Medication Entity Extraction using Transformer Models and Mistral7B: A Comparison

Frances Cue and **Ashok Sundararaman**
University of California, Berkeley
{fcue, ashoksun01}@berkeley.edu

Abstract

Clinical documentation frequently comprises of unstructured free-text notes filled with intricate medical data. Historically, transformer-based models have demonstrated significant efficacy in extracting entities from such medical notes. This paper evaluates the performance of current BERT-based models and compares their results with the more recent large language model architectures, specifically focusing on the Mistral-7B model's ability to identify drug-related events in clinical texts.

1 Introduction

At the heart of modern healthcare lies a paradox, in which the very tools designed to safeguard patient health have become sources of confusion and error. Electronic Health Records (EHRs), designed to streamline patient care and enhance safety, have inadvertently introduced complexities to the very goals they seek to achieve (Classen et al., 2023). This is especially alarming as EHR systems have been widely embraced across the U.S. healthcare landscape with over 95% of hospitals and more than 90% of ambulatory clinics implementing these systems to replace the traditional paper-based patient record (Classen et al., 2023; Ratwani, 2017).

EHRs are digital compilations of patients' health information, accumulated across multiple healthcare settings, encompassing patient demographics, medical history, medications, lab results, and much more (Modi and Feldman, 2022). These comprehensive reports facilitate the distribution of health information, support clinical and diagnostic decision-making, provide access to patient health portals, among other features (Modi and Feldman, 2022). Central to their creation was the aim to mitigate medication safety issues - the most frequent cause of preventable harm within healthcare settings (Classen et al., 2020).

Despite these goals and advantages over paper-based documentation, the implementation of EHR systems has introduced new challenges that may increase medication errors. Since EHRs are written by healthcare professionals, they are often "filled under time pressure and with low motivation due to the fact it takes time away from actual patient care" (Pomares Quimbaya et al., 2016). As a result, EHRs often suffer from poor quality, including variability in semantics, informal sentences, missing punctuation, words, misspellings, etc (Pomares Quimbaya et al., 2016). Additionally, a major complaint of physicians is the prevalence of "note bloat", an accumulation of extraneous patient information within medical charts that can obscure crucial data and impair comprehension, potentially leading to medication errors (Nijor et al., 2022).

Medication errors, defined as preventable events that may cause or lead to inappropriate medication use or patient harm, remain a significant problem (Tariq et al., 2024). Some common examples of medication errors are prescribing the improper dose, omitting a dose, mixing medications, and administering medication at the incorrect speed (Tariq et al., 2024). These errors can lead to Adverse Drug Events (ADEs), the most common types of inpatient errors, defined as injuries from a medication or a missed or inappropriately dosed medication, resulting in 700,000 emergency department visits and 100,000 hospitalizations annually as of 2019 (Tariq et al., 2024; for Healthcare Research and , AHRQ). Despite the promise of EHRs to enhance medication safety, studies have shown that these systems have not substantially improved medication safety performance, with 33% of known serious medication errors still not prevented in 2018 (Classen et al., 2023).

Given that physicians face considerable difficulties in extracting clinically relevant information from EHRs, which may cause clinical errors due to cognitive overload, a viable solution involves consolidating key information, specifically medication information, to combat the proliferation of EHR

data and ADEs (Nijor et al., 2022). Named Entity Recognition (NER), a natural language processing technique, emerges as a tool for extracting pertinent medication-related entities from the unstructured EHR data (Durango et al., 2023). By accurately identifying and categorizing key pieces of information, NER has the potential to significantly enhance the utility of EHRs and drastically reduce ADEs from occurring.

This paper proposes an application of NER technology aimed at extracting important medication-related entities from clinical records, addressing the issue of medication errors. To achieve this, we fine-tuned three distinct pretrained transformer models for token-level classification – BERT-base-cased, Microsoft's BiomedNLP-BiomedBERT, and Clinical-Longformer. Additionally, we extend our experimentation to include the fine-tuning of the Mistral-7B-v0.1 Large Language Model (LLM) to evaluate its performance relative to BERT models, exploring whether LLMs offer a superior solution.

