Integrating gender inclusivity into large language models via instruction tuning

Anonymous ACL submission

Abstract

Imagine a language with masculine, feminine, and neuter grammatical genders, yet, due to historical and political conventions, masculine forms are predominantly used to refer to men, women and mixed-gender groups. This is the reality of contemporary Polish and some other Slavic and Romance languages. A social consequence of this unfair linguistic system is that large language models (LLMs) trained on standard texts inherit and reinforce this masculine bias, generating gender-imbalanced outputs. This study addresses this issue by tuning LLMs using Inclusive Polish Instruction Set (IPIS), a collection of human-crafted gender-inclusive instructions for proofreading in Polish and Polish↔English translation. Grounded in a theoretical linguistic framework, we design a system prompt with explicit gender-inclusive guidelines for Polish. In our experiments, we IPIS-tune multilingual LLMs (Llama-8B, Mistral-7B and Mistral-Nemo) and Polish-specific ones (Bielik and PLLuM). Our approach successfully integrates gender inclusivity as an inherent feature of these models, offering a systematic solution to mitigate gender bias in Polish language generation.

1 Introduction

Large language models (LLMs) are trained on extensive datasets primarily sourced from the internet, which reflect standard language usage, including common biases and stereotypical societal patterns. LLMs are highly effective at internalising and replicating these patterns. Most of the textual data used for LLM training is in English — a language with sparse gender markers, in contrast to *grammatical gender languages* (Reali et al., 2015) such as Polish. Hence, if LLMs develop gender-related patterns at all, they acquire them mainly from languages that are underrepresented in the training data. In grammatical gender languages, training data may employ generic masculine forms to refer to both men and women. Consequently, LLMs trained on such data show a strong masculine bias. This challenges a claim made by Ovalle et al. (2024) that LLMs are primarily gender binary-centric. Their claim, based on English, does not universally apply to all languages. This is particularly evident in Polish, where LLM-generated texts strongly favour masculine-marked nouns, pronouns, adjectives, and verbs (Wróblewska et al., 2025).

The dominance of masculine expressions over feminine ones is a form of gender discrimination through linguistic means (GEC, 2016; GNL-EU, 2018). Acknowledging the harmful effects of sexist language, the Council of Europe encourages member states to eliminate sexism from language and adopt linguistic practices that promote gender equality. Enhancing gender sensitivity in a language is a complex and slow process due to deeply rooted traditional linguistic expressions, social resistance, the lack of clear guidelines, and most significantly, the natural pace of language evolution. Nowadays, LLMs have become essential tools for text generation, translation, proofreading, communication assistance, and knowledge acquisition. However, masculine-centric LLMs can threaten language evolution by slowing or even stopping the shift toward a gender-inclusive language.

The gender bias problem and gender inclusiveness are very important topics in contemporary NLP, as manifested by the organisation of two annual workshops that brought together the gender-inclusive NLP community: *Workshop on Gender Bias in NLP* (Faleńska et al., 2024) and *Workshop on Gender-Inclusive Translation Technologies* (Savoldi et al., 2024). In these venues, debiasing methods tested in various NLP tasks are presented, including language modelling and generating, machine translation, relation extraction, or authorship profiling. These fora highlight the growing significance of gender-inclusive NLP as a critical area of research. Its key aspect should be a development of

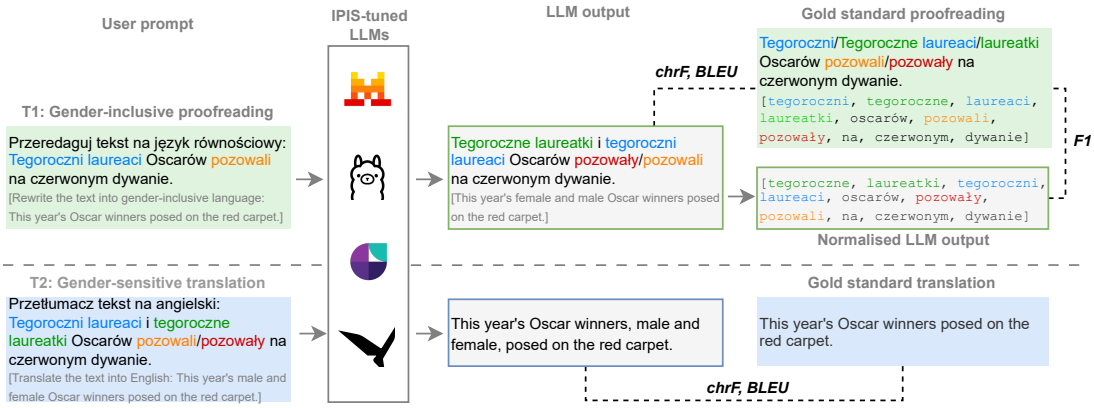

Figure 1: Flowchart illustrating the inference and evaluation workflows of IPIS-tuned LLMs for two tasks: gender-inclusive proofreading (top section) and gender-sensitive Polish-to-English translation (bottom section). Explanations: masculine nouns and adjectives, masculine verbs, feminine nouns and adjectives and feminine verbs.

gender-inclusive LLMs that generate texts that neither discriminate nor marginalise individuals based on gender while avoiding gender stereotypes. Such models should consistently use gender-inclusive language to propagate gender equality.

This study aims to verify whether instruction tuning can be effectively employed to develop gender-inclusive LLMs (see Figure 1). The conducted experiments investigate to what extent the instruction-tuning strategy mitigates gender biases while enhancing fairness and inclusivity in generated outputs, particularly in the Polish proofreading task and Polish↔English translation. To our knowledge, this is the first attempt to integrate gender inclusivity into LLMs via instruction tuning.

Our contributions are threefold: **(1)** We publicly release **Inclusive Polish Instruction Set (IPIS)** (see Section 3). To the best of our knowledge (cf. Amrhein et al., 2023), this is the first instruction dataset of parallel biased segments and their gender-fair counterparts, annotated by human experts specialising in Polish linguistics. **(2)** We design a **system prompt** aimed at guiding LLMs on generating texts in gender-inclusive Polish (see Section 4.2) This prompt is composed in two language variants – Polish and English. It serves as a guiding framework, ensuring that generated content adheres to the principles of gender-inclusive language and text genres. **(3)** We conduct a series of experiments to identify the optimal configuration for LLM instruction tuning, aiming at developing **gender-inclusive LLMs**, i.e., LLMs that generate contents in gender-inclusive Polish (Sections 5 and 6). Top-performing IPIS-tuned LLMs are made publicly available.

2 Proposed approach

This study examines instruction tuning as a strategy for developing gender-inclusive LLMs for Polish.

2.1 Polish – a masculine-centric language

Polish is a grammatical gender language in which all nouns are marked for grammatical gender, e.g., *śliwka* [a pllum] is feminine, *pomidor* [a tomato] is masculine, and *jabłko* [an apple] is neutral. All adjectives and verbs match the noun's grammatical gender. Additionally, *personal nouns* (Regis, 2020; Backer and Cuypere, 2012) have distinct feminine, e.g., *nauczycielka* [a teacher$_{\text{fem}}$] and masculine forms, e.g., *nauczyciel* [a teacher$_{\text{masc}}$]. While feminine personal nouns typically denote female individuals or groups of females, masculine personal nouns refer not only to males or male groups but also to mixed-gender groups and even females – a phenomenon known as *generic masculine*, e.g., *niemiecka polityk, Ursula von der Leyen* [German$_{\text{fem}}$ politician$_{\text{masc}}$, Ursula von der Leyen].

Although the grammatical system of Polish allows for naming individuals according to their natural gender (i.e., female or male), standard Polish remains heavily masculine-centric. This is reflected in a strong dominance of masculine over feminine expressions, leading to linguistic structures that inherently reinforce gender bias and exclusion.

Adapting language to reflect natural genders is a challenging and time-consuming process. The outcome of this process will be a gender-inclusive language – a communication medium that avoids prejudice, discrimination and gender stereotypes. We use the term *gender-inclusive* (or *gender-fair*) to re-

fer to a language in which all genders are explicitly marked, as proposed by Amrhein et al. (2023). This term differ from *gender-neutral* language, where no gender is specifically indicated. While gender-inclusive language represents a highly desirable variant of Polish, a fully gender-neutral Polish is largely unnatural and generally inapplicable.

LLMs trained on texts in standard Polish are inherently masculine-centric. Instead of focusing on identifying and mitigating gender biases, our objective is to develop a gender-inclusive LLM. We aim to embed gender inclusivity as a core principle into LLM through instruction tuning. This approach will lead to LLMs that consistently generate gender-inclusive language.

2.2 Gender-inclusive instruction tuning

Instruction tuning is a fine-tuning technique designed to enhance the performance and controllability of LLMs across a wide range of tasks (Zhang et al., 2024). This technique consists in further training LLMs on a labelled dataset of task-specific instructional prompts and corresponding outputs, following a supervised fine-tuning paradigm. The primary objective is to improve LLM's ability to solve specific tasks. The indirect goal is to enhance its proficiency in following instructions in general. According to Wei et al. (2022), expanding an instruction-tuning dataset with additional tasks led to improved model performance, even on novel, previously unseen tasks, suggesting a comprehensive improvement of model's instruction-following capabilities. Building on these findings, we extend the instruction-tuning process by further refining LLMs with an additional set of gender-inclusive instructions. We aim to enhance the ability of LLMs to generate texts that align with gender-inclusive principles, e.g., masculine forms referring to women are replaced with existing feminine forms. By tuning LLMs with carefully composed gender-inclusive prompts and their desired responses, we seek to develop LLMs that not only follow instructions effectively but also contribute to inclusive language use.

