

Exploring Data Augmentation Strategies for Biomedical NER: A Study on the GutBrainIE 2026 Shared Task

Notebook for the BioASQ Task GutBrainIE on Gut-Brain Interplay Information Extraction at CLEF 2026

Miguel Lourenço¹, Andre Lamurias^{1,*}

¹NOVA LINCS, NOVA School of Science and Technology

Abstract

This paper presents our participation in the GutBrainIE 2026 shared task, specifically in Subtask 6.1.1 (NER), in which we explored data augmentation strategies for Named Entity Recognition (NER) in the biomedical domain, with a focus on the gut-brain interplay. We investigate three augmentation approaches applied to the GutBrainIE dataset to fine-tune the NuNER Zero model: synonym substitution of non-entity tokens using WordNet, entity replacement using Word2Vec embeddings, and a combination of both strategies. Our development experiments show that moderate synonym substitution consistently outperforms more aggressive augmentation strategies, while the inclusion of Silver data alongside Gold improves macro-level generalization. On the official test set, our augmented runs consistently outperform our unaugmented baseline, with synonym-based augmentation proving the most effective overall. None of our runs surpassed the organizer baseline, yet our closest results were achieved using only Gold data with augmentation, whereas the baseline was trained on both Gold and Silver data, suggesting that well-chosen augmentation strategies can partially compensate for the absence of additional annotated data.

Keywords

Named Entity Recognition, Biomedical NLP, Data Augmentation, Gut-Brain Axis

1. Introduction

The gut-brain axis is a bidirectional communication system between the gastrointestinal tract and the central nervous system. Recently, new discoveries have shown the influence of the gut microbiome in a wide spectrum of neurological and psychiatric conditions, including depression, Alzheimer's and Parkinson's, making gut-brain research one of the most active frontiers in modern biomedicine.

As scientific literature on the gut-brain axis continues to increase, so does the need for robust information extraction systems that are capable of extracting and structuring the biomedical knowledge that is available. Named Entity Recognition (NER) is a fundamental task in biomedical information extraction that aims to identify key biomedical entities such as diseases, treatments or experimental variables.

One of the challenges of biomedical NER is data availability. In the biomedical domain, obtaining expertly annotated data is both time-consuming and expensive. This limitation is even more present in very specific areas like gut-brain research, with its unique vocabulary and conceptual landscape. Data augmentation is a promising approach to address the lack of data. By artificially expanding the amount of annotated data, it can improve model robustness and generalization without the need for additional manual annotation. Furthermore, state-of-the-art NER systems often rely on large language models or ensemble architectures that demand significant computational resources. Finding ways to improve the performance of more lightweight models through strategies such as data augmentation is an important consideration, as it allows for competitive results without the associated computational cost.

GutBrainIE is a Task that is part of the BioASQ CLEF Lab 2026 [1], with the focus of extracting structured information from abstracts related to the gut microbiota and its connections with neurological

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany

*Corresponding author.

✉ mo.lourenco@campus.fct.unl.pt (M. Lourenço); a.lamurias@fct.unl.pt (A. Lamurias)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

diseases. The task is divided into four sub-tasks:

- **Subtask 6.1.1 - Named Entity Recognition (NER):** classification of text spans into 13 established entity categories.
- **Subtask 6.1.2 - Named Entity Recognition and Disambiguation (NERD):** classification of text spans into 13 established entity categories, where each identified entity must be linked to a concept identifier in a provided biomedical reference resource.
- **Subtask 6.2.1 - Mention-Level Relation Extraction (M-RE):** identification of relations between entities mentioned in a document.
- **Subtask 6.2.2 - Concept-Level Relation Extraction (C-RE):** identification of relations between concepts that were linked from the entities, instead of their textual mentions.

Further information and details about the Task can be found in the extended GutBrainIE task overview [2].

This paper presents our approach to the first subtask, Subtask 6.1.1, where we focused on three data augmentation approaches:

- substituting non-entities with synonyms to introduce lexical variation in non-entity contexts
- replacing entities with semantically similar words identified through Word2Vec embeddings to create entity-level variations
- combining both strategies to maximize augmentation coverage

These approaches were applied to the training datasets, and the resulting datasets were used to fine-tune the NuNER Zero model ¹. All of the scripts used and mentioned can be accessed on our GitHub.

