Enhancing BERT Performance with LLMs: Structured Data Augmentation for Biomedical Entity Recognition

Ying Wei[1,2], Qi Li[1] and Jay Pillai[2,*]

[1]Iowa State University, Ames, Iowa, 50010, USA
[2]Truveta, Bellevue, WA 98004, USA

Abstract

Large Language Models (LLMs) have shown remarkable capabilities across many NLP tasks, but their performance on domain-specific named entity recognition (NER), such as in the biomedical field, remains limited. Meanwhile, BERT-based models continue to achieve strong results in biomedical NER but require substantial amounts of high-quality annotated data. In this work, we investigate how to harness LLMs to generate auxiliary annotation data for BERT-based NER models, offering a cost-effective alternative to manual annotation. We address three key research questions: (1) whether LLMs or fine-tuned BERT models provide more effective weak supervision for improving BERT-based NER, (2) how to best integrate augmented and gold-standard data during training, and (3) how factors such as data source and augmentation size affect downstream performance. In particular, we introduce a structured supervision framework where an LLM is fine-tuned to generate entity annotations in a context-rich JSON format, which are decoded into token-level labels for BERT training. Experimental results on the biomedical NER dataset show that LLM-generated auxiliary annotation data effectively enhances BERT performance. Our findings provide practical insights into designing hybrid systems that combine LLMs and BERT for scalable, high-quality biomedical NER.

Keywords

LLM, NER, Biomedical, Auxiliary annotations

1. Introduction

Named Entity Recognition (NER) remains a core task in biomedical natural language processing. Despite recent progress in large language models (LLMs), BERT-based models [1] continue to deliver state-of-the-art performance for biomedical NER [2, 3]. Their relatively small size makes them efficient to train and deploy in real-world systems. However, BERT's effectiveness is closely tied to the availability of high-quality labeled data, which is often scarce in specialized domains like healthcare [4].

In contrast, LLMs have demonstrated strong context understanding abilities and high diversity in generated text [5, 6]. This leads to strong performance in zero-shot and few-shot scenarios, particularly under instruction-based setups [7, 8, 9, 10]. However, their performance in standard biomedical NER tasks may be suboptimal compared to fine-tuned models like BERT [11, 12, 13]. Furthermore, LLMs are expensive to train and deploy, making them less practical for routine use.

Recent work has explored the use of LLMs for data augmentation to enrich training sets with more diverse contexts [14, 15, 16]. However, there is a lack of systematic studies assessing the effectiveness of these auxiliary annotations for downstream model performance in biomedical NER settings. In particular, it remains unclear how best to use LLM-generated data to improve BERT training.

In this work, we investigate how LLMs can be leveraged to augment training data for BERT-based NER models. Rather than replacing BERT, we explore whether LLMs can serve as effective auxiliary data annotator to enhance BERT performance, aiming to balance accuracy and efficiency, especially important for real-world biomedical and commercial applications [17].

Specifically, we aim to answer the following research questions:

Challenge and Workshop (BC9): Large Language Models for Clinical and Biomedical NLP, International Joint Conference on Artificial Intelligence (IJCAI), August 16–22, 2025, Montreal, Canada
*Corresponding author.
yingwei@iastate.edu (Y. Wei); qli@iastate.edu (Q. Li); jayadevp@truveta.com (J. Pillai)
0000-0001-7403-3495 (Y. Wei); 0000-0002-3136-2157 (Q. Li)


- RQ1: Which model, LLM or BERT, is more effective as a source for generating auxiliary annotation data to improve BERT performance?
- RQ2: Given augmented data from different models, what is the best strategy to combine them with the original labeled data during training?
- RQ3: What factors of the augmentation set influence the performance of the BERT-based NER model?

2. Related Work

2.1. BERT for Named Entity Recognition

BERT-based NER typically leverages the model's ability to produce contextualized embeddings for input tokens, which are then processed using task-specific labeling strategies. Two dominant paradigms have emerged for adapting BERT to NER:

Token-Level Encoding with BIO Tagging: In this widely adopted approach [18, 19], BERT generates contextual embeddings for each input token, which are then fed into a classifier—such as a linear layer or a Conditional Random Field (CRF)—to predict token-level labels using the BIO (Begin, Inside, Outside) tagging scheme.