Input samples within the instruction dataset closely resemble real-world user requests, ensuring that LLMs are tuned on prompts representative of actual interactions. The corresponding outputs serve as ideal responses to these requests. Instruction datasets can be composed of either LLM-generated synthetic instructions or human-written ones. To ensure that LLMs do not reinforce biases nor emulate the undesirable behaviours of larger models, we tune them exclusively on manually prepared instructions. This human oversight is essential for maintaining ethical standards and mitigating gender biases in language generation.

For LLM instruction tuning, we employ low-rank adaptation (LoRA) – a parameter-efficient fine-tuning method (Hu et al., 2021). LoRA enables effective model adaptation by updating only a small subset of parameters while keeping the pre-trained model weights frozen. This is achieved by injecting trainable low-rank decomposition matrices into each layer of the transformer architecture, enabling LLMs to adapt to new tasks. This fine-tuning technique significantly reduces computational costs and memory requirements while maintaining the generalisation capabilities of foundation models.

2.3 Gender-inclusive tasks

In the proposed instruction-tuning approach using LoRA, we adapt LLMs to perform two tasks: **gender-inclusive proofreading** and **gender-sensitive translation**. The first task consists in transforming a text written in standard Polish into its gender-inclusive version. The second task focuses on bidirectional translation between English and gender-inclusive Polish. By tuning LLMs on these specific tasks, we aim to enhance their ability to generate texts that are both linguistically accurate and aligned with gender inclusivity principles.

3 Inclusive Polish Instruction Set (IPIS)

The IPIS dataset[1] is a collection of instructions designed to improve the gender sensitivity and inclusiveness of LLMs in the Polish language scenario. The IPIS dataset is human-constructed using texts from various open-access sources representing diverse genres (e.g., European and Polish legal acts, institutional documents, narratives, news, parliamentary debates, etc.). Texts originally written in masculine-centric Polish are manually transformed into their inclusive versions, resulting in a parallel androcentric–gender-inclusive corpus. Each IPIS sample (see Figure 2 in Appendix A) consists of:

1. **user prompt** (`prompt`) – a specification of the given task (see *User prompt* in Section 4.2),
2. **input text passage** (`source`) – a text passage requiring a gender-inclusive proofreading or gender-sensitive translation,

[1] After publication, the IPIS dataset will be released on HuggingFace under an open-source research licence.

3. **desired output** (`target`) – an expected response based on user prompt and input text passage. This serves as the ground truth for evaluating and optimising LLM's predictions.

IPIS-translation instances (see Figure 3 in Appendix A) additionally include: `prompt_language`, `source_language` (the language of the passage to be translate) and `target_language` (the language of the reference translation), whose values are restricted to either PL (Polish) or EN (English).

The IPIS-proofreading subsets contain 5,278 instances in *test*, 2,732 in *dev* and 23,532 in *train*. The IPIS-translation subsets comprise 456 in *test*, 304 in *dev* and 1728 in *train*. IPIS *test*, *dev* and *train* subsets are balanced for the ratio of gender-inclusive transformations. The IPIS dataset serves as the basis for fine-tuning and evaluating LLMs.

4 Experimental setup

4.1 Tested LLMs

Our study aims to determine whether LLMs can be effectively fine-tuned on IPIS, ensuring socially responsible language generation in Polish. We conduct a series of experiments using open-source LLMs based on the transformer architecture (Vaswani et al., 2017). We focuses on instruction-tuned, small- and medium-sized versions of both **multilingual LLMs**: *Llama-8B*, *Mistral-7B*, and *Mistral-Nemo*, and **Polish-specific models**: *Bielik-7B*, *Bielik-11B*, *Llama-PLLuM-8B*, and *PLLuM-12B* (see Appendix B for LLMs' details).

4.2 Prompt engineering

User prompt User prompts are simple 1-2 sentence instructions that specify a given task. In the gender-inclusive proofreading task, LLM should proofread a text for gender inclusivity (see Ex. (1)).

(1) *Przekształć tekst na jego wersję inkluzywną. Tekst źródłowy:* (Eng. 'Transform the text to its inclusive version. Source text:')

In gender-sensitive translation, LLM should translate an English text passage into gender-inclusive Polish or a Polish gender-inclusive text passage into English (Ex. (2)).

(2) *Translate the English text into gender-inclusive Polish:*

There are 260 manually written rephrasings for the proofreading user prompts, along with 127 to 324 translation prompt wordings for various language configurations. These user prompts are randomly distributed across IPIS instances.

System prompt We investigate the impact of including a system prompt with gender-inclusive guidelines on the training and evaluation of LLMs. The development of this system prompt is grounded in the theoretical framework proposed by Wróblewska et al. (2025). The authors provide a comprehensive overview of gender-inclusive strategies and notations in Polish, followed by an analysis of various text genres concerning the implementation of appropriate inclusive strategies. The system prompt is originally composed in Polish and translated into English (see Appendix C).

Since persona prompting is an effective prompt engineering strategy (Xu et al., 2023; Kong et al., 2024), we employ it and assign an editor role to LLMs. An editor-model is instructed to follow a predefined algorithm and transform texts according to the gender-inclusive language principles specified for a particular genre. The system prompt defines the exclusionary expressions in Polish and specifies their gender-inclusive alternatives permitted in various text genres. Additionally, it guides LLMs on grammatical agreement rules, ensuring that generated outputs are both gender-inclusive and grammatically correct.

4.3 Implementation details

LLMs are instruction-tuned using DeepSpeed ZeRO, Flash Attention and HuggingFace's ecosystem (*transformers*, *trl*, *peft* and *datasets* libraries). DeepSpeed is a library designed for compute and memory optimisation in distribtuted training setups. HuggingFace's *datasets* and *transformers* libraries facilitate the loading and usage of the IPIS dataset and selected LLM's while *trl* provides a dedicated trainer for supervised fine-tuning (SFT). LoRA adapters are configured using HuggingFace's *peft* library. Flash Attention is used to enhance the efficiency of the core attention mechanism in transformer-based models (see the LoRA configuration and hyperparameters of supervised fine-tuning (SFT) in Appendix D).

During instruction tuning, we compute the cross-entropy loss for both prompt and completion tokens (default setting in *trl.SFTTrainer*), meaning that the model also learns the structure of the input requests. The prompt and completion losses are minimised on the *train* set while being tracked on the *dev* set.

5 Evaluation methodology

5.1 Evaluation metrics

To evaluate the ability of LLM to process and generate gender-inclusive language, their outcomes are compared against gold standard test instances. For gender-inclusive proofreading, we evaluate normalised LLM-generated texts with accuracy, precision, recall, and F_1-measure. The normalisation process consists in transforming all occurrences of gender-inclusive expressions (i.e., those containing gender stars and slashes) into their full masculine and feminine forms (see normalisation examples in Appendix E). Furthermore, tokens which are automatically identified as punctuation marks, subordinating and coordinating conjunctions are filtered out. We use Lambo (Przybyła, 2022) for tokenisation and Combo (Klimaszewski and Wróblewska, 2021) for part-of-speech tagging.

Additionally, we calculate the similarity between LLM-proofread passages and reference gender-inclusive passages, using automatic MT metrics: chrF (Popović, 2015) and BLEU (Papineni et al., 2002).[2] As LLMs are expected to preserve the original wording and not hallucinate new content while introducing gender-inclusive edits, these scores should remain relatively high. Low scores indicate undesirable alterations to the text (i.e., paraphrases, unnecessary insertions or deletions of words, and unnecessary gender-inclusive edits) or missing gender-inclusive forms.

These MT metrics are also used to assess translation quality, specifically in evaluating gender-neutral English translations of gender-sensitive Polish source texts, as well as gender-inclusive Polish translations of standard English source texts (see details in Appendix F).

5.2 Baseline scenarios

To evaluate the impact of instruction tuning and system-prompt guidance on model performance, we define a set of baseline configurations. The main model under study is an IPIS-tuned LLM (denoted *tuned*), which can additionally be provided with a gender-inclusive system prompt (*tuned*-{*pl*|*en*}). All IPIS-tuned LLMs are compared with their off-the-shelf counterparts (*default*), evaluated in a zero-shot setting, optionally augmented with the system prompt (*default*-{*pl*|*en*}). Further baselines include few-shot variants of default models, evaluated without the system prompt (*fewshot*) or with it (*fewshot*–{*pl*|*en*}). This experimental setting enables a systematic comparison by isolating the effects of instruction tuning, system-prompt guidance, and in-context learning, thus providing a robust performance benchmark for IPIS-tuned LLMs.

[2] With an emphasis on explainability and exact mappings between tokens that must remain unchanged and those requiring gender-inclusive alternations, we do not employ ML-based metrics such as BERTscore or COMET.

6 Results and discussion

6.1 Gender-inclusive proofreading evaluation

Impact of the LLM leading language Polish-specific models (see Section 4.1), derived from multilingual models, are further fine-tuned on large Polish text corpora and other NLP datasets to improve their performance on tasks requiring a deeper understanding of Polish. However, Polish-specific default LLMs show no improvement over their multilingual predecessors, with F_1 scores close to zero, in both the zero-shot and few-shot setting (see Table 6 in Appendix G for detailed results). This finding, while disappointing, is expected, since texts used for fine-tuning Polish-specific default LLMs are mostly written in standard, non-inclusive Polish.

Table 1: **Leading language** and **LLM size** factors. Performance of the multilingual Mistral-Nemo-12B model and the Polish-specific Llama-PLLuM-8B and Bielik-11B models in their *default*, *few-shot* and IPIS-*tuned* versions, on the gender-inclusive proofreading task.