2 Background

There have been many experiments and studies relating to extraction of medical entities, specifically medications from medical electronic records.

Chapman (2019) created a system that identifies mentions of symptoms and drugs in clinical notes and labels the relationship between them. Their system uses existing word embeddings and conditional random fields (CRFs) to perform NER and relation extraction (RE) . The system achieved good performance in all tasks, with an F1 score of 80.9% for NER, 88.1% for RE, and 61.2% for the integrated system. Overall, the study demonstrates the effectiveness of NLP for detecting ADEs.

Christopoulou et al. (2019) and colleagues employed a sophisticated ensemble of models to enhance the extraction of relations and drug entities from clinical texts. The researchers utilized an integration of bidirectional long short-term memory (BiLSTM) networks alongside CRFs to achieve an end-to-end extraction process. Their investigation extended to analyzing relations within sentences through the application of BiLSTM-based frameworks and examining inter-sentence relations via Transformer network architectures. By adopting an ensemble methodology, they aimed to refine the efficacy of the relation extraction procedure further. The findings from their research underscored the models' adeptness in recognizing intricate interactions, particularly those spanning within and between sentences, showcasing the models' competitive edge in the realm of relation extraction.

Narayanan (2020) achieved notable outcomes in their analysis of drug-related entities within the n2c2 Medication ADE dataset. Our methodology aligns closely with their study, which innovatively implemented sentence-level data augmentation at the time of prediction and capitalized on pre-trained medical transformers. Their approach not only utilized a suite of clinically fine-tuned models but also combined them through ensemble methods to enhance performance. Building on this foundation, our research adopts clinically refined models enriched with advanced embeddings and adapted to leverage the latest advancements through large language models.

LLMs have showcased remarkable capabilities across a variety of natural language processing (NLP) tasks, setting new benchmarks for performance. However, LLMs have not yet achieved the efficacy of supervised baselines, specifically for NER tasks. This shortfall is mainly due to the fundamental differences between the two areas: NER focuses on sequence labeling, while LLMs are inherently designed for text generation. Additionally, LLMs exhibit a propensity for "hallucinations", inaccurately identifying non-entity text as relevant entities, which further decreases optimal performance for NER. To navigate these challenges, two primary strategies are employed: fine-tuning, which involves further training a pre-existing model on a specific task and in-context learning, which utilizes carefully constructed prompts and few-shot examples to guide the model towards better task-specific performance. Both methods have been proven to significantly improve model performance (Wang et al., 2023). Building upon this knowledge, numerous studies have recently emerged, dedicated to applying NER to clinical data using LLMs.

Naguib et al. (2024) assessed the performance of LLMs for few-shot clinical entity recognition across English, Spanish, and French using a combination of eight clinical and six out-of-domain gold standard corpora. The research compared the performance of ten auto-regressive language models, limited to 100 annotated sentences for few-shot learning. Notably, among their highest-performing models, LLAMA-2-70B and Mistral-7B excelled on the n2c2-2019 dataset, which includes English discharge summaries from Partners

HealthCare and Beth Israel Deaconess Medical Center. LLAMA-2-70B achieved a micro-F1 score of 0.309, closely followed by Mistral-7B with a score of 0.291, demonstrating that few-shot learning using LLMs is not yet suitable for production-level NER in the clinical domain.

Hu et al. (2024) explored the capabilities of GPT-3.5 and GPT-4 using minimal training data for two specific clinical NER tasks: extracting medical problems, treatments, and tests from the MTSamples corpus and identifying nervous system disorder-related adverse events from the Vaccine Adverse Event Reporting System (VAERS). A novel, task-specific framework was developed to enhance the performance of these models incorporating baseline prompts, annotation guideline-based prompts, error analysis instructions, and annotated samples for few-shot learning. Utilizing all four prompt components led to substantial improvements over the baseline prompts alone, with GPT-3.5 and GPT-4 achieving relaxed-F1 scores of 0.794 and 0.861 for MTSamples, and 0.676 and 0.736 for VAERS, respectively. Despite trailing behind BioClinicalBERT, with an F1 of 0.901 for the MTSamples dataset and 0.802 for the VAERS, these models show promise for future advancements in clinical text NER tasks.

