Gender Swapping as a Data Augmentation Technique: Developing Gender-Balanced Datasets for Ukrainian Language Processing

Olha Nahurna
Ukrainian Catholic University
Lviv, Ukraine
onahurna@gmail.com

Mariana Romanyshyn
Grammarly
Kyiv, Ukraine
mariana.romanyshyn@grammarly.com

Abstract

This paper presents a pipeline for generating gender-balanced datasets through sentence-level gender swapping, addressing the gender-imbalance issue in Ukrainian texts. We select sentences with gender-marked entities, focusing on job titles, generate their inverted alternatives using LLMs and human-in-the-loop, and fine-tune Aya-101 on the resulting dataset for the task of gender swapping. Additionally, we train a Named Entity Recognition (NER) model on gender-balanced data, demonstrating its improved ability to recognize gendered entities. The findings unveil the potential of gender-balanced datasets to enhance model robustness and support more fair language processing. Finally, we make a gender-swapped version of NER-UK 2.0 and the fine-tuned Aya-101 model available for download and further research.

1 Introduction

The Ukrainian language has historically exhibited a lack of gender balance in professional titles, with masculine forms traditionally dominating. To address this imbalance, the 2019 revision of Ukrainian orthography[1] introduced official guidelines on the word formation of feminitives — feminine forms of personal nouns. Although these changes aim to promote more inclusive and gender-balanced language, their implementation remains relatively recent and, at times, controversial (Starko, 2024).

A study of trends in the usage of feminine personal nouns (Starko and Synchak, 2023) reveals that prior to 2019, their presence in Ukrainian corpora was minimal. Many texts used masculine forms even in contexts where grammatical gender agreement required a feminine equivalent. For example, in the sentence "Мені допомогла Оксана Миколаївна, вона найкраща лікар у місті." (en: *Oksana Mykolaivna helped me, she is the best doctor in the city.*). The word "лікар" (*male doctor*) should be replaced with "лікарка" (*female doctor*) for grammatical agreement.

The low representation of feminitives in existing corpora has resulted in limited availability of training data containing feminine personal noun forms. At the same time, there is a growing demand for NLP models that can accurately recognize, interpret, and generate gender-marked language in Ukrainian. To address this challenge, we propose a gender swapping pipeline designed to facilitate the creation of gender-balanced datasets through data augmentation.

The rest of the paper is organized as follows. In Section 2, we review existing research on gender bias in NLP and methods for achieving gender balance in data. Section 3 presents the gender-swapping pipeline for generating gender-parallel sentences. Section 4 covers two experiments: the first one focuses on fine-tuning a Large Language Model (LLM) using a gender-parallel dataset, and the second one investigates whether training a Named Entity Recognition (NER) model on gender-balanced data can improve the recognition of gendered entities. The paper ends with conclusions, limitations, and ethical considerations.

2 Related Work

This section reviews the existing research on gender bias in NLP and gives an overview of solutions for achieving gender balance in text corpora.

2.1 Gender Bias in NLP

NLP systems can inherit and reinforce different types of biases present in their training data, promoting societal inequalities associated with gender, religion, ethnicity, age, and other sensitive characteristics (Gallegos et al., 2023). Using gender-biased training data may perpetuate prevalent gen-

[1] https://mon.gov.ua/osvita-2/zagalna-serednya-osvita/ukrainskiy-pravopis

Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 147–161
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

der stereotypes and cause significant implications for fairness (Leong, 2024). A notable example is gender bias exhibited in recruitment processes and data, which is promoted to AI-driven recruitment systems trained on this data (Chang, 2023; Dikshit et al., 2024; Mujtaba and Mahapatra, 2024).

While a significant number of studies have been conducted on detecting and reducing gender bias in English (Chaloner and Maldonado, 2019; Nakanishi, 2024; Li and Zhang, 2024), addressing bias in morphologically rich languages remains underresearched.

Languages with notional gender—ones that do not mark grammatical gender—use straightforward solutions to address gender bias. One of the most widely used approaches in such languages is the creation of dictionaries containing gender-marked word pairs, typically consisting of masculine-feminine counterparts (Lund et al., 2023; Lu et al., 2019). However, such approaches can only be partially applied to inflected languages, where agreement with gender is crucial for forming grammatically correct sentences.

Morphologically rich languages, such as Spanish (Jain et al., 2021), Arabic (Habash et al., 2019), French (Gygax et al., 2012), and Slovenian (Ljubi et al., 2022), present new challenges. They use gender encoding not only for pronouns but also for verbs, nouns, and adjectives to ensure agreement across multiple parts of a sentence. As a result, mitigating gender bias in these languages needs advanced approaches that account for their linguistic features.

2.2 Methods for Achieving Gender Balance in Data

Ensuring gender balance in data reduces bias and promotes fairness in subsequent AI and NLP models. An effective strategy is to use **gender-fair** language (Sczesny et al., 2016), which minimizes manipulation with gender stereotypes and ensures equitable representation. Gender-fair language practices include gender neutralization and gender-marked data augmentation.

2.2.1 Gender-Neutralization

Gender-neutral forms are becoming increasingly useful, providing an effective alternative in contexts where specifying gender is unnecessary (Stanczak and Augenstein, 2021). Cetnarowska (2023) found that for people who learn English as a second language, gender-marked occupational terms such as *policeman* or *postman* can cause challenges in understanding the true meaning. The *"-man"* part may be misinterpreted as signifying that these professions are exclusively for men. In response to this challenge, Bartl and Leavy (2024) developed a catalog of 692 gender-exclusive terms along with gender-neutral variants, manually verified and further validated using sources such as WordNet, Wikidata, and Wikipedia. This catalog was subsequently used to construct a gender-inclusive fine-tuning dataset.

