# Structured LLM Augmentation for Clinical Information Extraction

Ying Wei[a,b] , Qi Li[a], and Jay Pillai[b]
[a] *Iowa State University*
[b] *Truveta*
ORCiD ID: Ying Wei https://orcid.org/0000-0001-7403-3495
ORCiD ID: Qi Li https://orcid.org/0000-0002-3136-2157

**Abstract.** Information extraction tasks, such as Named Entity Recognition (NER) and Relation Extraction (RE), are essential for advancing clinical research and applications. However, these tasks are hindered by the scarcity of labeled clinical documents due to privacy concerns and high annotation costs. This study introduces a novel framework combining Large Language Models (LLMs) for data augmentation with an adapted BERT model for clinical information extraction. The framework encodes entity and relational information within clinical note segments, enabling LLMs to generate diverse and contextually accurate augmentations while preserving structural integrity. Augmented data is used to train a segmentation-based BERT model, overcoming sequence length limitations and integrating global context via BiLSTM. Evaluations on public and proprietary datasets demonstrate significant performance improvements, highlighting the approach's potential to address data scarcity in clinical information extraction tasks.

**Keywords.** Information extraction, Large Language Model, Clinical informatics

## 1. Introduction

Information extraction tasks like Named Entity Recognition (NER) and Relation Extraction (RE) are critical in the clinical domain, serving as the foundation for many downstream applications [1]. However, the scarcity of labeled clinical data—due to privacy constraints and annotation costs—hinders progress. Data augmentation has emerged as a practical solution to address this challenge by generating synthetic data to enhance model training and performance. While Large Language Models (LLMs) offer a powerful tool for generating diverse and contextually rich synthetic data [2], their application to clinical data augmentation poses unique challenges [3]. These include handling domain-specific terminology, ensuring consistency in entity relationships, and effectively processing lengthy clinical notes. Addressing these challenges is essential to unlocking LLMs' full potential for clinical information extraction tasks [4].

This paper presents a novel framework combining an LLM-based data augmentation strategy with a modified BERT architecture to address NER and RE challenges in the clinical domain. The LLM generates diverse, contextually accurate synthetic data while preserving clinical relationships. To handle lengthy clinical documents, a segmentation-based approach with BiLSTM integration restores global context. The method achieves significant performance improvements on public and proprietary datasets, demonstrating its effectiveness in diverse clinical scenarios.

## 2. Methods

This study introduces an LLM-based augmenter for generating synthetic training data to enhance clinical information extraction tasks. Additionally, a BERT model is adapted to handle lengthy clinical notes for NER and RE.

### *2.1.* LLM-Augmenter

Since LLMs struggle to process lengthy clinical documents directly, we focus on processing note segments, defined as consecutive sentences containing a target entity or relation. After using the LLM to rewrite the target segment, we concatenate the original note content preceding and following the segment to form a new, augmented clinical note. Given a clinical note and its annotations, an LLM is employed to augment it through structured rewriting. As illustrated in Figure 1, this process consists of following steps:

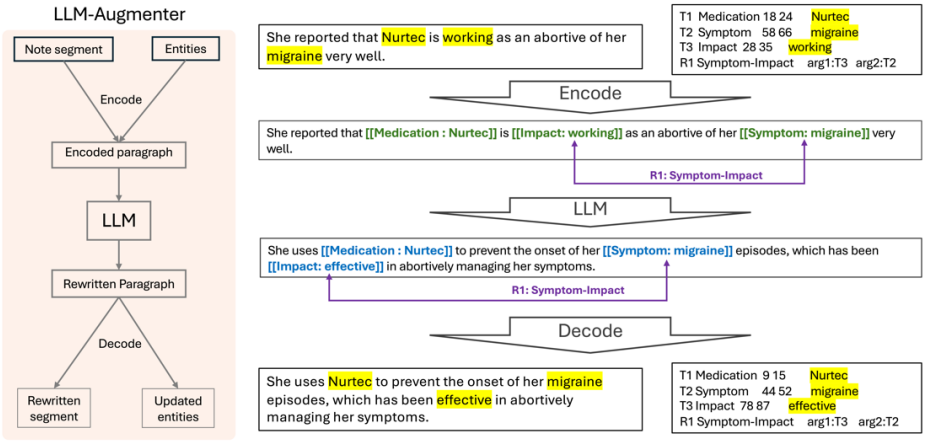


**Figure 1.** Illustration of augmentation process of LLM-Augmenter.

**Input Preparation***:* Each note segment and its corresponding annotations in BRAT format serve as inputs. For the note segment: "*She reported that Nurtec is working as an abortive of her migraine very well*," with three entities and one relation: Entity **T1**: Medication → Nurtec*,* Entity **T2**: Symptom → migraine*,* Entity **T3**: Impact → working*.* Relation: Symptom-Impact between Symptom (T2) and Impact (T3).

