Mitigating Gender Bias in English-Ukrainian Machine Translation Models

Pavels Ivanovs*, **Gina Welsh***, **Irini Selenica***
Department of Linguistics and Philology
Uppsala University, Sweden
{pavels.ivanovs.6536, gina.welsh.7005, irini.selenica.4908}
@student.uu.se

Abstract

This study investigates the presence and mitigation of gender bias in English-Ukrainian machine translation (MT). We evaluated two gender bias mitigation methods (gender tagging and Lapa LLM correction) on two pre-trained MT models (OPUS-MT and mBART) using a curated dataset of occupation-based sentences. Our results showed that both zero-shot models showed gender bias transfer, particularly for male-stereotyped occupations with female source gender. Gender tagging produced mixed results by over-assigning masculine forms to female-source sentences. LLM correction achieved the strongest mitigation, recovering female-labelled output to near source proportions. However, full morphological gender agreement remained a challenge for the LLM correction method. Our framework, which uses Ukrainian's overt gender morphology as a bias signal, is adaptable to other grammatically gendered languages.[1]

1 Introduction

One major challenge in the field of natural language processing (NLP) is the inadvertent propagation or amplification of cultural biases in their usage. The reinforcement of biases can occur in systems that depend on language corpora data originating from humans and therefore containing societal biases that are not necessarily of technical origin, including gender bias (Rudinger et al., 2017; Vanmassenhove et al., 2018; Sutton et al., 2018; Savoldi et al., 2021). In this study, gender bias is defined as the uneven assignment of nouns and adjectives associated with power and prestige for words marked with one particular gender (very often male) over female or non-binary genders; contrastively, female or non-binary genders can be marked with nouns and adjectives of lower prestige and power (Bolukbasi et al., 2016; Prates et al., 2020; Piergentili et al., 2023b). Neural machine translation can be an illustrative trigger of the inadvertent transfer of gender bias between two languages. If the source language's grammatical gender differs largely from the target language, as is the case with English and morphologically-rich, grammatically-gendered languages, this can result in mistranslation due to gender bias transfer to the target language (Moroz, 2025; Zmigrod et al., 2019).

Our study focuses on English-Ukrainian ma-

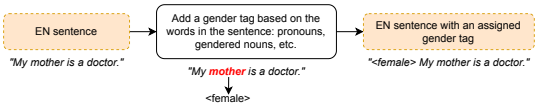

Figure 1: Translation pipeline using gender tagging process

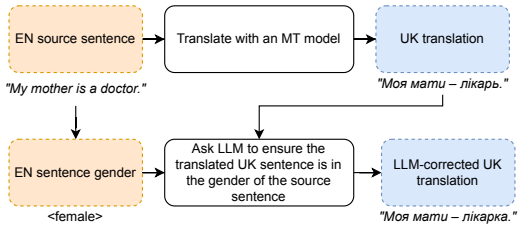

Figure 2: Translation pipeline using LLM-correction to mitigate EN-UKR gender bias.

*Equal contribution.
[1]Our code and data are available at https://github.com/pavelsivanovs/mitigating-gender-bias-in-en-uk-mt-models

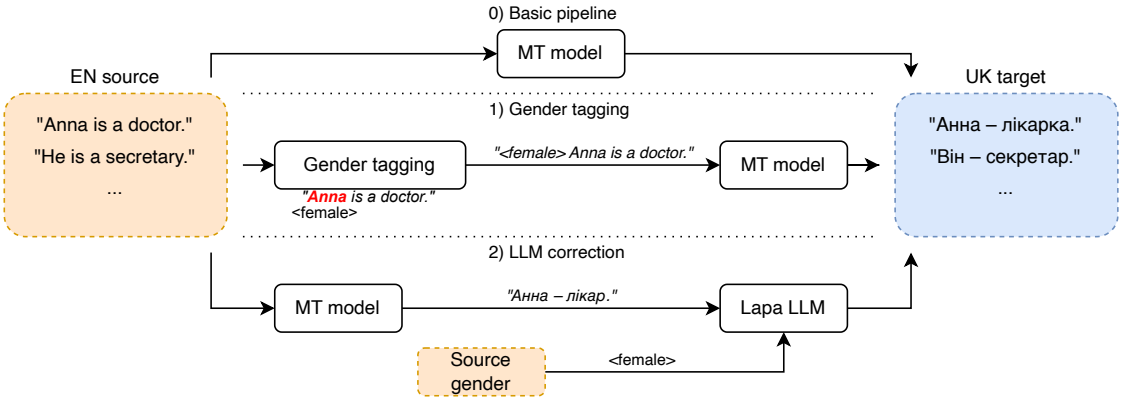

Figure 3: Schematic demonstration of the zero-shot translation pipeline (0), and our gender mitigating pipelines (1, 2): gender tagging and LLM correction.

chine translation gender bias within the professional domain for a few reasons. Firstly, Ukrainian overtly encodes gender in its word forms (see Figure 4 for an example), which can provide an obvious indication of gender bias transfer from English into Ukrainian. Secondly, professional roles in a language are a clear anchor for the investigation of gender bias, since these nouns often reflect cultural gender norms in the existence of male-female pairs within profession names. Finally, Ukrainian has a considerable amount of existing data available relative to low and very low-resourced languages, but the question of gender bias in machine translations regarding this language has been, so far, under-explored compared to high-resource languages such as English. Overall, this study provides insight into gender bias within machine translation models, particularly for languages that are not highly-resourced within the NLP field.

1.1 Contributions

The contributions of our study include:

1. A framework for evaluating gender bias transfer within English-Ukrainian MT models, using translation pairs with occupation names.
2. A demonstration of gender bias transfer within two pre-trained English-Ukrainian MT models (OPUS-MT and mBART).
3. An implementation of two gender bias mitigation pipelines that use gender tagging (see Figure 1) and large language model (LLM) correction (see Figure 2) as methods. These pipelines showed that both methods have an effect on gender assignment within the evaluation data, with gender tagging showing mixed results, and LLM-correction showing promising improvement on gender bias for English-Ukrainian translation pairs.

Overall, our study contributes a Ukrainian extension to gender bias mitigation studies and offers a MT gender bias mitigation framework that could be adapted to languages within the Slavic family, as well as lower-resource language projects with relatively less training data.