2. Related Work

NER is a task in Natural Language Processing that seeks to locate and classify named entities mentioned in unstructured text into predefined categories. In 2013, Tomas Mikolov, with the help of other researchers, developed Word2Vec, a word embedding model that represents words as high-dimensional vectors that attempt to capture meaning based on surrounding words [3]. Similar words have vectors that are also similar when measured using cosine similarity, which is the cosine of the angle between the vectors. For example, the words "walk" and "run" have similar vectors, as do "Portugal" and "Lisbon".

The introduction of transformer-based architectures, and in particular BERT [4], marked a significant shift in the field, as contextual word representations substantially improved NER performance across domains. In the biomedical domain specifically, models such as BioBERT [5], which uses BERT and pre-trains it on PubMed articles, demonstrated that domain-specific pre-training leads to meaningful gains in biomedical NER.

More recently, the field has moved towards flexible, generalist NER models capable of handling arbitrary entity types without retraining. GLiNER [6] introduced a compact approach based on bidirectional transformer encoders that frames NER as a matching problem between entity type prompts and textual spans, outperforming both fine-tuned LLMs and ChatGPT in zero-shot evaluations. Building on this architecture, NuNER Zero [7] extended GLiNER by pre-training on a large LLM-annotated dataset, achieving state-of-the-art compact zero-shot NER performance.

Data augmentation refers to methods used to increase the amount of training data by adding slightly modified copies of existing data or newly created synthetic data. Given the high cost of biomedical data annotation, it is an important tool for improving model robustness and generalization in low-resource scenarios. Traditional data augmentation methods apply simple transformations to existing annotated data, such as synonym replacement, random insertion or deletion of tokens, and back-translation [8]. Although these methods increase the size of the available data with low effort, they often result in

¹https://huggingface.co/numind/NuNER_Zero

limited linguistic variety and can introduce label inconsistencies when entity boundaries are altered. Due to this possibility, it is necessary to consider ways to avoid it when applying these techniques to NER.

More recently, generative language models have introduced new strategies for data augmentation, enabling the creation of entirely new sentences with coherent structure that incorporate specific named entities [9]. These generative approaches are particularly well-suited to NER tasks, as they allow for the controlled generation of domain-specific text containing underrepresented entity types, resulting in improved performance in few-shot settings. However, generative approaches are more computationally demanding, and prone to hallucinations, where the model generates seemingly normal but factually incorrect text. This can introduce noise into the training data and potentially hinder model performance.

3. Methodology

This section describes the methodology followed in our experiments. We begin by presenting the dataset used for training and evaluation, followed by the training setup and configuration. We then detail the three data augmentation strategies explored during development, and conclude with a description of the five runs submitted to the shared task.

3.1. Datasets

The GutBrainIE dataset consists of PubMed abstracts annotated across different quality tiers, those being expert-annotated (Gold, 639 abstracts), student-annotated (Silver, 1310 abstracts) and automatically annotated (Bronze, 2972 abstracts). In our experiments, we used the Gold dataset as our primary training set and the base for most of our data augmentations, as including the Silver dataset resulted in training times approximately three times longer than Gold-only runs with the computational resources available to us. We also included the Silver dataset in one run to assess whether additional original data could offset this computational cost through improved performance. All augmentations were applied exclusively to the training set. During development, we evaluated the runs on the official development set, using the task’s evaluation metrics.

3.2. Training

For all of our runs, we used NuNER Zero as our base model. Considering our objective was to compare different data augmentation strategies, we decided to use the same model as the one reported in the baseline results. The training for the model was performed on an Nvidia L4 GPU, with a batch size of 4 and 10 epochs. The rest of the hyperparameters were kept the same as the ones in the provided training script, accessible [here](#). To choose the best checkpoint, evaluation on the development set was carried out every 300 steps during training, using the F1 score as the metric for scoring. The predictions generated by the model are converted into the evaluation format using the conversion script that came with the training script.

3.3. Data augmentation

We explored three data augmentation strategies with the objective of artificially expanding the training dataset. All of our data augmentation was done using the `nlpaug` library’s augmenters ².