**Entity Encoding***:* The note segment is transformed to embed entity information directly within the text using double brackets, entity labels, and spans. For example, the encoded version of the segment is: "*She reported that [[Medication: Nurtec]] is [[Impact: working]] as an abortive of her [[Symptom: migraine]] very well.*"

**Augmentation with LLM***:* The encoded segment is fed into the LLM, accompanied by a prompt instructing it to rewrite the text while retaining the structural integrity of the entity labels and delimiters. The LLM is allowed to alter entity terms to introduce expression diversity while preserving contextual relationships. For example, the LLM output might be: "She uses [[Medication: Nurtec]] to prevent the onset of her [[Symptom: migraine]] episodes, which has been [[Impact: effective]] in aborting her symptoms."

**Output Decoding**: The rewritten segment is decoded to generate a new note segment and the updated annotations, incorporating the modified entities and relationships from the LLM output.

### *2.2.* BERT for Clinical Information Extraction

In this work, we use BERT [5] for NER and RE tasks, which has demonstrated exceptional performance in a variety of natural language processing tasks, including NER and RE, due to its ability to capture rich contextual representations. However, its maximum sequence length limitation (typically 512 tokens) poses a significant challenge [6] when applied directly to lengthy clinical notes, which often span thousands of tokens. Truncating or omitting parts of the text can lose critical contextual information.

To address this challenge, we propose a segmentation-based approach combined with a fusion mechanism. Clinical notes are divided into manageable segments, ensuring compatibility with BERT's sequence length constraints. Each segment is independently processed using a pre-trained BlueBERT-large model [7], fine-tuned on PubMed abstracts and the MIMIC-III clinical corpus [8], to generate token-level contextual embeddings. To restore the global context across segments, a BiLSTM [9] layer is employed to capture dependencies between them. The fused representations are then passed through two Multi-Layer Perceptrons (MLPs): one for predicting NER labels to identify entities and the other for predicting RE labels to capture relationships between entities. This approach harnesses BERT's strengths while mitigating its limitations, enabling robust and scalable information extraction from lengthy clinical notes.

### *2.3. Dataset*

The proposed approach is evaluated using three datasets: the publicly available i2b2-2012 and N2C2-2018 Track 2 datasets, as well as a proprietary dataset from Truveta.

- i2b2-2012 Dataset [10]: This dataset contains 310 de-identified clinical notes, split into 190 for training and 120 for testing. It is annotated with six entity types and serves as a benchmark for clinical information extraction tasks.
- N2C2-2018 Track 2 Dataset [11]: This dataset focuses on patient-level clinical information extraction and is well-suited for evaluating NER and RE tasks. It includes 505 de-identified discharge summaries annotated with nine entity types and eight relation types, divided into 242 summaries for training, 61 for development, and 202 for testing.
- Proprietary Dataset: To address the size limitations of the i2b2 and N2C2 datasets, we utilize a proprietary dataset comprising 2,306 de-identified clinical notes curated from real-world applications within Truveta. These notes are split into 1,679 for training, 313 for development, and 314 for testing. The dataset includes annotations for 29 entity types and 17 relation types, providing a comprehensive representation of real-world clinical scenarios

To enhance training data diversity, we augment 230 notes for the i2b2 dataset, 317 notes for the N2C2 dataset, and 420 notes for the proprietary dataset. These datasets provide a robust evaluation on benchmark datasets and real-world clinical applications.

## 3. Results

To evaluate the effectiveness of the proposed method, we use Precision, Recall, and F1-Score, which are standard metrics for information extraction tasks. These metrics are

computed under two evaluation settings: **Strict** and **Lenient**. In the Strict setting, an entity or relation is considered correct only if all aspects—type, span, and boundaries—match exactly. The Lenient setting relaxes these criteria by allowing partial matches, such as correct type recognition even if the boundaries are not exact. These metrics are applied to both NER and RE tasks to ensure a thorough evaluation.

**Table 1.** Evaluation results on the i2b2, N2C2 and proprietary datasets.

| | | NER (Strict) | | | RE (Strict) | | | NER (Lenient) | | | RE (Lenient) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Prec.** | **Recall** | **F1** | **Prec.** | **Recall** | **F1** | **Prec.** | **Recall** | **F1** | **Prec.** | **Recall** | **F1** |
| **i2b2** | No Aug | 81.31 | 81.10 | 81.20 | - | - | - | 88.00 | 87.13 | 87.56 | - | - | - |
| | With Aug | 83.39 | 80.27 | 81.80 | - | - | - | 89.72 | 85.91 | 87.77 | - | - | - |
| **N2C2** | No Aug | 89.40 | 89.61 | 89.50 | 82.65 | 82.99 | 82.82 | 95.14 | 91.07 | 93.06 | 90.42 | 90.36 | 90.39 |
| | With Aug | 91.82 | 88.16 | 89.95 | 87.62 | 80.94 | 84.15 | 96.04 | 92.10 | 94.03 | 94.58 | 87.24 | 90.76 |
| **Proprietary** | No Aug | 86.17 | 86.28 | 86.22 | 76.33 | 77.16 | 76.68 | 88.93 | 90.09 | 89.51 | 81.18 | 80.53 | 80.85 |
| | With Aug | 87.74 | 87.18 | 87.46 | 78.07 | 77.23 | 77.65 | 90.94 | 90.16 | 90.55 | 83.30 | 82.13 | 82.71 |

