BioSynNER: Synthetic Data Generation for Biomedical Named Entity Recognition

Anonymous ACL submission

Abstract

Named Entity Recognition (NER) is a crucial task in Natural Language Processing (NLP); however, achieving high NER performance in the biomedical domain remains a challenge due to the limited availability of annotated data. To tackle this issue, we propose a novel approach, BioSynNER, which utilizes synthetic data generation through large language models (LLMs). BioSynNER begins by mining key domain-specific attributes from seed sentences, which are then used to generate highly effective synthetic examples. Interestingly, we find that paraphrasing these seed sentences is more effective than generating data from scratch, as it preserves contextual and structural nuances that enhance Biomedical NER performance. Additionally, BioSynNER integrates the Unified Medical Language System (UMLS), a comprehensive yet noisy medical knowledge base, to address the complexity and diversity of biomedical entity types. This combined approach not only improves NER accuracy in biomedical texts but also provides a scalable framework for synthetic data generation applicable to other specialized domains. Experimental results confirm the effectiveness of BioSynNER, highlighting its potential to advance NER tasks significantly.

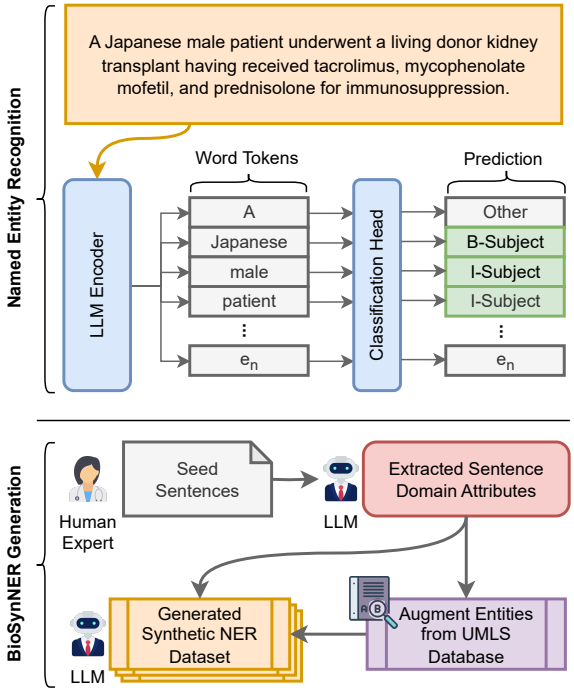


Figure 1: This figure illustrates the token-level Named Entity Recognition (NER) problem. Sentence Input: A sample medical sentence is provided for analysis. BioSynNER is an augmentation that augments the training datasets for this task.

1 Introduction

Named Entity Recognition (NER), a cornerstone task in Natural Language Processing (NLP), plays a vital role in various applications such as information extraction, machine translation, and question-answering systems. Despite its significance, achieving high performance in NER remains challenging due to the scarcity of annotated data, especially in specialized domains like medical text analysis.

One promising solution to overcome this issue is data generation, which can generate diverse and rich training data. Recent studies have shown the potential of LLMs in paraphrasing tasks, and their application in data synthesis has gained considerable attention. However, the application of LLM paraphrasing in enhancing NER performance, particularly in the context of biomedical text analysis, remains largely unexplored.

Moreover, the medical domain presents unique challenges for NER due to the complexity and variety of entity types. General domain LLMs do not have the expertise of highly specific biomedical knowledge. Thus, many NLP tasks make use of comprehensive medical terminologies like the Unified Medical Language System (UMLS), a common external knowledge base maintained by the National Institute of Health. This is a unique, highly comprehensive, and noisy knowledge, and its integration alone leads to additional challenges as well as benefits.

This paper aims to address these gaps by proposing a novel approach, BioSynNER, to improve token-level NER classification through LLM generation of existing NER datasets using UMLS, We believe that BioSynNER not only showcases how strong NER performance on highly specific biomedical domains but also demonstrates a potential framework of synthetic data generation to any specialized domain. We present experimental results to validate our approach and discuss potential directions for future research.

To summarize, our contributions are the following:

- **Novel NER Data Generation**: The paper introduces a novel approach (Figure 1) to enhance NER performance in specialized domains like biomedical text analysis through BioSynNER for data augmentation, particularly a 2-stage domain attribute and paraphrasing approach.
- **Biomedical-Specific Considerations**: We highlight the use of the Unified Medical Language System (UMLS), a comprehensive yet noisy medical knowledge base, integrated with NER to handle the complexity and variety of biomedical entity types, addressing unique challenges in the medical domain. Furthermore, we explore additional differences between general NER vs. biomedical NER dataset synthesis.
- **Empirical Validation**: BioSynNER quickly reaches close to the original performance, starting exponentially from 10 sentences. Models reach an average of 65.62% of the performance of being trained on the original data with only 20 seed sentences per label and 73.52% with 50 seed sentences per label.

2 Related Work

Previous work has attempted to generate NER datasets for general domain (Zhou et al., 2023; Heng et al., 2024)

For example, targeted distillation with mission-focused instruction tuning has been introduced to train student models for specific application domains. For instance, UniversalNER (Zhou et al., 2023), a model distilled from ChatGPT, demonstrates impressive performance in named entity recognition (NER) across 43 datasets. However, this model is still quite large, being based on llama-7b (Touvron et al., 2023) models, whereas the largest model we consider is DeBERTAa-V3 (He et al., 2021a) models less than 304M parameters. Furthermore, we demonstrate that for some datasets (Wang et al., 2023), there is no guarantee that a general LLM may be able to capture the specific annotator-specific tendencies.

ProgGen (Heng et al., 2024) is perhaps the most relevant work to ours, as we take inspiration from the domain-attribute-driven generation of the synthetic dataset. However, there are a few key differences between their work and ours. First, we demonstrate that simply generating sentences from scratch does not work well in highly specific domains like biomedical NER. Furthermore, we also augment with UMLS, an external dataset, whereas ProgGen solely utilizes GPT-4o.

Tangentially, related work shows that augmenting LLMs with definitions for biomedical NER is useful (Munnangi et al., 2024). However, we are able to avoid additional prompt complications through our approach generation of new data, which can be done without specific definitions (e.g. replacing one drug with another can be a direct find-and-replace operation). Furthermore, prompting LLMs for NER inference is still quite expensive compared to training a smaller encoder LLM.

3 BioSynNER

We introduce our approach leveraging LLM-based data augmentation and UMLS-driven entity augmentation to enhance model performance in biomedicine. The following sections will explore the explain the overall generation process as well as the UMLS entity augmentation.

Note that we utilize the Translation between Augmented Natural Languages (TANL) format (Paolini et al., 2021), as it seems to be a reasonable and interpretable format. All seed sentences and synthetic sentences will be structured in this format and will be converted to token-level metrics for encoder LLM training (and evaluation) using the IOB2 format (Sang and Veenstra, 1999).