- **Synonym Substitution (SYN):** To introduce lexical variation in non-entity tokens, we applied synonym-based augmentation using WordNet [10] as the synonym source. To ensure that the entity tokens are not changed during this augmentation, we replace each entity mention with a placeholder token, which is treated as a stopword by the augmentser. This process was also done to the formatting tags that came with the abstracts, which are `<i>` and `</i>`. The remaining tokens

²<https://github.com/makcedward/nlpaug>

are fed into the augmenter that substitutes them with a probability of 15%. For each training instance, we generate 10 augmented variations, each with a different seed, to ensure reproducible results. After the augmentation, the placeholders are restored to their original entity mentions. An example of this transformation is shown on Table 1.

Original	Restoring the intestinal and microbial homeostasis by beneficial bacteria and metabolites provide a potential therapeutic strategy to treat ALS.
Augmented	Reinstate the intestinal and microbial homeostasis by good bacterium and metabolites provide a potential curative scheme to treat ALS.

Table 1

Example of Synonym substitution (SYN) augmentation

- **Entity Replacement (EMB):** To introduce variation at the entity level, we replaced entity mentions with semantically similar words using the Common Crawl English 300-dimensional embeddings³. To achieve this, we start by feeding into the augmenter 15% of the entities. To ensure that the entities that are chosen get meaningfully altered, we set a minimum of 1 augmented word, and an augmentation percentage of 50%, which means that roughly half of the words that are part of an entity will get replaced. The augmenter then chooses replacement candidates from the top-20 most similar words in the embedding space. No filtering was applied to the replacement candidates. For each training instance, we generate 10 augmented variations, each with a different seed, to ensure reproducible results. An example of this transformation is shown on Table 2:

Original	F4 consumption reduced the elevated levels of serotonin (5-HT) in the brain observed in the CUS group.
Augmented	F4 consumption reduced the elevated levels of noradrenalin (5-HT) in the brain observed in the CUS group.

Table 2

Example of Entity replacement (EMB) augmentation

- **Combined Strategy (SYN+EMB):** We also experimented with combining both previous augmentation strategies. Starting with the 10 synonym-augmented variations generated using the method outlined in the SYN step, we apply entity replacement augmentation to each of the variations, only generating one new variation. This way, the resulting dataset introduces both non-entity lexical variation and entity-level variation. An example of this transformation is shown on Table 3:

Original	[...] and other related circuitry, which contribute to the development of both motor (bradykinesia , tremors, stiffness , abnormal gait) [...]
Augmented	[...] and other related circuitry, which impart _{SYN} to the development of both motor (dyskinesia _{EMB} , tremors, flexural _{EMB} , abnormal gait) [...]

Table 3

Example of Combined strategy (SYN+EMB) augmentation

SYN augmentation was applied exclusively to non-entity tokens, as the entities in this domain are highly specialized and unlikely to have meaningful synonyms in a general-purpose database such as WordNet. EMB augmentation was only applied to entity tokens, and extending this strategy to non-entity tokens remains unexplored and presents an opportunity for future work.

The generated data is in the same format as the training data. To use this data for training the model, we first run the `annotations_to_gliner_format` script. The converted data is then used in the training script.

³<https://fasttext.cc/docs/en/crawl-vectors.html>

3.4. Submitted runs

To assess the impact of each data augmentation strategy, we defined five experimental runs:

- **Run 1 (Baseline):** Fine-tuning on Gold data only, without any augmentation.
- **Run 2 (SYN-Gold):** Fine-tuning on Gold data and the Gold data augmentation variants obtained using synonym substitution.
- **Run 3 (SYN-Gold+Silver):** Fine-tuning on Gold data, Gold data augmentation variants, Silver data and Silver data augmentation variants obtained using synonym substitution.
- **Run 4 (EMB-Gold):** Fine-tuning on Gold data and the Gold data augmentation variants obtained using entity substitution based on embeddings.
- **Run 5 (SYN+EMB-Gold):** Fine-tuning on Gold data and the Gold data augmentation variants obtained using a combination of synonym substitution and entity substitution based on embeddings.

For every run, the model generated predictions with a confidence threshold of 75%, as this value consistently resulted in better performance during our testing.

4. Results

This section presents the results that were obtained on the evaluation set during development, and the final results on the test set.

4.1. Development dataset

To identify the most effective augmentation strategy before the official submission, we conducted a series of experiments on the development set, evaluating six configurations, as well as setting a baseline for comparisons. Table 4 reports macro and micro precision, recall, and F1 scores averaged over three training runs for each configuration, except where otherwise stated. The exceptions are due to computational constraints during development. Table 6 reports the individual run scores for the configurations that were evaluated on three different runs. Predictions are generated with a confidence threshold of 60%, the same as the organizer baseline.

