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, historically and politically entrenched norms have established masculine forms as the default for referring not only to men but also to women and mixed-gender groups. This is the reality of contemporary Polish and some other Slavic and Romance languages. As a result, large language models (LLMs) trained on standard corpora tend to reproduce androcentric patterns, generating gender-imbalanced outputs and impeding efforts towards gender-inclusive language.

In this study, we address the problem of default androcentrism in LLMs by adapting them to gender-inclusive behaviour in Polish. We introduce the Inclusive Polish Instruction Set (IPIS), a collection of human-crafted, gender-inclusive instructions for proofreading and Polish↔English translation. Grounded in linguistic theory, we design a system prompt with explicit guidelines for gender-inclusive Polish. We tune both multilingual and Polish-specific LLMs using IPIS and evaluate their outputs in terms of gender inclusivity.

Our results show that gender inclusivity can be effectively internalised through instruction tuning. We thus contribute a practical approach to mitigating gender bias in Polish and to supporting broader efforts to align LLMs with socially responsible linguistic practices.

1 Introduction

Large language models (LLMs) are trained on extensive datasets sourced primarily from the internet, which reflect standard language use, including socially embedded biases and stereotypes (Caliskan et al., 2017; Bender et al., 2021). LLMs are highly effective in internalising and replicating harmful patterns (Lucy and Bamman, 2021; Fleisig et al., 2023). In this study, we focus on one such phenomenon: *gender bias*. In English, gender bias is mainly lexical and conceptual (Stanczak and Augenstein, 2021), and LLMs tend to reflect a gender-binary orientation (Ovalle et al., 2024). In Polish, gender bias is structural and grammatical – masculine nouns are often used as a generic default to refer to men, women, and non-binary people (Koniuszaniec and Blaszkowska, 2003; Stahlberg et al., 2007; Gygax et al., 2019). Consequently, LLMs trained on Polish data acquire masculine-biased grammatical patterns and tend to favour a masculine orientation (Wróblewska et al., 2025).

The dominance of masculine expressions is a form of sociolinguistic gender discrimination (GEC, 2016; GNL, 2018). Acknowledging its harmful effects, the Council of Europe encourages member states to eliminate sexist language and promote gender-equal linguistic practices (CM/Rec, 2019).

Gender bias and inclusiveness are central concerns in contemporary NLP, as reflected in dedicated workshops GeBNLP and GITT. With the growing significance of gender-inclusive NLP, particularly in machine translation (MT, Vanmassenhove et al., 2018; Savoldi et al., 2021, 2025; Menis Mastromichalakis et al., 2025) and gender-neutral rewriting (Vanmassenhove et al., 2021; Sun et al., 2021), a key goal is to develop LLMs that reduce bias and use inclusive language.

This study takes an initial step toward this goal for Polish. We verify whether instruction tuning can produce gender-inclusive LLMs (see Figure 1), and evaluate their effectiveness in mitigating bias and improving inclusivity in Polish proofreading 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 release the Inclusive Polish Instruction Set (IPIS) (Section 3), the first expert-annotated dataset of parallel biased and gender-fair passages. **(2)** We design a system prompt to guide LLMs in generating gender-inclusive language (Section 4). **(3)** We evaluate how different instruction-tuning setups af-

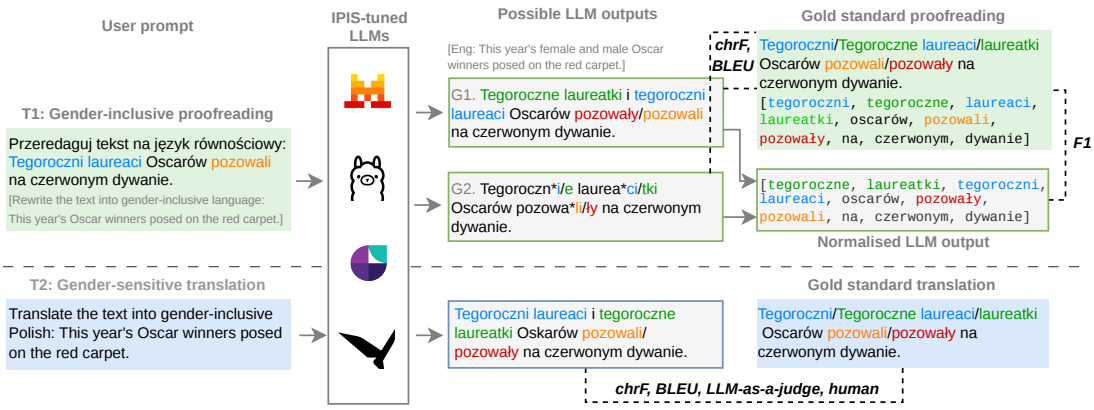

Figure 1: Flowchart of inference and evaluation workflows for IPIS-tuned LLMs for *gender-inclusive proofreading* (top) and Polish–English *gender-sensitive translation* (bottom; Polish direction shown). Explanations: masculine nouns and adjectives, masculine verbs, feminine nouns and adjectives and feminine verbs.

fect gender inclusivity in Polish proofreading and Polish↔English translation (Sections 5 and 6), and release the best-performing IPIS-tuned LLMs.