3.1 Synthetic Data Augmentation

To generate synthetic data for biomedical Named Entity Recognition (NER) classification using the provided framework, we start with a structured approach. First, we determine relevant attributes that are crucial for data generation given a set of seed sentences. This is provided by the human expert as seen in Figure 2. An LLM is then prompted to extract the following 6 main domain attributes based on the seed sentences. Note that we perform

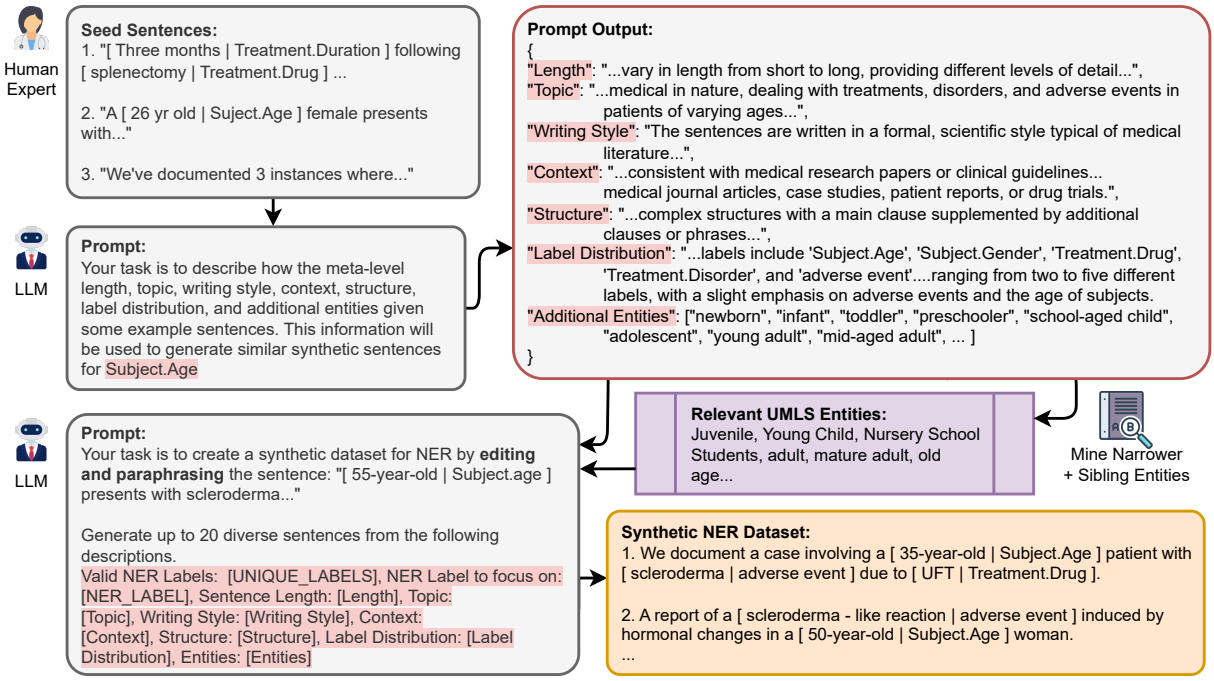


Figure 2: This figure illustrates the workflow for `BioSynNER`. Specifically, it demonstrates an example of how a biomedical NER data point is generated. First, human experts provide seed sentences, which are then analyzed by LLMs to extract meta-level domain attributes such as length, topic, writing style, and label distribution. These attributes guide the generation of synthetic sentences. UMLS entities are mined to diversify the generated sentences. Finally, an LLM is prompted to paraphrase the seed sentences to create the final synthetic dataset output.

Algorithm 1 `BioSynNER` Synthetic Dataset Generation Framework. Comments are shown in *# italics*. This is a formalized version of Figure 2.

Inputs: Initial set of seed sentences $\boldsymbol{S}$, number of sentences to generate per seed sentence N_{iters}
Outputs: Set of synthetic sentences $\tilde{\boldsymbol{S}}$

```
D ← [ ]
for S ∈ S do
    # Obtain meta-level sentence domain info
    D ← prompt_get_domain(S)
    D ← D + [D]
S̃ ← [ ]
for S ∈ S do
    for i ∈ N_iters do
        # Sample meta info from mined domains
        D̃ ~ D
        # Augment entities from UMLS database
        D̃_e ← get_UMLS_entities(D̃_e)
        # Paraphrase sentence S
        S̃ ← prompt_gen_sentence(D̃, S)
        S̃ ← S̃ + [S̃]
return S̃
```

this domain extract on multiple batches. I.e. if we are given 50 seed sentences, we might perform this extraction 50/batch_size times (in our experiments, we find that batch_size=3 is a reasonable choice). This repeated domain extraction is vital to improve the diversity and coverage of the domain attributes over the real dataset.

In total, we consider 6 main domain attributes:

1. **Length:** The length of the sentences, the amount of detail, etc
2. **Topic:** The main style of the seed sentences, such as formal, clinical, or scientific
3. **Context:** The main context in which the seed sentences seem to be taken. E.g. discharge summary, a case study, etc
4. **Structure:** This is a meta-level attribute that captures how the seed sentences are written. E.g. grammatical correctness, how many prepositions, etc.
5. **Label Distribution:** In theory, this should help ground the generation such that the label distribution of the synthetic sentences is not too skewed compared to the original label distribution.
6. **Additional Entities:** This represents alternative entities that may be used to diversify the synthetic sentences.

The next step involves generating NER datasets based on the filtered entities and attributes. We craft prompts to guide the language model in creating synthetic sentences and embedding entities from the pool into the text. An example prompt might be,

"Generate sentences about cardiology that include the terms 'Aspirin' and 'Heart Disease'." It is vital to list entities with their corresponding types, such as Gene, Disease, or Treatment, ensuring proper adherence to TANL format. Moreover, the generated text must adhere to domain-specific criteria, maintaining relevance to biomedical topics, reflecting appropriate writing styles, and incorporating specified biomedical terms.

Finally, using the generated dataset, we train a biomedical NER model. This model is then evaluated on real-world biomedical datasets, with adjustments made to enhance accuracy and performance. This methodology effectively leverages the framework to create diverse and relevant synthetic data for training robust biomedical NER models.

Prompts are shown in Appendix A.7.

3.2 UMLS Entity Augmentation

The entity augmentation process plays a crucial role in `BioSynNER` in enhancing the entity diversity (shown to be very beneficial in synthetic NER datasets (Heng et al., 2024)). To address this, the UMLS (Unified Medical Language System) ontology, a comprehensive resource containing detailed biomedical terms and relationships, is leveraged to expand the initial set of seed entities extracted from the original dataset.

The augmentation process begins by initializing an expanded set to include the original seed entities. The algorithm then proceeds in two key steps: First, it retrieves more specific entities, referred to as "children", from the UMLS for each entity in the seed set. These child entities represent finer-grained or narrower terms related to the seed entities, ensuring the model can capture more detailed variations of biomedical entities, which are often crucial in clinical contexts. For instance, if "disease" is the seed entity, its children might include specific diseases such as "diabetes" or "hypertension". By including these more specific terms, the model becomes more adept at recognizing highly detailed entities that might not be explicitly present in the initial dataset.

The second step involves obtaining "sibling" entities. To do this, the algorithm first queries UMLS for the "parents" of each entity in the seed set. It then retrieves the children of these parent entities, which are considered siblings of the original seed entity. Sibling entities share a common parent but may represent alternative or complementary terms in the same category. For example, if $e =$"toddler" is a seed entity, its sibling might be $e_{pc} =$"infant", as both are types of age classifications. Including sibling entities enriches the training data with broader contextual knowledge, helping the model understand a wider variety of terms related to a given domain.