The baseline model fine-tuned on Gold data alone achieves a Micro-F1 of 0.7886 and a Macro-F1 of 0.7249. Applying synonym substitution at a 15% probability (SYN-0.15) yields the best overall performance among the synonym-only configurations, improving Micro-F1 to 0.8025 and Macro-F1 to 0.7296. Increasing the substitution probability to 50% (SYN-0.5) leads to a drop in performance across all metrics, falling below the baseline on Macro-F1 (0.7188), suggesting that aggressive non-entity substitution introduces too much noise into the training data.

The AggressiveMix configuration, which combines synonym substitution (15%) with token swapping (10%) and token deletion (5%), also underperforms relative to the baseline, achieving the lowest Macro-F1 (0.7112) out of all configurations. This indicates that additional transformations do not contribute positively and may disrupt the syntactic structure that the model relies on for entity detection. Due to this poor result, we decided not to consider this configuration for our official submissions.

Expanding the training data by including Silver-tier annotations alongside Gold, both augmented with synonym substitution at 15% (SYN-Gold+Silver), results in the highest Macro-recall (0.7741) and Macro-F1 (0.7445) across all configurations, though its Micro-F1 (0.7970) falls slightly below SYN-0.15. This suggests that the additional Silver data improves the model’s ability to generalize across entity types, at a slight cost in overall token-level precision.

The embedding-based configurations, evaluated on a single run, show mixed results. EMB (Gold) achieves a Micro-F1 of 0.7964 and a Macro-F1 of 0.7234, performing comparably to the baseline, with a small improvement to the Micro-F1. The combined SYN+EMB-0.15 (Gold) configuration achieves the highest Micro-precision (0.8081) across all configurations, but its Macro-F1 (0.7152) falls below the

baseline, indicating that the combined strategy produces more precise predictions overall but struggles to generalize across all entity types equally.

Configuration	Macro-P	Macro-R	Macro-F1	Micro-P	Micro-R	Micro-F1
Organizer baseline (Gold+Silver)	0.7198	0.7765	<u>0.7438</u>	0.771	0.8278	<u>0.7984</u>
Our Baseline (Gold)	0.7221	0.7344	0.7249	0.7820	0.7956	0.7886
SYN-0.15 (Gold)	<u>0.7419</u>	0.7346	0.7296	<u>0.8017</u>	0.8034	0.8025
SYN-0.5 (Gold)	0.7348	0.7217	0.7188	0.7898	0.7792	0.7842
AggressiveMix (Gold)	0.7168	0.7191	0.7112	0.7857	0.7933	0.7895
SYN-0.15 (Gold+Silver)	0.7236	<u>0.7741</u>	0.7445	0.7803	<u>0.8144</u>	0.7970
EMB (Gold)*	0.7411	0.7226	0.7234	0.7944	0.7985	0.7964
SYN+EMB-0.15 (Gold)*	0.7456	0.7031	0.7152	0.8081	0.7882	0.7980

Table 4

Development set results for each augmentation configuration. Results are averaged across 3 runs except where indicated. *Single run, not averaged across 3 runs

4.2. Official test dataset

Table 5 presents the official results of our five submitted runs on the test set, alongside the organizer-provided baseline for reference. The organizer baseline, trained on both Gold and Silver datasets, outperforms all of our submitted runs on both Macro-F1 (0.7267) and Micro-F1 (0.7996), as well as on Micro-recall (0.8221). However, our augmented runs consistently outperform our own unaugmented baseline (Run 1) across most metrics, confirming that data augmentation provides a meaningful benefit over no augmentation at all.

