Prado Martin, D. V. Suero and H. Oosterhuis, Spanmarker for named entity recognition (2023).
23. S. K. Aryal, H. Prioleau, G. Washington and L. Burge, Evaluating ensembled transformers for

multilingual code-switched sentiment analysis, in *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2023.
24. S. K. Aryal, U. Shah, H. Prioleau and L. Burge, Ensembling and modeling approaches for enhancing alzheimer's disease scoring and severity assessment, in *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2023.
25. S. K. Aryal and H. Prioleau, Ad-hoc ensemble approach for detecting adverse drug events in electronic health records, *Journal of Computing Sciences in Colleges* **40**, 238 (2024).
26. S. Aryal and H. Prioleau, Howard university computer science at semeval-2023 task 12: A 2-step system design for multilingual sentiment classification with language identification, in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023.