3 Methodology

3.1 Dataset

The dataset, obtained from the n2c2 NLP Research Datasets available through the DBMI Portal at Harvard Medical School (University, 2018), has been extensively utilized to extract information on drugs, related drug entities, and adverse events from EHRs. The data, initially developed and manually annotated by participants in the i2b2 challenge project, is presented below, illustrating the entity distribution across training and testing corpora with relevant examples.

Table 1: Entity Distribution Across Training and Testing Corpora

Entity	Training	Testing	Example
Drug	16225	10575	Aspirin
Strength	6691	4359	500mg
Form	6651	4230	Tablet
Frequency	6281	4012	Daily
Route	5476	3513	Oral
Dosage	4221	2681	2 tablets
Reason	3855	2545	Pain relief
ADE	959	625	Nausea
Duration	592	378	7 days

3.2 Data Preprocessing

Our preprocessing approach utilizes a publicly available repository from ClinicalNLP-ADE provided by Narayanan (2020) to transform clinical data into a structured, word-level format using the ConLL BIO tagging scheme (ConLL). The BIO format, which stands for Beginning, Inside, Outside, is crucial for identifying the relational position of entities within sentences. This tagging strategy is particularly useful for accurately capturing medical information from unstructured text.

Table 2: BIO-Tag Example

Token	BIO-Tag
Prednisone	B-Drug
20	B-Dosage
mg	I-Dosage
B.I.D	B-Frequency
PO	B-Route
for	0
1	B-Duration
week	I-Duration
asthma	B-Reason

After converting the data, our initial step was to reassemble it into concise sentences using metadata such as filename and sentence number from the transformation. We intentionally avoided additional cleaning or preprocessing steps. This decision was made to preserve the authenticity of the medical data, which is often characterized by a high level of unstructured text and an abundance of medical abbreviations. For example, a common prescription notation like "Tylenol 325 mg PO BID" is typically documented in its abbreviated form rather than expanded to "Tylenol 325 milligrams by mouth twice a day." This approach ensures that our dataset closely mirrors the real-world conditions and challenges found in medical EHRs.

3.3 Data Augmentation

In our efforts to enhance the robustness of our machine learning models, we explored a variety of data augmentation strategies on a selected subset of the dataset. These included:

Synonym Augmentation: We attempted to enrich the dataset by substituting words with their synonyms. However, this technique occasionally introduced semantic discrepancies, where the substituted synonyms did not fully retain the original meaning of the sentences.

Rolling Sentence Augmentation: We implemented a method of concatenating sequences of

3-5 sentences, incrementally integrating content from preceding rows. The objective was to provide a flowing context from one sentence to the next. While conceptually promising, this approach unfortunately led to model over-fitting, indicating that the additional context was too prescriptive.

Sentence Concatenation: After evaluating the previous methods, we simplified our approach by directly concatenating more sentences within each data row. This simpler, yet effective method significantly improved the performance of our transformer-based models, striking the right balance between contextual richness and model generalizability.

3.4 Class Imbalance and Focal Loss

Our dataset suffers from class imbalance, in particular, there is an abundance of "Drug" entities and a small amount of "ADEs". Though this may be the case in the real world, we tried to remedy this by applying focal loss. Focal loss works by applying a modulating factor to the standard cross-entropy loss function. We built a custom focal loss function that takes into account weighting the classes of the minority group. Choosing focal loss improved our baseline model's ability to identify minority classes.

3.5 Transformer Models

For our baseline, we utilized the BERT-base-cased model, which recognizes casing—an essential feature in our EHR dataset. During the tokenization phase, we labeled only the first word if a word was split into subwords, simplifying the model's learning process. We fine-tuned the model with hyperparameters listed in Appendix A. We did not do hyperparameter tuning, and focused mostly on how to improve the quality of the data since the baseline result was already high. We evaluated performance on both short and long sentences, finding that longer sentences yielded better results. Additionally, class imbalance influenced our outcomes, prompting the use of Focal Loss, as detailed in the Class Imbalance section.