2 Background

2.1 Evaluating Gender Bias in Machine Translation

The invention of now widely-used neural methods has entailed a rise in research interest in the propagation of group stereotypes within NLP pipelines. Bolukbasi et al. (2016) highlighted gender bias within word embeddings as an upstream cause of bias that neural MT pipelines inherit. Zhao et al. (2018) introduced WinoBias, a benchmark for gender bias in coreference resolution using template sentences, showing coreference systems link gendered pronouns to pro-stereotypical occupation entities. Stanovsky et al. (2019) extended this approach with WinoMT, a protocol that uses English-source anaphoric resolution sentences containing professional roles. Their evaluation found that all six tested MT models produced gender bias to varying extents for all eight tested grammatically-gendered languages in the study (including Ukrainian). They additionally observed that adjectives with gender-stereotypical characteristics (e.g. *The pretty doctor asked the nurse to help her in the operation*) shifted gender assignment in the target output, an effect subse-

quently confirmed by Troles and Schmid (2021) across three commercial MT systems. Currey et al. (2022) introduced MT-GenEval to build on WinoMT through naturally-occurring Wikipedia sentences across eight target languages, although Ukrainian is not among them. Sun et al. (2019) and Prates et al. (2020) further demonstrated systemic gender bias in neural pipelines across several languages, particularly in the over-prediction of masculine forms for high-status professions such as doctors or engineers.

Gender bias is also shown to affect translation quality across modalities, with Bentivogli et al. (2020) demonstrating gender bias within the speech translation domain. For non-binary gender accuracy, Piergentili et al. (2023a) and Piergentili et al. (2023b) developed the theoretical foundations for gender-neutral translation and introduced a benchmark for gender-neutral English-Italian translation. Ukrainian's morphology makes neutrality particularly difficult for singular nouns that refer to people (see Section 2.4). This paradigm represents an important parallel research direction in gender bias evaluation.

The over-prediction of masculine forms by neural MT methods can be attributed to data imbalances (Vanmassenhove et al., 2019; Savoldi et al., 2021) and MT pipeline architectural choices. Costa-jussà et al. (2022) demonstrated that within multilingual MT, shared encoder-decoder architectures retain less gender information in source embeddings and produce more concentrated attention patterns than language-specific encoder-decoders, thereby disfavouring the generation of feminine forms in the target language due to narrower attention defaulting to male forms. Thus, the phenomenon of gender bias should be assessed not only as a data imbalance issue, but also as an architectural issue across all languages treated by MT pipelines.

2.2 Mitigation approaches

Vanmassenhove et al. (2018) prepended a speaker-gender token to source sentences during training (gender tagging), enabling the model to condition target-side gender realisation using this signal at the decoding phase; this showed measurable improvements in gender accuracy for translations. Stafanovičs et al. (2020) extended this gender-tagging approach, handling cases containing multiple referents of different genders, which showed accuracy gains for MT systems for English to Latvian, Lithuanian, and other morphologically rich languages. Saunders and Byrne (2020) fine-tuned a pre-trained neural MT model on a small hand-crafted, gender-balanced set of profession-related sentences with two countermeasures (Elastic Weight Consolidation and lattice rescoring) without degrading BLEU scores. After the release of transformer-based large language models (Brown et al., 2020), more studies have integrated LLMs into gender debiasing in MT. Sant et al. (2024) showed that careful LLM prompting can substantially mitigate gender bias, and Sánchez et al. (2024) showed that LLMs can generate gender-controlled translations without fine-tuning for Spanish, Italian, French and Portuguese. Nunziatini and Diego (2024) implemented an English-Spanish MT pipeline that included automatic post-editing with LLMs (GPT-4 and PaLM2) to fix gender bias in the MT output, finding that GPT-4 corrected 80% of the gender-bias issues in the Spanish output using an English prompt. However, post-edits were also shown to introduce unnecessary changes and inconsistencies that affected overall translation output. Overall, our study is guided by this existing work on gender-tagging and LLM correction and is, to our knowledge, the first to adapt these methods specifically to English-to-Ukrainian MT.

2.3 Low-resource and Ukrainian-specific work

A subfield of machine translation investigates gender bias for mid- to low-resource languages. Examples of such studies include Sewunetie et al. (2024), which constructs gender bias benchmark datasets for three low-resource languages; and Wairagala et al. (2022), which explores gender bias in Luganda-English machine translation, finding a tendency of MT models producing gendered target outputs for this language pair, despite Luganda containing exclusively gender-neutral pronouns.

For Ukrainian, Moroz (2025) emphasises the need for gender bias research in MT, given the language's overt gender encoding. Recent NLP work on Ukrainian includes OPUS-MT's English to Ukrainian MT pipeline (Tiedemann et al., 2024)[2] and Lapa LLM (Paniv et al., 2025),[3] an open-source large language model adapted to Ukrainian.

[2] https://huggingface.co/Helsinki-NLP/opus-mt-uk-en

[3] https://github.com/lapa-llm/lapa-llm

Buleshnyi et al. (2025) found that embedding debiasing alone is insufficient for morphologically rich languages like Ukrainian, and Nahurna and Romanyshyn (2025) developed gender-balanced datasets to address gender-imbalance issues for Ukrainian textual data. Our study extends on this work to the specific domain of gender bias in machine translation to Ukrainian.

2.4 Ukrainian gender morphology

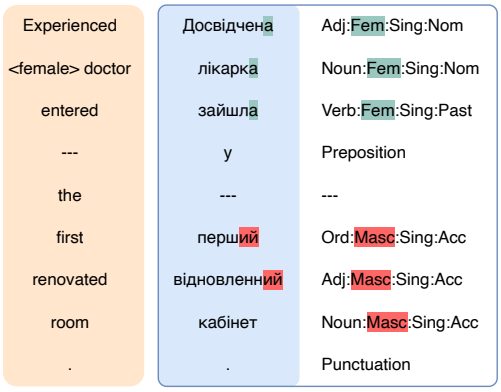

Figure 4: Example of gender encoding in the Ukrainian language. Example sentence *"Досвідчена лікарка зайшла у перший відновлений кабінет."* (*"Experienced <female> doctor entered the first renovated room."*) demonstrates that every word has gender information encoded: words that depend on the noun (adjectives, verbs, ordinals, etc.) must agree on the gender with the noun they depend on.

Ukrainian is a highly-inflectional and explicitly-gendered language. A single noun lemma can be inflected with 14 forms encoding 7 grammatical cases in both singular and plural (Nesset, 2003; Lesiuk, 2022). Each noun must be assigned either a feminine, masculine, or neuter grammatical gender; as a result, it is difficult to maintain gender-neutrality in Ukrainian nouns denoting singular individuals, since they must be inflected in either the feminine or masculine form, but not the neutral form.[4] Overall, the masculine gender is used as the default marking in the singular form, which is prevalent in words referring to professional roles. The Ukrainian plural form, as seen in *ми приїхали* (*my pryikhaly* "we arrived"), does not convey gendered information for the first-person plural form "we". However, in the case of nouns indicating professions, gender marking is overt even for plural forms referring to groups of females. For instance, English noun *programmers* can be translated into Ukrainian as either *програмістки* (*programistky*, feminine plural) or *програмісти* (*programisty*, masculine plural).[5] If a group of professionals contains at least one man, then grammatically the "default" male gender-marking is used to name their profession, even if the remaining members of the group are women.