2 Proposed approach

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

2.1 Polish – an androcentric language

Polish is a grammatical gender language (Reali et al., 2015) in which all nouns are marked for *grammatical gender*:[1] *śliwka* [plum] is feminine, *pomidor* [tomato] is masculine, and *jabłko* [apple] is neutral. Adjectives and verbs show grammatical agreement with nouns. Some *personal nouns* (Backer and Cuypere, 2012) are realised as gendered dyads, e.g., *prawnik/prawniczka* [lawyer$_{\text{masc/fem}}$]. Feminine personal nouns refer to women, whereas masculine nouns can refer to men, mixed or unspecified groups, and even women (*generic masculine*), e.g., *niemiecka polityk, Ursula* [German$_{\text{fem}}$ politician$_{\text{masc}}$, Ursula$_{\text{fem}}$].

Although its grammatical system allows marking *referential gender*,[2] Polish is heavily masculine-oriented, with a dominance of masculine expression and reinforcement of gender bias. Adapting Polish to reflect referential genders is a challenging process. Its goal is *gender-inclusive* (or *gender-fair*) language with all genders explicitly marked (Amrhein et al., 2023), as opposed to *gender-neutral* language, which avoids marking genders.

As LLMs trained on standard Polish corpora reproduce androcentric patterns (Wróblewska et al., 2025), we investigate whether targeted instruction tuning can enable gender-sensitive generation. We focus on well-established feminine forms and masculine–feminine dyads, leaving room for future inclusion of non-binary forms (not yet standardised). Our goal is to make gender inclusivity a core modelling principle, enabling consistent and context-appropriate use of gender-fair Polish.

2.2 Gender-inclusive instruction tuning

Instruction tuning is a fine-tuning technique that improves the versatility and controllability of LLMs across a wide range of tasks (Zhang et al., 2024). It consists in training LLMs on instruction–response pairs, following a supervised learning paradigm.

We use instruction tuning to adapt pre-trained LLMs to gender-inclusive language use by exposing them to a set of instructions with gender-inclusive responses (see Section 3). For this, we employ *low-rank adaptation* (LoRA, Hu et al., 2021), a parameter-efficient fine-tuning method. LoRA updates only a small subset of parameters by injecting low-rank decomposition matrices into transformer layers while keeping the original weights frozen. This reduces computational and memory requirements, enables controlled gender-inclusive adaptation without full retraining, and preserves model's general linguistic competence. This well-established method also supports reproducibility.

2.3 Proposed gender-inclusive tasks

Gender-inclusive proofreading This task consists in transforming standard Polish text into its gender-inclusive version. It goes beyond surface-level paraphrasing, as models must reason about referential gender and apply linguistically informed

[1] *Grammatical gender* is a category that classifies nouns into declension groups and governs agreements in a sentence.

[2] *Referential gender* refers to the biological sex (female or male) or gender identity of a living referent.

transformations. Specifically, LLMs must distinguish personal nouns from non-personal ones (including collective nouns, e.g., *komitet*$_{\text{masc}}$ [committee]); modify generic masculine nouns, but preserve masculine forms that refer genuinely to men or male-only groups; preserve epicene nouns (e.g., *rodzic*$_{\text{masc}}$ [parent]) and inherently gender-specific nouns (e.g., *wujek*$_{\text{masc}}$ [uncle]). LLMs must generate linguistically and socially accepted gender-inclusive forms and correctly adjust all grammatical agreement (i.e., adjectives, verbs and pronouns), including implicit references in pro-drop contexts.

Gender-sensitive translation This task focuses on bidirectional translation between English and gender-inclusive Polish, with the goal of producing linguistically accurate translations that adhere to gender inclusivity principles (see examples of translation challenges in Appendix A).

3 Inclusive Polish Instruction Set (IPIS)

The IPIS dataset[3] is a collection of instructions for improving gender sensitivity and inclusiveness of LLMs in Polish. It contains human-annotated texts from diverse open-access sources (legal acts, institutional texts, narratives, news, transcripts, etc.). Originally androcentric texts are manually rewritten into inclusive versions, resulting in a parallel androcentric–to–gender-inclusive corpus (annotation details and examples in Apps. B.1 and B.2).

Each IPIS-proofreading sample consists of:

1. **user prompt** (`prompt`) – a 1-2 sentence instruction that specifies a given task and resembles a real-world request (e.g., *Rewrite...*),
2. **input text passage** (`source`) – an androcentric passage requiring proofreading,
3. **desired output** (`target`) – an expected gender-inclusive response based on `prompt` and `source`, which is used as ground truth for evaluating and optimising LLM's predictions.

IPIS-translation instances additionally include `(prompt|source|target)_language` with values restricted to PL (Polish) or EN (English). IPIS examples appear in Figure 1 and Appendix B.3.

The monolingual IPIS-proofreading subsets contain 5,278 instances in *test*, 2,732 in *dev*, and 23,532 in *train*. The Polish–English IPIS-translation dataset is provided in two variants. The **human-annotated** variant comprises 456 samples in *test*, 304 in *dev*, and 1728 in *train*. The **mixed-annotated** variant includes all gender-inclusive IPIS-proofreading texts paired with English translations. To mitigate the risk of MT anomalies and inconsistencies – given that LLMs are predominantly trained on androcentric Polish corpora – we translate the IPIS-proofreading source passages rather than gender-inclusive ones. English translations are generated using Gemma-3-27B (Kamath et al., 2025), which performs superior in our preliminary evaluations (see Table 8, Appendix G, last row). The IPIS *test*, *dev*, and *train* splits are balanced for the ratio of gender-inclusive transformations. IPIS is used for LLM tuning and evaluation.