Span-Based NER: An alternative paradigm involves enumerating and classifying all possible spans within a sentence (up to a predefined maximum length) as either entities or non-entities [20, 21, 22]. This approach is particularly advantageous for handling nested entities and reduces dependence on token-level prediction accuracy. However, it typically incurs higher computational costs due to the exhaustive nature of span enumeration.

2.2. LLM for Named Entity Recognition

Recent work has explored the use of LLMs for NER through prompting. A common approach is to prompt an LLM to directly extract entity mentions from text [23], typically returning only the surface forms of entities without positional information. While this is suitable for certain applications, it limits usability in domains such as biomedical or clinical NER, where the precise location of entities in the text is often essential for downstream tasks like relation extraction or clinical decision support.

To address this limitation, some studies [11, 24] have proposed prompting LLMs to output token-level labels using the BIO tagging scheme, enabling recovery of entity positions. However, this method presents challenges when applied to long documents common in the biomedical domain—such as clinical notes, due to context window limitations, performance degradation over long sequences, and high computational cost.

2.3. Model-based Auxiliary Annotations for Named Entity Recognition

To reduce the reliance on costly manual annotations, recent work has explored generating auxiliary training data using existing models. These approaches aim to improve model performance in low-resource settings or to expand training datasets with minimal human effort.

Self-Training: Self-training techniques use a strong model (teacher) to generate pseudo-labels on unlabeled data, which are then used to train a student model. For example, Gao et al. [25] and Li et al. [26] apply this paradigm to NER, showing that high-quality pseudo-labels can significantly boost performance, particularly when combined with confidence-based filtering. However, such methods may reinforce the biases of the teacher model and offer limited diversity in the generated annotations.

LLMs for Generating Annotations: LLMs have recently been used to generate auxiliary annotations for downstream tasks, including NER [16, 15, 27]. These studies show that LLM-generated annotations can effectively enhance the performance of smaller, task-specific NER models [24].

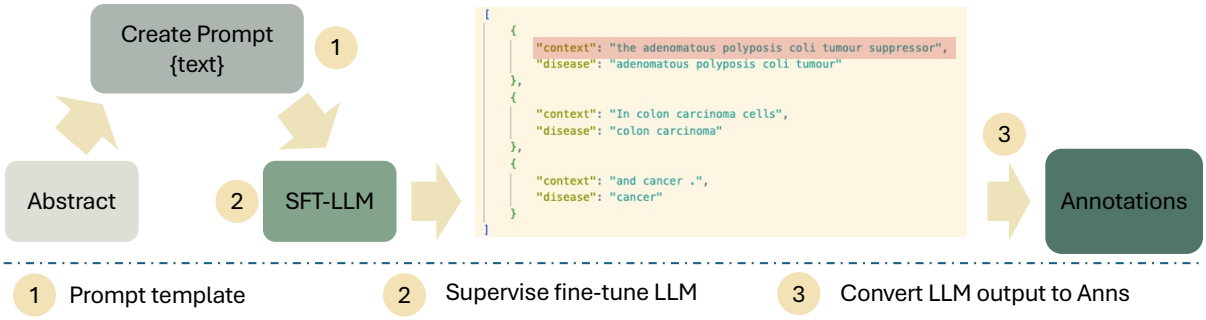

Figure 1: Illustration of SFT LLM phrase.

3. Methods

We propose LLM-Guided Synthetic Annotation, a framework to enhance BERT-based biomedical NER models by leveraging synthetic labels generated from a supervised fine-tuned (SFT) LLM.

3.1. Span-based NER with BERT

For named entity recognition, we adopt a span-based classification architecture built upon BERT. Given an input document $D = [t_1, t_2, ..., t_n]$, we first encode it using a pre-trained BERT model to obtain contextualized token embeddings $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n]$, where $\mathbf{h}_i \in \mathbb{R}^d$ is the representation of token t_i.

We then enumerate all candidate spans $s = (i, j)$ such that $1 \leq i \leq j \leq n$ and $j - i + 1 \leq L$, where L is the maximum span length. For each candidate span, we compute a fixed-length span representation $\mathbf{s}_{i,j}$ by applying a pooling operation (e.g., max pooling) over the token embeddings within the span: $\mathbf{s}_{i,j} = \text{Pool}([\mathbf{h}_i, \mathbf{h}_{i+1}, ..., \mathbf{h}_j])$ The resulting span representation is passed through a multi-layer perceptron (MLP) to classify whether the span corresponds to an entity and, if so, predict its entity type: $\hat{y}_{i,j} = \text{MLP}(\mathbf{s}_{i,j})$. Spans predicted with non-null entity types are retained as final NER outputs. This architecture enables flexible modeling of variable-length entities and supports overlapping spans.