3 Methodology

3.1 Experimental Set-up

3.1.1 Pipelines

In our experiments, we implemented pipelines (see Figure 3) to test gender bias transfer from English to Ukrainian:

0. Baseline: two MT pipelines comprised of zero-shot OPUS-MT and mBART (Liu et al., 2020) EN-UKR models;

1. Gender tagging: two pipelines fine-tuned on the OPUS-MT and mBART baselines with gender tagging as a mitigation strategy;

2. LLM correction: two pipelines set up for OPUS-MT and mBART pipelines using Lapa LLM as a corrective tool for the baseline output.

The pipelines were designed to test for gender bias already existing within widely-used MT models, as well as the efficacy of the gender bias mitigations outlined in 1) and 2) for these models. The mBART pre-trained model [6] contains approximately 680 million parameters (Liu et al., 2020), and the number of parameters for the OPUS-MT model [7] is estimated to be 300 million parameters. [8]

3.1.2 Dataset

For our source dataset, we extracted all English sentences containing profession names from the Tatoeba dataset (Tiedemann, 2020). We focused on sentences containing professions, under the hypothesis that these word forms could trigger gender bias transfer within the MT pipelines, through mistranslation of gender-neutral English profession names to gender-biased Ukrainian outputs.

[4] People who identify themselves as non-binary, in general, use the plural forms of pronouns, which are genderless: *ми* (*my* "we"), *ви* (*vy* "you" pl.), or *вони* (*vony* "they")

[5] All possible inflections are shown in Appendix A (Table B).

[6] `facebook/mbart-large-50-many-to-many-mmt`

[7] `https://huggingface.co/Helsinki-NLP/opus-mt-uk-en`

[8] `https://aihub.qualcomm.com/models/opus_mt_es_en`

ID	Template	Gender
1	I used to work as a <OCCUPATION>.	neutral
2	I am a poor <OCCUPATION>.	neutral
3	I am a nice <OCCUPATION>.	neutral
4	The <OCCUPATION> called me today.	neutral
5	Anne is a <OCCUPATION>.	female
6	Anne is an experienced <OCCUPATION>.	female
7	Tom is a <OCCUPATION>.	male
8	Tom is a poor <OCCUPATION>.	male
9	Jessie is a nice <OCCUPATION>.	neutral
10	Jessie is a poor <OCCUPATION>.	neutral

Table 1: Test sentence templates

For the evaluation of all pipelines, we composed a dataset of 810 English source sentences representing a total of 81 professions, based on the template of Table 1. Each sentence contained an occupation and at least one pronoun or participant, in accordance with the Winogender schema (Rudinger et al., 2018). Overall, the template was designed to test for gender bias by comparing the translations of minimal sentences with profession names. Sentences 1-4 contain gender neutral pronouns in the English source as a point of comparison with the grammatically-marked verbs, adjectives and occupational nouns in the translated Ukrainian output. Sentences 5-6 and 7-8 contain gender-specific names 'Anne' and 'Tom'. Sentences 9-10 include a gender-neutral name 'Jessie', as the US Social Security Administration data shows a comparable gender distribution for this name throughout history.[9]

Selected adjectives (poor, nice, experienced) were added to some templates as modifiers, following Stanovsky et al. (2019) and Troles and Schmid (2021) who found that adjectives can trigger gender bias transfer. The adjectives were chosen to probe prestige connotations: poor signals low prestige, nice signals agreeableness (a feminine stereotype), and experienced signals high prestige or authority.

3.2 Gender Bias Mitigation Methods

3.2.1 Gender Tagging

We fine-tuned both the OPUS-MT and mBART models with the gender tagging mitigation method on our 1287-sentence Tatoeba data subset (Tiedemann, 2020). A Python script automatically assigned each source English sentence a gender tag. The tags "female" or "male" were prepended to sentences containing gendered nominals such as *mother*, *wife*, *father*, *husband*, or pronouns like *she*, *he* (see Appendix C). Sentences without gendered words were prepended with a "neutral" tag (see Figure 1). The OPUS-MT and mBART models were fine-tuned on a gender-tagged version of the 1287-sentence professions data subset, with a learning rate of 5e-5, a batch size of 8, and 3 training epochs as hyperparameters. After training, both models were evaluated on our evaluation set, with an accuracy score of how often the English source sentence's gender (based on the subject's pronouns) aligned with the translated Ukrainian sentence's gender (based on the subject's adjectives and nouns).

3.2.2 Gender bias LLM-correction

For the LLM correction pipeline, we decided to experiment with a Large Language Model (LLM) to correct translations that were assigned Ukrainian gender markings incongruous with the English source sentence. We used Lapa LLM, an open Ukrainian Large Language Model based on Google's Gemma 3 model and trained on 25 datasets of varying domains. The LLM system was assigned the role: "You are a linguistic assistant for Ukrainian text editing." The user prompt was: "Edit the following Ukrainian sentence by ensuring the occupation is in the `TARGET_GENDER` form, keeping the structure of the sentence unchanged: `SENTENCE`, where `TARGET_GENDER` is the gender of a source sentence and `SENTENCE` is a zero-shot translation." Due to hardware constraints, we used the quantised version of the model, with each model's weight reduced by 75%.

3.3 Evaluation

For our gender-accuracy evaluation, two Ukrainian speakers annotated the gender of 4,860 output sentences (810 sentences per pipeline for 6 pipelines) in accordance with our annotation guidelines (see Appendix A). Each pipeline's Ukrainian output sentences were compared with their respective English source sentence, and a gender-accuracy measurement was calculated by the proportion of times the English source sentence's gender matched the gender of its Ukrainian translated output. For this, two types of accuracy measures were calculated: one assumed that English gender-neutral source sentences should match a masculine-gendered Ukrainian output, since this is considered the 'default' gender in Ukrainian; the second accuracy measure did not include this as-

[9] `https://www.ssa.gov/oact/babynames/limits.html`

sumption. Inter-annotator agreement was calculated using Cohen's kappa, with a very high value of 91.67%. For our translation quality metric, we calculated BLEU (Papineni et al., 2002) and COMET scores (Rei et al., 2020). Results were additionally broken down by gender-stereotyped occupation category and prestige tier.

4 Results

Our results, shown in Table 2, indicate that both the OPUS-MT and mBART English-Ukrainian models, to varying extents, demonstrate evidence of gender bias transfer, with mBART showing a higher gender-label accuracy score than OPUS-MT. The automatic evaluation results reveal interesting patterns across the three mitigation methods.

4.1 Overall accuracy by mitigation strategy

Zero-shot setting. In the zero-shot setting, mBART outperforms OPUS-MT on both accuracy (81.81% vs 76.39%) and BLEU (48.74 vs 42.30), though OPUS-MT scores are slightly higher on COMET (86.20 vs 85.81).