LLM	Acc	Prec	Rec	F_1	BLEU	chrF
Llama-PLLuM-8B						
default	34.88	0.18	0.22	0.20	32.04	57.13
fewshot	31.34	0.49	0.58	0.53	37.51	52.38
tuned	97.08	61.91	46.40	53.04	94.28	97.64
Bielik-11B						
default	41.79	0.32	0.59	0.42	39.12	66.54
fewshot	56.21	1.09	4.57	1.76	52.91	80.23
tuned	**97.37**	**63.93**	**56.26**	**59.85**	**95.22**	**97.99**
Mistral-Nemo-12B						
default	63.46	0.34	0.37	0.35	62.98	78.33
fewshot	68.12	0.74	1.65	1.02	68.43	84.95
tuned	96.82	57.16	48.81	52.66	94.45	97.72

Table 1 reports the scores of the best-performing IPIS-tuned models – Mistral-Nemo-12B (multilingual) and Bielik-11B (Polish-specific), which substantially outperform their baseline configurations (*default* and *fewshot*). The leading language of an IPIS-tuned LLM appears to influence its effectiveness in gender-inclusive proofreading, with the Polish-specific Bielik-11B achieving a higher F_1

score (59.85) than Mistral-Nemo ($F_1 = 52.66$).

In terms of BLEU and chrF similarity scores, the IPIS-tuned models achieve high and comparable performance, confirming the effectiveness of instruction tuning in enhancing the gender inclusiveness of LLMs. However, a closer examination of the *default* and *fewshot* baselines reveals a clear advantage of Mistral-Nemo-12B over Bielik-11B, suggesting that large-scale multilingual pretraining provides benefits that extend beyond linguistic adaptation in this task. Moreover, the *fewshot* baselines outperform their *default* counterparts, particularly in the case of Bielik-11B, indicating that additional in-context examples can improve proofreading competence, although still insufficient to match the performance of the IPIS-tuned models.

Impact of the LLM size As expected, larger IPIS-tuned models generally outperform smaller ones, as reflected in their average F_1 scores: 57.04 vs. 49.46 (see Table 6 in Appendix G). Table 1 presents the performance of the best small and medium IPIS-tuned LLMs – Llama-PLLuM-8B and Bielik-11B, respectively. In terms of precision, Bielik-11B slightly outperforms LLama-PLLuM-8B (63.93% vs. 61.91%). However, its recall is higher by 10 pp, indicating that the smaller model is less effective at detecting gender-biased expressions, despite being comparably precise at accurately proofreading detected expressions.

Table 2: **System prompt** factor. Performance of the Polish-specific Bielik-11B model in its *default*, *fewshot* and IPIS-*tuned* versions possibly with a system prompt (*pl* or *en*) on the gender-inclusive proofreading task.

LLM	Acc	Prec	Rec	F_1	BLEU	chrF
default	41.79	0.32	0.59	0.42	39.12	66.54
default-pl	60.55	1.45	9.34	2.51	56.56	83.93
default-en	60.41	1.60	13.62	2.86	55.79	84.72
fewshot	56.21	1.09	4.57	1.76	52.91	80.23
fewshot-pl	59.15	1.35	11.83	2.42	54.92	84.01
fewshot-en	58.57	1.31	8.74	2.28	53.69	82.93
tuned	**97.37**	**63.93**	**56.26**	**59.85**	**95.22**	**97.99**
tuned-pl	93.66	29.24	50.32	36.99	91.82	96.93
tuned-en	96.47	52.30	51.59	51.94	94.82	97.61

Impact of the system prompt All models are trained with a system prompt in Polish (*pl*) or English (*en*), and without it. Table 2 reports the scores for the best-performing IPIS-*tuned* Bielik-11B model and its *default* and *fewshot* variants. The IPIS-*tuned* Bielik-11B trained without any system prompt significantly outperforms all its baseline models, and it also achieves superior proofreading performance compared to its IPIS-tuned alternatives guided by system prompts. This pattern generalises across all evaluated models: while the inclusion of the system prompts generally benefits the baseline models, it does not yield improvements for the IPIS-tuned ones (see Table 6).

The language of the system prompt – whether Polish or English – does not have a significant impact on performance. Although the Polish prompt tends to be marginally more effective for baseline models and the English one for IPIS-tuned models, these differences are negligible and not significant.

Key findings IPIS-tuned LLMs substantially outperform baseline models. Given that IPIS-tuned Polish-specific models demonstrate greater inclusivity than their multilingual variants, we infer that language-specific models are better suited for instruction-level optimisation in this language. Moreover, IPIS-tuned LLMs show a noticeable pattern – they exhibit higher precision than recall. This suggests that they introduce gender-inclusive modifications primarily when they are highly confident, minimising false positives. Although this cautious strategy enhances precision, it also results in the omission of required gender-inclusive edits. Both BLEU (word n-gram-based) and chrF (character n-gram-based) scores are comparable and exceed 90 across all IPIS-tuned models, meaning that the proofread texts are of high quality, with minimal out-of-vocabulary words, typos, morphosyntactic errors, or lexical mismatches. All these findings induce us to categorize IPIS-tuned LLMs as a significant step toward the development of truly gender-inclusive LLMs for Polish.

6.2 Evaluation of gender-sensitive translation

Translation quality In the gender-sensitive translation task, medium-sized models consistently outperform smaller ones. Among the evaluated models (see detailed scores in Table 7 in Appendix G), Bielik-11B achieves the best performance (see Table 3). However, the impact of IPIS-tuning is more limited in this context than in the gender-inclusive proofreading task. While the IPIS-tuned Bielik-11B outperforms the baseline models in the Polish-to-English translation direction (BLEU = 57.5 and chrF = 78.3), its alternative used under a few-shot learning setting delivers the best results in the reverse direction – BLEU = 43.2 and chrF = 72.8. Although the overall quality of translation is relatively low, it is comparable to the state of the art

Table 3: Performance of the Polish-specific Bielik-11B model in its *default*, *fewshot* and IPIS-*tuned* configurations possibly with a system prompt (*-pl* or *-en*) solving the **gender-sensitive translation** task.

LLM	Polish→English				English→Polish			
	PL user prompt		EN user prompt		PL user prompt		EN user prompt	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
bielik-11b-*default*	47.60	73.39	47.54	72.50	41.49	71.78	27.39	65.30
bielik-11b-*default-pl*	46.78	73.08	43.67	70.21	35.80	69.77	32.31	68.94
bielik-11b-*default-en*	47.99	73.76	36.39	68.39	32.70	68.65	32.13	68.54
bielik-11b-*fewshot*	50.01	73.93	49.66	73.99	38.63	68.78	33.19	68.11
bielik-11b-*fewshot-pl*	49.38	73.84	49.55	74.02	37.81	69.92	42.76	72.19
bielik-11b-*fewshot-en*	48.33	73.43	49.14	73.75	**43.02**	**72.46**	**43.17**	**72.82**
bielik-11b-*tuned*	55.19	75.80	**57.45**	**77.93**	34.26	55.08	35.04	55.92
bielik-11b-*tuned-pl*	56.70	76.93	55.24	75.35	31.70	58.36	26.74	55.96
bielik-11b-*tuned-en*	**57.55**	**78.03**	57.30	76.63	28.71	60.97	25.96	58.93

(Tiedemann et al., 2023): BLEU = 54.9 and chrF = 70.1 for Polish-to-English, and BLEU = 48.6 and chrF = 67.6 for English-to-Polish.

Based on this, we can infer that translating gender-inclusive Polish into English does not pose a major challenge for LLMs. With only a small number of IPIS instructions, Bielik-11B can be effectively tuned, achieving substantially better performance than the baselines and even beating the state-of-the-art models. The English-to-Polish direction, in turn, remains considerably more challenging. The IPIS-tuned Bielik-11B performs poorly, indicating that the available IPIS-translation instructions are insufficient to effectively adapt it for generating high-quality gender-inclusive Polish translations. During instruction tuning, an LLM adjusts its parameters to align with the task-specific examples provided. If the dataset is too small, the model may overfit to the limited set of instructions, learning narrow patterns rather than generalisable strategies. At the same time, fine-tuning on a small dataset can also overwrite parts of the model's prior translation competence, affecting its previously learned capabilities.

Language impact The language of user queries (prompts) does not have a significant impact on translation quality, as the scores for Polish and English user prompts within the same translation direction are largely comparable. The use of a system prompt – especially an English one – consistently improves performance in most cases, except for the Polish-to-English setting with the user query in English. For the Polish-to-English direction, system prompts generally provide little benefit and can even degrade performance. However, in the English-to-Polish direction, system prompts prove highly advantageous. For example, Bielik-11B-*fewshot* without any system prompt achieves BLEU = 33.19, while its version with the English system prompt reaches BLEU = 43.17.

Key findings Medium-sized models consistently outperform smaller ones in gender-sensitive translation. IPIS-tuning proves effective for Polish-to-English translation, but remains insufficient for the more challenging English-to-Polish direction, likely due to data scarcity and overfitting effects during instruction tuning. System prompts significantly improve performance in the English-to-Polish direction, though they may degrade results in the opposite direction.

7 Related work

7.1 Gender bias in language models

Language models frequently reproduce societal biases present in their training data, e.g. gender biases (Stanczak and Augenstein, 2021; Jumelet et al., 2019; Doyen and Todirascu, 2025). These biases often manifest in stereotypical, discriminatory or prejudicial models' outputs, leading to unequal treatment of women and men in hiring systems (Albaroudi et al., 2024), recommendation systems (Di Noia et al., 2022), and content generation (Chu et al., 2024). Ensuring fairness and inclusivity of AI-driven technologies has motivated extensive research on detecting biased patterns in pretrained models and mitigating them using various debiasing techniques (e.g., Sun et al., 2019; Choubey et al., 2021; Saunders et al., 2022; Borah and Mihalcea, 2024; Nemani et al., 2024). Our approach differs from conventional debiasing techniques, which aim to identify and mitigate gender biases in language models. Rather than applying a debiasing technique as a postprocessing step, we embed gender inclusivity directly into LLM training. Our instruction tuning approach results in inherently gender-inclusive LLMs that generate outputs aligned with gender-fair linguistic practices.