The decision to augment the dataset with both children and sibling entities stems from the goal of increasing the breadth and depth of the entity set. Children entities ensure that the model is equipped to handle more specific, detailed terms, which are essential for high-precision tasks in domains like healthcare. Sibling entities, on the other hand, broaden the model's understanding by introducing alternate or related terms, thereby improving its generalization across different but related entities. Together, this augmentation ensures that the NER model can more effectively recognize and classify both common and rare entities, improving performance in specialized tasks where entity variety is a significant challenge.

Algorithm 2 `BioSynNER` Unified Medical Language System (UMLS) Entity Augmentation. This is represented as `get_UMLS_entities()` in Algorithm 1

Inputs: Initial set of seed entities $\boldsymbol{e}$
Outputs: Set of expanded entities $\tilde{\boldsymbol{e}}$

```
ẽ ← e
for e ∈ e do
    # Obtain more specific entities
    for e_c ∈ children(e) do
        ẽ ← ẽ + [e_c]
    # Obtain "sibling" entities
    for e_p ∈ parents(e) do
        for e_pc ∈ children(e_p) do
            ẽ ← ẽ + [e_pc]
return ẽ
```

4 Experiments

4.1 Datasets

The 5 datasets used to evaluate `BioSynNER` include a variety of biomedical and drug-related NER datasets. The GENIA dataset (Kim et al., 2003) consists of 1,999 Medline abstracts, annotated with multiple layers of linguistic and semantic information, aimed at supporting information extraction in molecular biology. The PHEE dataset (SUN et al., 2022) is a large pharmacovigilance corpus, containing over 5,000 annotated medical events, specifically designed for drug safety and adverse event analysis. The NCBI Disease corpus (Dogan et al., 2014) comprises 793 PubMed abstracts, with

6,892 disease mentions annotated and mapped to unique disease concepts using MeSH or OMIM identifiers. The BC2GM corpus (Kocaman and Talby, 2020) focuses on identifying gene mentions, with over 13,500 gene annotations manually reviewed for consistency and accuracy. Lastly, the BC5CDR dataset (Li et al., 2016) contains 1,500 PubMed articles, annotated with 4,409 chemicals, 5,818 diseases, and 3,116 chemical-disease interactions, providing rich data for studying chemical-disease relationships. PHEE and GENIA were obtained through InstructUIE (Wang et al., 2023), and the rest were obtained via Huggingface datasets (Lhoest et al., 2021).

Table 1: Table of Dataset Statistics

Dataset Name	# Train	# Valid	# Test	# Ner Tags
PHEE	2,898	961	968	14
GENIA	15,023	1,669	1,854	5
ncbi_disease	5,433	924	941	1
bc2gm_corpus	12,500	2,500	5,000	1
bc5cdr	5,228	5,330	5,865	2

Further details regarding the datasets, including label frequency distribution and additional background are described in Appendix A.4.

4.2 Baselines

We utilize 3 main models for our experiments RoBERTa-Large (Liu et al., 2019), Biomedical RoBERTa (base) (Gururangan et al., 2020), and DeBERTa-v3-Large (He et al., 2021b). We also utilized SapBERT (Liu et al., 2021), BioBERT (Lee et al., 2020), and BioClinicalBERT (Alsentzer et al., 2019), but we found that these BERT-based models generally performed worse compared to the RoBERTa and DeBERTa-based models, so we conduct our main experiments using the 3 best-performing models.

4.3 Evaluation

In accordance with existing literature (Tjong Kim Sang and De Meulder, 2003), we utilize `seqeval` to obtain entity mention-level metrics for precision, recall, and F1 respectively. We utilize TANL (Paolini et al., 2021) format for the representation of entities in the sentences, as seen in **Seed Sentences** in Figure 1. We convert the raw unstructured text to token-level predictions using a simple word level token splitting and labeling entity tokens in the IBO2 format (Sang and Veenstra, 1999).

4.4 Results

Table 2 compares the performance of various models on NER tasks using original training data versus data augmented by `BioSynNER` $_{400}$(400 synthetic paraphrased sentences per label). Key observations include the following.

For most datasets (PHEE, GENIA, ncbi_disease, bc2gm_corpus), models trained with BioSynNER's synthetic data exhibit lower performance (precision, recall, and F1 scores) compared to the original training data. This reflects the challenge of maintaining the same level of performance when relying heavily on synthetic examples, despite the data augmentation.

In most cases, around 90% of the performance is preserved. Notable exceptions include the bc2gm_corpus, which notably contains many highly complex and specific Gene-related entities (e.g. "IgE", "CPK", "NF (H) promoter', 'beta - galactosidase reporter gene"), which may make it more difficult to mine related entities.

Furthermore, different models exhibit varying levels of sensitivity to synthetic data augmentation. DeBERTa-v3-L performs better than Biomed. RoBERTa and RoBERTa-Large across multiple datasets, particularly in cases where synthetic data is used. This suggests that larger or more robust models like DeBERTa-v3-L may better generalize from augmented data.

Synthetic Data seems to be most effective in maintaining recall. Across datasets, recall tends to be maintained at a higher percentage of the original training data's performance compared to precision, implying that the synthetic data generation may help in identifying more potential entities, though at the cost of precision (increased false positives). Models trained on synthetic data generally show a more pronounced drop in precision, while recall remains closer to the original values. This indicates that while synthetic data helps in identifying more entities, it may lead to more misclassifications, hence lowering precision.

While BioSynNER's synthetic data helps to expand training data and maintain reasonable performance in certain cases, particularly in the recall, it does not entirely replace the efficacy of original annotated data in terms of precision and overall F1 scores across all datasets. The choice of model also plays a significant role in how effectively it leverages synthetic data.

Table 2: Results on unseen test data from our best performing method of data synthesis: BioSynNER $_{400}$, where we paraphrase up to 400 initial seed sentences from each label in the dataset, ran on the empirically best performing NER classification encoder models. The green subscripts denote performance as a percentage of the original data. Bootstrapped standard deviations are shown in the Appendix.