Replacing gender-marked words with gender-neutral forms can enhance clarity, promote inclusivity for both binary and non-binary individuals, and reduce gender bias in NLP systems (Sobhani et al., 2023). However, this approach is not applicable to languages that mandate the use of masculine or feminine grammatical gender for person nouns and contextual grammatical agreement.

2.2.2 Gender-Marked Data Augmentation

Gender-marked data augmentation means creating additional variations of sentences to reflect different grammatical genders.

Counterfactual Data Augmentation (CDA) is an approach that augments training data by altering gender-marked terms to their counterparts (e.g., replacing "he" with "she"). This approach aims to disrupt perpetual associations for gender-marked words (Lu et al., 2019).

Initially, CDA techniques focused on rule-based gender swapping, relying on dictionaries of masculine-feminine word pairs. However, this approach has two main limitations: (1) bounded dictionaries, which are usually unable to cover all gender-marked words in the language, and (2) non-preservation of grammatical agreement with the replacement word. Unlike English, in morphologically rich languages, the default method of word swapping without contextual grammatical agreement would often yield grammatically incorrect structures. To address this issue, Zmigrod et al. (2019) proposed an approach that uses Markov random field with an optional neural parameterization to correct agreement after word swapping. This method has been successfully applied to create Spanish, Hebrew, French, and Italian datasets.

Another improvement of CDA proposed by Lund et al. (2023) includes part-of-speech (POS) tagging and the resolution of agreement issues with the help of a dependency parser. This solution was developed to implement augmentation for the sin-

gular "they" in English.

CDA with LLMs addresses the limitations of the classical CDA and can be used to facilitate the creation of gender-balanced data for morphologically rich languages. The core principle remains unchanged: an LLM is prompted with an original sentence and instructions to generate a gender-swapped equivalent. LLMs were designed to perform text generation tasks, which makes this approach promising, but their performance may be hindered by hallucinations and reduced accuracy, particularly in low-to-mid-resource languages. To mitigate these challenges, fine-tuning the LLM on a gender-parallel dataset can significantly improve its ability to produce correct and contextually appropriate gender-swapped forms (Bartl and Leavy, 2024).

Fairflow is a low-resource method designed to overcome the limitations of the rule-based CDA and the lack of parallel data for LLM fine-tuning (Tokpo and Calders, 2024). Fairflow proposes an end-to-end pipeline that begins with identifying gender-marked words in text using a pre-trained BERT embedding model (Devlin et al., 2019). It then employs a Disentangling Invertible Interpretable Network (DIIN) (Esser et al., 2020) to generate counterfactual equivalents for each word. Finally, an error correction scheme is applied to generate parallel data that maintains correct structure and agreement. However, this method has been developed and tested only for the English language.

3 Proposed Solution

Since Ukrainian is a morphologically rich language with the category of grammatical gender, we focus on using CDA with LLMs to create gender-balanced datasets. To the best of our knowledge, no prior research has explored this approach for the Ukrainian language.

This section describes the key components of the proposed gender swapping pipeline applicable to morphologically rich languages (see Figure 1 for visualization).

3.1 Dataset Selection

The first step involves compiling a dataset of sentences with gendered entities. The set of gendered entities depends on the language and may include person names, pronouns, and gendered personal nouns that describe a person's occupation, ethnicity, political views, character, etc. The detection of such entities in the pre-selected sentences can be performed manually, automatically via dictionaries of gendered terms or POS taggers, or with the help of a NER system, if available.

3.2 Gender-Swapping with LLM

The next step is prompting an LLM to perform sentence-level gender swapping on the collected sentences with gendered entities. The prompt should instruct the model to switch masculine entities with feminine ones and vice versa while ensuring that gender-neutral entities remain unchanged and the related words are updated for grammatical agreement.

To ensure that the required entities are addressed in the generation, we propose feeding the annotated entities to the prompt. Additionally, to minimize potential bias in person names generated by the model, we propose adding a list of random male and female names in the target language to the prompt.

3.3 Human in the Loop

The LLM-generated gender-swapped sentences likely contain errors, such as misspelled words, inconsistent grammatical agreement, and incorrectly swapped entities, rendering this dataset of "bronze quality". To address this, we propose adding a human-in-the-loop step, where human judges can review the generated data and accept, correct, or dismiss the generated output, which results in a "silver-quality" dataset.

3.4 LLM Fine-Tuning

Finally, the parallel sentence dataset can be split into train and test subsets, and the train part can be used for LLM fine-tuning. It is essential to choose a model that is suitable for the target language and capable of handling instruction-based tasks.

3.5 Evaluation

To assess the model's quality, we propose the following metrics:

- **Exact Match**: Measures the fraction of exact matches between the test set and LLM-generated sentences.
- **BLEU** (Papineni et al., 2002): Evaluates the n-gram overlap between the test and LLM-generated sentences. It provides insight into

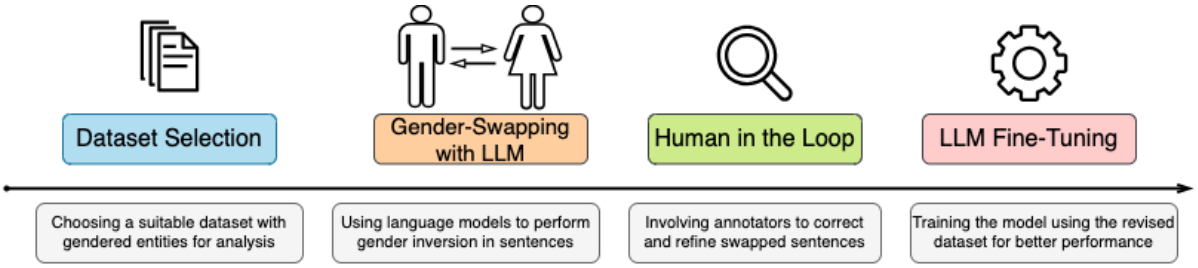

Figure 1: Gender-Swapping Pipeline

how well the model preserves sentence structure and meaning while swapping gendered entities.