Overall, the most successful data augmentation approach was synonym-based augmentation. Run 2 (SYN-Gold) achieves the highest Macro-F1 (0.6964) and the second highest Micro-F1 (0.7945) of all Gold-only augmented runs, while Run 3 (SYN-Gold+Silver) achieves the highest Macro-F1 (0.7025) and Macro-recall (0.6997) of all our submissions, suggesting that the additional Silver data helps the model generalize better across all entity types. Run 4 (EMB-Gold) achieves the highest Micro-F1 (0.7952) and the second highest Micro-precision (0.8056) among our runs. Run 5 (SYN+EMB-Gold) has the highest Micro-precision out of all the submitted runs (0.8163), but at the cost of Macro-F1 (0.6738) and Macro-recall (0.6444), the lowest among our augmented runs. This is consistent with the pattern observed during development where the combined strategy tends to produce more precise predictions at the cost of coverage, particularly across less frequent entity types. Run 1 (Baseline), with no augmentation, performs the worst across almost all metrics, with the lowest Macro-F1 (0.6526) and Micro-recall (0.7220) of all our submissions.

Run	Macro-P	Macro-R	Macro-F1	Micro-P	Micro-R	Micro-F1
Organizer baseline	0.7114	0.7480	0.7267	0.7782	0.8221	0.7996
Run 1 (Baseline)	<u>0.7194</u>	0.6205	0.6526	0.7935	0.7220	0.7561
Run 2 (SYN-Gold)	0.7175	0.6839	0.6964	0.8033	<u>0.7858</u>	0.7945
Run 3 (SYN-Gold+Silver)	0.7085	<u>0.6997</u>	<u>0.7025</u>	0.7911	0.7746	0.7828
Run 4 (EMB-Gold)	0.7086	0.6733	0.6848	<u>0.8056</u>	0.7850	<u>0.7952</u>
Run 5 (SYN+EMB-Gold)	0.7305	0.6444	0.6738	0.8163	0.7706	0.7928

Table 5

Official test dataset results for each submitted run

While no runs surpassed the organizer baseline in Micro-F1, the reference metric, the runs achieving the closest results to the baseline were the Gold training data only runs. Run 2 (SYN-Gold, Micro-F1: 0.7945) and Run 4 (EMB-Gold, Micro-F1: 0.7952) were both within approximately 0.005 Micro-F1 points of the baseline (0.7996). Considering that the baseline was trained on both the Gold and Silver datasets and our closest runs trained on Gold only with augmentations, we find this to be a promising result. It

suggests that the appropriate selection of data augmentation techniques can partially mitigate the loss of additional annotated data and that the improvement of these techniques can close this gap further.

Also of note is the fact that our predicted values were generated at a confidence threshold of 75%, as mentioned in section 3, compared to the baseline’s threshold of 60%. A higher confidence threshold will produce higher precision, but lower recall, which explains why our runs consistently score higher than the baseline with respect to Micro-precision, while they fall below the baseline in respect to Micro-recall. Run 5 specifically, achieves a much higher Macro-precision (0.7305), even when comparing to our other runs, while Micro-precision (0.8163) is still higher than the baseline, but closer to the rest of the runs. This trade-off might be useful for biomedical information extraction. Although missed entities might result in incomplete knowledge extraction, false positive predictions can propagate errors to downstream applications. Depending on the target application, our observed increase in precision can be advantageous.

5. Conclusion

In this paper, we presented our approach to the GutBrainIE 2026 shared task, exploring three data augmentation strategies for biomedical Named Entity Recognition: synonym substitution (SYN), entity replacement based on Word2Vec embeddings (EMB), and a combination of both (SYN+EMB). These strategies were applied to the NuNER Zero model, fine-tuned on the Gold tier of the GutBrainIE dataset.

Our development experiments showed that moderate synonym substitution (SYN-0.15) consistently outperformed more aggressive augmentation strategies, and that adding Silver-tier data alongside Gold with synonym augmentation yielded the best macro-level generalization. Aggressive transformations such as higher substitution probabilities and token deletion or swapping consistently degraded performance, suggesting that preserving syntactic structure is important for NER augmentation.

On the official test set, our augmented runs consistently outperformed our unaugmented baseline, confirming the benefit of data augmentation for this task. Synonym-based augmentation proved to be the most effective strategy overall, with Run 2 (SYN-Gold) and Run 3 (SYN-Gold+Silver) achieving the best balance across metrics. While none of our runs surpassed the organizer baseline, the closest results were achieved by Gold-only augmented runs, which is encouraging given that the organizer baseline was trained on a larger dataset combining both Gold and Silver data.