		Original			AUG		
Dataset	Model	Precision	Recall	F1	Precision	Recall	F1
PHEE	RoBERTa-Large	0.764	0.779	0.772	$0.693_{90.68\%}$	$0.689_{88.41\%}$	$0.691_{89.54\%}$
	Biomed. RoBERTa	0.735	0.688	0.711	$0.695_{94.53\%}$	$0.672_{97.73\%}$	$0.683_{96.16\%}$
	DeBERTa-v3-L	0.77	0.766	0.768	$0.696_{90.35\%}$	$0.698_{91.18\%}$	$0.697_{90.77\%}$
GENIA	RoBERTa-Large	0.798	0.794	0.796	$0.707_{88.65\%}$	$0.773_{97.35\%}$	$0.739_{92.81\%}$
	Biomed. RoBERTa	0.781	0.78	0.781	$0.716_{91.57\%}$	$0.765_{98.14\%}$	$0.740_{94.74\%}$
	DeBERTa-v3-L	0.809	0.802	0.805	$0.709_{87.63\%}$	$0.770_{96.00\%}$	$0.738_{91.64\%}$
ncbi_disease	RoBERTa-Large	0.869	0.893	0.881	$0.787_{90.54\%}$	$0.850_{95.19\%}$	$0.817_{92.77\%}$
	Biomed. RoBERTa	0.821	0.885	0.852	$0.730_{88.98\%}$	$0.798_{90.17\%}$	$0.763_{89.55\%}$
	DeBERTa-v3-L	0.873	0.902	0.887	$0.759_{86.89\%}$	$0.846_{93.83\%}$	$0.800_{90.17\%}$
bc2gm_corpus	RoBERTa-Large	0.813	0.816	0.814	$0.560_{68.91\%}$	$0.744_{91.20\%}$	$0.639_{78.48\%}$
	Biomed. RoBERTa	0.79	0.803	0.797	$0.601_{76.01\%}$	$0.704_{87.71\%}$	$0.648_{81.40\%}$
	DeBERTa-v3-L	0.835	0.84	0.837	$0.646_{77.43\%}$	$0.762_{90.74\%}$	$0.699_{83.54\%}$
bc5cdr	RoBERTa-Large	0.883	0.896	0.889	$0.806_{91.30\%}$	$0.862_{96.25\%}$	$0.833_{93.69\%}$
	Biomed. RoBERTa	0.857	0.89	0.873	$0.801_{93.53\%}$	$0.864_{97.09\%}$	$0.832_{95.24\%}$
	DeBERTa-v3-L	0.89	0.9	0.895	$0.828_{93.03\%}$	$0.866_{96.22\%}$	$0.847_{94.58\%}$

4.4.1 Ablation: Without UMLS Entity Augmentation

Table 3: Test results *without* UMLS entity augmentation.

Dataset	Model	Precision	Recall	F1
PHEE	RoBERTa-Large	$0.705_{92.33\%}$	$0.681_{87.35\%}$	$0.693_{89.80\%}$
	Biomed. RoBERTa	$0.722_{98.18\%}$	$0.673_{97.91\%}$	$0.697_{98.04\%}$
	DeBERTa-v3-L	$0.703_{91.26\%}$	$0.691_{90.28\%}$	$0.697_{90.77\%}$
GENIA	RoBERTa-Large	$0.699_{87.65\%}$	$0.759_{95.59\%}$	$0.728_{91.46\%}$
	Biomed. RoBERTa	$0.700_{89.54\%}$	$0.743_{95.24\%}$	$0.721_{92.30\%}$
	DeBERTa-v3-L	$0.706_{87.26\%}$	$0.762_{95.10\%}$	$0.733_{91.03\%}$
ncbi disease	RoBERTa-Large	$0.725_{83.38\%}$	$0.820_{91.84\%}$	$0.769_{87.35\%}$
	Biomed. RoBERTa	$0.710_{86.55\%}$	$0.782_{88.37\%}$	$0.745_{87.42\%}$
	DeBERTa-v3-L	$0.733_{83.95\%}$	$0.846_{93.81\%}$	$0.785_{88.52\%}$
bc2gm corpus	RoBERTa-Large	$0.514_{63.30\%}$	$0.704_{86.25\%}$	$0.594_{72.99\%}$
	Biomed. RoBERTa	$0.551_{69.70\%}$	$0.655_{81.50\%}$	$0.598_{75.09\%}$
	DeBERTa-v3-L	$0.549_{65.78\%}$	$0.698_{83.18\%}$	$0.615_{73.44\%}$
bc5cdr	RoBERTa-Large	$0.797_{90.27\%}$	$0.850_{94.94\%}$	$0.823_{92.53\%}$
	Biomed. RoBERTa	$0.811_{94.62\%}$	$0.857_{96.25\%}$	$0.833_{95.41\%}$
	DeBERTa-v3-L	$0.833_{93.58\%}$	$0.836_{92.84\%}$	$0.834_{93.21\%}$

In analyzing the results of the model without UMLS entity augmentation in Table 3, we observe a consistent decrease in performance across all five datasets, with F1 scores generally lower by 1-2 points compared to the augmented model. This drop, although moderate in most cases, highlights the importance of UMLS entity augmentation in improving model generalization and entity recognition.

One notable case is the bc2gm corpus dataset, which experiences a more significant performance decline. The F1 score for this dataset drops considerably, suggesting that the absence of UMLS-driven entity expansion particularly affects models dealing with complex biomedical corpora like bc2gm. The likely explanation for this sharper decline is that the GPT-generated entities alone, without UMLS augmentation, do not provide sufficient diversity. The limited variety of unique entities generated by GPT might fail to cover the full spectrum of entity types necessary for accurately identifying and classifying entities in such specialized datasets. Without the additional specific and sibling terms from UMLS, the model lacks the necessary depth to generalize across various entity representations, resulting in diminished performance.

This pattern across the datasets reinforces the value of using external knowledge sources like UMLS to enrich the entity set, particularly in domains where the complexity of entity types can significantly influence NER model performance.

4.5 Ablation: Generating from Scratch Instead of Paraphrasing

The results in Table 4 highlight the impact of generating synthetic data from scratch rather than using paraphrasing, as is done in BioSynNER. Across all datasets, generating from scratch yields noticeably

Table 4: Test results when prompted to *generate sentences from scratch*.

Dataset	Model	Precision	Recall	F1
PHEE	roberta-large	$0.514_{67.29\%}$	$0.554_{71.08\%}$	$0.533_{69.08\%}$
	biomed_roberta_base	$0.525_{71.40\%}$	$0.537_{78.09\%}$	$0.531_{74.72\%}$
	deberta-v3-large	$0.516_{66.99\%}$	$0.555_{72.50\%}$	$0.535_{69.67\%}$
GENIA	roberta-large	$0.473_{59.27\%}$	$0.623_{78.51\%}$	$0.538_{67.61\%}$
	biomed_roberta_base	$0.474_{60.65\%}$	$0.615_{78.88\%}$	$0.535_{68.54\%}$
	deberta-v3-large	$0.495_{61.21\%}$	$0.622_{77.59\%}$	$0.551_{68.44\%}$
ncbi disease	roberta-large	$0.467_{53.74\%}$	$0.438_{49.03\%}$	$0.452_{51.31\%}$
	biomed_roberta_base	$0.430_{52.39\%}$	$0.438_{49.47\%}$	$0.434_{50.95\%}$
	deberta-v3-large	$0.334_{38.26\%}$	$0.492_{54.55\%}$	$0.398_{44.86\%}$
bc2gm corpus	roberta-large	$0.267_{32.85\%}$	$0.270_{33.10\%}$	$0.269_{33.04\%}$
	biomed_roberta_base	$0.322_{40.75\%}$	$0.328_{40.84\%}$	$0.325_{40.80\%}$
	deberta-v3-large	$0.281_{33.66\%}$	$0.223_{26.56\%}$	$0.249_{29.74\%}$
bc5cdr	roberta-large	$0.396_{44.87\%}$	$0.433_{48.35\%}$	$0.414_{46.57\%}$
	biomed_roberta_base	$0.483_{56.38\%}$	$0.528_{59.32\%}$	$0.504_{57.73\%}$
	deberta-v3-large	$0.480_{53.93\%}$	$0.523_{58.09\%}$	$0.501_{55.97\%}$

lower F1 scores compared to other synthetic data generation strategies. For instance, on the PHEE dataset, the F1 scores are only 69% to 75%, of the original performance which is around 15% lower than paraphrasing, which obtained a minimum of 89%. This trend of underperformance can be attributed to the lack of domain-specific contextual nuances when generating sentences from scratch, *even when given examples of seed sentences*, making it harder to capture the necessary diversity and structure of biomedical entities. The models struggle particularly in datasets like bc2gm_corpus, where generating from scratch results in a dramatic decrease in F1 scores, compared to much higher paraphrasing results.