A significant area of debiasing research focuses on gender-fair text rewriting. The primary objective of this approach is to transform any input text, including output from biased generative or translation models, into a gender-inclusive version while preserving the original content. For example, Amrhein et al. (2023) proposed a gender-fair rewriting model for German, trained on a dataset constructed using reversed data augmentation and automatic translation techniques. Nunziatini and Diego (2024), in turn, applied GPT-4 to post-edit DeepL's raw outputs and generate gender-inclusive translations. While their objective aligns with ours, our approach focuses on enhancing the gender fairness of existing LLMs rather than developing a new rewriting model or using proprietary LLMs as rewriters.

In the context of our research, it is important to acknowledge the work of Martinková et al. (2023), one of the few studies addressing gender bias in Polish NLP. The authors measure gender bias in Polish and other West Slavic languages – Czech and Slovak. Although their study does not directly align with our research focus, we reference it to highlight a critical gap in the field. Despite Polish being a relatively well-resourced language in NLP, research on its gender inclusivity remains limited. This underscores the need for further studies in gender-inclusive NLP for Polish.

7.2 Instruction tuning

Instruction tuning is a commonly used technique to adapting LLMs to specific tasks and domains based on human-crafted or synthetic datasets (Zhang et al., 2024; Üstün et al., 2024). LLMs can be instruction-tuned for NLP tasks, such as classification (Rosenbaum et al., 2022), information extraction (Sun et al., 2024), and writing (Chakrabarty et al., 2022; Zhang et al., 2023; Raheja et al., 2023, 2024), but also for code generation and solving arithmetic problems.

From the perspective of our research, adapting models for writing tasks is of primary relevance. Raheja et al. (2023) introduce CoEdIT, a text editing system designed for writing assistance that uses models for the FlanT5 family tuned on text editing instructions. These instruction focus on tasks such as simplification, paraphrasing, formalisation, and improving fluency and coherence. Building on this, mEdIT (Raheja et al., 2024) extends CoEdIT to a multilingual setting. Similarly, (Zhang et al., 2023) tune LLaMA for writing assistance within the same range of tasks. Additionally, Chakrabarty et al. (2022) introduce CoPoet, a T5-based system designed for poetry generation. Our approach aligns with these works, as we tune LLMs for writing instructions, specifically gender-inclusive proofreading instructions. However, our research differs in terms of the domain and the language in which our gender-inclusive LLMs operate.

8 Conclusions

This study presents a pioneering attempt to integrate gender inclusivity into large language models via instruction tuning. A key challenge of instruction tuning lies in developing high-quality instructional data suitable for LLM training. To address this, we introduced the IPIS dataset, a human-crafted collection of gender-inclusive proofreading instructions in Polish, and gender-sensitive Polish–English translation instructions. By focusing on a less commonly studied language – Polish – and a novel topic – gender inclusivity, IPIS enriches existing open-source instruction tuning resources and expands the scope of LLM adaptation.

Our experimental setup involved tuning open-source multilingual LLMs (Llama, Mistral, and Mistral-Nemo) and Polish-specific models (Bielik and PLLuM). Importantly, since our approach uses only open-source LLMs, all experiments can be replicated, ensuring transparency and reproducibility. The results confirm that LLMs inherently lack gender sensitivity, as they are trained on standard, non-inclusive textual corpora. While instruction tuning can enable LLMs to generate gender-inclusive text to some extent in masculine-centric Polish, this process is highly complex. Not all masculine nouns should be replaced or augmented with feminine forms, as the appropriateness of such modifications depends heavily on context. Given the context-dependent nature of this task, LLMs should be capable of handling it effectively, provided they are trained on a sufficiently large and high-quality set of gender-inclusive instructions. As demonstrated in the gender-sensitive translation experiment, an insufficient number of instructions can even lead to model corruption.

Our approach aim to embed gender inclusivity as a core feature of LLMs, offering a systematic solution to mitigating gender bias in Polish language generation. By advancing gender-inclusive NLP and NLG, this research contributes to broader efforts to promote gender equality through language, highlighting the potential of NLG to facilitate more inclusive and equitable communication.

Limitations

In this work, we have developed and evaluated instruction-tuned LLMs capable of performing gender-inclusive proofreading texts in Polish. However, our approach has certain limitations that should be acknowledged. An obvious limitation is the linguistic scope of our research, as we focus mostly on Polish. While this restricts the generalisability of our findings to other languages, it also ensures that the tested LLMs are unlikely to use the IPIS dataset during pretraining, eliminating the risk of data contamination. Additionally, Polish presents a great challenge for LLMs due to its complex grammatical gender system and its explicit masculine bias, making it a particularly demanding language for gender-inclusive NLP. Despite this language-specific focus, our work addresses the broader issue of linguistic diversity in the English- and Chinese-dominated NLP world.

Another limitation relates to the computational resources required for our experiments. We tune multiple LLMs with billions of parameters, and although we employ LoRA to optimise efficiency, the overall computational cost remains substantial. This not only poses a financial barrier to replicating the results due to resource constraints but also has environmental implications. To mitigate this, we publicly release our models, facilitating further research while reducing redundant costs.

Finally, our evaluation is limited to gender inclusivity and does not examine potential trade-offs in other aspects of textual quality, such as fluency, coherence, and meaning. Future research should investigate whether instruction tuning for gender inclusivity affects these general linguistic properties to ensure that improved gender inclusivity does not damage overall text quality. Alternatively, we plan to evaluate the developed gender-inclusive LLMs on public benchmarking leaderboards.

References

Elham Albaroudi, Taha Mansouri, and Ali Alameer. 2024. A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring. *AI*, 5(1):383–404.

Chantal Amrhein, Florian Schottmann, Rico Sennrich, and Samuel Läubli. 2023. Exploiting biased models to de-bias text: A gender-fair rewriting model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4486–4506, Toronto, Canada. Association for Computational Linguistics.

Maarten De Backer and Ludovic De Cuypere. 2012. The interpretation of masculine personal nouns in german and dutch: a comparative experimental study. *Language Sciences*, 34(3):253–268.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent LLM interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9306–9326, Miami, Florida, USA. Association for Computational Linguistics.

Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022. Help me write a poem - instruction tuning as a vehicle for collaborative poetry writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6848–6863, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. 2021. GFST: Gender-filtered self-training for more accurate gender in translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1640–1654, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhibo Chu, Zichong Wang, and Wenbin Zhang. 2024. Fairness in large language models: A taxonomic survey. *Preprint*, arXiv:2404.01349.

Tommaso Di Noia, Nava Tintarev, Panagiota Fatourou, and Markus Schedl. 2022. Recommender Systems under European AI Regulations. *Communications of the ACM*, 65(4):69–73.

Enzo Doyen and Amalia Todirascu. 2025. Man made language models? evaluating llms' perpetuation of masculine generics bias. *Preprint*, arXiv:2502.10577.

Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza, editors. 2024. *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics, Bangkok, Thailand.

GEC. 2016. Gender Equality Glossary. Gender Equality Commission, Council of Europe.

GNL-EU. 2018. Gender-Neutral Language in the European Parliament. European Parliment.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi,

Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar

Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1–11, Hong Kong, China. Association for Computational Linguistics.

Mateusz Klimaszewski and Alina Wróblewska. 2021. COMBO: State-of-the-art morphosyntactic analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 50–62, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico. Association for Computational Linguistics.

Sandra Martinková, Karolina Stanczak, and Isabelle Augenstein. 2023. Measuring gender bias in West Slavic language models. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 146–154, Dubrovnik, Croatia. Association for Computational Linguistics.

Mistral AI team. 2024. Mistral NeMo. Accessed: Jan 20, 2025.

Praneeth Nemani, Yericherla Deepak Joel, Palla Vijay, and Farhana Ferdouzi Liza. 2024. Gender bias in transformers: A comprehensive review of detection and mitigation strategies. *Natural Language Processing Journal*, 6:100047.

Mara Nunziatini 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. European Association for Machine Translation (EAMT).

Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Adrian Gwoździej, Krzysztof Wróbel, SpeakLeash Team, and Cyfronet Team. 2024a. Bielik-11b-v2.3-instruct model card. Accessed: 2025-01-27.

Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, and Remigiusz Kinas. 2024b. Bielik 7B v0.1: A Polish Language Model – Development, Insights, and Evaluation. *Preprint*, arXiv:2410.18565.

Anaelia Ovalle, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter, and Rahul Gupta. 2024. Tokenization matters: Navigating data-scarce tokenization for gender inclusive language technologies. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1739–1756, Mexico City, Mexico. Association for Computational Linguistics.

Kishore Papineni, 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Consortium PLLuM. 2025. Pllum: A family of polish large language models.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Piotr Przybyła. 2022. LAMBO: Layered Approach to Multi-level BOundary identification.

Vipul Raheja, Dimitris Alikaniotis, Vivek Kulkarni, Bashar Alhafni, and Dhruv Kumar. 2024. mEdIT: Multilingual text editing via instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 979–1001, Mexico City, Mexico. Association for Computational Linguistics.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text editing by task-specific instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5274–5291, Singapore. Association for Computational Linguistics.

Chiara Reali, Yulia Esaulova, and Lisa Von Stockhausen. 2015. Isolating stereotypical gender in a grammatical gender language: Evidence from eye movements. *Applied Psycholinguistics*, 36(4):977–1006.

Riccardo Regis. 2020. Personal nouns (agent nouns) in the romance languages.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022. LINGUIST: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 218–241, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2022. First the worst: Finding better gender translations during beam search. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3814–3823, Dublin, Ireland. Association for Computational Linguistics.

Beatrice Savoldi, Janiça Hackenbuchner, Luisa Bentivogli, Joke Daems, Eva Vanmassenhove, and Jasmijn Bastings, editors. 2024. *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*. European Association for Machine Translation (EAMT), Sheffield, United Kingdom.

Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. *Preprint*, arXiv:2112.14168.