- **ROUGE-L** (Lin, 2004): Assesses the longest common subsequence between the test and LLM-generated sentences. It measures the overall similarity of the sentences, ensuring that the meaning and structure are retained after gender inversion.
- **BERTScore (F1)** (Zhang et al., 2020): Based on contextual embeddings, measures the semantic similarity between the test and LLM-generated sentences. In contrast to BLEU and ROUGE-L, this metric produces a high similarity score for different forms of the same word.
- **Token Count Match Rate**: Measures the consistency in token length between the test and LLM-generated sentences. It helps ensure that the sentence length remains the same after gender swapping.

4 Experiments

This section provides a deep dive into the experimental part of our research. First, in Section 4.1, we apply the proposed gender-swapping pipeline to build a gender-swapped dataset and a gender-swapping model for the Ukrainian language. We specifically focus on applying this data augmentation technique to the Ukrainian language job titles, considering the recent shift in the use of feminitives in Ukrainian and their low representation in Ukrainian language corpora. Then, in Section 4.2, we use the developed dataset to train a NER model for Ukrainian.

4.1 Gender Swapper UK

4.1.1 Dataset Selection for Ukrainian

We selected the NER-UK 2.0 corpus (Chaplynskyi and Romanyshyn, 2024) as our data source because it is the largest dataset manually annotated for the named entity recognition task in the Ukrainian language. The corpus contains annotations for such gendered entities as person names (PERS) and job titles (JOB), which creates a solid basis for our research. NER-UK 2.0 consists of two subcorpora: Nashi Groshi, entity-rich news texts on the Ukrainian economy and anticorruption efforts, and multi-genre BRUK. The corpus was annotated for thirteen entity types. See Appendix A for the distribution of all entity labels in the corpus.

We extracted all sentences with JOB entities from NER-UK 2.0 to form our dataset for gender swapping. This sampling resulted in 1,513 sentences with 1,982 JOB entities and 1,384 PERS entities.

Additionally, in collaboration with GenderGid[2], we released a Ukrainian gender-paired dictionary of 1,102 entries[3]. Compiled and validated by domain experts, this resource provides high-quality, linguistically accurate masculine–feminine word pairs. The dictionary serves as the foundation for tools that automate gender classification and validation.

Subsequently, we classified all JOB and PERS entities in the dataset by grammatical gender (*masculine*, *feminine*, or *common*[4]) using lemmatization with pymorphy3[5], syntax parsing with stanza[6], and lookup to the gender-paired dictionary. All entities that could not be reliably assigned a grammatical gender were marked as *unknown*. The code for gender classification, along with all other scripts used in this research, is available in our GitHub

[2] https://gendergid.org.ua/pro-nas/

[3] https://github.com/lang-uk/uk-gender-word-mapper

[4] Common-gender words in Ukrainian agree in grammatical gender with masculine, feminine, and, in some cases, neuter word forms. Examples: суддя (en: *judge*), голова (en: *head*), листоноша (en: *mailperson*).

[5] https://pypi.org/project/pymorphy3/

[6] https://stanfordnlp.github.io/stanza/

Dataset	Total	Masculine		Feminine		Common		Unknown	
		Count	Fraction	Count	Fraction	Count	Fraction	Count	Fraction
Nashi Groshi (JOB)	1,344	1,135	84.4%	27	2.0%	170	12.6%	12	0.9%
BRUK (JOB)	638	511	80.0%	49	7.6%	53	8.3%	25	3.9%
Total (JOB)	1,982	1,646	**83%**	76	**3.8%**	223	**11.3%**	37	**1.8%**
Nashi Groshi (PERS)	1,058	526	49.7%	180	17.0%	0	0%	352	33.3%
BRUK (PERS)	326	141	43.2%	51	15.6%	0	0%	134	41.1%
Total (PERS)	1,384	667	**48.2%**	231	**16.7%**	0	0%	486	**35.1%**

Table 1: Gender composition of JOB and PERS entities in the initial dataset sampled from NER-UK 2.0.

repository[7]. A detailed distribution of gendered entities in the dataset is presented in Table 1.

The grammatical gender classification revealed a severe gender bias in JOB entities, with 82.9% masculine and 3.7% feminine forms, although the fraction of female names in the dataset is higher: 48.2% male vs. 16.7% female names. The high fraction of PERS entities of unknown gender is due to PERS-labeled surnames that can be morphologically ambiguous in Ukrainian. Broader context analysis could mitigate this issue, and we leave it for future work.

4.1.2 Gender-Swapping with LLM

We selected GPT-4o-mini (OpenAI, 2024) as the language model for generating gender-parallel sentence pairs. To enhance generation quality, we engineered a prompt that included clear instructions, transformation rules, and constraints. We employed a few-shot learning approach by providing several manually designed gender-swapped examples.

To reduce potential bias in name generation, each prompt was supplemented with a set of three male and three female names. These names were randomly sampled from the frequency dictionary of Ukrainian names[8].

The input data for the prompt consisted of a sentence and a list of gendered entities in this sentence, together with their types.

We provide the resulting prompting template on GitHub[9].

4.1.3 Human in the Loop

We invited sixteen native speakers of Ukrainian from the Ukrainian NLP community to review the generated sentences. We provided the annotators with the original sentence, the GPT-generated gender-swapped sentence, and target entities, and asked them to *accept*, *correct*, or *dismiss* the generated output. The *dismiss* category covered complex cases where it was challenging to identify an appropriate gender-inverse counterpart for the target word or to determine whether the sentence required any modification at all.

The annotators received detailed instructions outlining the step-by-step revision process, including grammatical constraints, links to external dictionaries, and examples, to support accurate and consistent judgments. We publish the annotation guidelines in English and Ukrainian on our GitHub[10].