Gender tagging. BLEU drops for both models, from 42.30 to 28.48 for OPUS-MT and from 48.74 to 22.21 for mBART, likely reflecting surface-level divergence from the reference translations introduced by the tagging process. However, COMET scores remain stable or even slightly improve, suggesting that the translation quality in terms of general meaning is preserved or even enhanced. This difference between BLEU and COMET is consistent with the known penalisation of BLEU when vocabulary or length differs in translation.

LLM correction. LLM correction demonstrates the strongest results overall, with mBART achieving the highest accuracy 86.87% and a substantial BLEU improvement 63.93, while both systems converge on similar COMET scores (around 86.43 to 86.45). This suggests evidence that LLM post-editing successfully recovers fluency and surface similarity to the reference while maintaining semantic quality.

4.2 Accuracy by template, occupation type and gender stereotype

Accuracy by template sentence. The results in Table 4 show uneven accuracy across templates, driven by the gender name, adjective and occupation in each sentence. First-person neutral templates (1-4) score consistently high (82.5-93.4%), reflecting the Ukrainian 'masculine as neutral' default. There is strong asymmetry between male and female name templates: Tom templates (7-8) have the highest accuracy overall (94.0%, 95.3%) and Anne templates have the lowest (40.9%, 38.1%); with very low scores for the Gender Tagging mitigation pipeline (8.6% and 0.0%). LLM partially recovers the Anne templates (58.0% - 70.4%) but this does not reach the same accuracy range as the Tom templates. The templates' adjectives also show an effect: for example, *Anne is an experienced <OCC>* shows the lowest accuracy in the entire table (38.1%; 0.0% for mBART gender tagging pipeline), suggesting that prestige framing through adjectives appears to suppress the female name signal for the gender tagging method. The adjective *poor* has a negligible effect on the Tom template, but lowers some accuracies for the Jessie template (*nice* - 85.2-97.5%; *poor* - 65.4 - 96.3%). These findings indicate that adjectives by themselves can affect gender accuracy, in keeping with findings from Stanovsky et al. (2019) and Troles and Schmid (2021).

Accuracy by occupation category and prestige. The variation across occupational groups shown in Table 7 is moderate at zero-shot, but widens under the gender-tagging pipeline. The occupation category *Building and Grounds Cleaning and Maintenance* shows the largest zero-shot model gap (43.3% - 93.3%), potentially reflecting feminine lexical defaults for cleaning roles. Cleaning roles sit at the intersection of low prestige and strong female stereotyping in the BLS data (see Table 12). Overall, it appears that OPUS-MT defaults more strongly to feminine forms for these roles, while mBART defaults masculine across the board. Occupations with high prestige (shown in Table 3) have higher accuracy levels than those with lower prestige for the OPUS-MT model, but not so much for mBART. These findings could reflect how architectural differences can handle gender differently, as discussed in Costa-jussà et al. (2022). LLM correction raises accuracy scores across all groups (occupational category and prestige level), showing overall robustness across all occupation types.

Gender stereotype Table 6 organises occupations into female-stereotyped, male-stereotyped and gender-neutral based on U.S. Bureau of Labor Statistics (2018) (see Appendix Table 12 for

	Accuracy (%)			BLEU			COMET		
Model	Zero-Shot	Gender-tagging	LLM	Zero-Shot	Gender-tagging	LLM	Zero-Shot	Gender-tagging	LLM
OPUS-MT	76.39	77.59	84.58	42.30	28.48	47.22	86.20	86.40	**86.45**
mBART	81.81	77.47	**86.87**	48.74	22.21	**63.93**	85.81	84.71	86.43

Table 2: Overall pipeline gender-label accuracies, BLEU, and COMET scores, with gender-neutral English source sentences considered correctly masculine in Ukrainian output.

	OPUS-MT Acc. (%)			mBART Acc. (%)		
Occ. Prestige	**Low**	**Mid**	**High**	**Low**	**Mid**	**High**
Zero-shot	74.3	74.7	81.1	82.5	**82.1**	80.5
Gender Tag	72.1	79.4	**81.1**	75.7	77.4	80.0
LLM Corr.	81.8	84.1	87.9	86.8	86.5	**88.4**

Table 3: Gender accuracy (%) by occupation prestige tier (Ganzeboom et al., 1992) across pipeline conditions for two translation models (OPUS-MT and mBART).

further details). Across all pipelines and mitigation strategies, sentences with a male source gender showed a high accuracy regardless of stereotype, with mBART hitting 100% for this category across all pipelines. Sentences with a female source gender show the lowest accuracy values in the entire evaluation across all pipelines. Notably, the gender-tagging method lowers the accuracy value considerably for both mBART and OpusMT (1.4% and 4.3%). LLM correction partially recovers the rate, but cannot match parity with the male source sentences. Overall, the LLM correction method is the only pipeline reliably achieving above-baseline accuracy for female source gender across all three stereotype categories, for both models.

4.3 WinoMT comparison

In Table 5, we provide an explicit comparison between our results and WinoMT results by Stanovsky et al. (2019), which included English-Ukrainian gender bias evaluation in MT systems. At the time of this study, the best commercial system was Microsoft Translator, which reached 41.3% accuracy. In our study, our zero-shot baselines exceed this by a large margin (76.39% for OPUS-MT and 81.81% for mBART). With LLM post-editing, mBART reaches 86.87%, more than doubling the accuracy of the strongest commercial system reported by Stanovsky et al. (2019). Some factors to take into account for these findings are: 1) our studies include newer MT systems, 2) our evaluation test sets are different, and 3) we include the explicit assumption that neutral English source gender must match masculine-coded Ukrainian output to masculine target output, which is not so clear in Stanovsky's study. Therefore, while direct comparison should be interpreted with caution given differences in test sets and evaluation setups, our results contribute further to English-Ukrainian MT studies by suggesting that open-source NMT models, combined with targeted post-editing, can substantially reduce gender bias in Ukrainian translation.

5 Discussion

5.1 Gender tagging as gender bias mitigation

Our results show that gender tagging improved accuracy for gender-neutral source sentences (templates 1-4: 82.5%-93.4%) but failed to demonstrate a global mitigation effect in our data. The proportion of female-labelled sentences dropped from 15.4% to 5.4%, with Anne templates collapsing to 8.6% and 0.0% accuracy under mBART, indicating the gender-tagging method over-assigns masculine forms. The "experienced" adjective compounded this effect, consistent with Stanovsky et al. (2019) and Troles and Schmid (2021). This may suggest that gender tags provide clearer source-side gender cues, but the models do not consistently integrate these cues into their decoding process, which highlights the need for model architectures that better respect explicit tags (Costa-jussà et al., 2022).