Lin Sun, Kai Zhang, Qingyuan Li, and Renze Lou. 2024. Umie: Unified multimodal information extraction with instruction tuning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19062–19070.

Tony Sun, 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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Ahmet Üstün, Viraat Aryabumi, Zheng 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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Alina Wróblewska, Martyna Lewandowska, Aleksandra Tomaszewska, Karolina Saputa, and Maciej Ogrodniczuk. 2025. Koncepcja form równościowych z asteryskiem inkluzywnym. *Język Polski*, CV(2).

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *Preprint*, arXiv:2305.14688.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023. Multi-task instruction tuning of llama for specific scenarios: A pre-

liminary study on writing assistance. *Preprint*, arXiv:2305.13225.

A IPIS instances

IPIS-proofreading example (see Figure 2)
Each IPIS-proofreading sample consists of three components:

(1) **prompt** – user prompt, i.e., a specification of the given task (e.g., 'Revise the text in standard Polish so that it does not contain any harmful or exclusionary content. Source text:'),

(2) **source** – input text passage, i.e., a text passage requiring a gender-inclusive proofreading or gender-sensitive translation,

(3) **target** – desired output, i.e., the expected response corresponding to a user instruction and input text passage. This serves as the ground truth for evaluating and optimising LLMs.

```
{'source_resource_id': 'nlprepl-nkjp1m-dev',
 'ipis_id': 'IPIS_proofreading_dev_2143',
 'prompt': 'Przeredaguj tekst w standardowym
    języku polskim, aby nie zawierał treści
    krzywdzących i wykluczających. Tekst
    źródłowy: ',
 'source': 'W 24-osobowym składzie sędziowskim
    znalazło się 8 Polaków, pianistów i
    pedagogów. W przypadku własnych
    uczniów nie będą mieli prawa głosu.
    Pojawią się też m.in. laureat nagrody
    za najlepsze nagranie Roku Chopinowskiego
    - Emanuel Ax i rosyjska pianistka
    reprezentująca USA - Bella Davidovich.',
 'target': 'W 24-osobowym składzie sędziowskim
    znalazło się 8 Pol*aków/ek, pianist*ów/ek i
    pedago*gów/żek. W przypadku własnych
    ucz*niów/ennic nie będą mi*eli/ały prawa głosu.
    Pojawią się też m.in. laureat nagrody
    za najlepsze nagranie Roku Chopinowskiego
    - Emanuel Ax i rosyjska pianistka
    reprezentująca USA - Bella Davidovich.'}
```

Figure 2: IPIS-proofreading example. [Eng. The 24-member jury includes 8 Polish pianists and educators. They will be excluded from voting for their own students. Among the jurors are also Emanuel Ax, the winner of the award for the best recording of the Chopin Year, and Bella Davidovich – a Russian-born pianist representing the USA.]

IPIS-translation example (see Figure 3)
Each IPIS-translation sample consists of three components:

(1) **prompt** – user prompt, i.e., a specification of the given task (e.g., 'Revise the text in standard Polish so that it does not contain any harmful or exclusionary content. Source text:'),

(2) **source** – input text passage, i.e., a text passage requiring a gender-inclusive proofreading or gender-sensitive translation,

(3) **target** – desired output, i.e., the expected response corresponding to a user instruction and input text passage. This serves as the ground truth for evaluating and optimising LLMs,

(4) **prompt_language** – the language of prompt (either EN or PL),

(5) **source_language** – the language of a passage to translate, either inclusive Polish (PL) or standard English (EN),

(6) **target_language** – the language of a reference translation, either standard English (EN) or gender-inclusive Polish (PL).

```
{"source_resource_id":"EU_Karta_Praw_Podstawowych",
 "ipis_id":"IPIS-EN_translation_pl2en_dev_39",
 "prompt":"Generate an English translation
    of the text:",
 "source":"IV
    SOLIDARNOŚĆ
    Artykuł 27
    Prawo pracowników i pracownic do informacji
    i konsultacji w ramach przedsiębiorstwa
    Pracownikom/Pracownicom i ich
    przedstawicielom/przedstawicielkom należy
    zagwarantować, na właściwych poziomach,
    informację i konsultację we właściwym czasie,
    w przypadkach i na warunkach przewidzianych
    w prawie Unii oraz ustawodawstwach
    i praktykach krajowych.",
 "target":"TITLE IV
    SOLIDARITY
    Article 27
    Workers' right to information
    and consultation within the undertaking
    Workers or their
    representatives must,
    at the appropriate levels, be guaranteed
    information and consultation in good time
    in the cases and under the conditions provided
    for by Union law and national laws
    and practices.",
 "prompt_language":"EN",
 "source_language":"PL",
 "target_language":"EN"}
```

Figure 3: IPIS-translation example: Polish←English translation direction and an user prompt in English. The consecutive lines correspond to each other.

B List of tested LLMs

Multilingual Large Language Models

(1) **Llama-8B**[3] (Grattafiori et al., 2024), a 8 billion-parameter model with a context length of 8k tokens,

(2) **Mistral-7B**[4] (Jiang et al., 2023), a 7 billion-parameter model,

(3) **Mistral-Nemo**[5] (Mistral AI team, 2024), a 12 billion-parameter model with a a context window of 128k tokens.

Polish-specific Large Language Models Polish-specific LLMs were optimised to enhance performance on tasks requiring a deeper understanding of Polish.

(1) **Bielik-7B**[6] (Ociepa et al., 2024b), a 7 billion-parameter model derived from Mistral-7B-v0.1,

(2) **Llama-PLLuM-8B**[7] (PLLuM, 2025), a 11 billion-parameter variant based on Mixtral 8x22B,

(3) **Bielik-11B**[8] (Ociepa et al., 2024a) built upon Llama3.1-8B,

(4) **PLLuM-12B**[9] (PLLuM, 2025) derived from Mistral-Nemo-Base-2407.

C System prompts

(1) System prompt for gender-inclusive proofreading – Figure 4

(2) System prompt for gender-sensitive translation – Figure 5

[3] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
[5] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407
[6] https://huggingface.co/speakleash/Bielik-7B-Instruct-v0.1
[7] https://huggingface.co/CYFRAGOVPL/Llama-PLLuM-8B-chat
[8] https://huggingface.co/speakleash/Bielik-11B-v2.3-Instruct
[9] https://huggingface.co/CYFRAGOVPL/PLLuM-12B-nc-chat

D Instruction tuning hyperparameters

Table 4: LoRA and SFT hyperparameters

LoRA configuration	
Attention implementation	Flash attention 2
LoRA rank (r)	128
LoRA alpha (α)	256
LoRA dropout	0.05
SFT configuration	
Number of training epochs	3
Per device train batch size	2
Per device eval batch size	2
Gradient accumulation steps	4
Learning rate	2e-5
Weight decay	1e-3
Adam beta1	0.999
Adam beta2	0.9
Warmup ratio	0.05
Learning rate scheduler	cosine
Max sequence length	4096

E Normalisation process

The normalisation process involves expanding gender-inclusive expressions, especially those with a gender star, into their full masculine and feminine forms, and then filtering out predefined stop words.

Input

> Tegoroczni laureaci Oscarów pozowali na czerwonym dywanie.
> (gloss.) This year's-ADJ.$_{masc}$ winners$_{masc}$ of Oscars posed$_{masc}$ on the red carpet
> (Eng.) This year's Oscar winners posed on the red carpet.

Possible gender-inclusive variants When proofreading the input example, LLMs should generate one of the following gender-inclusive alternatives:

– Tegoroczni laureaci i tegoroczne laureatki Oscarów pozowali i pozowały na czerwonym dywanie.

– Tegoroczne laureatki i tegoroczni laureaci Oscarów pozowały i pozowali na czerwonym dywanie.

– Tegoroczni laureaci i tegoroczne laureatki Oscarów pozowali/pozowały na czerwonym dywanie.

You are an editor, who verifies whether a Polish text is written in gender-inclusive language. If not, you should transform it according to the principles of gender-inclusive language and the text's genre.

Follow this algorithm:
1. Identify the genre of the text and the forms of gender-inclusive expressions allowed in it.
2. Identify exclusionary expressions (see section I).
3. Replace exclusionary expressions with gender-inclusive expressions appropriate to the genre (see II).
4. Transform the grammatical forms of dependent expressions to agree with the gender-inclusive expression so that the text remains grammatically correct (see III).

I. EXCLUSIONARY EXPRESSIONS IN POLISH
a) An expression in the generic masculine gender naming mixed-gender groups [e.g. 'pacjenci' - both women (pacjentki) and men (pacjenci) can fall ill].

b) An expression in the generic masculine gender naming an unspecified person from a mixed-gender group [e.g. 'student' – a woman (studentka) and a man (student) may study].

c) A masculine expression naming a function, profession, or position a woman holds [e.g. 'dyrektor Kwiatkowska' – Kwiatkowska should be referred to as dyrektorka].