As a result, we obtained the following evaluation statistics:

1. *to Accept:* 58.5% of the generated sentences did not need any correction.
2. *to Correct:* 37.6% of the generated sentences were updated by annotators.
3. *to Dismiss:* 3.9% of examples were dismissed as complex or ambiguous.

After manually reviewing the sentences marked as *to Correct*, we identified several types of gender-swapping mistakes: unnecessary changes to names or unrelated nouns, incorrect swapping of plural job titles, and cases where the original sentence already had mismatched gender forms. We also observed hallucinated or rare name substitutions, failures in gender agreement, and invalid or non-existent feminine forms of job titles. Refer to Appendix B for examples of the mistakes.

To further assess the consistency of gender-swapped outputs, we calculated the token count match rate between original, GPT-generated, and manually reviewed sentences (see Table 2). The results demonstrate that the majority of generated sentences closely follow the original token structure, suggesting reliable performance in maintaining sentence structure during gender inversion. However,

[7] https://github.com/linndfors/ner_for_fem

[8] https://github.com/lang-uk/name_freq_dict_uk

[9] https://github.com/linndfors/ner_for_fem/blob/main/data/prompt.txt

[10] https://github.com/linndfors/ner_for_fem/blob/main/annotation_project/annotation_instruction.txt

the need for manual corrections in over one-third of cases highlights the complexity of the gender-swapping task for LLMs.

Dataset Pair	Token Count Match
Original vs GPT	0.97
GPT vs Annotated	0.96
Original vs Annotated	0.95

Table 2: Token count consistency across original, GPT-generated, and corrected sentences.

To ensure dataset consistency, we removed all pairs marked as *Dismiss* and filtered out 4% of duplicates where the generated sentence was annotated as *Correct*, but matched the original without gender modifications. After the filtering, the final dataset contained 1,403 parallel sentence pairs, with 1,733 JOB entities and 1,282 PERS entities.

Additionally, we evaluated the correctness of gender-swapped JOB entities, using the above-mentioned GenderGid dictionary of gendered word pairs. Specifically, we extracted all JOB entity pairs from the parallel sentences, formed candidate pairs, and checked whether these pairs were present in the dictionary. The results showed that 83% of pairs could be found in the dictionary, which we consider a good indicator of the data quality.

In Figure 2, we provide an example of an original and gender-swapped sentence pair. After gender swapping, the job title Черговий лікар (en: *male doctor on duty*) changes to Чергова лікарка (en: *female doctor on duty, feminine form*), and the connected verb поінформував (en: *informed , masculine form*) changes to поінформувала (en: *informed, feminine form*).

Original Sentence:

Черго**вий** лік**ар** ще вночі ґрунтовно поінформув**ав** про перспективи одужання.

*(en: At night, the **doctor on duty** (masculine form) thoroughly **informed** (masculine form) me about the prospects for recovery.)*

Gender-Swapped Sentence:

Черго**ва** лік**арка** ще вночі ґрунтовно поінформув**ала** про перспективи одужання.

*(en: At night, the **doctor on duty** (feminine form) thoroughly **informed** (feminine form) me about the prospects for recovery.)*

Figure 2: An example of a gender-swapped sentence.

4.1.4 LLM Fine-Tuning

We selected Aya-101 (Üstün et al., 2024) for further experiments on fine-tuning. Aya-101 is a multilingual instruction-tuned model supporting 101 languages, including Ukrainian. Its instruction-based architecture makes it particularly well-suited for the gender-swapping task. Additionally, Aya-101 has been previously successfully applied to other text editing tasks in Ukrainian (Saini et al., 2024).

We fine-tuned Aya-101 with two instructions:

- Перефразуй це речення, змінивши його гендерні сутності на протилежні (чоловічий <-> жіночий) (en: *Perform gender inversion on the sentence below by swapping gender-marked entities (masculine <-> feminine)*). We split our parallel gender-swapped dataset to train and test sets, and used the train set examples as input for this instruction.
- Перефразуй це слово, змінивши його гендер на протилежний (чоловічий <-> жіночий) (en: *Perform gender inversion on the word below (masculine <-> feminine)*). Here, we used random word pairs from the GenderGid gendered word pair dictionary.

We additionally mixed the order of sentences in the parallel train dataset and the order of words in the gendered word pairs to balance them and avoid bias, as most of the samples were originally rewritten from masculine to feminine. The training process was conducted using a Parameter-Efficient Fine-Tuning (PEFT) framework (Mangrulkar et al., 2022) with the Quantized Low-Rank Adapter (QLoRA) technique (Dettmers et al., 2023) applied to the base model Aya-101 (13B). The training was performed on an A100 GPU using Google Colab Pro+, with a batch size of 4, a learning rate of 5e-5, and the `AdamW` optimizer (Loshchilov and Hutter, 2019). The model and the corresponding dataset of parallel sentence pairs are publicly available via Hugging Face[11].

4.1.5 Gender-Swapping Model Evaluation

We complement the metrics proposed in Section 3.5 with two more task-specific metrics:

- **JOB Match**: Evaluates the fraction of matched job titles in the test and LLM-generated sentences.

[11] https://huggingface.co/linndfors/uk-gender-swapper-aya-101

Metric	Aya-101 original	Aya-101 fine-tuned	GPT-4o-mini
Exact Match	0.15	0.44	**0.45**
Exact Match w/o PERS	0.22	**0.64**	0.62
JOB Match	0.30	**0.82**	0.65
BLEU	0.65	**0.82**	**0.82**
ROUGE-L	0.21	0.21	**0.22**
BERTScore (F1)	0.97	**0.99**	**0.99**
Token Count Match	0.69	**0.92**	0.91

Table 3: Evaluation of LLMs performing the gender-swapping task on the parallel gender-swapped test set.

- **Exact Match w/o PERS**: Measures the fraction of exact matches between the test set and LLM-generated sentences, ignoring person names, which are randomly generated by the model.