One potential explanation is that both models have been pretrained on large-scale data containing typical English-to-Ukrainian translation gender biases, but introducing explicit tags could conflict with these learned patterns and fail to mitigate them. Our study has a small fine-tuning set (n = 1287) due to Ukrainian data limitations, which

Template	Model	Zero-shot	Gender Tag.	LLM Corr.
I used to work as a [OCC].	mBART	96.3	**97.5**	96.3
	OPUS-MT	79.0	**93.8**	79.0
I am a poor [OCC].	mBART	92.6	**97.5**	92.6
	OPUS-MT	90.1	**95.1**	88.9
I am a nice [OCC].	mBART	90.1	**97.5**	90.1
	OPUS-MT	93.8	**96.3**	92.6
The [OCC] called me today.	mBART	**91.4**	84.0	85.2
	OPUS-MT	88.9	**92.6**	82.7
Anne is a [OCC].	mBART	42.0	8.6	**82.7**
	OPUS-MT	13.6	19.8	**79.0**
Anne is an experienced [OCC].	mBART	58.0	0.0	**70.4**
	OPUS-MT	37.0	4.9	**58.0**
Tom is a [OCC].	mBART	91.4	**97.5**	95.1
	OPUS-MT	91.4	**95.1**	93.8
Tom is a poor [OCC].	mBART	96.3	97.5	**98.8**
	OPUS-MT	91.4	92.6	**95.1**
Jessie is a nice [OCC].	mBART	95.1	**97.5**	93.8
	OPUS-MT	86.4	**91.4**	85.2
Jessie is a poor [OCC].	mBART	65.4	**96.3**	65.4
	OPUS-MT	88.9	**91.4**	87.7

Table 4: Gender-label accuracy (%) by sentence template, model, and pipeline. Neutral source gender treated as masculine (Ukrainian default). Colour coding represents best (green) and worst (red) performing accuracies.

System	Acc. (%)
SYSTRAN[10]	28.9
Google Translate[11]	38.4
Microsoft Translator[12]	41.3
OPUS-MT (zero-shot)	76.39
mBART (zero-shot)	81.81
OPUS-MT + LLM correction	84.58
mBART + LLM correction	**86.87**

Table 5: Gender accuracy on English-to-Ukrainian compared to commercial systems evaluated by Stanovsky et al. (2019).

could have reduced the explanatory power of the mixed results. Future studies could include more gender-tagged data in their fine-tuning process to better measure its effect on bias, or explore gender tag refining strategies for languages with lower amounts of data, such as multi-token or context-aware gender tags.

5.2 LLM correction as gender bias mitigation

The LLM correction produced the strongest results overall (mBART: 86.87% accuracy, BLEU 63.93, COMET 86.43), with the female-labelled proportion rising from 15.4% to 22.9%, closely matching the 20% female source proportion. We observed that for the sentences of the template types "Anne is a …" and "Anne is an experienced …", 112 out of 200 sentences incorrectly generated in the masculine form were rewritten to the correct feminine form. This does not necessarily mean that the LLM we used is aware of the gender stereotypes and how to combat them. But it shows, to some degree, the capability to rewrite the given sentence in the requested gender. It is worth noting that the time period of the Ukrainian training data for Lapa LLM might be a contributing factor, since there is an increased visibility of women-specific counterparts for professional names in Ukrainian in recent years. To a different extent, we could observe the disagreement in gender encoding between the occupation and the related verb and/or adjective. Male-stereotyped occupations with female source gender remained most resistant (mBART 62.9%, OPUS-MT 52.9%). Partial agreement failures were observed where adjectives were correctly feminised but occupational nouns were not. We assume that here, the MT model calculates that the sentence subject is feminine, but it fails to change the form of the occupation itself to demonstrate it properly. Overall, this demonstrates that full morphological agreement remains a challenge for LLM correction in Ukrainian.

5.3 Other Observed Biases

Defaulting to Russian. We observed that the OPUS-MT model occasionally translated English source text to Russian instead of Ukrainian. Two technical explanations are plausible. First, the training data may have contained Russian text,

		mBART Acc. %			OPUS-MT Acc. %		
Stereotype	**Source Gender**	**Zero-shot**	**Tagged**	**LLM**	**Zero-shot**	**Tagged**	**LLM**
Female-stereotyped	Female	77.8	9.3	**87.0**	51.9	25.9	**81.5**
	Male	81.5	**92.6**	90.7	87.0	83.3	92.6
	Neutral	72.8	**89.5**	72.8	78.4	82.1	76.5
Male-stereotyped	Female	30.0	1.4	**62.9**	10.0	4.3	52.9
	Male	**100.0**	**100.0**	**100.0**	97.1	98.6	97.1
	Neutral	96.7	98.1	94.8	95.7	**99.5**	92.9
Gender-neutral	Female name	47.4	2.6	86.8	15.8	7.9	78.9
	Male	**100.0**	**100.0**	**100.0**	86.8	**100.0**	92.1
	Neutral	95.6	**97.4**	93.9	86.8	98.2	86.8

Table 6: Gender-label accuracy (%) by occupation gender stereotype, source gender, model, and pipeline. See Appendix Table 12 for gender stereotype categorisation details. Colour coding represents best (green) and worst (red) performing accuracies.

	OPUS-MT Acc. (%)			mBART Acc. (%)		
Occupation Group	**Zero-shot**	**Gender Tagging**	**LLM Correction**	**Zero-shot**	**Gender Tagging**	**LLM Correction**
Management	80.0	80.0	90.0	80.0	80.0	95.0
Business and Financial Operations	78.2	80.9	88.2	87.3	80.0	92.7
Computer and Mathematical	80.0	80.0	90.0	90.0	80.0	90.0
Architecture and Engineering	80.0	80.0	85.0	85.0	80.0	85.0
Life, Physical, and Social Science	80.0	83.3	90.0	80.0	80.0	90.0
Community and Social Service	80.0	80.0	90.0	90.0	80.0	90.0
Legal	65.0	85.0	70.0	85.0	75.0	95.0
Educational Instruction and Library	82.5	82.5	85.0	80.0	82.5	77.5
Arts, Design, Entmt., Sports, and Media	75.0	70.0	90.0	95.0	75.0	90.0
Healthcare Practitioners and Technical	76.8	78.4	84.2	76.8	75.8	82.6
Protective Service	77.5	80.0	87.5	82.5	77.5	92.5
Food Preparation and Serving Related	72.5	77.5	90.0	85.0	80.0	90.0
Building and Grounds Cleaning & Maint.	43.3	50.0	53.3	83.3	80.0	93.3
Personal Care and Service	75.0	70.0	80.0	73.3	73.3	75.0
Sales and Related	80.0	80.0	95.0	85.0	80.0	100.0
Office and Administrative Support	84.0	72.0	88.0	70.0	68.0	80.0
Construction and Extraction	78.0	80.0	78.0	94.0	78.0	92.0
Installation, Maintenance, and Repair	80.0	80.0	80.0	90.0	80.0	90.0
Production	30.0	70.0	50.0	90.0	80.0	80.0
Transportation and Material Moving	80.0	80.0	100.0	80.0	80.0	95.0
Unclassifiable	90.0	80.0	90.0	90.0	80.0	90.0