3.2. Using LLM to Support BERT for NER

While LLMs demonstrate strong language understanding, they often struggle to produce structured outputs with accurate span boundaries. These limitations make LLMs less effective when directly applied to NER tasks, motivating the need for alternative strategies. To address this, we propose a structured context-based output format that enables LLMs to serve as high-quality annotators for training BERT-based NER models.

Supervised Fine-Tuning (SFT) Phase: We fine-tune an LLM on human-labeled biomedical datasets to serve as a structured annotator. Each example consists of a biomedical sentence and its corresponding NER annotation in a standardized JSON format, enabling the model to learn structured entity extraction.

LLMs perform well in language understanding but struggle with precise span boundaries. To mitigate this, we adopt a context-based JSON format that reframes span detection as context matching rather than character-level indexing. Each entity is represented as a JSON object with:

- Entity span: The surface form of the entity (e.g., “metastatic melanoma”).
- Entity context: A surrounding text window including the entity (e.g., “Patients with metastatic melanoma often exhibit”).
- Entity label: The predicted biomedical type (e.g., Disease, Drug, Symptom).

This format enables accurate span recovery through string matching within context, avoiding brittle index prediction and leveraging the LLM’s strength in coherent text generation. During supervised fine-tuning (e.g., using LLaMA-3), the model is trained to generate the correct structured JSON. Token-level loss encourages exact reproduction of the ground-truth output, ensuring both syntactic validity and semantic fidelity. Figure 1 provides an illustration of this process.

Table 1
Model performance comparison using different auxiliary data annotators.

Model	NCBI Dataset				Proprietary Dataset			
	Axuiliary Data	Prec.	Recall	F1	Training Data	Prec.	Recall	F1
LLM	-	85.38	83.86	84.62	-	82.44	81.58	82.01
BERT	-	87.95	88.96	88.45	-	84.38	86.72	85.52
BERT-L	BC5CDR (LLM)	89.30	90.42	89.86	Proprietary-val (LLM)	85.05	88.01	86.50
BERT-B	BC5CDR (BERT)	89.17	89.17	89.17	Proprietary-val (BERT)	84.38	87.13	85.73

LLM Inference Phase: After supervised fine-tuning, the LLM is applied to unlabeled biomedical texts to generate synthetic NER annotations in a structured JSON format. To recover character-accurate spans, we match each predicted entity within its surrounding context back to the original sentence. This context-based matching avoids reliance on error-prone start/end index predictions.

If an entity span cannot be confidently aligned due to hallucination or formatting inconsistencies, it is discarded to maintain label quality. This approach allows a small set of human-labeled examples to be scaled into a much larger pseudo-labeled corpus. The structured generation and context-grounded decoding ensure high-precision annotations, making them effective for generating auxiliary annotations.

4. Experimental Studies

We investigate our research questions through NER tasks in the biomedical domain.

4.1. Datasets and Experimental Setups

Datasets. We evaluate our method on one biomedical dataset and one proprietary clinical notes dataset, each with different auxiliary data sources:

Train-Test Datasets:

- **NCBI Disease Dataset:** A benchmark dataset for disease named entity recognition, consisting of 793 PubMed abstracts (593 train / 100 val / 100 test), with 7,311 sentences and 185,322 tokens.
- **Proprietary Dataset:** A de-identified clinical notes dataset from real-world Truveta applications, including 2,306 notes (1,679 train / 313 val / 314 test), with 347,371 sentences and 4,168,463 tokens.

Auxiliary Datasets:

- **BC5CDR Corpus:** In-domain for NCBI; contains 1,500 PubMed abstracts, 16,423 sentences, and 334,598 tokens.
- **BioRED Corpus:** In-domain for NCBI and out-of-domain for Proprietary; includes 5,935 PubMed abstracts and 1,980,273 tokens.
- **NCBI Validation Set:** Used as auxiliary data in some settings for the NCBI dataset.
- **Unlabeled Clinical Notes:** Sampled from the proprietary corpus for generating pseudo-labels.