Overall, this ablation study underscores the advantage of `BioSynNER`'s approach of using paraphrasing to expand the original dataset. Paraphrasing retains key linguistic structures and contextual information, leading to better generalization and higher performance in named entity recognition tasks across biomedical domains.

4.5.1 Ablation: Varying the Number of Seed Sentences

Although we obtain the best results with higher numbers of seed sentences (since it is more representative of the entire dataset), we also experiment with smaller numbers of seed sentences.

Figure 3 shows that all datasets and models exhibit lower F1 scores when using small synthetic datasets (k=10, k=20), indicating that minimal synthetic data may not be enough to generalize well to unseen test data. As the size of the synthetic data increases, F1 scores tend to improve significantly, often approaching or even surpassing original training data performance.

PHEE shows relatively modest gains in F1 scores over all models. GENIA, ncbi_disease, and bc5cdr show some of the best overall performance, with all models improving rapidly as synthetic data is added. bc2gm_corpus seems to be the hardest dataset for models to improve upon, which makes sense as it contains highly specific Gene entities that are difficult to augment.

Across all datasets, DeBERTa-v3 exhibits strong performance similar to RoBERTa-Large, especially with larger synthetic datasets, indicating its robustness when trained with sufficient synthetic data. Biomed_roberta_base, while initially weaker on small datasets, benefits significantly from the addition of synthetic data, especially in datasets like GENIA and bc5cdr, where it nearly matches the other models at k=400. Synthetic data generation with `BioSynNER` proves to be highly effective in enhancing NER performance, particularly as the amount of generated data increases. However, smaller datasets are less effective in boosting model performance, indicating the need for sufficient diversity and quantity of synthetic data to achieve meaningful improvements in generalization and accuracy.

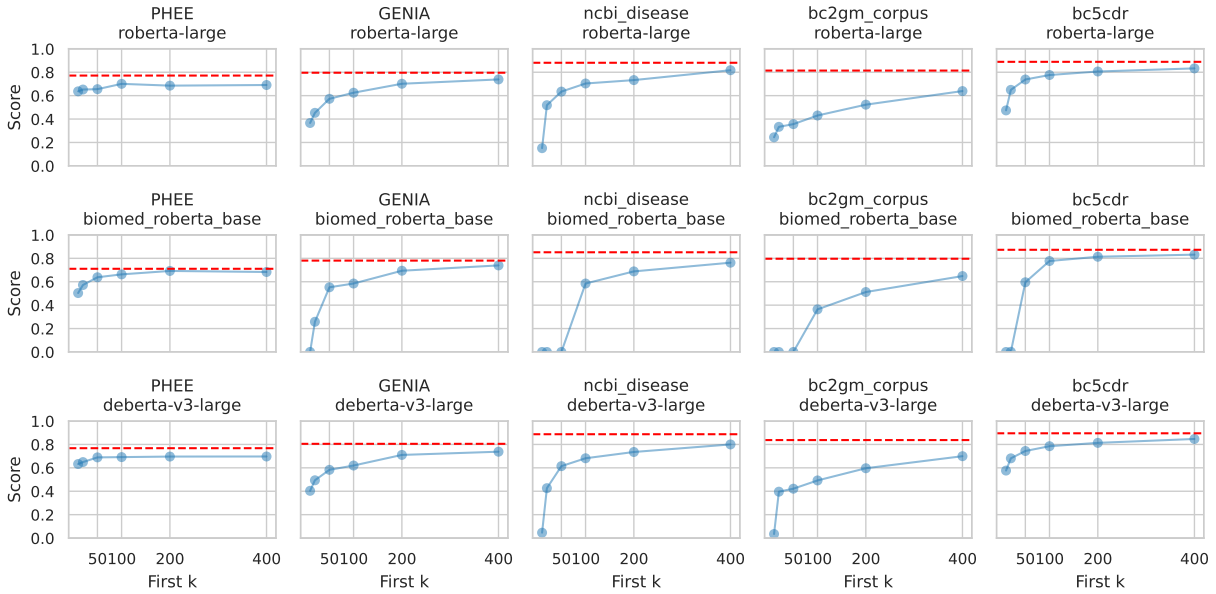
4.5.2 Ablation: Other Base Models

Table 5: Performance of Other Base Models on the Original Dataset

Dataset	Model	Precision	Recall	F1
PHEE	SapBERT	0.736	0.571	0.643
	BioBERT	0.701	0.703	0.702
	BioClinicalBERT	0.724	0.661	0.691
GENIA	SapBERT	0.785	0.782	0.784
	BioBERT	0.78	0.773	0.776
	BioClinicalBERT	0.779	0.764	0.772
ncbi_disease	SapBERT	0.847	0.854	0.85
	BioBERT	0.841	0.875	0.858
	BioClinicalBERT	0.8	0.844	0.822
bc2gm_corpus	SapBERT	0.811	0.822	0.817
	BioBERT	0.789	0.805	0.797
	BioClinicalBERT	0.766	0.783	0.774
bc5cdr	SapBERT	0.843	0.885	0.863
	BioBERT	0.841	0.865	0.853
	BioClinicalBERT	0.818	0.854	0.836

In this set of results, we observe the performance of three BERT-based models—SapBERT, BioBERT, and BioClinicalBERT—across five biomedical NER datasets. These models show relatively consistent performance, with minor varia-

Figure 3: Effect of different numbers of initial seed sentences on the F1 performance of various models across five biomedical datasets. The red dashed line denotes test F1 of a model trained on the original dataset, and the blue line denotes test F1 of each first k value through BioSynNER respectively.



tions across datasets, but several key trends stand out, particularly when compared to the RoBERTa-based and DeBERTa models used in our main experiments. Across the datasets, SapBERT generally exhibits higher precision and recall in certain datasets, such as GENIA and bc5cdr, but these performance differences are relatively narrow. BioBERT and BioClinicalBERT also show comparable performance, with F1 scores in a similar range (e.g., in PHEE and ncbi_disease). The models perform reasonably well, but the F1 scores hover only in the mid-0.70s to mid-0.80s across all datasets. This reflects the general capability of BERT-based models in NER tasks, but with some limitations in terms of robustness across diverse datasets. In certain datasets, like ncbi_disease, BioBERT tends to outperform the other BERT variants, particularly in terms of recall (0.875). However, in more complex datasets like bc2gm_corpus, none of the BERT-based models achieve very high performance, struggling on more complex biomedical corpora that require better generalization.

5 Conclusion

In conclusion, BioSynNER offers a promising solution to the challenges of Named Entity Recognition (NER) in specialized domains such as biomedical text analysis, where annotated data is scarce. By combining two key innovations—synthetic data augmentation using large language models (LLMs) and entity augmentation via the Unified Medical Language System (UMLS)—BioSynNER enhances both the diversity and generalizability of training data. The first step in the approach involves mining key domain-specific attributes from seed sentences, which are then used to generate diverse synthetic examples, addressing the need for more varied training data. The integration of UMLS further strengthens the model by expanding entity types with both more specific and sibling entities, thus improving its ability to recognize complex biomedical terms.