Table 7: Gender accuracy breakdown by occupation group across pipeline conditions for two translation models (OPUS-MT and mBART). Gender-neutral English source sentences considered correctly masculine in Ukrainian output. Occupations classified by 2018 U.S. Bureau of Labor Statistics (BLS) Standard Occupational Classification (SOC) major groups (U.S. Bureau of Labor Statistics, 2018).

leading the model to conflate the two languages. Second, the model may have been developed using transfer learning from a pre-trained Russian model, a common approach for low-resource languages where parallel data is scarce. The second explanation carries political implications. Ukrainian is not a variant of Russian, but a distinct language that has historically been subjected to deliberate political suppression. The notion that Russians and Ukrainians are 'brothers' together with centuries of russification policies of Ukraine led to Russian-language dominance (Shulzhenko, 2023), which can impact the development and composition of NLP tools, resources and data. Therefore, it is important for NLP practitioners to treat Ukrainian and Russian as separate languages, both in data curation and in

modelling decisions, rather than defaulting to Russian as a convenient approximation.

Translating Person Names. We observed that some names are inconsistently either translated or transliterated across models. One example is the female name "Anne", which can be seen both translated as "Анна" (transliteration) and "Енн" (transcription). Although this does not directly mean that the gendered translation did not work as intended; it still needs to be examined further, as it could be interfering with the output of a correctly gendered translation in general.

6 Conclusion

Our study demonstrates three key findings. First, both OPUS-MT and mBART show evidence of gender bias in their pre-trained forms, particularly for male-stereotyped occupations and female source gender sentences. Second, gender tagging affected gender reassignment but produced mixed results by over-assigning masculine forms to female-name templates, suggesting a conflict between explicit tags and pretrained biases. Third, LLM correction via Lapa demonstrated the strongest mitigation overall, recovering female-labelled output to near source-data proportions. However, the LLM correction method still had difficulties with full morphological agreement across word classes. Overall, the results of our study point to LLM correction as a promising direction for morphologically rich, mid-resource languages like Ukrainian. We believe our evaluation framework, using overt gender morphology as a bias signal, is adaptable to other grammatically gendered languages, particularly within the Slavic family, and offers a replicable foundation for future gender bias mitigation work in lower-resource MT settings.

7 Ethical Considerations and Limitations

The main limitation of this experiment is the fact that it was carried out with only Ukrainian as a target language, and for sentences with professional names. However, we would argue that our overall framework, that is, using overt gender-marking to detect gender bias transfer, can be reused and adapted to other grammatically gendered languages, such as Polish, French, or Arabic. Moreover, the Lapa model was evaluated in its quantised form due to GPU memory constraints. While quantisation is widely used in practice and generally preserves model behaviour, it may subtly affect the reported results and the performance of the full-precision model. Future work with access to higher-memory hardware could validate these findings using the unquantized version.

It is also important to extend this work to other contexts to make it more inclusive. One major limitation of this work is that it focuses on only one dimension of gender (male–female) and within a very specific context (professions). We argue that future research should examine biases affecting people across the full spectrum of genders. Nevertheless, our findings highlight a broader underlying issue and demonstrate clear room for improvement, which makes this contribution both timely and relevant.

For the russification bias, we would encourage further work on improving datasets and MT pipelines for the Slavic languages, to avoid mistranslation into Russian or other higher resource languages of the same language family.

8 Sustainability Statement

To conduct our study, we ran code both locally and using GPU power from Google Colab (Tesla T4). Given that our project used a relatively small dataset, the environmental impact of our experiments is rather small. However, we do believe it is important to take into consideration the environmental impact of such studies in the future and to try to minimise it as much as possible.

Acknowledgments

We express our whole-hearted gratitude to *Meriem Beloucif*. If it were not for her, this work would have never been published. The feedback she provided is invaluable. We also thank the anonymous reviewers for their insightful feedback on this work.

References

Bentivogli, Luisa, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? Evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online, July. Association for Computational Linguistics.

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016.

Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 4349–4357.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 1877–1901.

Buleshnyi, Mykhailo, Maksym Buleshnyi, Marta Sumyk, and Nazarii Drushchak. 2025. Gender bias evaluation and mitigation for ukrainian large language models. In Romanyshyn, Mariana, editor, *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 64–72, Vienna, Austria (online), July. Association for Computational Linguistics.

Costa-jussà, Marta R., Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2022. Interpreting gender bias in neural machine translation: Multilingual architecture matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11855–11863.

Currey, Anna, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Ganzeboom, Harry B. G., Paul M. De Graaf, and Donald J. Treiman. 1992. A standard international socio-economic index of occupational status. *Social Science Research*, 21(1):1–56.

Lesiuk, Mykola. 2022. Formation of grammatical forms of full-meaning parts of speech in ukrainian and polish languages. *Philological Review*.

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Moroz, Maryna. 2025. Reproduction of gender stereotypes in machine translation systems (based on the material of the Ukrainian and English languages). *Transcarpathian Philological Studies*, 2(39):136.

Nahurna, Olha and Mariana Romanyshyn. 2025. Gender swapping as a data augmentation technique: Developing gender-balanced datasets for Ukrainian language processing. In Romanyshyn, Mariana, editor, *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 147–161, Vienna, Austria (online), July. Association for Computational Linguistics.

Nesset, Tore. 2003. Gender assignment in ukrainian: Language specific rules and universal principles.

Nunziatini, Mara and Sara Diego. 2024. Implementing gender-inclusivity in MT output using automatic post-editing with LLMs. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 580–589, Sheffield, UK, June. European Association for Machine Translation (EAMT).

Paniv, Yurii, Bogdan Didenko, and Mykola Haltiuk. 2025. Lapa LLM.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Piergentili, Andrea, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023a. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 71–83, Tampere, Finland, June. European Association for Machine Translation.

Piergentili, Andrea, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023b. Hi guys or hi folks? benchmarking gender-neutral machine translation with the GeNTE corpus. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14124–14140, Singapore, December. Association for Computational Linguistics.

Prates, Marcelo O. R., Pedro H. Avelar, and Luís C. Lamb. 2020. Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, 32:6363–6381.

Rei, Ricardo, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. *CoRR*, abs/2009.09025.

Rudinger, Rachel, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *EthNLP@EACL*.

Rudinger, Rachel, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, New Orleans, Louisiana, June. Association for Computational Linguistics.