We opted for Microsoft's BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext (we will call it BBB for short), trained on a substantial corpus of PubMed abstracts and full texts. This domain-specific model was chosen for its relevance to the biomedical field, and its training implementation follows that of the standard BERT-base-uncased model.

EHRs are often unstructured and lengthy. The LongFormer, capable of processing extended sequences, uses a unique global attention mechanism. This feature allows designated global tokens to attend to all other tokens in the sequence. Concurrently, local tokens employ windowed attention to concentrate on nearby words, enhancing efficiency for long texts (Li et al., 2022). We adapted the Clinical-LongFormer, pretrained on extensive medical documents, to our needs. We passed one EHR record per row, instead of breaking it apart into sentences. The max sequence length was set to 4096. We experimented with and without global attention. Setting the global attention mask to 1 for the initial token significantly improved the model's ability to identify ADEs. Further attempts to enhance performance by filtering documents containing ADEs did not yield better results.

3.6 Data Preprocessing Mistral-7B

Our preprocessing approach for the Mistral-7B model diverges from that used for BERT models, as we no longer require the BIO tagging scheme. To accommodate the requirements of the LLM, we transform our initial dataset, originally in CoNLL format, into JSON. This results in a JSONL file where each line is a distinct JSON object. These objects represent individual sentences, including its full text and a dictionary of identified entities, categorized by entity type. This format optimizes our dataset for directly fine-tuning the Mistral-7B model.

Listing 1: JSON representation of Mistral-7B Input Format

```
{
  "text": "The patient is
     prescribed 20 mg of
     Prednisone B.I.D PO x 1 week
      for asthma.",
  "entities": {
    "Drug": ["Prednisone"],
    "Duration": ["1 week"],
    "Dosage": ["20 mg"],
    "Frequency": ["B.I.D"],
    "Strength": [],
    "Form": [],
    "Route": ["PO"],
    "Reason": ["asthma"],
    "ADE": []
  }
}
```

Table 3: Performance Metrics Comparison Across Transformer Models

Entity	BERT-base-cased			Microsoft BBB			Clinical-Longformer		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
ADE	0.45	0.39	0.42	0.43	0.53	0.48	0.54	0.54	**0.54**
Dosage	0.87	0.91	0.89	0.89	0.93	**0.91**	0.89	0.88	0.89
Drug	0.88	0.94	0.91	0.91	0.94	**0.92**	0.93	0.91	**0.92**
Duration	0.55	0.68	0.61	0.62	0.74	**0.68**	0.67	0.68	**0.68**
Form	0.94	0.91	**0.92**	0.93	0.92	**0.92**	0.92	0.91	0.91
Frequency	0.78	0.82	0.80	0.83	0.85	**0.84**	0.80	0.81	0.80
Reason	0.47	0.64	0.54	0.54	0.66	0.59	0.62	0.58	**0.60**
Route	0.94	0.94	**0.94**	0.93	0.93	0.93	0.93	0.92	0.92
Strength	0.94	0.96	**0.95**	0.95	0.96	**0.95**	0.93	0.94	0.93

3.7 Large Language Models

For our experiment, we chose Mistral-7B as our LLM, given its superior performance over the previously acclaimed Llama 2 across all tasks and tested benchmarks (Jiang et al., 2023). Mistral-7B employs advanced mechanisms like group-query-attention (GQA) and sliding window attention (SWA), enhancing inference speed while minimizing memory usage during decoding (Jiang et al., 2023). Additionally, SWA's capability to efficiently handle longer sequences offers significant computational savings (Jiang et al., 2023). To evaluate Mistral-7B's capabilities for our task, we experimented with three distinct configurations: the Mistral-7B model in a few shot learning setting, a zero-shot learning setting with fine-tuning, and a combined approach of few-shot learning with fine-tuning. Our instruction to the model is as follows:

> Prompt
>
> "Extract the entities for the specified labels from the given medical text and provide the results in JSON format - Entities must be extracted precisely as they appear in the text. - Return each entity under its label without creating new labels. - Provide a list of entities for each label. If no entities are found for a label, return an empty list. - Prioritized accuracy and relevance in the identification of entities.
>
> Here are the entity labels and their descriptions: 1. Drug: Extract any mentioned medications or drugs. 2. Duration: Extract the duration of treatment or medication usage. 3. Dosage: Extract dosages related to medications, including units. 4. Frequency: Extract how often the medication or treatment is to be taken or administered. 5. Strength: Extract the concentration or potency of the medication. 6. Form: Extract the form in which the medication is to be used. 7. Route: Extract the method of administration for a medication. 8. Reason: Extract the reason or condition the medication is prescribed for. 9. ADE: Extract adverse drug events or side effects mentioned.
>
> Make sure to go through the text carefully and extract all entities mentioned above if they are present. Do not create fictitious data."

4 Results

4.1 Transformer Models

BERT-base-cased served as a satisfactory baseline, achieving an F1 score of 88% with longer sentences. However, due to class imbalance, the performance on the minority class was notably lower, with an F1 score of only 38% for ADE. By employing focal loss, we improved ADE detection to 42%, though the overall F1 score slightly decreased to 86%.

Microsoft's BBB model demonstrated strong performance, achieving an 88% F1 score with the use of both longer sentences and focal loss. This model efficiently handled the challenges posed by class imbalances.

The Clinical-Longformer, set with a global attention mask and without additional data cleaning or preprocessing, also showed high performance. This model was notably easy to implement and most effective in detecting ADEs, demonstrating its suitability for handling complex EHR data.

Overall, Microsoft's BBB appears to have the best performance across most entity types, achieving the highest F1 score for Dosage, Drug, Duration, Form, Frequency, Reason, and Strength. Clinical-Longformer achieved the highest F1 score for ADE. BERT-base-cased had the highest F1 score for Route entity. Detailed scores are outlined in Table 3 and 4 for each entity and model.

Model	Precision	Recall	Micro Avg F1
BERT-base-cased	0.84	0.88	0.86
MicrosoftBBB	0.89	0.86	0.88
Clinical-Longformer	0.87	0.86	0.87

Table 4: Final Results of Transformer Models

4.2 Mistral-7B Model

The outputs from the Mistral-7B models often included additional text beyond the expected JSON

Table 5: Model Performance on Entire Test Set and Without Empty Predictions

	Full Test Set			Without Empty Predictions		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Mistral-7B (Few-Shot)	0.17	0.18	0.15	0.44	0.47	0.41
Mistral-7B (Zero-Shot, Fine-Tuned)	**0.44**	**0.46**	**0.44**	**0.66**	**0.68**	**0.66**
Mistral-7B (Few-Shot, Fine-Tuned)	0.36	0.37	0.34	0.45	0.45	0.42

object containing the extracted entities. To address this, we developed a function to isolate and store only the first JSON object outputted, which we assigned as the final prediction. While analyzing the model outputs, we noticed numerous instances where no predictions were generated for the provided input text. This ambiguity in the LLM's output, distinguishing between a genuine absence of a prediction due to "hallucinations" and a prediction devoid of a JSON object, posed interpretive challenges. To navigate this complexity, we implemented two distinct evaluation pipelines: one retaining all model-generated predictions and another excluding rows lacking a prediction.

The results presented in Table 5 highlights a significant discrepancy in model performance based on whether rows lacking predictions are included or excluded from the evaluation. This difference aligns with expectations, as the presence of empty predictions impact evaluation metrics. It is important to note our assessment criteria focuses on achieving an exact match between the model's predictions and the true values. Upon close analysis of the results, we observed instances where the LLM identified semantically equivalent words for entities rather than extracting the exact terms used in the input text. For instance, the model interpreted "fevers at home of 103.8" capturing "fever" as a "Reason" entity, illustrating its tendency to generalize specific inputs to broader concepts.