Experimental Setups. We evaluate our framework on the biomedical NER task, using domain-specific models to ensure strong baselines. For the biomedical dataset, we use BioBERT, a widely adopted BERT variant trained on large-scale biomedical corpora. As our LLM backbone, we use LLaMA-3, a general-purpose language model that is fine-tuned to generate structured biomedical annotations.

The entity annotations are converted to a unified format to support consistent processing across datasets. We fine-tune our LLM-based extractor using supervised learning on each training set separately and evaluate performance on their corresponding test sets. All experiments were performed using a single NVIDIA A100 GPU with 80GB memory.

4.2. Experimental Study for Research Question 1

In this section, we study our first research question: Given that LLMs under-perform compared to BERT on NER tasks, which model, LLM or BERT, is more effective as a source for generating auxiliary

Table 2
Comparison with different strategies of combining auxiliary annotations and human annotations.

Model	Training set in stage 1	Training set in stage 2	Prec.	Recall	F1
BERT-L	LLM-labeled NCBI-val (30 epochs)	Human-labeled NCBI-train (30 epochs)	87.05	91.04	89.00
BERT-L	Human-labeled NCBI-train LLM-labeled NCBI-val (30 epochs)	–	86.62	91.04	88.78
BERT-L	Human-labeled NCBI-train (20 epochs)	Human-labeled NCBI-train LLM-labeled NCBI-val (10 epochs)	87.06	89.69	88.35
BERT-L	Human-labeled NCBI-train LLM-labeled NCBI-val (20 epochs)	Human-labeled NCBI-train (10 epochs)	86.21	91.15	88.61

Table 3
Model performance comparison results of different factors on BERT performance.

Model	Training data	Testing data	Prec.	Recall	F1
BERT	NCBI-train	NCBI-test	87.95	88.96	88.45
BERT	Proprietary-train	Proprietary-test	84.38	86.72	85.52
BERT-L	NCBI-train + LLM-labeled NCBI-val	NCBI-test	87.05	91.04	89.00
BERT-L	NCBI-train + LLM-labeled BC5CDR	NCBI-test	89.30	90.42	89.86
BERT-L	NCBI-train + LLM-labeled BioRED	NCBI-test	89.84	91.15	90.49
BERT-L	Proprietary-train + LLM-labeled Proprietary-val	Proprietary-test	85.05	88.01	86.50
BERT-L	Proprietary-train + LLM-labeled BioRED	Proprietary-test	84.78	86.69	85.72

annotation data to improve BERT performance?

To answer this question, we compare the downstream NER performance of BERT models trained with different annotation sources:

- BERT-L models: BERT models trained on auxiliary annotation data generated by an *LLM* trained on the NCBI/Proprietary training set.
- BERT-B models: BERT models trained on auxiliary annotation data generated by a *BERT* model trained on the NCBI/Proprietary training set.

The comparison results of the LLM, the original BERT and the augmented BERT models are summarized in Table 1. As shown in the table, both the BERT-L and BERT-B models are first trained using auxiliary annotations from the BC5CDR corpus and then fine-tuned with the original NCBI training data. When trained with NCBI-train + auxiliary-labeled BC5CDR, BERT-L reaches 89.86 F1 outperforms that of BERT-B reaches 89.17.

Table 1 summarizes the performance of the LLM, the original BERT model, and the BERT models trained with auxiliary annotations. In both the NCBI and proprietary settings, augmenting BERT with LLM-generated annotations (BERT-L) consistently improves performance over using no augmentation and yields better F1 scores than BERT-generated annotations (BERT-B). For example, on NCBI-test, BERT-L achieves an F1 of 89.86, marginally higher than BERT-B (89.17). Similarly, on the proprietary test set, BERT-L reaches 86.50, outperforming BERT-B (85.73).

These results suggest that LLMs, despite weaker direct performance on biomedical NER, are more effective in generating auxiliary annotation data. We attribute this to their decoder-based architecture and extensive pretraining on diverse, open-domain corpora, which likely enhance their generative fluency and contextual coverage. This enables LLMs to produce higher-quality and more varied annotations—providing greater benefits during augmentation than the more accurate but narrower

predictions made by BERT models.