We evaluated the original Aya-101, the fine-tuned Aya-101, and GPT-4o-mini on the test part of our parallel gender-swapped dataset. Table 3 demonstrates that the fine-tuned model showed a substantial performance improvement over the original Aya-101 and achieved results compatible with the much larger GPT-4o-mini model.

Additionally, we conducted an experiment to evaluate whether performing a round-trip gender swapping (i.e., swapping the gendered entities forth and back) would reconstruct the original sentence. The results are consistent with the one-way gender swapping performance (see Appendix C: Table 8).

Name Generation Bias in LLM: While generating the initial gender-swapped sentences via GPT, we used the frequency dictionary of Ukrainian names to provide name options to the model and minimize the potential name bias. As a result, the name distribution in the dataset reflected the frequency distribution of Ukrainian names. However, during the evaluation of Aya-101 fine-tuned on our dataset, we discovered a significant distributional bias in female name generation. Specifically, the name Наталія (en: *Natalia*) accounted for 25% of all generated female names. This suggests that the model exhibits name bias, presumably inherited from the pretraining data. The model also occasionally hallucinates non-existent names. Addressing these issues remains an area for future work.

4.2 Enhancing Gender-Marked Entity Recognition

NER models are known to exhibit demographic bias because they are trained on imbalanced datasets. Even when tested on synthetic data representing different ethnicities and genders, the best-recognized names are predominantly *"white male names"* (Mishra et al., 2020). Moreover, Mehrabi et al. (2020) discovered that female names are more frequently missed or misclassified as LOCATION by NER models compared to male names. Such examples emphasize the challenge posed by the lack of gender-balanced training data in NER models.

To assess the potential of the generated gender-swapped dataset we obtained, we used it to train a NER model and compare it to the current state of the art. The goal of this evaluation was to determine whether the updated training data leads to improved recognition and classification of gender-marked entities, thereby enhancing the model's overall accuracy and robustness.

4.2.1 Gender-Balancing NER-UK 2.0

After inverting the original sentences from NER-UK 2.0, we used them to construct a gender-swapped NER-UK 2.0 subset, with the corresponding entity annotations carried over from the original text files. Since the swapped sentences contained changes in both the gendered entities and the forms of related words, which impacted the character-level sentence length, we recalculated the positions of entities in the swapped sentences. For easy tracking and future references, we also saved .meta files with sentence IDs of the original NER-UK 2.0 sentences that were used to create the gender-swapped NER-UK 2.0 subset. Finally, we preserve the train/test split from the original NER-UK 2.0. We make the gender-swapped NER-UK 2.0 subset accessible via GitHub[12].

Next, we merged the original NER-UK 2.0 dataset with the gender-swapped NER-UK 2.0. As a result, the dataset size increased. The number of JOB titles grew, but not exactly doubled, as some modified sentences were filtered out previously. Other entity types increased proportionally, as each gender-swapped sentence was included alongside its original.

Table 4 provides details about the distribution

[12] https://github.com/lang-uk/ner-uk/tree/master/v2.0-swapped

of entities in the original, gender-swapped, and augmented NER-UK 2.0.

Entity Type	Original	Gender-Swapped	Augmented
ART	635	48	683
DATE	2,047	374	2,421
DOC	142	18	160
JOB	**1,982**	**1,733**	**3,715**
LOC	3,000	341	3,341
MISC	515	35	550
MON	943	108	1,051
ORG	5,213	1,267	6,480
PCT	263	48	311
PERIOD	596	88	684
PERS	**6,235**	**1,282**	**7,517**
QUANT	382	40	422
TIME	40	3	43
Total	**21,993**	**5,385**	**27,378**

Table 4: Entity type distribution in the original, gender-swapped, and augmented NER-UK 2.0.

The augmented dataset contains a significantly better gender distribution across key entity types. The initial imbalance of 83% masculine vs. 3.8% feminine JOB entities was reduced to 49.2% masculine vs. 37.4% feminine. Similarly, for PERS entities, the distribution shifted from 34.0% masculine vs. 20.6% feminine to a more balanced 30.2% masculine vs. 26.8% feminine (see Appendix D: Tables 9 and 10).

4.2.2 NER Model Training

As our baseline for benchmarking, we use **uk_ner_web_trf_13class**, the current state-of-the-art NER model for Ukrainian[13] published with the NER-UK 2.0 paper. For fair comparison, we followed the configuration and training pipeline outlined in the paper. Specifically, we trained a classifier based on the Ukrainian version of the RoBERTa-large model (Minixhofer et al., 2022), using the spaCy[14] framework for implementation. We used the *augmented* NER-UK 2.0 train set for training.

4.2.3 NER Evaluation

Finally, we evaluated the two NER models — uk_ner_web_trf_13class (**Original NER**) and our newly trained gender-balanced NER model (**Gender-Balanced NER**)[15] — on three test sets: the original NER-UK 2.0 test set, the gender-swapped NER-UK 2.0 test set, and the augmented test set that combines them both. We provide the evaluation results for the JOB and PERS entity categories in Table 5 and detailed results on all entity types in Appendix E.

Focusing specifically on the JOB entity, the results show that the Gender-Balanced NER model improves performance on the gender-swapped test set, demonstrates slight gains on the augmented set, but exhibits a decline on the original set. In contrast, for PERS-labeled entities, no significant performance changes were observed likely due to their sufficient representation for both genders in the original dataset, which provided a strong foundation for learning.