Experimental results confirm that BioSynNER significantly boosts NER performance across various biomedical datasets, with particularly strong improvements seen in cases where the complexity and diversity of entities are high. Without UMLS entity augmentation, the performance drops, highlighting its critical role in expanding the entity set and enhancing model generalization. Our ablation studies further reveal that generating data from scratch, rather than using paraphrasing, leads to significantly lower performance across all datasets. This reinforces the importance of paraphrasing in preserving contextual and structural nuances when generating synthetic data. Overall, BioSynNER not only improves NER performance but also provides a scalable and adaptable framework for addressing data scarcity in other highly specialized domains.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Yuzhao Heng, Chunyuan Deng, Yitong Li, Yue Yu, Yinghao Li, Rongzhi Zhang, and Chao Zhang. 2024. Proggen: Generating named entity recognition datasets step-by-step with self-reflexive large language models. *arXiv preprint arXiv:2403.11103*.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 Suppl 1:i180–2.

Veysel Kocaman and David Talby. 2020. Biomedical named entity recognition at scale. *ArXiv*, abs/2011.06315.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Monica Munnangi, Sergey Feldman, Byron C Wallace, Silvio Amir, Tom Hope, and Aakanksha Naik. 2024. On-the-fly definition augmentation of llms for biomedical ner. *arXiv preprint arXiv:2404.00152*.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *arXiv preprint arXiv:2101.05779*.

Erik F Sang and Jorn Veenstra. 1999. Representing text chunks. *arXiv preprint cs/9907006*.

ZHAOYUE SUN, Jiazheng Li, Gabriele Pergola, Byron C. Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. Phee: A dataset for pharmacovigilance event extraction from text. In *Conference on Empirical Methods in Natural Language Processing*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

Contents

A Appendix **11**
- A.1 Limitations 11
- A.2 Ethics and Reproducibility 11
- A.3 Human Impact 11
- A.4 Datasets 11
- A.5 Main Results with Standard Deviations 12
- A.6 Precision and Recall Plots 12
- A.7 Prompts 12

A Appendix

A.1 Limitations

While BioSynNER demonstrates considerable improvements in Named Entity Recognition (NER) performance, it is not without its limitations. One potential drawback is the reliance on large language models (LLMs) for synthetic data generation. Although LLMs can produce diverse and contextually relevant data, the quality of the generated examples can vary, particularly in highly specialized biomedical domains where subtle contextual nuances are critical. This variability may result in the generation of unrealistic or irrelevant examples, which could potentially mislead the model during training. One must always exercise caution when using LLM output.

Additionally, while the integration of the Unified Medical Language System (UMLS) enriches the diversity of entity types by introducing specific and sibling entities, the noisy nature of UMLS could introduce inconsistencies in the entity labels, particularly when ambiguous or less well-defined medical terms are used. This might reduce the clarity of annotations, impacting model performance. For example, one such case might occur when mining siblings that are too different / too broad, despite sharing the same parent concept. Additionally, BioSynNER 's reliance on UMLS for entity expansion may limit its applicability to other domains where such structured knowledge bases are unavailable or less comprehensive.

A.2 Ethics and Reproducibility

We utilized GPT-4-32k from Microsoft Azure, with knowledge up to Sep 2021. The code will be released later upon further polishing. All experiments were run on a machine with 2 NVIDIA V100, and took a total of around 5 days. Each every-based model was finetuned for a total of 3 epochs with a learning rate of 5e-5.

GPT-4o was utilized in writing this paper.

A.3 Human Impact

BioSynNER has the potential to create significant positive impacts on various human-centered applications, particularly in the biomedical and healthcare sectors. Improving NER directly benefits researchers, clinicians, and healthcare professionals by making the extraction of critical information from vast biomedical texts—such as scientific literature, clinical trial reports, or patient records—more efficient and reliable.

The enhanced ability to identify complex medical entities, such as diseases, drugs, and interactions, can accelerate drug discovery, improve patient safety through faster recognition of adverse drug reactions, and facilitate better decision-making in clinical settings. For instance, BioSynNER 's approach can be applied in pharmacovigilance tasks, helping to monitor and identify harmful side effects or interactions, which could lead to earlier interventions and better patient outcomes.

Additionally, the framework can extend its impact beyond biomedicine by addressing data scarcity issues in other specialized domains, enabling more effective automation in areas like legal text analysis, cybersecurity, or environmental science. By generating synthetic data through paraphrasing and entity augmentation, BioSynNER provides a scalable solution that could aid in data mining, reducing human effort and improving consistency.

A.4 Datasets

We describe the 5 main datasets here more in detail.

- GENIA is a collection of 1,999 1,999 Medline abstracts, selected through a PubMed search using the MeSH terms "human," "blood cells," and "transcription factors." It is annotated with multiple layers of linguistic and semantic information. It was developed to aid in the creation and evaluation of information extraction and text mining systems specific to the field of molecular biology. The following Table 6 shows the percentage that each ner tag occurs over the number of all annotated sentences.

- PHEE is a large pharmacovigilance dataset with over 5,000 annotated events from medical case reports and literature. The following Table 7

Table 6: GENIA Label Prevalence

NER Tag	Frequency	Raw Count
DNA	27.95%	5184
RNA	4.18%	776
cell line	14.97%	2,777
cell type	27.2%	5,044
protein	69.9%	12,964

shows the percentage that each ner tag occurs over the number of all annotated sentences.

Table 7: PHEE Label Prevalence

NER Tag	Frequency	Raw Count
Subject.Age	13.9%	671
Subject.Disorder	6.75%	326
Subject.Gender	11.5%	555
Subject.Population	8.72%	421
Subject.Race	1.33%	64
Treatment.Disorder	31.72%	1,531
Treatment.Dosage	8.51%	411
Treatment.Drug	98.96%	4,777
Treatment.Duration	3.13%	151
Treatment.Freq	2.09%	101
Treatment.Route	11.79%	569
Treatment.Time_elapsed	6.17%	298
adverse event	90.57%	4,372
potential therapeutic event	9.34%	451

- ncbi_disease is comprised of 793 PubMed abstracts, divided into training (593), development (100), and test (100) subsets. It is fully annotated with disease mentions using concept identifiers from MeSH or OMIM, and two annotators manually labeled disease mentions and linked them to the appropriate concepts, ensuring high inter-annotator agreement. The final corpus includes 6,892 disease mentions mapped to 790 unique disease concepts, with 88% linked to MeSH and the remainder to OMIM. Table 8 shows the prevalence of the labels. We note that not all sentences have a label.

Table 8: ncbi_disease Label Prevalence

NER Tag	Frequency	Raw Count
Disease	54.23%	3,958

- bc2gm_corpus focuses on identifying gene mentions within sentences. BioCreative II includes over 13,500 GENE and ALTGENE annotations, with all annotations manually reviewed for accuracy to improve consistency across gene mentions. Table 9 shows the prevalence of the labels. We note that not all sentences have a label.

Table 9: bc2gm_corpus Label Prevalence

NER Tag	Frequency	Raw Count
GENE	51.12%	10,223

- bc5cdr is composed of 1,500 PubMed articles, featuring 4,409 annotated chemicals, 5,818 diseases, and 3,116 chemical-disease interactions. Table 10 shows the prevalence of the labels.