Our current approach has some limitations worth noting: the Word2Vec embeddings used for entity replacement are trained on a general web corpus (Common Crawl), which may result in entity substitutions that are semantically inappropriate in a biomedical context. Additionally, the EMB and SYN+EMB configurations were evaluated on a single training run rather than averaged over multiple runs, which reduces the reliability of those specific results. As future work, we intend to explore more sophisticated augmentation strategies, including generative approaches using large language models, which have shown promise in producing diverse and domain-coherent training examples. This work serves as a foundation for our broader research on data augmentation for biomedical information extraction.

Declaration on Generative AI

During the preparation of this work, the authors used Claude and Grammarly in order to: Grammar and spelling check, and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

Acknowledgments

This work is supported by NOVA LINCS (UID/04516/2025) with the financial support of FCT/IP.

References

- [1] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodríguez-López, N. Loukachevitch, I. Rozhkov, E. Tutubalina, D. Dimitriadis, V. Patsiou, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2026: The fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, volume Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026) of *Lecture Notes in Computer Science*, Springer, 2026.
- [2] M. Martinelli, V. Bonato, G. M. Di Nunzio, G. Faggioli, N. Ferro, O. Irrera, B. Kantz, S. Marchesin, L. Menotti, S. Merlo, R. Michelotto, G. Tussardi, F. Vezzani, P. Waldert, G. Silvello, Overview of GutBrainIE@CLEF 2026: Gut-Brain Interplay Information Extraction, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
- [3] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, 2013. URL: <http://arxiv.org/abs/1301.3781>. doi:10.48550/arXiv.1301.3781, arXiv:1301.3781 [cs].
- [4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. URL: <http://arxiv.org/abs/1810.04805>. doi:10.48550/arXiv.1810.04805, arXiv:1810.04805 [cs].
- [5] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* 36 (2020) 1234–1240. URL: <https://doi.org/10.1093/bioinformatics/btz682>. doi:10.1093/bioinformatics/btz682.
- [6] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 5364–5376. URL: <https://aclanthology.org/2024.naacl-long.300/>. doi:10.18653/v1/2024.naacl-long.300.
- [7] S. Bogdanov, A. Constantin, T. Bernard, B. Crabbé, E. Bernard, Nuner: Entity recognition encoder pre-training via llm-annotated data, 2024. URL: <https://arxiv.org/abs/2402.15343>. arXiv:2402.15343.
- [8] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A Survey of Data Augmentation Approaches for NLP, 2021. URL: <http://arxiv.org/abs/2105.03075>. doi:10.48550/arXiv.2105.03075, arXiv:2105.03075 [cs].
- [9] A. Dao, H. Teranishi, Y. Matsumoto, F. Boudin, A. Aizawa, Overcoming data scarcity in named entity recognition: Synthetic data generation with large language models, in: Proceedings of the 24th Workshop on Biomedical Language Processing, 2025, pp. 328–340.
- [10] G. A. Miller, WordNet: A lexical database for English, in: Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994, 1994. URL: <https://aclanthology.org/H94-1111/>.

A. Appendix

Configuration	Run	Macro-P	Macro-R	Macro-F1	Micro-P	Micro-R	Micro-F1
Our Baseline (Gold)	Run 1	0.7121	0.7398	0.7221	0.7805	0.8040	0.7921
	Run 2	0.7190	0.7329	0.7226	0.7715	0.8036	0.7873
	Run 3	0.7353	0.7304	0.7301	0.7939	0.7791	0.7864
SYN-0.15 (Gold)	Run 1	0.7381	0.7335	0.7280	0.8007	0.8017	0.8012
	Run 2	0.7586	0.7266	0.7320	0.8093	0.7961	0.8026
	Run 3	0.7291	0.7437	0.7289	0.7950	0.8124	0.8036
SYN-0.5 (Gold)	Run 1	0.7148	0.7324	0.7130	0.7842	0.8100	0.7969
	Run 2	0.7460	0.7133	0.7192	0.8055	0.7985	0.8020
	Run 3	0.7437	0.7193	0.7243	0.7798	0.7291	0.7536
AggressiveMix (Gold)	Run 1	0.7224	0.7205	0.7156	0.7944	0.7925	0.7935
	Run 2	0.7193	0.7116	0.7077	0.7844	0.7921	0.7882
	Run 3	0.7087	0.7253	0.7103	0.7783	0.7953	0.7867

Table 6

Individual run results for each averaged configuration. The individual runs of the "SYN-0.15 (Gold+Silver)" run are not shown because they were not saved.