II. INCLUSIVE EXPRESSIONS AND TEXT GENRE

a) Masculine and feminine nouns with a coordinating conjunction
- a feminine and a masculine noun connected by a coordinating conjunction
- e.g. 'pacjenci i pacjentki'
- speeches and oral statements, expressions to an addressee in written documents, narrative texts, scientific and press articles

b) Masculine and feminine nouns with a slash
- a feminine and a masculine form connected by a slash (/)
- e.g. 'student/studentka'
- forms and documents, legal acts, speeches and oral statements, narrative texts, scientific and press articles

c) Abbreviated gender-inclusive form with an inclusive asterisk
- the expression consists of a root (i.e., the shared part of masculine and feminine forms), an inclusive asterisk (*), and gender-specific suffixes connected by a slash in the predefined male/female order or only the female suffix if the male suffix is null
- e.g. 'pracowni*ków/czek'
- forms and documents, legal acts, narrative texts, scientific and press articles

d) Osoba-form
- the noun 'osoba' (Eng. person) and its attribute
- 'osoby uczestniczące w spotkaniu' instead of 'uczestnicy spotkania'
- forms, documents, legal acts

d) Neutral form
- a gender-neutral collective noun
- 'personel medyczny' instead of 'lekarze i pielęgniarki'
- forms, documents, legal acts

III. GRAMMATICAL AGREEMENTS

a) You MUST transform the grammatical gender of the modifier to agree with the gender of the governing gender-inclusive expression [e.g. pracowni*cy/czki naukow*i/e].

b) If the gender-inclusive expression is the subject of a sentence, you MUST adjust the grammatical gender of the predicate to agree with the gender of that subject [e.g. rolniczki/rolnicy strajkowały/strajkowali].

c) You MUST transform the grammatical gender of pronouns referring to the gender-inclusive expression throughout the text [e.g. Student*ka ma obowiązki. Jego/jej absencja jest nieakceptowana].

d) You MUST use collective numerals when referring to mixed-gender groups [e.g. pięcioro kandydat*ów/ek].

It is NOT ALLOWED to paraphrase or edit parts of the text that are not exclusionary expressions.

It is NOT ALLOWED to add new text that does not result from inclusive transformations.

Figure 4: System prompt in English for the gender-inclusive proofreading task.

You are a Polish translator.

To translate texts written in English into gender-inclusive Polish, follow this algorithm:
1. Specify the genre of the text and the forms of gender-inclusive expressions allowed in it.
2. Translate English text into gender-inclusive Polish.
3. If you identify exclusionary expressions (see section I) in the translation:
- replace exclusionary expressions with gender-inclusive expressions appropriate to the genre (see II).
- transform the grammatical forms of dependent expressions to agree with the gender-inclusive expression so that the text remains grammatically correct (see III).

To translate gender-inclusive Polish texts into English, follow this algorithm:
1. Specify the genre of the text and the forms of Polish gender-inclusive expressions allowed in it (see II).
2. Identify gender-inclusive expressions in the Polish source text.
3. Translate gender-inclusive expressions into their corresponding English expressions [e.g. 'ambitni studenci i ambitne studentki' -> 'ambitious students'].

I. EXCLUSIONARY EXPRESSIONS IN POLISH

a) An expression in the generic masculine gender naming mixed-gender groups [e.g. 'pacjenci' - both women (pacjentki) and men (pacjenci) can fall ill].

b) An expression in the generic masculine gender naming an unspecified person from a mixed-gender group [e.g. 'student' – a woman (studentka) and a man (student) may study].

c) A masculine expression naming a function, profession, or position a woman holds [e.g. 'dyrektor Kwiatkowska' – Kwiatkowska should be referred to as dyrektorka].

II. INCLUSIVE EXPRESSIONS AND TEXT GENRE

a) Masculine and feminine nouns with a coordinating conjunction
- a feminine and a masculine noun connected by a coordinating conjunction
- e.g. 'pacjenci i pacjentki'
- speeches and oral statements, expressions to an addressee in written documents, narrative texts, scientific and press articles

b) Masculine and feminine nouns with a slash
- a feminine and a masculine form connected by a slash (/)
- e.g. 'student/studentka'
- forms and documents, legal acts, speeches and oral statements, narrative texts, scientific and press articles

c) Abbreviated gender-inclusive form with an inclusive asterisk
- the expression consists of a root (i.e., the shared part of masculine and feminine forms), an inclusive asterisk (*), and gender-specific suffixes connected by a slash in the predefined male/female order or only the female suffix if the male suffix is null
- e.g. 'pracowni*ków/czek'
- forms and documents, legal acts, narrative texts, scientific and press articles

d) Osoba-form
- the noun 'osoba' (Eng. person) and its attribute
- 'osoby uczestniczące w spotkaniu' instead of 'uczestnicy spotkania'
- forms, documents, legal acts

d) Neutral form
- a gender-neutral collective noun
- 'personel medyczny' instead of 'lekarze i pielęgniarki'
- forms, documents, legal acts

III. GRAMMATICAL AGREEMENTS

a) You MUST transform the grammatical gender of the modifier to agree with the gender of the governing gender-inclusive expression [e.g. pracowni*cy/czki naukow*i/e].

b) If the gender-inclusive expression is the subject of a sentence, you MUST adjust the grammatical gender of the predicate to agree with the gender of that subject [e.g. rolniczki/rolnicy strajkowały/strajkowali].

c) You MUST transform the grammatical gender of pronouns referring to the gender-inclusive expression throughout the text [e.g. Student*ka ma obowiązki. Jego/jej absencja jest nieakceptowana].

d) You MUST use collective numerals when referring to mixed-gender groups [e.g. pięcioro kandydat*ów/ek].

It is NOT ALLOWED to paraphrase or edit parts of the text that are not exclusionary expressions.

It is NOT ALLOWED to add new text that does not result from inclusive transformations.

Figure 5: System prompt in English for the gender-sensitive translation task.

– Tegoroczni laureaci/Tegoroczne laureatki Oscarów pozowali/pozowały na czerwonym dywanie.

– Tegoroczne laureatki/Tegoroczni laureaci Oscarów pozowały/pozowali na czerwonym dywanie.

– Tegoroczni/Tegoroczne laureaci/laureatki Oscarów pozowali/pozowały na czerwonym dywanie.

– Tegoroczni/Tegoroczne laureaci/laureatki Oscarów pozowa*li/ły na czerwonym dywanie.

– Tegoroczn*i/e laurea*ci/tki Oscarów pozowa*li/ły na czerwonym dywanie.

Normalisation All of the gender-inclusive variants listed above can be normalised to the Polish tokens presented in Table 5.

Polish	English glosses
tegoroczni	this year's$_{masc}$
laureaci	winners$_{masc}$
tegoroczne	this year's$_{fem}$
laureatki	winners$_{fem}$
oscarów	oscar
pozowali	posed$_{masc}$
pozowały	posed$_{fem}$
na	on
czerwonym	red
dywanie	carpet

Table 5: Normalisation of all possible gender-inclusive variants in the running example. Colour legend: masculine NPs, feminine NPs, masculine verb, and feminine verb. All tokens are lowercased. Punctuation marks and coordinating conjunctions *i* ['and'] are filtered out.

```
{'source': 'Tegoroczni laureaci Oskarów pozowali
    na czerwonym dywanie.',
 'target': 'Tegoroczn*i/e laurea*ci/tki Oscarów
    pozowa*li/ły na czerwonym dywanie.',
 'normalised_target': [['tegoroczni', 'tegoroczne',
    'laureaci', 'laureatki', 'oscarów', 'pozowali',
    'pozowały', 'na', 'czerwonym', 'dywanie']]}
```

Figure 6: The gold standard of the running example.

Gold standard comparison Since the manually corrected reference text is normalised using the same procedure (see Figure 6), the normalised LLM outputs can be directly compared to the reference tokens. Word order is irrelevant, as normalised tokens are treated as bags of tokens rather than coherent sentences.

F Translation examples

Gender-sensitive translating is not a trivial task. Since *English↔gender-inclusive Polish* datasets hardly exist, LLMs are typically trained on standard parallel corpora, reducing their chances to acquire differences in gender encoding.

F.1 Polish-to-English translation

In Polish, gender-inclusive expressions are often encoded as doubled nouns, pronouns, adjectives, and verbs. These double forms are generally translated into gender-neutral expressions in English, resulting in significant differences in token count between Polish and English texts, cf. (1) and (1-a). Moreover, given its unnatural sounding, it is questionable whether the gender distinctions present in the original should be explicitly included in the English translation, see (1-b) and (2-b), or whether a more neutral version is preferable, see (1-a) and (2-a).

(1) *Szczyt Women In Tech zgromadził liczne uczestniczki i licznych uczestników, które/którzy uczestniczyły/uczestniczyli w różnorodnych prelekcjach.*
[gloss.] Woman In Tech Summit attracted numerous$_{fem}$ attendees$_{fem}$ and numerous$_{masc}$ attendees$_{masc}$[10] who$_{fem}$/who$_{masc}$ participate$_{fem}$/participate$_{masc}$ in a variety of lectures

a. Women In Tech Summit attracted numerous attendees who participated in a variety of lectures.

b. ?Women In Tech Summit attracted numerous female attendees and numerous male attendees who participated in a variety of lectures.

c. *Women In Tech Summit attracted numerous attendees and numerous attendees who participated in a variety of lectures.

(2) *W piątek policja aresztowała pracownika i*

[10] Both women and men participate in Women In Tech Summits.

pracownicę banku.
[gloss.] On Friday the police arrested employee$_{masc}$ and employee$_{fem}$ of bank

a. On Friday, the police arrested two bank employees.

b. ?On Friday, the police arrested a male and a female bank employee.

The aim of the Polish-to-English translation experiments is to examine the ability of LLMs to generate high-quality English translations in the non-standard context of gender-inclusive Polish source texts. Despite the use of double forms for nouns, pronouns, adjectives, and verbs in the gender-inclusive Polish source texts, the expected output is a gender-neutral English translation, which is free from explicit gender markers, redundant repetitions, or other unnecessary linguistic phenomena.

F.2 English-to-Polish translation

Depending on the context, English gender-neutral expressions should be translated into Polish either as double forms, see (3-a) and (4-a), or as a single gendered expression, see (5-a) and the feminine translation of 'Rector' in (5-a). Translations with generic masculine forms are not acceptable, see (3-b), (4-b), and (5-b). Translations that contradict world knowledge are also not acceptable, see (4-c) and (5-c).