4.3. Experimental Study for Research Question 2

In this section, we study our second research question: Given augmented data from different models, what is the best strategy to combine them with the original labeled data during training?

To explore this, we evaluate multiple two-stage training strategies that integrate auxiliary annotations generated by an LLM into BERT model training (Table 2). All strategies demonstrate performance improvements over the baseline BERT trained only on human-labeled data, confirming the value and robustness of incorporating auxiliary annotations.

Among the strategies, joint training on the NCBI training set and LLM-labeled validation data in Stage 1 achieves the highest F1 score (88.78), followed closely by the two-stage finetuning strategy that begins with LLM-labeled data and then fine-tunes on the training set (89.00 F1). These results suggest that exposing the model to both data sources early—either jointly or sequentially—can be beneficial, as long as human-labeled data is emphasized in the latter stages.

In contrast, the setup that begins with training only on the human-labeled data and then fine-tunes with a mix of human and auxiliary annotations yields slightly lower performance (F1 = 88.35), indicating that late-stage mixing may be less effective. Overall, our findings suggest that the timing and method of incorporating auxiliary annotations matter, with early integration either through joint or warm-up strategies, yielding the best results.

4.4. Experimental Study for Research Question 3

In this section, we investigate the third research question: what factors of the augmentation set influence the performance of BERT-based NER models? Specifically, we examine whether increasing the amount of auxiliary annotations improves performance, and how this effect interacts with domain relevance.

We consider three auxiliary sources for the NCBI dataset: NCBI-val (100 abstracts), BC5CDR (1,500 abstracts), and BioRED (5,935 abstracts). While all three are in-domain, BioRED is substantially larger. As shown in Table 3, augmenting NCBI-train with BC5CDR improves F1 to 89.86, and using the larger BioRED set yields further gains (F1: 90.49), indicating that annotation volume has a positive effect on model performance.

To assess the impact of data source domain, we compare augmenting the proprietary training set with LLM-labeled data from two sources: the in-domain proprietary validation set and the out-of-domain BioRED corpus (sampled to match token count). As shown in Table 3, in-domain data leads to higher performance (F1: 86.50) than out-of-domain data (F1: 85.72), suggesting that domain relevance is critical when annotation volume is held constant.

5. Conclusion

This work investigates the effectiveness of using LLMs to generate auxiliary annotation data for biomedical NER, comparing their performance to that of traditional fine-tuned models such as BERT. We systematically study how label quality, data quantity, and domain relevance influence downstream performance across various auxiliary data generation strategies. Our results show that LLMs can produce auxiliary annotation data that are competitive with those generated by fine-tuned BERT models. In particular, combining a small amount of gold-labeled data with LLM-labeled auxiliary annotation data yields strong performance gains. Additionally, while in-domain data (e.g., NCBI-val) offers the best contextual alignment, larger out-of-domain corpora (e.g., BioRED) lead to even greater improvements, highlighting that scale can effectively mitigate domain mismatch when the labeler is reliable. Taken together, our findings suggest that LLM-generated weak supervision is a scalable and effective alternative to manual annotation, especially in low-resource biomedical domains. As LLMs continue to improve, they may become increasingly viable as universal annotators, capable of adapting to new tasks and domains with minimal human input.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 for grammar and spelling check. After using these tool(s)/service(s), the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[2] Q. Lu, R. Li, A. Wen, J. Wang, L. Wang, H. Liu, Large language models struggle in token-level clinical named entity recognition, arXiv preprint arXiv:2407.00731 (2024).

[3] C. Sun, Z. Yang, L. Wang, Y. Zhang, H. Lin, J. Wang, Biomedical named entity recognition using bert in the machine reading comprehension framework, Journal of Biomedical Informatics 118 (2021) 103799.

[4] Z. Gouliev, R. R Jaiswal, Efficiency of llms in identifying abusive language online: A comparative study of lstm, bert, and gpt, in: Proceedings of the 2024 Conference on Human Centred Artificial Intelligence-Education and Practice, 2024, pp. 1–7.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[6] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools, Advances in Neural Information Processing Systems 36 (2023) 68539–68551.

[7] Y. Zhang, D. G. Vlachos, D. Liu, H. Fang, Rapid adaptation of chemical named entity recognition using few-shot learning and llm distillation, Journal of Chemical Information and Modeling (2025).