To understand why the Gender-Balanced NER model shows lower results on the original test set but higher results on the gender-swapped test set, which predominantly contains feminine JOB entities, we conducted a follow-up evaluation in which these entities were split by gender.

$$\text{Recall}_g = \frac{|\text{TP}_g|}{|\text{TP}_g| + |\text{FN}_g|} \quad \text{for } g \in \{\text{male, female, common}\} \quad (1)$$

To evaluate the model's ability to recognize JOB entities, we used the Recall metric, which quantifies the proportion of actual entities correctly identified. Specifically, we extracted all True Positive (TP) and False Negative (FN) JOB entities, classified them by gender using the method described earlier, and calculated recall for each gender category using Formula 1. As shown in Table 6, when compared to the Original NER model, Gender-Balanced NER demonstrated a significant improvement in recognizing feminine JOB entities, maintained comparable performance for common gender titles, but exhibited a notable decline in recall for masculine entities. This inconsistency may stem from the altered gender distribution, a larger training corpus, and the original model's focus on masculine entities, which could reduce recognition of masculine job titles in Gender-Balanced NER. Future work will focus on optimizing configuration parameters to better align the model with the configuration characteristics of the revised dataset and improve performance across all gender categories.

Across the remaining NER classes, we observed overall performance improvements, with only minor exceptions where slight declines occurred. These variations may be attributed to overfitting introduced during dataset augmentation, particularly in cases where specific labeled entities were duplicated. To address this, future work could enhance the gender-swapping methodology by shifting from sentence-level to document-level transformations,

[13] https://huggingface.co/dchaplinsky/uk_ner_web_trf_13class

[14] https://spacy.io/

[15] https://huggingface.co/linndfors/ner-uk_for_gender-balanced_dataset

Test Set	Original NER				Gender-Balanced NER			
	Entity Type	P	R	F1	Entity Type	P	R	F1
Original	JOB	74.39	65.45	**69.64**	JOB	75.05	59.06	66.10
	PERS	96.20	96.60	**96.40**	PERS	97.01	95.18	96.08
Gender-swapped	JOB	89.08	71.26	79.18	JOB	90.63	80.78	**85.42**
	PERS	98.60	98.60	98.60	PERS	98.60	98.88	**98.74**
Augmented	JOB	80.53	67.75	73.59	JOB	82.51	68.43	**74.81**
	PERS	96.58	97.00	**96.79**	PERS	97.15	95.62	96.38

Table 5: Performance comparison of Original NER and Gender-Balanced NER for JOB and PERS entities across different test sets.

Category	Original NER	Gender-Balanced NER
Feminine recall	0.69	**0.80**
Masculine recall	**0.64**	0.59
Common-gender recall	0.85	**0.87**

Table 6: Recall comparison by gender category between Original NER and Gender-Balanced NER.

thereby fostering greater contextual diversity and consistency in the training data.

5 Conclusions

In this paper, we introduced a sentence-level gender-swapping pipeline that utilizes gender-marked data. Using this approach, we fine-tuned the Aya-101 model on a Ukrainian gender-parallel corpus, achieving substantial performance gains over the original Aya-101 and performance parity with GPT-4o-mini.

Furthermore, we trained a NER model on an augmented gender-balanced dataset, which led to improved recognition of feminine JOB entities. However, performance declined on the Original set, which predominantly contains masculine entities. These results highlight the potential of gender-balanced data to improve NER performance for underrepresented gender categories, while also revealing the difficulty of preserving consistent accuracy across differing gender distributions.

As part of this research, we have made several key contributions available to the community: (1) a dataset of parallel gender-swapped sentences, (2) a gender-swapped NER-UK 2.0 subset of sentences with job titles, and (3) a fine-tuned Aya-101 model capable of gender swapping sentences in the Ukrainian language.

6 Limitations and Future Work

The method presents the following limitations:

1. Our method currently works at the sentence level, which is contextually limited. Future work will focus on developing a more robust method for document-level gender swapping that takes into account broader context and minimizes errors.

2. We used a proprietary GPT-4o-mini model for the initial data generation, which may impact the reproducibility of our results.

3. Currently, the model has a significant bias in generated female names and may produce non-existent names. Therefore, future work will focus on developing a solution capable of selecting from a list of valid name variants, ensuring a close-to-life distribution of names in the gender-swapped sentences.

4. We focused our work on Ukrainian feminitives that denote occupations. Future work may validate the proposed approach on other gendered entities in the Ukrainian language, like personal nouns denoting ethnicity, religion, political views, character, etc. We also continue to explore alternative LLMs and refine training configurations to further improve performance and adaptability.

7 Ethical Considerations

The current model was trained on all available gender-marked sentences, enabling it to perform gender swapping on any sentence identified as gender-marked. However, this approach does not account for the broader contextual nuances, which may result in hallucinations and misinformation when an entity is not suitable for swapping (e.g., when the original sentence contains facts about public figures). In the future, we aim to enhance the model's ability to classify and manage cases where gender swapping is inappropriate or contextually incorrect.

Acknowledgments

We express our gratitude to the team of annotators who generously contributed their time and effort in

creating and validating the dataset used in this research. We thank the Faculty of Applied Sciences at the Ukrainian Catholic University for providing the computational resources necessary for the model training. Furthermore, we sincerely thank GenderGid for the gender-pair dictionary essential to our gender swapping method.

References

Marion Bartl and Susan Leavy. 2024. From 'showgirls' to 'performers': Fine-tuning with gender-inclusive language for bias reduction in LLMs. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 280–294, Bangkok, Thailand. Association for Computational Linguistics.

Bozena Cetnarowska. 2023. The use of gender-marked and gender-neutral forms: The importance of linguistic corpora in increasing the linguistic awareness of l2 learners of english. *Roczniki Humanistyczne*, 71:61–77.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. Association for Computational Linguistics.

Xinyu Chang. 2023. Gender bias in hiring: An analysis of the impact of amazon's recruiting algorithm. *Advances in Economics, Management and Political Sciences*, 23:134–140.

Dmytro Chaplynskyi and Mariana Romanyshyn. 2024. Introducing NER-UK 2.0: A rich corpus of named entities for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 23–29, Torino, Italia. ELRA and ICCL.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Malika Dikshit, Houda Bouamor, and Nizar Habash. 2024. Investigating gender bias in STEM job advertisements. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 179–189, Bangkok, Thailand. Association for Computational Linguistics.