4 Experimental setup

Tested LLMs Our study investigates 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 focus 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 D for model details).

System prompt We study the effect of including a system prompt with gender-inclusive guidelines during training and evaluation. The system prompt, grounded in the theoretical framework developed by Wróblewska et al. (2025), is originally written in Polish and translated into English (Appendix C).

Using persona prompting (Xu et al., 2023; Kong et al., 2024), we assign LLMs the role of an editor or translator. They are instructed to follow an algorithm and proofread/translate texts according to gender-inclusive principles. The system prompt defines exclusionary expressions, provides inclusive alternatives, and specifies agreement rules to ensure grammatically correct, gender-inclusive outputs.

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 distributed training setups. HuggingFace's *datasets* and *transformers* libraries facilitate the loading and usage of the IPIS dataset and selected LLMs while *trl* provides a dedicated trainer for supervised fine-tuning (SFT). LoRA adapters are configured us-

[3] https://gitfront.io/r/anonymous-researcher/QswJdpbHaanK/Inclusive-LLMs/

ing HuggingFace's *peft* library. Flash Attention enhances the efficiency of the core attention mechanism in transformer-based models. Details on the LoRA configuration, SFT hyperparameters, and the libraries used are in Appendix F.

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

Evaluation metrics For gender-inclusive proofreading, we evaluate normalised LLM-generated outputs against gold-standard test instances using accuracy, precision, recall, and F_1-measure. Normalisation includes lowercasing, removing punctuation and conjunctions, and transforming all gender-inclusive expressions into full masculine and feminine forms. For automatic punctuation and conjunction detection, we use Lambo tokeniser (Przybyła, 2022) and Combo (Klimaszewski and Wróblewska, 2021) for part-of-speech tagging. Figure 1 illustrates normalisation of possible LLM outcomes, e.g., the G2 generation with gender stars and slashes; further normalisation examples are provided in Appendix E.

Additionally, we measure similarity between LLM-proofread and reference gender-inclusive passages, using the MT metrics chrF (Popović, 2015) and BLEU (Papineni et al., 2002). Since LLMs should improve inclusivity while preserving structure and content, and not hallucinate, these metrics are complementary adequacy indicators.

BLEU and chrF are also used to evaluate bidirectional translations between English and gender-sensitive Polish. Since these metrics may not capture gender-inclusive changes, especially in English-to-Polish translation, we additionally apply an LLM-as-a-judge evaluation paradigm and compare the results with human ratings under the same scoring protocol. Following the G-Eval-style approach (Liu et al., 2023), we carefully design an evaluation prompt that instructs the judge to identify translation errors and assign scores on a 1–5 scale for the MT-related criteria (*completeness*, *clarity*, *structure*, *correctness*) and a dedicated *gender inclusivity* one (prompt in Appendix Figure 6).

Baseline scenarios We evaluate the impact of instruction tuning and system-prompt guidance using several 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 in a zero-shot setting (*default*), optionally augmented with the system prompt (*default*-{*pl*|*en*}). Further baselines include few-shot variants of default models, evaluated without (*fewshot*) or with the system prompt (*fewshot*-{*pl*|*en*}). This experimental setting enables a systematic comparison by isolating the effects of instruction tuning, system-prompt guidance, and in-context learning.

6 Results and discussion

6.1 Gender-inclusive proofreading evaluation

Impact of the LLM leading language Polish-specific models (see Section 4), derived from multilingual LLMs and further fine-tuned on large Polish corpora, are intended to improve performance on tasks requiring a deeper understanding of Polish. However, in both zero-shot and few-shot settings, they show no improvement over their multilingual predecessors, achieving F_1 scores close to zero (see all results in Table 7 in Appendix G). This finding, while disappointing, is expected, since texts used for tuning Polish-specific LLMs are mostly written in standard, non-inclusive Polish.

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: **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*, *fewshot* and IPIS-*tuned* versions, on the gender-inclusive proofreading task.

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 effec-

tiveness 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 (Appendix G, Table 7). Table 1 compares 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 in proofreading the detected expressions.

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

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.

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 (Appendix G, Table 7).

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 improves 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 proofread texts are of high quality, with minimal out-of-vocabulary words, typos, morphosyntactic errors, or lexical mismatches. All these findings induce us to categorise IPIS-tuned LLMs as a significant step toward the development of truly gender-inclusive LLMs for Polish.

6.2 Evaluation of gender-sensitive translation

Preliminary experiments We fine-tune and evaluate LLMs on the small **human-annotated** variant of the IPIS-translation dataset. Medium-sized models outperform smaller ones (Appendix G, Table 8), with Bielik-11B performing best overall. In this preliminary setting, benefits of IPIS tuning are mixed (see *hum-tuned* rows in Table 3). Although the IPIS-tuned Bielik-11B performs best for Polish-to-English translation (BLEU = 57.55; chrF = 78.03), its few-shot version yields the best English-to-Polish results (BLEU = 43.02; chrF = 72.46).