[8] X. Zhu, F. Dai, X. Gu, B. Li, M. Zhang, W. Wang, Gl-ner: Generation-aware large language models for few-shot named entity recognition, in: International Conference on Artificial Neural Networks, Springer, 2024, pp. 433–448.

[9] F. Villena, L. Miranda, C. Aracena, llmner:(zero| few)-shot named entity recognition, exploiting the power of large language models, arXiv preprint arXiv:2406.04528 (2024).

[10] T. Hiltmann, M. Dröge, N. Dresselhaus, T. Grallert, M. Althage, P. Bayer, S. Eckenstaler, K. Mendi, J. M. Schmitz, P. Schneider, et al., Ner4all or context is all you need: Using llms for low-effort, high-performance ner on historical texts. a humanities informed approach, arXiv preprint arXiv:2502.04351 (2025).

[11] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, Gpt-ner: Named entity recognition via large language models, arXiv preprint arXiv:2304.10428 (2023).

[12] M. Monajatipoor, J. Yang, J. Stremmel, M. Emami, F. Mohaghegh, M. Rouhsedaghat, K.-W. Chang, Llms in biomedicine: A study on clinical named entity recognition, arXiv preprint arXiv:2404.07376 (2024).

[13] G. Tolegen, A. Toleu, R. Mussabayev, Enhancing low-resource ner via knowledge transfer from llm, in: International Conference on Computational Collective Intelligence, Springer, 2024, pp. 238–248.

[14] S. Sharma, A. Joshi, Y. Zhao, N. Mukhija, H. Bhathena, P. Singh, S. Santhanam, When and how to paraphrase for named entity recognition?, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 7052–7087.

[15] J. Ye, N. Xu, Y. Wang, J. Zhou, Q. Zhang, T. Gui, X. Huang, Llm-da: Data augmentation via large language models for few-shot named entity recognition, arXiv preprint arXiv:2402.14568 (2024).

[16] Y. Naraki, R. Yamaki, Y. Ikeda, T. Horie, K. Yoshida, R. Shimizu, H. Naganuma, Augmenting ner

datasets with llms: towards automated and refined annotation, arXiv preprint arXiv:2404.01334 (2024).
[17] W. Xu, R. Dang, S. Huang, Llm's weakness in ner doesn't stop it from enhancing a stronger slm, in: Proceedings of the Second Workshop on Ancient Language Processing, 2025, pp. 170–175.
[18] K. Hakala, S. Pyysalo, Biomedical named entity recognition with multilingual bert, in: Proceedings of the 5th workshop on BioNLP open shared tasks, 2019, pp. 56–61.
[19] U. Naseem, M. Khushi, V. Reddy, S. Rajendran, I. Razzak, J. Kim, Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–7.
[20] M. G. Sohrab, M. S. Bhuiyan, Span-based neural model for multilingual flat and nested named entity recognition, in: 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), IEEE, 2021, pp. 80–84.
[21] M. Zuo, Y. Zhang, A span-based joint model for extracting entities and relations of bacteria biotopes, Bioinformatics 38 (2022) 220–227.
[22] Y. Tang, J. Yu, S. Li, B. Ji, Y. Tan, Q. Wu, Span classification based model for clinical concept extraction, in: The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, Springer, 2020, pp. 1880–1889.
[23] J. Santoso, P. Sutanto, B. Cahyadi, E. Setiawan, Pushing the limits of low-resource ner using llm artificial data generation, in: Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 9652–9667.
[24] M. S. Obeidat, M. S. A. Nahian, R. Kavuluru, Do llms surpass encoders for biomedical ner?, arXiv preprint arXiv:2504.00664 (2025).
[25] S. Gao, O. Kotevska, A. Sorokine, J. B. Christian, A pre-training and self-training approach for biomedical named entity recognition, PloS one 16 (2021) e0246310.
[26] Z.-z. Li, D.-w. Feng, D.-s. Li, X.-c. Lu, Learning to select pseudo labels: A semi-supervised method for named entity recognition, Frontiers of Information Technology & Electronic Engineering 21 (2020) 903–916.
[27] Y. Wei, Q. Li, J. Pillai, Structured llm augmentation for clinical information extraction, in: MEDINFO 2025—The Future Is Accessible, IOS Press, 2025.