Sánchez, Eduardo, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2024. Gender-specific machine translation with large language models. In *Proceedings of the 4th Workshop on Multilingual Representation Learning (MRL 2024)*. Association for Computational Linguistics.

Sant, Aleix, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of prompts: Evaluating and mitigating gender bias in MT with LLMs. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–139, Bangkok, Thailand, August. Association for Computational Linguistics.

Saunders, Danielle and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online, July. Association for Computational Linguistics.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Sewunetie, Walelign Tewabe, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Hellina Hailu Nigatu, Gashaw Kidanu, Zewdie Mossie, Hussien Seid, Eshete Derb, and Seid Muhie Yimam. 2024. Evaluating gender bias in machine translation for lowresource languages. In *5th Workshop on African Natural Language Processing*.

Shulzhenko, Daria. 2023. How Russia has attempted to erase Ukrainian language, culture throughout centuries. *The Kyiv Independent*, May. `https://kyivindependent.com/how-russia-has-attempted-to-erase-ukrainian-language-culture-throughout-centuries/` [Accessed: 2026-03-13].

Stafanovičs, Artūrs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online, November. Association for Computational Linguistics.

Stanovsky, Gabriel, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In Korhonen, Anna, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July. Association for Computational Linguistics.

Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In Korhonen, Anna, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July. Association for Computational Linguistics.

Sutton, Adam, Thomas Lansdall-Welfare, and Nello Cristianini. 2018. Biased embeddings from wild data: Measuring, understanding and removing. *ArXiv*, abs/1806.06301.

Tiedemann, Jörg, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2024. Democratizing neural machine translation with opus-mt. *Language Resources and Evaluation*, 58(2):713–755.

Tiedemann, Jörg. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual mt. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online, November. Association for Computational Linguistics.

Troles, Jonas-Dario and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives.

U.S. Bureau of Labor Statistics. 2018. Standard occupational classification (soc) system. Technical report, U.S. Bureau of Labor Statistics.

Vanmassenhove, Eva, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November. Association for Computational Linguistics.

Vanmassenhove, Eva, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland, August. European Association for Machine Translation.

Wairagala, Eric Peter, Jonathan Mukiibi, Jeremy Francis Tusubira, Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, and Ivan Ssenkungu. 2022. Gender bias evaluation in Luganda-English machine translation. In Duh, Kevin and Francisco Guzmán, editors, *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 274–286, Orlando, USA, September. Association for Machine Translation in the Americas.

Zhao, Jieyu, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, October-November. Association for Computational Linguistics.

Zmigrod, Ran, Sabrina J. Mielke, Hanna M. Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *ArXiv*, abs/1906.04571.

Appendices

Appendix A Annotation guidelines

Objective: classify the *gender* encoded in the translated Ukrainian *sentence*.

- If the sentence contains *exclusively* masculine markings on its occupation name, occupation-related adjectives and occupation-related verbs, label the sentence as `male`.
  - "Бармен зателефонував мені сьогодні." -> `male` (occupation and verb are masculine; no feminine markings)
- If the sentence contains *any* feminine markings on its occupation name, occupation-related adjectives or occupation-related verbs, label the sentence as `female`.
  - "Анна є *досвідченою* науковцем." -> `female` (adjective is feminine)
  - "Я колись *працювала* кухарем." -> `female` (verb is feminine)
  - "*Кухня* зателефонував мені сьогодні." -> `female` (occupation name is wrong, though its gender is feminine)
- If there is no occupation name, adjective, or verb that could help identify the gender in the sentence, mark the sentence as `none`.
  - "Том погано продає." -> `none`

Appendix B All Possible Forms of the Word *Programmer* Translated to Ukrainian

	Singular		Plural	
Case	**Masc.**	**Fem.**	**Masc.**	**Fem.**
Nominative	програміст	програмістка	програмісти	програмістки
Genitive	програміста	програмістки	програмістів	програмісток
Dative	програмісту / програмістові	програмістці	програмістам	програмісткам
Accusative	програміста	програмістку	програмістів	програмісток
Instrumental	програмістом	програмісткою	програмістами	програмістками
Locative	програмісту	програмістці	програмістах	програмістках
Vocative	програмісте	програмістко	програмісти	програмістки

Table 8: All inflection forms of *programmer* in Ukrainian

Appendix C Examples of Gender-Tagged Translations

Tagged Source	**Ukrainian Output**
<female> What's her teacher's name?	Як звати її вчителя?
<male> When he saw the police officer, he ran away.	Коли він побачив поліцейського, він втік.
<neutral> When I grow up, I want to be a doctor.	Коли я виросту, я хочу бути лікарем.

Table 9: Example gender-tagged sentences

Appendix D Complete List of Occupations Used for Generating Test Sentences

Accountant, administrator, advisor, agent, appraiser, architect, assistant, auditor, babysitter, baker, bartender, beautician, broker, builder, carpenter, cashier, chef, chemist, cleaner, clerk, cosmetologist, cook, counselor, dentist, designer, dietitian, dispatcher, doctor, driver, educator, electrician, engineer, examiner, firefighter, florist, gynecologist, hairdresser, hairstylist, housekeeper, hygienist, inspector, instructor, investigator, janitor, lawyer, librarian, machinist, manager, mechanic, nurse, nutritionist, officer,

painter, paralegal, paramedic, pathologist, pediatrician, pharmacist, physician, pilot, planner, plumber, practitioner, programmer, psychologist, receptionist, salesperson, scientist, secretary, specialist, soldier, stylist, supervisor, surgeon, teacher, technician, therapist, veterinarian, worker.

Appendix E Complete Lists of Words Used for Gender Tags

Female Gender-Tagged Words. *Pronouns:* she, her, hers, herself. *General gendered nouns:* woman, women, lady, ladies, female, girl, girls, mother, mom, mommy, mum, mama, grandmother, grandma, granny, daughter, niece, sister, aunt, queen, princess, duchess. *Gendered roles/professions:* actress, waitress, stewardess, goddess, heroine, policewoman, chairwoman, businesswoman, saleswoman, spokeswoman, craftswoman, congresswoman, sportswoman, countrywoman, housewife, midwife, nun, ballerina, maid, nanny, seamstress, bridesmaid. *Relationship terms:* wife, girlfriend, fiancée, bride, mistress, widow, bachelorette, stepmother, stepdaughter, mother-in-law, sister-in-law, auntie. *Common female first names:* anna, mary, maria, elizabeth, sophia, emma, isabella, olivia, ava, mia, amelia, grace, claire, lucy, ella, anne.