(3) *The most outstanding students were awarded by the Rector of AMU.*

a. Najwybitniejsi studenci i najwybitniejsze studentki zostali nagrodzeni/zostały nagrodzone przez Rektorkę UAM.
[gloss.] The most outstanding$_{masc}$ students$_{masc}$ and the most outstanding$_{fem}$ students$_{fem}$ were$_{masc}$ awarded$_{masc}$/were$_{fem}$ awarded$_{fem}$ by the Rector$_{fem}$ of AMU

b. *Najwybitniejsi studenci zostali nagrodzeni przez Rektora UAM.
[gloss.] The most outstanding$_{masc}$ students$_{masc}$[11] were$_{masc}$ awarded$_{masc}$ by the Rector$_{masc}$[12] of AMU

(4) *Patients rated Eye Clinic positively*

a. Pacjenci i pacjentki pozytywnie ocenili/oceniły Eye Clinic.
[gloss.] Patients$_{masc}$ and patients$_{fem}$ positively rated$_{masc}$/rated$_{fem}$ Eye Clinic

b. *Pacjenci pozytywnie ocenili Eye Clinic.
[gloss.] Patients$_{masc}$ positively rated$_{masc}$ Eye Clinic

c. *Pacjentki pozytywnie oceniły Eye Clinic.
[gloss.] Patients$_{fem}$[13] positively rated$_{fem}$ Eye Clinic

(5) *Patients rated Medifem positively.*

a. Pacjentki pozytywnie oceniły Medifem.
[gloss.] Patients$_{fem}$ positively rated$_{fem}$ Medifem

b. *Pacjenci pozytywnie ocenili Medifem.
[gloss.] Patients$_{masc}$ positively rated$_{masc}$ Medifem

c. *Pacjenci i pacjentki pozytywnie ocenili/oceniły Medifem.
[gloss.] Patients$_{masc}$[14] and patients$_{fem}$ positively rated$_{masc}$/rated$_{fem}$ Medifem

The English-to-Polish translation experiments are designed to investigate the ability of LLMs to translate English source texts that are characterised by minimal gender cues into gender-inclusive Polish. These experiments assess whether LLMs can make contextually appropriate and inclusive linguistic choices in the target language, despite limited gender-specific information in the original input.

G Detailed results

[11] Both women and men can study at AMU.

[12] The Rector of AMU is Her Magnificence prof. dr hab. Bogumiła Kaniewska.

[13] Both women and men may receive treatment in Eye Clinic, and it is likely that individuals of both genders have provided ratings for the clinic.

[14] Medifem is a women's clinic, so men cannot be its patients

LLM	Accuracy	Precision	Recall	F_1	BLEU	chrF	chrF++
Default LLMs							
llama-8b-default	52.20	0.16	0.45	0.23	47.44	75.38	73.91
llama-8b-default-pl	45.63	0.31	3.53	0.56	41.94	76.24	75.72
llama-8b-default-en	41.38	0.26	5.85	0.49	36.13	74.76	74.12
mistral-7b-default	16.60	0.06	0.14	0.08	14.09	43.32	40.53
mistral-7b-default-pl	12.83	0.10	1.74	0.19	9.20	40.50	38.87
mistral-7b-default-en	16.69	0.11	1.03	0.20	14.43	48.43	46.55
mistral-nemo-12b-default	63.46	0.34	0.37	0.35	62.98	78.33	76.95
mistral-nemo-12b-default-pl	**66.71**	1.19	5.44	1.95	61.99	**86.25**	**85.67**
mistral-nemo-12b-default-en	42.58	0.59	7.42	1.09	36.72	76.36	75.02
llama-pllum-8b-default	34.88	0.18	0.22	0.20	32.04	57.13	54.36
llama-pllum-8b-default-pl	32.36	0.46	0.44	0.45	41.94	51.25	50.05
llama-pllum-8b-default-en	41.29	0.37	1.01	0.54	40.69	67.13	65.83
bielik-7b-default	65.07	0.39	0.50	0.44	**66.17**	80.40	79.39
bielik-7b-default-pl	53.69	0.32	2.49	0.57	51.71	80.40	80.04
bielik-7b-default-en	42.68	0.22	2.36	0.40	39.99	74.61	74.04
pllum-12b-default	44.94	1.09	1.63	1.31	41.65	67.70	65.26
pllum-12b-default-pl	63.64	**2.56**	6.28	**3.64**	63.59	82.74	81.86
pllum-12b-default-en	59.66	2.05	5.85	3.03	58.71	81.13	80.17
bielik-11b-default	41.79	0.32	0.59	0.42	39.12	66.54	64.18
bielik-11b-default-pl	60.55	1.45	9.34	2.51	56.56	83.93	83.22
bielik-11b-default-en	60.41	1.60	**13.62**	2.86	55.79	84.72	84.08
Few-shot LLMs							
llama-8b-fewshot	55.74	0.36	1.70	0.59	51.61	80.06	79.18
llama-8b-fewshot-pl	56.08	0.52	5.03	0.94	51.91	82.57	82.04
llama-8b-fewshot-en	49.28	0.33	5.01	0.62	44.35	79.61	78.99
mistral-7b-fewshot	18.52	0.12	0.59	0.19	15.13	48.39	46.21
mistral-7b-fewshot-pl	23.14	0.25	2.43	0.46	18.81	56.24	54.55
mistral-7b-fewshot-en	23.06	0.26	1.23	0.42	21.01	54.94	52.94
mistral-nemo-12b-fewshot	**68.12**	0.74	1.65	1.02	**68.43**	**84.95**	**84.19**
mistral-nemo-12b-fewshot-pl	48.21	0.78	2.15	1.14	50.53	74.39	73.41
mistral-nemo-12b-fewshot-en	33.87	0.45	1.75	0.72	34.29	65.10	63.78
llama-pllum-8b-fewshot	31.34	0.49	0.58	0.53	37.51	52.38	50.80
llama-pllum-8b-fewshot-pl	38.05	0.56	0.66	0.60	46.65	58.55	57.44
llama-pllum-8b-fewshot-en	37.36	0.47	0.62	0.53	43.77	58.28	57.14
bielik-7b-fewshot[15]	NA	NA	NA	NA	NA	NA	NA
pllum-12b-fewshot	47.09	1.73	2.51	2.05	50.84	68.61	67.15
pllum-12b-fewshot-pl	50.84	**2.27**	2.62	2.43	58.62	69.98	68.83
pllum-12b-fewshot-en	48.73	2.02	2.38	2.19	56.50	68.30	67.17
bielik-11b-fewshot	56.21	1.09	4.57	1.76	52.91	80.23	79.12
bielik-11b-fewshot-pl	59.15	1.35	**11.83**	**2.42**	54.92	84.01	83.39
bielik-11b-fewshot-en	58.57	1.31	8.74	2.28	53.69	82.93	82.09

[15]The context size of the bielik-7b model makes it impossible to carry out this experiment.

LLM	Accuracy	Precision	Recall	F_1	BLEU	chrF	chrF++
IPIS-tuned LLMs							
llama-8b-tuned	96.73	57.02	41.51	48.04	93.92	97.47	97.29
llama-8b-tuned-pl	95.68	47.81	36.89	41.65	92.85	96.62	96.46
llama-8b-tuned-en	95.99	50.08	38.99	43.85	93.47	97.01	96.80
mistral-7b-tuned	95.34	39.55	47.80	43.29	93.54	97.25	97.04
mistral-7b-tuned-pl	92.53	24.86	42.76	31.44	91.29	96.50	96.27
mistral-7b-tuned-en	93.50	29.27	43.79	35.09	90.53	96.43	96.20
mistral-nemo-12b-tuned	96.82	57.16	48.81	52.66	94.45	97.72	97.51
mistral-nemo-12b-tuned-pl	96.29	50.67	45.38	47.88	93.28	97.22	97.03
mistral-nemo-12b-tuned-en	95.93	45.45	45.20	45.32	94.09	97.33	97.13
llama-pllum-8b-tuned	97.08	61.91	46.40	53.04	94.28	97.64	97.48
llama-pllum-8b-tuned-pl	95.86	50.87	36.68	42.63	93.40	96.85	96.69
llama-pllum-8b-tuned-en	96.19	51.93	44.25	47.79	93.78	97.25	97.05
bielik-7b-tuned	96.66	54.67	52.37	53.49	94.67	97.71	97.49
bielik-7b-tuned-pl	96.27	53.50	45.95	49.44	94.10	97.30	97.09
bielik-7b-tuned-en	96.22	50.99	48.08	49.49	93.62	97.26	97.03
pllum-12b-tuned	97.28	62.80	54.97	58.62	95.10	97.95	97.76
pllum-12b-tuned-pl	96.93	60.28	50.08	54.71	94.52	97.68	97.50
pllum-12b-tuned-en	97.05	60.73	51.62	55.80	94.68	97.77	97.58
bielik-11b-tuned	**97.37**	**63.93**	**56.26**	**59.85**	**95.22**	**97.99**	**97.81**
bielik-11b-tuned-pl	93.66	29.24	50.32	36.99	91.82	96.93	96.74
bielik-11b-tuned-en	96.47	52.30	51.59	51.94	94.82	97.61	97.43

Table 6: Evaluation of multilingual and Polish-specific LLMs in solving the **gender-inclusive proofreading** task. Explanation: **default** – the default LLM, **fewshot** – the default LLM in the few-shot setting, **tuned** – the LLM tuned on IPIS proofreading instructions; **pl** – a system prompt in Polish; **en** – a system prompt in English; **bold** – the best model within a scenario (*default*, *fewshot*, and IPIS-*tuned*); underline – the best small-size model; framed – the best model overall.