Table 10: bc5cdr Label Prevalence

NER Tag	Frequency	Raw Count
Chemical	56.56%	9,289
Disease	49.9%	8,195

A.5 Main Results with Standard Deviations

Table 11 contains the same results as Table 2, only with the 100 sample bootstrapped standard deviations.

A.6 Precision and Recall Plots

Figure 4 and Figure 5 are the test performance of models trained on `BioSynNER` with different numbers of seed examples. Similar to Figure 3, we see an exponential increase in performance that approaches the performance of the model trained on the human-annotated training set.

A.7 Prompts

Table 12 and Table 13 demonstrate how `BioSynNER` uses LLMs to extract the domain attributes and generate new synthetic data by paraphrasing, respectively.

Table 11: Results on unseen test data from our best performing method of data synthesis: BioSynNER $_{400}$, where we paraphrase up to 400 initial seed sentences from each label in the dataset, ran on the empirically best performing NER classification encoder models. The standard deviations are in subscripts.

		Original			AUG		
Dataset	Model	Precision	Recall	F1	Precision	Recall	F1
PHEE	RoBERTa-Large	$0.764_{0.005}$	$0.779_{0.005}$	$0.772_{0.004}$	$0.693_{0.007}$	$0.689_{0.006}$	$0.691_{0.006}$
	Biomed. RoBERTa	$0.735_{0.005}$	$0.688_{0.004}$	$0.711_{0.004}$	$0.695_{0.007}$	$0.672_{0.007}$	$0.683_{0.007}$
	DeBERTa-v3-L	$0.770_{0.007}$	$0.766_{0.005}$	$0.768_{0.006}$	$0.696_{0.006}$	$0.698_{0.006}$	$0.697_{0.006}$
GENIA	RoBERTa-Large	$0.798_{0.003}$	$0.794_{0.005}$	$0.796_{0.003}$	$0.707_{0.004}$	$0.773_{0.004}$	$0.739_{0.003}$
	Biomed. RoBERTa	$0.781_{0.005}$	$0.780_{0.004}$	$0.781_{0.003}$	$0.716_{0.005}$	$0.765_{0.005}$	$0.740_{0.005}$
	DeBERTa-v3-L	$0.809_{0.005}$	$0.802_{0.005}$	$0.805_{0.004}$	$0.709_{0.005}$	$0.770_{0.006}$	$0.738_{0.005}$
ncbi_disease	RoBERTa-Large	$0.869_{0.009}$	$0.893_{0.005}$	$0.881_{0.004}$	$0.787_{0.010}$	$0.850_{0.005}$	$0.817_{0.006}$
	Biomed. RoBERTa	$0.821_{0.004}$	$0.885_{0.005}$	$0.852_{0.003}$	$0.730_{0.010}$	$0.798_{0.008}$	$0.763_{0.008}$
	DeBERTa-v3-L	$0.873_{0.007}$	$0.902_{0.005}$	$0.887_{0.005}$	$0.759_{0.009}$	$0.846_{0.010}$	$0.800_{0.009}$
bc2gm_corpus	RoBERTa-Large	$0.813_{0.003}$	$0.816_{0.003}$	$0.814_{0.002}$	$0.560_{0.005}$	$0.744_{0.003}$	$0.639_{0.004}$
	Biomed. RoBERTa	$0.790_{0.002}$	$0.803_{0.003}$	$0.797_{0.002}$	$0.601_{0.004}$	$0.704_{0.003}$	$0.648_{0.003}$
	DeBERTa-v3-L	$0.835_{0.003}$	$0.840_{0.003}$	$0.837_{0.002}$	$0.646_{0.004}$	$0.762_{0.005}$	$0.699_{0.003}$
bc5cdr	RoBERTa-Large	$0.883_{0.002}$	$0.896_{0.001}$	$0.889_{0.001}$	$0.806_{0.003}$	$0.862_{0.001}$	$0.833_{0.002}$
	Biomed. RoBERTa	$0.857_{0.002}$	$0.890_{0.003}$	$0.873_{0.002}$	$0.801_{0.003}$	$0.864_{0.003}$	$0.832_{0.003}$
	DeBERTa-v3-L	$0.890_{0.002}$	$0.900_{0.003}$	$0.895_{0.002}$	$0.828_{0.003}$	$0.866_{0.003}$	$0.847_{0.002}$

Figure 4: Effect of different numbers of initial seed sentences on the precision performance of various models across five biomedical datasets. The red dashed line denotes test precision of a model trained on the original dataset, and the green line denotes test precision of each first k value through BioSynNER respectively.

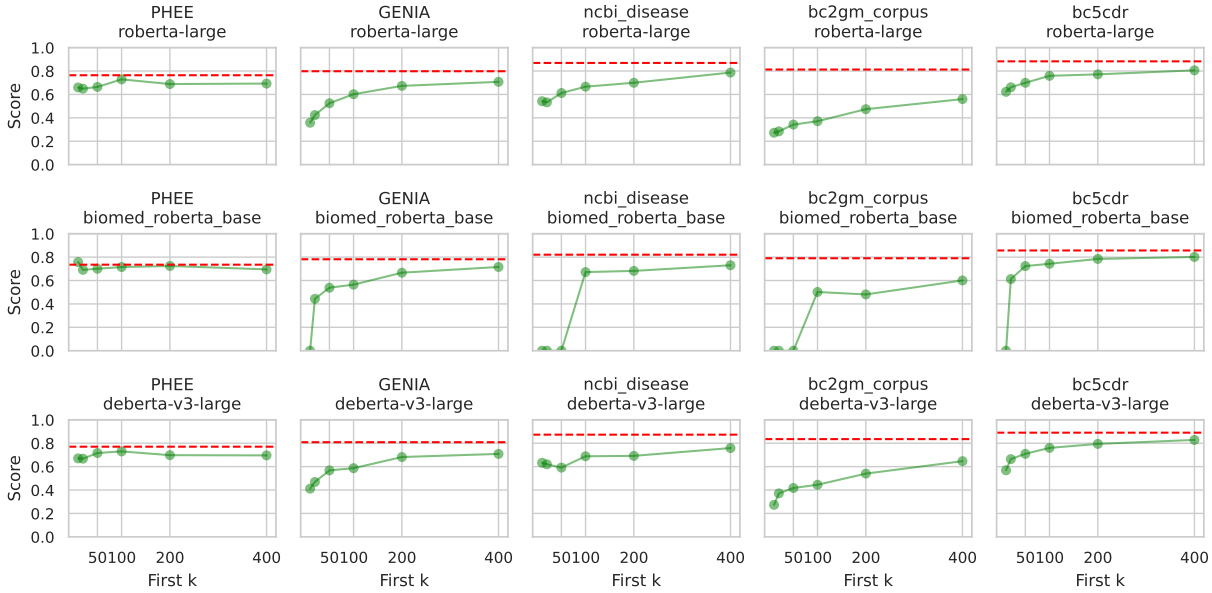


Figure 5: Effect of different numbers of initial seed sentences on the recall performance of various models across five biomedical datasets. The red dashed line denotes test recall of a model trained on the original dataset, and the orange line denotes test recall of each first k value through `BioSynNER` respectively.