Male Gender-Tagged Words. *Pronouns:* he, him, his, himself. *General gendered nouns:* man, men, male, boy, boys, gentleman, gentlemen, father, dad, daddy, grandfather, grandpa, son, nephew, brother, uncle, king, prince, duke. Gendered roles/professions: actor, waiter, steward, god, hero, policeman, chairman, businessman, salesman, spokesman, craftsman, congressman, sportsman, countryman, fireman, mailman, fisherman, repairman, workman, watchman, deliveryman, hunter, soldier, monk, priest, carpenter, blacksmith. Relationship terms: husband, boyfriend, fiancé, groom, widower, bachelor, stepfather, stepson, father-in-law, brother-in-law, uncle. *Common male first names:* john, michael, james, robert, david, william, alexander, daniel, matthew, joseph, george, thomas, charles, henry, andrew, peter, paul, mark, lucas, benjamin, tom.

Appendix F Occupation prestige categories

Prestige	**Occupations**
High (ISEI 65+)	surgeon, doctor, physician, psychologist, pharmacist, veterinarian, pathologist, dentist, gynecologist, pediatrician, dermatologist, lawyer, engineer, architect, chemist, scientist, pilot, specialist
Mid (ISEI 40-64)	manager, supervisor, administrator, accountant, auditor, planner, programmer, technician, specialist, advisor, broker, appraiser, agent, designer, educator, teacher, instructor, librarian, paramedic, nurse, therapist, counselor, dietitian, nutritionist, practitioner, hygienist, paralegal, inspector, examiner, investigator, dispatcher, officer, firefighter, chef, pharmacist, machinist
Low (ISEI below 40)	mechanic, electrician, secretary, plumber, carpenter, bartender, cook, baker, hairdresser, hairstylist, stylist, beautician, cosmetologist, receptionist, clerk, cashier, salesperson, janitor, cleaner, housekeeper, babysitter, florist, painter, builder, driver, soldier, worker, assistant

Table 10: Occupations organised by prestige tier, based on the International Socio-Economic Index of Occupational Status (ISEI) (Ganzeboom et al., 1992).

Appendix G Occupation type categories

SOC Major Group	Occupations
11-0000 Management	manager, supervisor
13-0000 Business and Financial Operations	auditor, accountant, administrator, planner, broker, appraiser, agent, advisor, examiner, inspector, specialist
15-0000 Computer and Mathematical	programmer
17-0000 Architecture and Engineering	engineer, architect
19-0000 Life, Physical, and Social Science	chemist, scientist, psychologist
21-0000 Community and Social Service	counselor
23-0000 Legal	lawyer, paralegal
25-0000 Educational Instruction and Library	teacher, instructor, librarian, educator
27-0000 Arts, Design, Entertainment, Sports, and Media	designer, florist
29-0000 Healthcare Practitioners and Technical	paramedic, surgeon, practitioner, dietitian, doctor, therapist, pharmacist, physician, nutritionist, veterinarian, pathologist, hygienist, nurse, dentist, gynecologist, pediatrician, dermatologist, technician
33-0000 Protective Service	firefighter, investigator, officer, soldier
35-0000 Food Preparation and Serving Related	bartender, chef, cook, baker
37-0000 Building and Grounds Cleaning and Maintenance	janitor, cleaner, housekeeper
39-0000 Personal Care and Service	hairdresser, babysitter, beautician, cosmetologist, hairstylist, stylist
41-0000 Sales and Related	salesperson, cashier
43-0000 Office and Administrative Support	secretary, receptionist, clerk, dispatcher, assistant
47-0000 Construction and Extraction	electrician, plumber, carpenter, painter, builder
49-0000 Installation, Maintenance, and Repair	mechanic
51-0000 Production	machinist
53-0000 Transportation and Material Moving	driver, pilot
Unclassified	worker

Table 11: Occupations organised by SOC major group (U.S. Bureau of Labor Statistics, 2018).

Table 12: Occupations by Gender Stereotype and Prestige Tier

Occupation	Prestige (ISEI)	% Women	Total Employed (thousands)
Panel A: Male-stereotyped occupations			
Surgeons	High	25.9	73
Physicians (other)	High	41.7	929
Dentists	High	38.6	192
Lawyers	High	42.9	1,146
Engineers (all other)	High	17.9	685
Architects	High	29.8	254
Chemists and materials scientists	High	31.1	106
Aircraft pilots and flight engineers	High	7.0	201
Managers (all other)	Mid	37.7	5,471
Computer programmers	Mid	18.5	381
Construction and building inspectors	Mid	12.6	120
Police officers	Mid	16.4	705
Firefighters	Mid	5.1	352
Chefs and head cooks	Mid	26.4	516
Machinists	Low	5.8	309
Automotive service technicians and mechanics	Low	1.8	901
Electricians	Low	3.5	1,063
Plumbers, pipefitters, and steamfitters	Low	3.1	624
Carpenters	Low	3.1	1,178
Painters and paperhangers	Low	11.1	485
Driver/sales workers and truck drivers	Low	7.7	3,583
Panel B: Female-stereotyped occupations			
Psychologists	High	71.3	152
Pharmacists	High	62.7	365
Veterinarians	High	70.9	91
Postsecondary teachers	Mid	52.3	1,058
Elementary and middle school teachers	Mid	79.2	3,511
Librarians and media collections specialists	Mid	84.9	195
Registered nurses	Mid	87.3	3,528
Therapists (all other)	Mid	78.1	323
Counselors (all other)	Mid	74.5	257
Dietitians and nutritionists	Mid	91.7	136
Dental hygienists	Mid	95.0	203
Paralegals and legal assistants	Mid	82.4	486
Secretaries and administrative assistants	Low	91.9	1,491
Hairdressers, hairstylists, and cosmetologists	Low	92.0	747
Receptionists and information clerks	Low	89.6	1,247
Cashiers	Low	68.5	2,490
Maids and housekeeping cleaners	Low	86.4	1,392
Childcare workers	Low	93.2	1,106
Medical assistants	Low	90.8	680
Panel C: Gender-neutral occupations			
Accountants and auditors	Mid	59.1	1,766
Designers (other)	Mid	45.6	379
Paramedics	Mid	28.5	117
Dispatchers (exc. police, fire)	Mid	61.7	180
Bartenders	Low	57.2	432
Cooks	Low	43.7	2,099
Bakers	Low	62.5	277
Retail salespersons	Low	47.0	2,661
Janitors and building cleaners	Low	36.2	2,341

Note: Gender stereotype panels are assigned based on the proportion of women in each occupation: male-stereotyped (<35%), female-stereotyped (>65%), and gender-neutral (35–65%). Prestige tiers follow the International Socio-Economic Index (ISEI): High (≥65), Mid (40–64), Low (<40). Total employed figures are in thousands and reflect annual 2025 averages for the civilian population aged 16 and over.

Source: U.S. Bureau of Labor Statistics. (2025). *Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity* (Table 11). Current Population Survey. `https://www.bls.gov/cps/cpsaat11.htm`