These results suggest that translating gender-inclusive Polish into English is not challenging for LLMs, and even limited IPIS tuning improves

LLM	Polish→English		English→Polish	
	BLEU	chrF	BLEU	chrF
human-annotated IPIS-translation-test				
default	47.60	73.39	41.49	71.78
default-pl	46.78	73.08	35.80	69.77
default-en	47.99	73.76	32.70	68.65
fewshot	50.01	73.93	38.63	68.78
fewshot-pl	49.38	73.84	37.81	69.92
fewshot-en	48.33	73.43	**43.02**	**72.46**
hum-tuned	55.19	75.80	34.26	55.08
hum-tuned-pl	56.70	76.93	31.70	58.36
hum-tuned-en	**57.55**	**78.03**	28.71	60.97
mix-tuned	57.27	77.90	34.29	63.93
mix-tuned-pl	54.67	75.82	29.84	61.16
mix-tuned-en	53.76	75.08	33.26	63.86
nllb-default	45.26	65.97	32.43	63.31
mixed-annotated IPIS-translation-test				
mix-tuned	**60.31**	**77.66**	**46.57**	**70.60**
mix-tuned-pl	55.89	75.04	42.81	68.54
mix-tuned-en	54.94	75.16	43.12	68.69
nllb-default	25.90	48.77	21.68	50.71

Table 3: Performance of Bielik-11B on the **gender-sensitive translation** task across *default*, *fewshot* and IPIS-*tuned* settings, with Polish user prompts and optional system prompts (*-pl*, *-en*). The model is fine-tuned on either human-annotated (*hum-*) or mixed-annotated (*mix-*) data; corresponding test sets are used for evaluation (top vs. bottom). Results from the NLLB-200-distilled-1.3B model are reported for comparison.

Bielik-11B over the baselines. The English-to-Polish direction, in turn, remains more difficult. The IPIS-tuned Bielik-11B underperforms its few-shot counterpart, indicating that the size and diversity of the available human-annotated IPIS data are insufficient to reliably support high-quality gender-inclusive Polish translation.

Follow-up experiments During instruction tuning, an LLM updates its parameters to reflect task-specific examples; however, instruction tuning on small datasets can lead to overfitting, limiting generalisation (Shi et al., 2024). A common mitigation strategy is data augmentation. We thus tune Bielik-11B on the full **mixed-annotated** IPIS-translation set (*mix-tuned* rows in Table 3). The *mix-tuned* model without a system prompt consistently outperforms the *hum-tuned* counterpart, achieving higher chrF scores in both translation directions (77.9 vs. 75.8 for PL→EN, and 63.93 vs. 55.08 for EN→PL). Moreover, when evaluated on the mixed-annotated IPIS-translation test set (bottom part of Table 3), the *mix-tuned* model performs substantially better, indicating improved generalisation compared to the preliminary setting.

Language impact The language of user queries (prompts) does not significantly impact 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 the English one – improves performance in some cases (e.g., in the *hum-tuned* Polish-to-English setting), but generally provides marginal benefits and may even degrade performance.

SOTA comparison To contextualise our results, we compare them with the state-of-the-art translation model – *NLLB-200-distilled-1.3B* (Costa-jussà et al., 2022, see *nllb-default* rows in Table 3). All Bielik-11B configurations – both baselines and IPIS-tuned – significantly outperform NLLB in both translation directions and on both test sets. NLLB performs particularly poorly on the mixed-annotated test set, indicating limited robustness to gender-inclusive language and confirming the effectiveness of our IPIS-based adaptation.

LLM-as-a-judge and human evaluation Although all IPIS-tuned Bielik-11B models outperform the SOTA NLLB model, BLEU and chrF do not capture gender-inclusive changes, especially in English-to-Polish translation. We therefore complement automatic evaluation with LLM-as-a-judge and human evaluation on 114 human-annotated IPIS-translation instructions. We use *Qwen3-30B-A3B-Thinking-2507* (Qwen Team, 2025) to judge Polish translations generated by Bielik-11B.

Across all criteria, the LLM-judge assigns lower scores than the human evaluator, suggesting its stricter scoring behaviour (see details and scores in Appendix H, Table 9). For MT-related criteria, especially adequacy-related aspects, both evaluators assign higher scores to the highest-BLEU subset than to the lowest-BLEU translations (for *completeness*, LLM: 2.632 vs. 1.281, Human: 4.053 vs. 1.333). In contrast, the difference between the highest- and lowest-BLEU subsets is significantly smaller for *gender inclusivity* (LLM: 3.263 vs. 2.947, Human: 4.246 vs. 3.579). This indicates that gender-inclusive language choices are only weakly associated with BLEU-based lexical similarity, and n-gram metrics may fail to capture non-standard, socially relevant linguistic properties. At the same time, correlation between LLM and human inclusivity scores is moderate and statistically significant (Pearson $r = 0.517$; Spearman $\rho = 0.539$), suggesting that LLM-based evaluation approximates human judgments.

These results suggest that *gender inclusivity* behaves as an independent dimension of translation quality, i.e., inclusive expressions are still present in low-quality translations (see additional analysis in Appendix H.3). As LLM can generate both high- and low-quality translations that include inclusive expressions, adding gender-inclusive guidance during its tuning does not harm its overall translation ability. Instead, the model learns to use inclusive language without compromising translation quality. Therefore, targeted gender-inclusive instruction tuning is a practical way to improve the inclusivity of LLMs without sacrificing overall performance.