Patrick Esser, Robin Rombach, and Björn Ommer. 2020. A disentangling invertible interpretation network for explaining latent representations. *Preprint*, arXiv:2004.13166.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, and et al. 2023. Bias and fairness in large language models: A survey.

Pascal Gygax, Ute Gabriel, Arik Lévy, Pool, Grivel, and Pedrazzini. 2012. The masculine form and its competing interpretations in french: When linking grammatically masculine role names to female referents is difficult. *Journal of Cognitive Psychology*, 24:395–408.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Nishtha Jain, Maja Popović, Declan Groves, and Eva Vanmassenhove. 2021. Generating gender augmented data for NLP. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 93–102, Online. Association for Computational Linguistics.

Sung A. Leong, K. 2024. Gender stereotypes in artificial intelligence within the accounting profession using large language models. page 11.

Yingjie Li and Yue Zhang. 2024. Pro-woman, anti-man? identifying gender bias in stance detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3229–3236, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. page 10.

Jasna Mikić Ljubi, Andra Matkovi, Jurij Bon, and Aleksandra Kanjuo Mrela. 2022. The effects of grammatical gender on the processing of occupational role names in slovene: An event-related potential study. *Frontiers in Psychology*, 13.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2019. Gender bias in neural natural language processing. *Preprint*, arXiv:1807.11714.

Gunnar Lund, Kostiantyn Omelianchuk, and Igor Samokhin. 2023. Gender-inclusive grammatical error correction through augmentation. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 148–162, Toronto, Canada. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2020. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, page 231–232, New York, NY, USA. Association for Computing Machinery.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing demographic bias in named entity recognition. *Preprint*, arXiv:2008.03415.

Dena F. Mujtaba and Nihar R. Mahapatra. 2024. Fairness in ai-driven recruitment: Challenges, metrics, methods, and future directions. *Preprint*, arXiv:2405.19699.

Takafumi Nakanishi. 2024. Detection of latent gender biases in data and models using the approximate generalized inverse method. In *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, pages 191–196.

OpenAI. 2024. Gpt-4o system card. OpenAI. https://cdn.openai.com/gpt-4o-system-card.pdf.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation.

Aman Saini, Artem Chernodub, Vipul Raheja, and Vivek Kulkarni. 2024. Spivavtor: An instruction tuned Ukrainian text editing model. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 95–108, Torino, Italia. ELRA and ICCL.

Sabine Sczesny, Magdalena Formanowicz, and Franziska Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in Psychology*, 7.

Nasim Sobhani, Kinshuk Sengupta, and Sarah Jane Delany. 2023. Measuring gender bias in natural language processing: Incorporating gender-neutral linguistic forms for non-binary gender identities in abusive speech detection. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1121–1131, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *Preprint*, arXiv:2112.14168.

Vasyl Starko. 2024. Problematic cases of forming personal feminine nouns in Ukrainian corpora and dictionaries. *Language: classic - modern - postmodern*, pages 99–117.

Vasyl Starko and Olena Synchak. 2023. Feminine personal nouns in Ukrainian: Dynamics in a corpus. In *International Conference on Computational Linguistics and Intelligent Systems*.

Ewoenam Kwaku Tokpo and Toon Calders. 2024. Fairflow: An automated approach to model-based counterfactual data augmentation for nlp. *Preprint*, arXiv:2407.16431.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

A NER-UK 2.0 Entity Type Distribution

Entity Type	Nashi Groshi	BRUK	Total
ART	319	316	635
DATE	1,496	551	2,047
DOC	108	34	142
JOB	1,344	638	1982
LOC	1,380	1,620	3,000
MISC	102	413	515
MON	897	46	943
ORG	4,431	782	5,213
PCT	186	77	263
PERIOD	341	255	596
PERS	1,820	4,415	6,235
QUANT	276	106	382
TIME	4	36	40
Total	**12,704**	**9,289**	**21,993**

Table 7: Distribution of entity types in the NER-UK 2.0 subcorpora.

B Examples of Most Common Gender-Swapping Mistakes Made by Few-Shot GPT-4o-mini

1. Changes word forms not directly linked to the JOB entity.
Example:
Син директо**ра**. → **Донька** директо**рки**.

*(en: **Director's** (masculine form) son. → **Director's** (feminine form) **daughter**.)*

2. Swaps job titles in the plural form even for both gender entities.
Example:
Письменн**ики**: Євген Маланюк, Наталя Лівицька-Холодна, ... → Письменн**иці**: Євгенія Маланюк, Сергій Лівицький-Холодний, ...

*(en: **Writers** (masculine form), Yevhen Malanyuk, Natalia Livytska-Kholodna. → **Writers** (femine form), Yevhenia Malanyuk, Serhiy Livytskyi-Kholodnyi.)*

3. Feminine subject originally paired with a masculine job title.
Example:
Вважав **її** хорош**им** педаго**гом**. → Вважав **його** хорош**ою** педаго**гинею**.

*(en: Considered **her** a **good teacher** (masculine form) → Considered **him** a **good teacher** (feminine form))*

4. Hallucinated or rare name.
Example:
Мит**ець Станіслав**. → Мистк**иня Станіслава**.

*(en: **Artist** (masculine form) **Stanislav** (male name) → **Artist** (feminine form) **Stanislava** (very rare female name).)*

5. Gender-agreement failure.
Example:
Вигадав **він**. → Вигадав **вона**.

*(en: **he** invented (masculine form). → **she** invented (masculine form).)*

6. Invalid job-title swap.
Example:
Він – найбагатш**ий барон**. → **Вона** – найбагат**ша баронка**.
*(en: **He** is the **richest baron** (masculine form).→ **She** is the **richest baronka** (incorrect feminine form of baron).)*

Note**: Bold** words (parts of the word) indicate those that were changed during gender-swapping, while **underlined** words indicate those that caused the error.