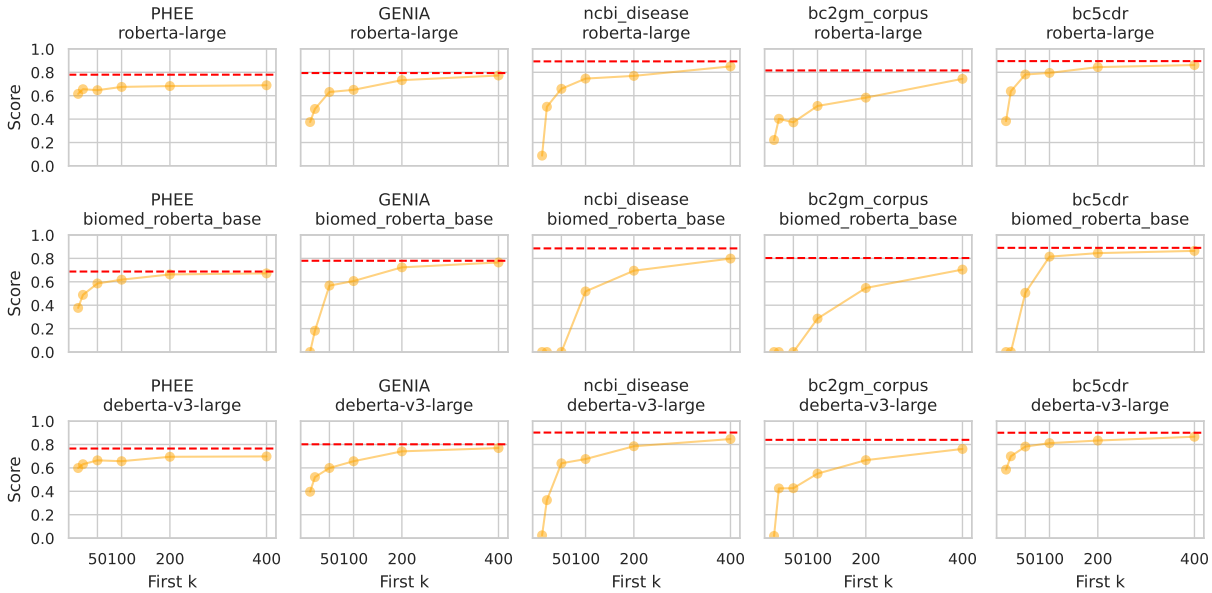


Table 12: Prompt to generate the domain attributes

```
Your task is to describe how the meta-level length, topic, writing style, context, structure, label
distribution, and additional entities given example sentences. This information will be used to
**generate similar synthetic sentences** later for **NER Label to focus on**
Notes:
- Sentences
are given in a format such that [ entity | NER label ] is the NER label for each a given entity.
-
There may be more than one kind of NER label in the sentences. Pay attention mainly on **NER Label
to focus on**.
- The output should be a valid JSON with the keys "Length", "Topic", "Writing
Style", "Context", "Structure", "Label Distribution", "Entities" and the values are lists of
string descriptions. E.g. ["description paragraph 1", "description paragraph 2", ...]
- For Label
Distribution, describe the number of times that ner tags occur in the sentences **in general** for
the purpose of generating similar sentences
- For "Entities", generate a list of up do 20 entities
that could be more examples of **NER Label to focus on** in the following format: "Entities":
['Entity 1', 'Entity 2', ...]

# Example: /*
NER Label to focus on: Subject.Age
1. An evaluation
of ovarian structure and function should be considered in [ women | Subject.Gender ] of [
reproductive age | Subject.Age ] being treated with [ valproate | Treatment.Drug ] for [ epilepsy
| Treatment.Disorder ] , especially if they [ develop | adverse event ] menstrual cycle
disturbances during treatment .
2. [ Phenobarbital | Treatment.Drug ] hepatotoxicity [ in |
adverse event ] an [ 8 - month - old infant | Subject.Age ] .
3. A case of heatstroke is [
reported | adverse event ] in a [ 32 - year - old | Subject.Age ] [ man | Subject.Gender ]
diagnosed with [ schizophrenia | Treatment.Disorder ] and on [ clozapine | Treatment.Drug ]
monotherapy .

Domain Attributes:
"'json
{
"Length": ["The sentences vary in length from short to
long, providing different levels of detail. This variety can be mimicked in the extended dataset,
creating a range of sentences from succinct to detailed."],
"Topic": ["The sentences are medical
in nature, dealing with treatments, disorders, and adverse events in patients of varying ages. The
dataset can be extended by creating similar sentences that cover a broader range of medical
scenarios, conditions, treatments, and adverse events."],
"Writing Style": ["The sentences are
written in a formal, scientific style typical of medical literature. This style should be
maintained when extending the dataset, while also introducing new medical terminologies and
diverse sentence structures."],
"Context": ["The context of the sentences is consistent with
medical research papers or clinical guidelines. The extended dataset can include sentences
suitable for a variety of contexts like medical journal articles, case studies, patient reports,
or drug trials."],
"Structure": ["The sentences have complex structures with a main clause
supplemented by additional clauses or phrases. Similar complex sentence structures can be used in
the extended dataset, introducing different medical scenarios and information."],
"Label
Distribution": ["In the provided sentences, there is a diverse distribution of NER labels. The
labels include 'Subject.Age', 'Subject.Gender', 'Treatment.Drug', 'Treatment.Disorder', and
'adverse event'. Each sentence contains multiple labels, ranging from two to five different
labels, with a slight emphasis on adverse events and the age of subjects.]
"Entities": ["newborn",
"infant", "toddler", "preschooler", "school-aged child", "adolescent", "young adult", "mid-aged
adult", "elderly", "octogenarian", "nonagenarian", "centenarian", "teenager", "juvenile", "middle-
aged", "senior", "2-year-old", "40-year-old", "70-year-old", "90-year-old"]
}
"'
*/

NER Label to
focus on: [NER_LABEL]
Sentences:
[TANL_SENTENCES]

Domain Attributes:
```

Table 13: Prompt to generate the synthetic data given the domain attributes.

```
Your task is to create a synthetic dataset for NER by editing and parapgrasing a given sentence.
Generate up to [SENTENCES_TO_GEN] **diverse** sentences from the following descriptions.
Notes:
-
Sentences should be generated in a format such that [ entity | NER label ] is the NER label for
each a given entity. Ensure there is a space between the and the entity, |, and ner labels.
-
While generating sentences, generate at least one of **NER Label to focus on**. However, NER
labels in **Valid NER Labels** can also be generated.
- Ensure that generated sentences follow the
suggested domain attributes (Length, Topic, Writing Style, Context, Structure, and Label
Distribution).
- Do **not** generate other entity labels that are not **Valid NER Labels**
-
Entities are suggested entities to use for the synthesis, and may not be in correct format. Choose
relevant entities and format them in such such that they make more sense in context. Use Example
Sentences as examples of good entity placement and formatting.

**Valid NER Labels**:
[UNIQUE_LABELS]
**NER Label to focus on**: [NER_LABEL]
Domain Attributes:
- Sentence Length:
[Sentence Length]
- Topic: [Topic]
- Writing Style: [Writing Style]
- Context: [Context]
-
Structure: [Structure]
- Label Distribution: [Label Distribution]
- Entities: [Entities]

Input
Sentence:
[TANL_SENTENCES]

Synthetic Sentences:
```