Key findings Gender-sensitive translation depends on translation direction and training data scale. Translating gender-inclusive Polish into English is relatively straightforward for LLMs, with Polish-specific Bielik-11B performing well even if fine-tuned on a small human-annotated IPIS set. In contrast, generating gender-inclusive Polish from English is more challenging and benefits from larger, more diverse training data. Instruction tuning on the full mixed-annotated IPIS dataset significantly improves performance and generalisation, particularly in the English-to-Polish direction. Bielik-11B consistently outperforms the SOTA NLLB model, showing the effectiveness of language-specific adaptation for gender-sensitive translation. Our results also show that gender inclusivity is not captured by standard MT metrics and can be improved through instruction tuning without reducing overall translation quality.

6.3 General capabilities of IPIS-tuned LLM

To evaluate the effect of IPIS-tuning on general knowledge capabilities, we use *LLMzSzŁ* (Jassem et al., 2025), a comprehensive, publicly available Polish benchmark comprising 18.8k closed-ended questions from 154 domains. We compare the *default* Bielik-11B against its IPIS-tuned variants on proofreading and translation tasks (Table 4).

LLM	Accuracy	Delta
default	56.56	–
tuned (proofreading)	55.74	-0.82
tuned (translation)	55.87	-0.69

Table 4: Performance of IPIS-tuned Bielik-11B variants vs. the default model on *LLMzSzŁ*.

The default model achieves accuracy of 56.56%, while IPIS-tuned variants show only marginal drops: 55.74% for the proofreading model (-0.82%) and 55.87% for the translation model (-0.69%). Such minor degradation is expected in task-specific fine-tuning and indicates that IPIS datasets can be readily incorporated into broader instruction corpus without meaningfully affecting downstream performance, while improving model inclusivity.

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.

7.2 Instruction tuning

Instruction tuning is a fine-tuning technique designed to enhance the performance and controllability of LLMs (Zhang et al., 2024; Üstün et al., 2024). LLMs can be instruction-tuned for NLP and NLG 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. 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 (Ouyang et al., 2022; Mishra et al., 2022). Expanding an instruction-tuning dataset with additional tasks led to improved model performance, even on novel, previously unseen tasks, suggesting an improvement of the model's instruction-following capabilities (Wei et al., 2022).

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 instructions 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 research area – 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 we used only open-source LLMs, all experiments can be replicated, ensuring transparency and reproducibility.

The results confirmed that off-the-shelf LLMs inherently lack gender sensitivity because they are trained on non-inclusive textual corpora. The tasks examined in this study – gender-inclusive proofreading and gender-sensitive translation – proved highly complex and context-dependent. They require LLMs to reason about referential gender, linguistic structure, and world knowledge in order to identify generic masculine forms, distinguish personal from non-personal nouns, and generate grammatically and socially appropriate inclusive forms. Our experiments showed that LLMs can perform these tasks effectively, if they are tuned on a high-quality and sufficiently large set of instructions. In particular, 23k IPIS-proofreading instructions were sufficient to tune LLMs to generate gender-inclusive Polish texts. Limited training data in the preliminary translation experiment, in turn, led to overall model degradation. Bielik-11B fine-tuned on the full mixed-annotated IPIS-translation set (23k) outperformed both its preliminary variant and the strong multilingual NLLB model. LLM-as-a-judge and human evaluation further showed that gender inclusivity is only weakly correlated with BLEU and chrF scores. It behaves as an independent dimension of translation quality that can be improved through targeted instruction tuning without reducing overall translation performance.

Our approach embeds gender inclusivity as a core feature of LLMs, offering a systematic solution to mitigate gender bias in Polish language generation. By advancing gender-inclusive NLP and NLG, this work contributes to broader efforts to promote gender equality through more inclusive and equitable communication.

Limitations

In this work, we developed and evaluated instruction-tuned LLMs for gender-inclusive proofreading in Polish and bidirectional translation between gender-inclusive Polish and English. However, our approach has several limitations that should be acknowledged. An obvious limitation is the linguistic scope of our research, as we focus mostly on Polish. While this limits the generalisability of our findings to other languages, it also reduces the risk that the tested LLMs encountered the IPIS dataset during pretraining, minimising potential data contamination. Additionally, Polish is also particularly challenging for LLMs because of its complex grammatical gender system and strong masculine bias, making it a particularly demanding language for gender-inclusive NLP. Despite this language-specific focus, our work contributes to broader efforts to improve linguistic diversity in the largely English- and Chinese-dominated NLP world.

Our research focuses on binary gender inclusion. While the Polish generic masculine can refer to all individuals, including non-binary individuals, we target only feminine forms, masculine-feminine dyads, and neutral expressions. We do not model non-binary forms, as they are not yet standardised and remain socially emergent, unlike well-established feminine forms. Following a descriptive rather than prescriptive (normative) approach to language use, we do not introduce non-binary forms into our dataset. We acknowledge that this does not constitute full inclusivity, but consider the systematic inclusion of feminine forms as a linguistically grounded and socially relevant starting point. It is also worth noting that, from a broad repertoire of abbreviated gender-inclusive notations, we adopt a notation proposed by Wróblewska et al. (2025), e.g., *urzędni*k/czka/cze* [official], which encodes masculine (*urzędnik*), feminine (*urzędniczka*), and a candidate non-binary form (*urzędnicze*). While we adapt its version with only masculine and feminine forms, this notation remains extensible, and we plan to incorporate non-binary forms once they become standardised.