Our baseline Mistral-7B, utilizing few-shot prompting, exhibited subpar performance with an F1 score of 0.15 for the entire test set and 0.41 when excluding rows lacking predictions. There are several potential factors contributing to these results. The quality and relevance of the examples significantly influence LLM performance, as inadequate examples may mislead the model (Zer). Furthermore, Mistral-7B's limited context size restricts the variety and number of examples, impeding the model's ability to generalize effectively and potentially causing over-fitting, thus negatively affecting performance (Zer).

When fine-tuned on clinical data and employing zero-shot prompting, Mistral-7B significantly improved, achieving an F1 score of 0.44 across the entire test set and 0.66 after omitting rows without predictions. This improvement is largely due to the model's domain-specific training, which enhances its understanding and accuracy in generating contextually relevant predictions (Souai). Fine-tuning facilitates pattern recognition and comprehension of clinical text nuances, improving generalization to unseen data (Souai). Although this is a significant improvement over the baseline, it still falls short of BERT-based models' performance.

In conclusion, our experiments indicate that Mistral-7B achieves its optimal performance for our task when it is fine-tuned on clinical data and operates under zero-shot prompting. Although the performance is not yet production-ready or on par with BERT models, it highlights its potential for future use.

4.3 Error Analysis

Extracting adverse events due to drugs have been particularly difficult for the Transformer models. An example would be:

> *Patient had been given a script for Lisinopril. He started to feel his lips , tongue and face swelling and it progressively worsened to include his throat.*

All three models failed to detect the ADE, despite proximity to the drug that caused the reaction.

Mistral-7B had a tendency to misclassify the "Reason" entity as "ADE". In instances such as the sentence,

> *"It was felt that the patient's seizures were caused by the combination of Ritalin and thalidomide"*

the LLM incorrectly identified "seizures" as a "Reason" for medication rather than an "ADE".

Mistral-7B also has difficulty distinguishing between medication jargon and blood products as

Drugs. Unlike the Transformer models, which correctly classify “PRBCs” (Packed Red Blood Cells) and “pressors”, short for Vasopressors, as "Drug" entities, Mistral-7B overlooks these terms, indicating a need for more medical data to improve recognition of medication terminology for downstream tasks.

In addition, Mistral-7B encountered challenges distinguishing between “Dosage” and “Strength” entities, oftentimes getting the two confused for one another. For instance, in the sentence,

> *Aspirin 325 mg Tablet Sig : One (1) Tablet PO DAILY (Daily)*

"325 mg" was misclassified both as a “Dosage” and a “Strength”. Contextually, the two entities are similar, so this confusion can also occur in human analysis, indicating a need for additional training data that more clearly differentiates one from the other

5 Limitations

The dataset employed in this study originates from a challenge conducted by i2b2 and relies on human annotations, which inherently carry the risk of errors. Our exploratory data analysis (EDA) revealed instances of potential oversights, such as the antibiotic “Cefepime” being overlooked on three occasions. While a comprehensive review of all annotations was beyond the scope of our current work, further scrutiny is essential to ensure data integrity

Additionally, our investigation was limited to a handful of transformer models and only one LLM. It’s conceivable that there exist alternative models that could outperform those we evaluated. Moreover, Mistral-7B has shown tendencies to produce hallucinated data; hence, human oversight and having a human in the loop, remains indispensable when implementing LLMs.

Future research should focus on evaluating a broader array of LLMs to establish a more comprehensive understanding of their baseline capabilities. Such investigations are crucial for discerning the variations in performance across different architectures and training paradigms. Additionally, the practice of prompt engineering, which is critical for optimizing LLM outputs, merits deeper exploration. This involves refining the art of prompt design to not only enhance model performance but also to fully comprehend the implications of prompt-based fine-tuning on LLMs behavior. Continued efforts in this area are essential for advancing the mastery of effective LLM utilization and achieving more reliable and nuanced interactions with these models

6 Conclusion

Historically, transformer models have demonstrated their excellence in NER tasks. As LLMs advance, they signal a shift toward a new standard in NER. It is imperative to deepen our understanding of LLMs and refine their capabilities for entity identification, as they are poised to redefine the benchmarks in this domain.