Table 1 summarizes the results on the i2b2, N2C2 and proprietary datasets for NER and RE tasks. Results are compared between models trained without and with LLM-based data augmentation. On the i2b2 dataset, the augmentation approach shows notable improvements for NER. Under strict evaluation, the F1 score increases from 81.20% without augmentation to 81.80% with augmentation. Similarly, in lenient evaluation, the F1 score improves from 87.56% to 87.77%. For the N2C2 dataset, in strict NER, the model with augmentation achieved an F1 score of 89.95%, compared to 89.50% without augmentation, demonstrating a slight improvement in precision and a minor decrease in recall. For strict RE, the augmented model achieved an F1 score of 84.15%, outperforming the non-augmented model. Under lenient evaluation, the NER F1 score improved from 93.06% to 94.03%, while the RE F1 score increased from 90.39% to 90.76% with augmentation. This demonstrates the ability of the LLM-Augmenter in smaller datasets. For the proprietary dataset, in strict NER, the augmented model achieved an F1 score of 87.46%, compared to 86.22% without augmentation. For strict RE, the augmented model scored 77.65%, slightly higher than 76.68%. Under lenient evaluation, the NER F1 score rose from 89.51% to 90.55%, and the RE F1 score improved from 80.85% to 82.71% with augmentation.

## 4. Discussion

The results highlight the benefits of using LLM-based data augmentation for clinical information extraction tasks, particularly in low-resource settings. Across both datasets, models trained with augmented data consistently outperformed those trained without augmentation, with notable improvements in F1 scores for both NER and RE tasks. For the i2b2 and N2C2 datasets, the augmented model exhibited substantial gains in RE performance under both strict and lenient settings, demonstrating the value of synthetic data in capturing complex entity relationships. The slight decrease in recall for strict NER suggests that while augmented data increases precision, it may introduce variability that impacts sensitivity. For the proprietary dataset, the improvements were more pronounced under lenient evaluation, suggesting that the diversity of augmented data effectively complements the larger dataset size. However, the relatively modest gains under strict evaluation indicate that further refinements to the augmentation process could enhance alignment with real-world data distributions. Overall, these findings validate the efficacy of LLM-based augmentation in enhancing the robustness and generalizability of clinical

information extraction models, particularly for tasks with limited annotated data. Future work could focus on augmentation strategies to balance precision and recall effectively.

## 5. Conclusions

This study presents a novel LLM-based data augmentation approach to enhance information extraction tasks in the clinical domain. By addressing data scarcity and leveraging structured rewriting, the method generates diverse and contextually accurate synthetic data via LLM. Combined with a BERT-based model that overcomes sequence length limitations, the approach demonstrates significant performance gains on both public datasets and a proprietary dataset.

## Acknowledgements

This work is partially supported by the National Science Foundation under Grant No. 2152117 and NSF-CAREER 2237831. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## References

[1] N. Perera, M. Dehmer and F. Emmert-Streib, "Named entity recognition and relation detection for biomedical information extraction," in Frontiers in cell and developmental biology, 2020.
[2] Z. Meng, T. Liu, H. Zhang, K. Feng and P. Zhao, "CEAN: Contrastive Event Aggregation Network with LLM-based Augmentation for Event Extraction," in Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, 2024.
[3] M. Zhang, G. Jiang, S. Liu, J. Chen and M. Zhang, "LLM–assisted data augmentation for chinese dialogue–level dependency parsing," in Computational Linguistics, 2024.
[4] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson and A. Jain, "Structured information extraction from scientific text with large language models," in Nature Communications, 2024.
[5] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," in arXiv preprint, 2018.
[6] I. Beltagy, M. E. Peters and A. Cohan, "Longformer: The long-document transformer," in arXiv preprint arXiv:2004.05150, 2020.
[7] Y. Peng, S. Yan and Z. Lu, "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets," 2019.
[8] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi and R. G. Mark, "MIMIC-III, a freely accessible critical care database," in Scientific data, 2016.
[9] S. Siami-Namini, N. Tavakoli and A. S. Namin, "The performance of LSTM and BiLSTM in forecasting time series," in 2019 IEEE International conference on big data (Big Data), 2019.
[10] 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," in Journal of the American Medical Informatics Association, 2020.
[11] 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," in Journal of the American Medical Informatics Association, 2020.