We rely on BLEU and chrF as reference-based indicators of translation quality; however, both metrics have limitations in the context of gender-inclusive language. In particular, they are insensitive to gender ambiguity and penalise valid gender-inclusive or neutral reformulations that diverge from single-reference translations. This limitation is especially pronounced for English-to-Polish translation, where morphosyntactic gender must often be made explicit, leading to lower scores even when outputs are linguistically well-formed and socially motivated. Consequently, BLEU and chrF should not be interpreted as direct measures of bias mitigation or inclusivity, but rather as coarse indicators of surface-level correspondence with reference translations. Due to the open-ended nature of the translation task, standard metrics are not effective for measuring gender inclusivity. Therefore, we evaluated English-to-Polish translation using an LLM-as-a-judge approach and a manual review of a sample of 114 translations (all English-to-Polish translations with English user prompts from the human-annotated IPIS-translation test set).

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.

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.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

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.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

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.

CM/Rec. 2019. Recommendation CM/Rec(2019)1 of the Committee of Ministers to member States on preventing and combating sexism. Council of Europe, Committee of Ministers.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *Preprint*, arXiv:2207.04672.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. *Preprint*, arXiv:2205.14135.

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.

Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach. 2023. FairPrism: Evaluating Fairness-Related Harms in Text Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6231–6251, Toronto, Canada. Association for Computational Linguistics.

GeBNLP. 2025. Workshop on Gender Bias in Natural Language Processing. `https://aclanthology.org/venues/gebnlp`. Workshop series 2019–2025.

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

GITT. 2025. Workshop on Gender-Inclusive Translation Technologies. `https://aclanthology.org/venues/gitt`. Workshop series 2023–2025.

GNL. 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, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Pascal Marc Gygax, Daniel Elmiger, Sandrine Zufferey, Alan Garnham, Sabine Sczesny, Lisa von Stockhausen, Friederike Braun, and Jane Oakhill. 2019. A Language Index of Grammatical Gender Dimensions to Study the Impact of Grammatical Gender on the Way We Perceive Women and Men. *Frontiers in Psychology*, 10(1604).

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.

Krzysztof Jassem, Michał Ciesiółka, Filip Graliński, Piotr Jabłoński, Jakub Pokrywka, Marek Kubis, Monika Jabłońska, and Ryszard Staruch. 2025. LLMzSzŁ: a comprehensive LLM benchmark for Polish. *Preprint*, arXiv:2501.02266.

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.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 196 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

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.

Jan Kocoń, Maciej Piasecki, Arkadiusz Janz, Teddy Ferdinan, Łukasz Radliński, Bartłomiej Koptyra, Marcin Oleksy, Stanisław Woźniak, Paweł Walkowiak, Konrad Wojtasik, Julia Moska, Tomasz Naskręt, Bartosz Walkowiak, Mateusz Gniewkowski, Kamil Szyc, Dawid Motyka, Dawid Banach, Jonatan Dalasiński, Ewa Rudnicka, and 80 others. 2025. Pllum: A family of polish large language models. *Preprint*, arXiv:2511.03823.

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.

Gabriela Koniuszaniec and Hanka Blaszkowska. 2003. *Polish. Language and gender in Polish*, volume 3, pages 259–285. John Benjamins Publishing Company.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. 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.

Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Maria Symeonaki, and Giorgos Stamou. 2025. Assumed Identities: Quantifying Gender Bias in Machine Translation of Gender-Ambiguous Occupational Terms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32233–32249, Suzhou, China. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Mistral AI team. 2024. Mistral NeMo. Accessed: May 19, 2026.

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: 2026-05-19.

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.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

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.

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.

Qwen Team. 2025. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

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.

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, Giuseppe Attanasio, Eleonora Cupin, Eleni Gkovedarou, Janiça Hackenbuchner, Anne Lauscher, Matteo Negri, Andrea Piergentili, Manjinder Thind, and Luisa Bentivogli. 2025. Mind the Inclusivity Gap: Multilingual Gender-Neutral Translation Evaluation with mGeNTE. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13698–13720, Suzhou, China. Association for Computational Linguistics.

Beatrice Savoldi, 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.

Zhengyan Shi, Adam X. Yang, Bin Wu, Laurence Aitchison, Emine Yilmaz, and Aldo Lipani. 2024. Instruction Tuning With Loss Over Instructions. *Preprint*, arXiv:2405.14394.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the Sexes in Language. In Klaus Friedler, editor, *Social communication*, Frontiers of Social Psychology, pages 163–187. Psychology Press.

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.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, Them, Theirs: Rewriting with Gender-Neutral English. *Preprint*, arXiv:2102.06788.

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.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender Neutral Alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eva Vanmassenhove, 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. 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, Karol 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 Preliminary Study on Writing Assistance. *Preprint*, arXiv:2305.13225.

A Translation examples

Gender-sensitive translating is not a trivial task. Since *English↔gender-inclusive Polish* datasets are extremely scarce, LLMs are usually trained on standard parallel corpora and therefore rarely learn how gender is encoded differently across the two languages.

A.1 Polish-to-English translation

In Polish, gender-inclusive language is often expressed through noun, pronoun, adjective, and verb dyads. These forms are generally translated into gender-neutral English expressions, which can lead to significant differences in token count between Polish and English texts, cf. (1) and (1-a). Moreover, because explicitly marked gender-inclusive forms may sound unnatural in English, it remains unclear whether such distinctions should be preserved in translation, see (1-b) and (2-b), or replaced with more natural gender-neutral alternatives, 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$_{masc}$[4] 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 pracownicę banku.*
[gloss.] On Friday the police arrested employee$_{masc}$ and employee$_{fem}$ of bank

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

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

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

The aim of the Polish-to-English translation experiments is to evaluate whether LLMs can generate high-quality English translations from gender-inclusive Polish source texts. Although the Polish inputs contain gendered dyads, the expected outputs are natural gender-neutral English translations without explicit gender marking, redundant repetitions, or other unnecessary linguistic phenomena.