Figure 3: Mistakes observed during sentence-level gender swapping.

C Model Performance on Round-Trip Gender Swapping

Metric	Aya-101 original	Aya-101 fine-tuned	GPT-4o-mini
Exact Match	0.21	**0.52**	0.51
Exact Match w/o PERS	0.34	**0.73**	0.70
JOB Match	0.76	**0.87**	0.62
BLEU	0.79	**0.87**	0.85
ROUGE-L	0.21	0.21	**0.22**
BERTScore (F1)	0.97	**0.99**	**0.99**
Token Count Match	0.64	**0.93**	0.91

Table 8: Evaluation results after **round-trip** gender swapping on the test set.

D Gender Composition of the Test Sets

Dataset	**Total**	**Masculine**		**Feminine**		**Common**		**Unknown**	
		Count	Fraction	Count	Fraction	Count	Fraction	Count	Fraction
Original	1,982	1,646	83%	76	3.8%	223	11.3%	37	1.8%
Augmented	3,715	1,828	49.2%	1,392	37.4%	393	10.5%	102	2.7%

Table 9: Gender composition of JOB entities for **Original** and **Augmented** NER-UK 2.0 datasets.

Dataset	**Total**	**Male**		**Female**		**Unknown**	
		Count	Fraction	Count	Fraction	Count	Fraction
Original	6,235	2,120	34.0%	1,286	20.6%	2,829	45.4%
Augmented	7,517	2,276	30.2%	2,016	26.8%	3,225	42.9%

Table 10: Gender composition of PERS entities for **Original** and **Augmented** NER-UK 2.0 datasets.

E Performance Comparison of Original NER and Gender-Balanced NER Across Different Test Sets

Test Set	Original NER				Gender-Balanced NER			
	Entity Type	**P**	**R**	**F1**	**Entity Type**	**P**	**R**	**F1**
Original	JOB	74.39	65.45	**69.64**	JOB	75.05	59.06	66.10
	PERS	96.20	96.60	**96.40**	PERS	97.01	95.18	96.08
	LOC	93.27	88.14	**90.63**	LOC	92.19	88.02	90.06
	ORG	90.89	90.71	90.80	ORG	92.89	89.93	**91.38**
	MISC	36.13	30.28	32.95	MISC	48.42	32.39	**38.82**
	QUANT	81.00	91.01	85.71	QUANT	89.66	87.64	**88.64**
	DATE	85.32	91.62	88.35	DATE	92.65	88.02	**90.28**
	PERIOD	76.92	70.27	73.45	PERIOD	80.25	70.27	**74.93**
	TIME	66.67	60.00	63.16	TIME	66.67	60.00	63.16
	ART	73.87	69.20	71.46	ART	70.52	79.75	**74.85**
	DOC	64.29	45.00	52.94	DOC	63.64	52.50	**57.53**
	MON	95.48	91.08	93.23	MON	97.07	91.69	**94.30**
	PCT	95.70	98.89	97.27	PCT	100.00	98.89	**99.44**
	Weighted avg.	89.12	87.17	88.13	**Weighted avg.**	90.89	86.08	**88.42**
Gender-swapped	JOB	89.08	71.26	79.18	JOB	90.63	80.78	**85.42**
	PERS	98.60	98.60	98.60	PERS	98.60	98.88	**98.74**
	LOC	90.53	92.47	**91.49**	LOC	92.31	90.32	91.30
	ORG	92.76	93.28	93.02	ORG	95.70	93.56	**94.62**
	MISC	33.33	9.09	14.29	MISC	80.00	36.36	**50.00**
	QUANT	85.71	75.00	80.00	QUANT	100.00	75.00	**85.71**
	DATE	92.47	93.48	**92.97**	DATE	94.19	88.04	91.01
	PERIOD	75.00	83.33	**78.95**	PERIOD	65.22	83.33	73.17
	TIME	0.00	0.00	0.00	TIME	100.00	100.00	**100.00**
	ART	53.33	61.54	57.14	ART	52.63	76.92	**62.50**
	DOC	33.33	20.00	**25.00**	DOC	20.00	20.00	20.00
	MON	96.97	96.97	96.97	MON	100.00	100.00	**100.00**
	PCT	100.00	100.00	100.00	PCT	100.00	100.00	100.00
	Weighted avg.	92.17	85.81	88.87	**Weighted avg.**	93.33	89.16	**91.20**
Augmented	JOB	80.53	67.75	73.59	JOB	82.51	68.43	**74.81**
	PERS	96.58	97.00	**96.79**	PERS	97.15	95.62	96.38
	LOC	93.65	89.02	**91.28**	LOC	92.42	88.36	90.35
	ORG	91.34	91.19	91.26	ORG	93.17	90.66	**91.90**
	MISC	36.59	29.41	32.61	MISC	49.49	32.03	**38.89**
	QUANT	81.31	89.69	85.29	QUANT	90.32	86.60	**88.42**
	DATE	86.10	91.91	88.91	DATE	92.86	87.69	**90.20**
	PERIOD	77.42	70.94	74.04	PERIOD	78.80	71.43	**74.94**
	TIME	60.00	54.55	57.14	TIME	70.00	63.64	**66.67**
	ART	72.69	69.20	70.90	ART	69.34	79.60	**74.12**
	DOC	61.29	42.22	50.00	DOC	57.89	48.89	**53.01**
	MON	95.34	91.34	93.30	MON	97.36	92.74	**94.99**
	PCT	96.43	99.08	97.74	PCT	100.00	99.08	**99.54**
	Weighted avg.	89.76	86.97	88.34	**Weighted avg.**	91.31	86.60	**88.89**

Table 11: Evaluation of NER models on the three test set variations.