References

Zeroshot and few-shot prompting | prompt engineering | generative ai | hyperskill. (n.d.).

Peters Cohan Beltagy, Matthew. 2020. Longformer: The long-document transformer.

Peterson K. S. Alba P. R. DuVall S. L. Patterson O. V. Chapman, A. B. 2019. Detecting adverse drug events with rapidly trained classification models. *Drug Safety*, 42(1):147–156.

Fenia Christopoulou, Thy Thy Tran, Sunil Kumar Sahu, Makoto Miwa, and Sophia Ananiadou. 2019. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, 27(1):39–46.

David C. Classen, A. Jay Holmgren, Zoe Co, Lisa P. Newmark, Diane Seger, Melissa Danforth, and David W. Bates. 2020. National Trends in the Safety Performance of Electronic Health Record Systems From 2009 to 2018. *JAMA Network Open*, 3(5):e205547–e205547.

David C. Classen, Christopher A. Longhurst, Taylor Davis, Julia Adler Milstein, and David W. Bates. 2023. Inpatient EHR User Experience and Hospital EHR Safety Performance. *JAMA Network Open*, 6(9):e2333152–e2333152.

ConLL. Conll-u format - universal dependencies.

María C. Durango, Ever A. Torres-Silva, and Andrés Orozco-Duque. 2023. Named entity recognition in electronic health records: A methodological review. *Healthcare Informatics Research*, 29(4):286–300.

Agency for Healthcare Research and Quality (AHRQ). 2022. Medication errors and adverse drug events.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2019. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences.

S. Modi and S. S. Feldman. 2022. The value of electronic health records since the health information technology for economic and clinical health act: Systematic review. *JMIR medical informatics*, 10(9).

Marco Naguib, Xavier Tannier, and Aurélie Névéol. 2024. Few shot clinical entity recognition in three languages: Masked language models outperform llm prompting.

Rajan Rangan Narayanan, Mannam. 2020. Evaluation of transfer learning for adverse drug event (ADE) and medication entity extraction.

S. Nijor, G. Rallis, N. Lad, and E. Gokcen. 2022. Patient safety issues from information overload in electronic medical records. *Journal of patient safety*, 18(6).

Branislav Pecher, Ivan Srba, and Maria Bielikova. 2024. Fine-tuning, prompting, in-context learning and instruction-tuning: How many labelled samples do we need? *cs.CL*.

Alexandra Pomares Quimbaya, Laura Montes-Escudero, Ana Maria Suarez-Cetrulo, and Carmenza Moreno-Tanguis. 2016. Named entity recognition for electronic health records using conditional random fields. *Procedia Computer Science*, 100:55–61.

Ratwani. 2017. Electronic health records and improved patient care: Opportunities for applied psychology. *Current directions in psychological science*, 26(4):359–365.

Wiem Souai. Fine-tuning mistral 7b for named entity recognition (ner).

Rayhan A. Tariq, Rishik Vashisht, Ankur Sinha, and et al. 2024. Medication dispensing errors and prevention. *StatPearls [Internet]*.

Harvard University. 2018. 2018 challenge | national nlp clinical challenges (n2c2).

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models.

A Appendix

BERT-base-cased Hyperparameters

Parameter	Value
MAX_LEN	512
TRAIN_BATCH_SIZE	4
VALID_BATCH_SIZE	2
TEST_BATCH_SIZE	32
EPOCHS	3
LEARNING_RATE	1e-05
MAX_GRAD_NORM	10

Clinical-Longformer Hyperparameters

Parameter	Value
MAX_LEN	4096
TRAIN_BATCH_SIZE	1
VALID_BATCH_SIZE	1
TEST_BATCH_SIZE	2
EPOCHS	10
LEARNING_RATE	1e-05
MAX_GRAD_NORM	10