A.2 English-to-Polish translation

Depending on the context, English gender-neutral expressions should be translated into Polish either as gendered dyads, see (3-a) and (4-a), or as a single gendered form, see (5-a) and the feminine translation of 'Rector' in (3-a). Translations with generic masculine forms are not acceptable, see (3-b), (4-b), and (5-b), nor are translations that contradict world knowledge, 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$_{masc}$[5] were$_{masc}$ awarded$_{masc}$ by the Rector$_{masc}$[6] 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$_{fem}$[7] 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$_{masc}$[8] and patients$_{fem}$ positively rated$_{masc}$/rated$_{fem}$ Medifem

The English-to-Polish translation experiments examine whether LLMs can translate English texts with minimal gender cues into gender-inclusive Polish by making contextually appropriate and inclusive choices in the target language.

B Details on the IPIS dataset

B.1 Human annotation of IPIS

This section describes the human annotation process used to create the IPIS underlying corpus

Annotators. The annotation was performed by trained Polish linguists (native speakers of Polish) employed full-time in a project. The team included two postdoctoral researchers and four PhD students.

Annotation procedure. Each text was independently transformed by two annotators. A third annotator reviewed both versions and could accept them, select one, or revise one or both proposals.

Gender-inclusive strategies. Annotators used two main inclusive strategies: feminisation (introducing feminine forms) and neutralisation (using impersonal forms or collective nouns). No single strategy was enforced; annotators selected the most appropriate strategy depending on the context.

Mixing strategies: Annotators were allowed to combine different strategies within the same text.

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

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

[7] 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.

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

This reflects natural language use, where such variation is common. As a result, we did not require consistent use of a single strategy, nor did we ask to create multiple versions of a text for each strategy. Instead, the annotators produced a single fluent version adapted to the genre.

Annotation guidelines. Annotators followed several general guidelines:

- Different inclusive strategies and notation variants could be combined within a single text if appropriate for the genre (e.g., gender-star notation should be avoided in conversation transcripts).
- Gender-star notation was the only allowed abbreviated form.
- Annotators were encouraged to follow the principle of linguistic economy, e.g., using a full gendered dyad at first mention and an abbreviated form later.
- Grammatical agreement between nouns, verbs, and adjectives had to be consistently adjusted; gender-star expressions are preferred in such cases.

Task difficulty. Annotators did not create new lexical items but used existing, attested forms. As native speakers, they generally did not find the task of introducing gender-inclusive or gender-neutral nouns, or adjusting grammatical agreement, particularly difficult.

Limitations. The annotation process has several limitations:

- Because multiple formulations may be equally valid, token-level inter-annotator agreement cannot be reliably measured. Semantically equivalent expressions may differ in form, while standard agreement metrics rely on surface similarity.
- The annotation process is not fully deterministic, since various inclusive forms and notations are acceptable. However, identifying which elements require gender-related changes remains consistent.
- We did not conduct user studies on perceived inclusiveness. Nevertheless, some forms (e.g., feminine forms and gendered dyads) are generally more accepted in Polish than strongly neutralised constructions such as *osoba studencka* [a student person].

B.2 Examples illustrating gender-inclusive strategies

Feminisation Introducing feminine forms either as gendered dyads (see 1 – 3) or as replacements for generic masculine forms (see 4).

1. Masculine and feminine nouns connected with a coordinating conjunction, e.g., *pacjenci i pacjentki* [patients$_{\text{masc}}$ and patients$_{\text{fem}}$],
2. Masculine and feminine nouns connected with a slash, e.g., *student/studentka* [student$_{\text{masc}}$/student$_{\text{fem}}$],
3. Abbreviated gender-inclusive form with an asterisk (gender star), e.g., *pracowni*ków/czek* [employees$_{\text{masc/fem}}$],
4. A feminine form denoting a function, profession, or position held by a woman, e.g., *dyrektorka Kwiatkowska* [director$_{\text{fem}}$ Kwiatkowska$_{\text{fem}}$] instead of *dyrektor Kwiatkowska* [Director$_{\text{masc}}$ Kwiatkowska$_{\text{fem}}$].

Neutralisation Using impersonal expressions or collective nouns.

1. The noun *osoba* [person] and its modifiers, e.g., *osoby uczestniczące w spotkaniu* [people participating in the meeting] instead of *uczestnicy spotkania* [meeting participants$_{\text{masc}}$],
2. A gender-neutral collective noun, e.g., *personel medyczny* [medical staff] instead of *lekarze i pielęgniarki* [doctors$_{\text{masc}}$ and nurses$_{\text{fem}}$].

B.3 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,

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 main components and four translation-specific ones:

1. **prompt** – user prompt, i.e., a specification of the given task (e.g., 'Generate an English translation of the text:'),

2. **source** – input text passage, i.e., a text passage requiring a 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 an user prompt (either EN or PL),

5. **source_language** – the language of a passage to translate, either gender-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),

7. **human** – the component of the full mixed-annotated set indicating whether the English text passage was written by a human (`true`) or not (`false`) (i.e. it was translated by the Gemma-3-27B model).

```
{"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",
 "human":true}
```

Figure 3: IPIS-translation example: English→gender-inclusive Polish translation direction and an user prompt in English.