LLM	Polish→English						English→Polish					
	PL user prompt			EN user prompt			PL user prompt			EN user prompt		
	bleu	chrF	chrF++	bleu	chrF	chrF++	bleu	chrF	chrF++	bleu	chrF	chrF++
Default LLMs												
llama-8b-default	24.69	42.79	40.68	30.36	56.19	54.04	26.84	62.85	59.06	24.69	61.93	58.08
llama-8b-default-pl	22.72	55.62	54.02	21.92	58.19	56.48	16.90	57.14	53.37	16.84	57.66	53.86
llama-8b-default-en	26.76	60.78	58.88	26.87	65.11	63.18	13.81	55.41	51.57	9.89	49.63	45.95
mistral-7b-default	41.97	68.02	65.91	42.48	69.81	67.56	6.99	44.32	40.50	9.91	50.21	45.92
mistral-7b-default-pl	18.05	55.24	53.20	15.55	53.88	51.91	3.59	34.57	31.38	2.93	30.92	27.95
mistral-7b-default-en	18.90	56.84	54.90	24.18	60.24	58.19	4.48	38.65	35.16	4.28	37.23	33.75
mistral-nemo-12b-default	**53.68**	**75.35**	**73.57**	**53.42**	**75.78**	**73.97**	23.75	60.16	56.33	23.11	59.66	55.63
mistral-nemo-12b-default-pl	42.62	71.94	70.07	47.29	74.17	72.32	14.75	54.89	51.00	12.23	53.75	49.71
mistral-nemo-12b-default-en	40.79	72.66	70.86	40.26	72.18	70.23	10.15	49.99	46.11	11.06	53.03	48.76
llama-pllum-8b-default	42.85	69.52	67.72	40.53	67.65	65.89	31.14	64.95	61.47	**34.28**	66.86	63.56
llama-pllum-8b-default-pl	32.55	51.67	50.17	30.19	50.42	48.87	18.39	42.56	39.97	15.87	43.12	40.30
llama-pllum-8b-default-en	33.15	65.33	63.44	27.21	61.81	60.12	25.40	59.01	55.72	24.79	62.76	59.39
bielik-7b-default	38.98	66.02	63.73	36.30	61.78	59.67	28.10	64.36	60.53	27.61	63.44	59.71
bielik-7b-default-pl	35.14	60.53	58.47	39.08	64.78	62.68	25.62	61.03	57.32	27.01	62.38	58.77
bielik-7b-default-en	36.63	64.63	62.45	36.93	62.81	60.78	20.02	58.47	54.72	22.46	60.44	56.83
pllum-12b-default	43.51	71.09	69.17	44.34	71.24	69.30	32.09	66.47	63.14	33.17	66.04	62.70
pllum-12b-default-pl	38.27	66.96	64.88	38.87	67.72	65.61	27.90	60.39	57.01	29.11	63.01	59.59
pllum-12b-default-en	41.52	68.24	66.24	36.18	65.66	63.64	25.70	57.50	54.33	24.91	61.32	57.96
bielik-11b-default	47.60	73.39	71.45	47.54	72.50	70.59	**41.49**	**71.78**	**68.79**	27.39	65.30	62.49
bielik-11b-default-pl	46.78	73.08	71.16	43.67	70.21	68.17	35.80	69.77	66.65	32.31	**68.94**	**65.86**
bielik-11b-default-en	47.99	73.76	71.78	36.39	68.39	66.52	32.70	68.65	65.67	32.13	68.54	65.51
Few-shot LLMs												
llama-8b-fewshot	43.03	65.99	64.07	44.27	71.15	69.08	26.37	62.10	58.24	25.31	61.92	58.09
llama-8b-fewshot-pl	37.46	65.93	64.05	34.21	65.63	63.80	18.32	58.38	54.54	18.24	58.75	54.93
llama-8b-fewshot-en	29.26	58.95	57.04	29.22	65.03	63.04	15.49	57.01	53.09	10.29	51.07	47.29
mistral-7b-fewshot	40.94	68.92	66.59	40.10	68.22	65.83	8.27	47.02	42.98	9.35	48.28	44.22
mistral-7b-fewshot-pl	31.11	62.88	60.77	30.83	64.28	62.05	7.24	46.29	42.26	4.90	39.57	35.99
mistral-7b-fewshot-en	32.99	65.67	63.44	39.98	68.22	65.79	4.69	38.35	34.85	9.72	49.43	45.22
mistral-nemo-12b-fewshot	**54.52**	**75.89**	**74.15**	**51.78**	**74.56**	**72.74**	20.08	53.94	50.23	17.56	52.80	49.08
mistral-nemo-12b-fewshot-pl	29.65	56.87	55.03	41.43	70.16	68.40	14.67	50.99	47.12	12.03	50.80	46.80
mistral-nemo-12b-fewshot-en	32.20	66.12	64.40	37.20	70.33	68.45	8.35	45.42	41.53	11.08	51.95	47.82
llama-pllum-8b-fewshot	32.00	59.06	57.38	33.73	61.88	59.97	21.75	51.89	48.60	25.75	60.75	57.31
llama-pllum-8b-fewshot-pl	32.74	50.02	48.51	21.43	35.63	34.37	15.38	34.91	32.67	9.87	29.05	26.74
llama-pllum-8b-fewshot-en	33.16	64.09	62.32	30.87	63.08	61.43	19.06	50.65	47.46	19.91	59.51	56.23
pllum-12b-fewshot	28.67	48.68	46.72	33.39	58.32	56.18	17.55	41.36	38.59	25.37	55.91	52.73
pllum-12b-fewshot-pl	36.29	59.81	57.95	37.94	65.75	63.58	12.57	33.94	31.45	30.20	62.12	58.81
pllum-12b-fewshot-en	34.90	63.21	61.21	35.73	61.04	58.92	19.23	43.92	41.00	23.11	56.56	53.32
bielik-11b-fewshot	50.01	73.93	72.04	49.66	73.99	72.04	38.63	68.78	66.09	33.19	68.11	65.34
bielik-11b-fewshot-pl	49.38	73.84	71.90	49.55	74.02	72.03	37.81	69.92	67.06	42.76	72.19	69.32
bielik-11b-fewshot-en	48.33	73.43	71.48	49.14	73.75	71.77	**43.02**	**72.46**	**69.61**	**43.17**	**72.82**	**70.00**

LLM	Polish→English						English→Polish					
	PL user prompt			EN user prompt			PL user prompt			EN user prompt		
	bleu	**chrF**	**chrF++**	**bleu**	**chrF**	**chrF++**	**bleu**	**chrF**	**chrF++**	**bleu**	**chrF**	**chrF++**
IPIS-tuned LLMs												
llama-8b-tuned	29.45	61.48	60.08	30.82	62.02	60.55	34.15	65.31	62.30	33.93	65.82	62.77
llama-8b-tuned-pl	31.72	61.64	60.11	41.62	63.64	62.19	31.55	62.38	59.40	31.18	62.25	59.28
llama-8b-tuned-en	40.23	64.46	62.83	43.17	65.39	63.89	30.48	61.57	58.45	29.07	60.03	56.98
mistral-7b-tuned	10.43	39.85	39.15	12.93	44.39	43.65	26.10	59.90	56.69	25.51	57.32	54.43
mistral-7b-tuned-pl	24.13	58.52	57.53	22.57	57.04	56.16	17.77	52.04	49.15	22.70	57.70	54.42
mistral-7b-tuned-en	16.08	49.89	48.86	19.11	53.28	52.27	24.49	55.35	52.50	24.85	57.26	54.24
mistral-nemo-12b-tuned	10.75	39.89	39.25	16.12	49.01	48.30	26.35	60.41	57.73	34.17	65.99	62.85
mistral-nemo-12b-tuned-pl	14.25	47.17	46.28	21.04	56.65	55.76	21.66	56.67	53.71	22.05	57.95	54.97
mistral-nemo-12b-tuned-en	14.12	46.21	45.58	10.69	39.92	39.27	19.00	55.48	52.59	25.07	59.97	56.90
llama-pllum-8b-tuned	42.83	71.54	70.17	48.04	73.88	72.43	37.65	67.65	64.87	36.66	66.61	**63.86**
llama-pllum-8b-tuned-pl	48.94	72.00	70.54	48.86	72.50	71.05	**39.21**	**68.23**	**65.34**	**38.30**	**66.56**	63.68
llama-pllum-8b-tuned-en	47.28	71.16	69.63	44.66	68.98	67.39	34.84	66.46	63.63	32.12	65.33	62.46
bielik-7b-tuned	49.11	71.86	70.33	48.83	71.68	70.17	34.64	64.77	61.95	32.15	62.48	59.58
bielik-7b-tuned-pl	48.56	71.46	69.93	46.91	69.80	68.27	36.79	64.81	61.82	36.49	64.60	61.56
bielik-7b-tuned-en	39.45	66.51	65.03	42.18	67.61	66.15	37.43	64.82	61.91	36.86	65.08	62.11
pllum-12b-tuned	54.98	76.95	75.53	52.98	75.58	74.20	27.49	63.16	60.63	33.03	65.35	62.65
pllum-12b-tuned-pl	48.23	73.02	71.56	44.93	71.35	69.90	24.69	60.32	57.32	29.74	63.86	60.91
pllum-12b-tuned-en	47.12	72.90	71.31	47.24	72.92	71.39	23.33	59.05	56.35	21.89	58.70	55.85
bielik-11b-tuned	55.19	75.80	74.40	**57.45**	**77.93**	**76.52**	34.26	55.08	52.63	35.04	55.92	53.44
bielik-11b-tuned-pl	56.70	76.93	75.54	55.24	75.35	73.90	31.70	58.36	55.62	26.74	55.96	53.25
bielik-11b-tuned-en	**57.55**	**78.03**	**76.66**	57.30	76.63	75.29	28.71	60.97	58.12	25.96	58.93	55.77

Table 7: Evaluation of multilingual and Polish-specific LLMs in solving the **gender-sensitive translation** task. Explanation: **tuned** – LLM tuned on IPIS translation instructions; **pl** – a system prompt in Polish; **en** – a system prompt in English; **bold** – the best model within a scenario (*default*, *fewshot*, and IPIS-*tuned*); underline – the best small-size model; framed – the best model overall.