C System and evaluation prompts

1. System prompt for gender-inclusive proofreading – Figure 4

2. System prompt for gender-sensitive translation – Figure 5

3. Evaluation prompt for LLM-as-a-judge evaluation – Figure 6

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

e) 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

e) 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.

You are an expert evaluator of translation quality. Your task is to assess a machine-generated Polish translation by comparing it with a reference Polish translation.

Evaluate whether the generated translation adheres to standard Polish language conventions, including appropriate terminology, grammar, and stylistic and rhetorical norms. Where applicable, also assess compliance with gender-inclusive language conventions used in the reference – applying inclusive forms where necessary, but avoiding their unjustified addition.

—

Evaluation Procedure

Step 1 – Identify Issues

Carefully compare the machine translation with the reference. For each issue found, provide:
- a brief explanation of the problem
- the relevant fragment from both the translation and the reference

Use the following issue types:

- **Omission** – missing clauses, sentences, or details
- **Addition** – information present in the translation but absent from the reference
- **Meaning error** – distorted or incorrect meaning relative to the reference
- **Terminology** – incorrect Polish terminology or mistranslations
- **Structure** – changes in paragraph structure or enumeration layout
- **Style** – deviations from standard Polish style or awkward phrasing
- **Gender inclusivity** – missing, incorrect, or inconsistent use of gender-inclusive forms

—

Step 2 – Assign Scores

Score each criterion from **1** (very poor) to **5** (excellent).

1. Completeness
Does the translation include all information present in the reference?
- Missing sentences, clauses, or phrases
- Unjustified additions not present in the reference
- Redundant repetitions

2. Clarity and Logical Coherence
Is the translation of the Polish text clear and logically coherent?
- Written entirely in Polish
- Logically structured at both the sentence and text levels
- Free of incorrect or misleading information

3. Structural Fidelity
Does the translation preserve the structural organisation of the reference?
- Paragraph and enumeration layout
- References to other texts

4. Linguistic Correctness
Is the translation grammatically correct and stylistically appropriate?
- Polish orthography, punctuation, and grammar
- Stylistic and phraseological accuracy
- Appropriate lexical choices and phrasing

5. Gender-Inclusive Language
Does the translation correctly reflect the gender-inclusive conventions of the reference?
- Preservation of inclusive forms (e.g., *członkowie i członkinie*, *członk*owie/inie*)
- Consistency in applying inclusive forms
- Absence of unjustified gender bias or unjustified inclusive forms

—

Step 3 – Output

Return your evaluation using the following JSON structure only – no additional commentary:

Figure 6: LLM-as-a-judge prompt for evaluation of English-to-Polish translation.

D List of tested LLMs

Tested multilingual LLMs

1. **Llama-8B**[9] (Grattafiori et al., 2024), a 8 billion-parameter model with a context length of 8k tokens,

2. **Mistral-7B**[10] (Jiang et al., 2023), a 7 billion-parameter model,

3. **Mistral-Nemo**[11] (Mistral AI team, 2024), a 12 billion-parameter model with a a context window of 128k tokens.

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

1. **Bielik-7B**[12] (Ociepa et al., 2024b), a 7 billion-parameter model derived from Mistral-7B-v0.1,

2. **Llama-PLLuM-8B**[13] (Kocoń et al., 2025), a 11 billion-parameter variant based on Mixtral 8x22B,

3. **Bielik-11B**[14] (Ociepa et al., 2024a) built upon Llama3.1-8B,

4. **PLLuM-12B**[15] (Kocoń et al., 2025) derived from Mistral-Nemo-Base-2407.

Reference translation model For translation experiments, *NLLB-200-distilled-1.3B*[16] (Costa-jussà et al., 2022) is used as the state-of-the-art reference model.

LLM-as-a-judge evaluation model For LLM-as-a-judge evaluation, *Qwen/Qwen3-30B-A3B-Thinking-2507*[17] (Qwen Team, 2025) is used with the following hyper-parameters: `temperature` of 0.6, `top_p` of 0.95 and `top_k` of 20.

[9] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[10] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
[11] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407
[12] https://huggingface.co/speakleash/Bielik-7B-Instruct-v0.1
[13] https://huggingface.co/CYFRAGOVPL/Llama-PLLuM-8B-chat
[14] https://huggingface.co/speakleash/Bielik-11B-v2.3-Instruct
[15] https://huggingface.co/CYFRAGOVPL/PLLuM-12B-nc-chat
[16] https://huggingface.co/facebook/nllb-200-distilled-1.3B
[17] https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507

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.
- 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 gendered dyads 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 7: The gold standard of the running example.

Gold standard comparison Since the manually corrected reference text is normalised using the same procedure (see Figure 7), 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 Instruction tuning hyper-parameters

Frameworks and libraries used in training:

- DeepSpeed: https://github.com/microsoft/DeepSpeed,
- FlashAttention (Dao et al., 2022): https://github.com/Dao-AILab/flash-attention,
- HuggingFace's ecosystem:
  - transformers: https://github.com/huggingface/transformers,
  - datasets: https://github.com/huggingface/datasets,
  - trl: https://github.com/huggingface/trl,
  - peft: https://github.com/huggingface/peft
- LoRA (Hu et al., 2021).

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

Table 6: LoRA and SFT hyperparameters

G Detailed results

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[18]	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

[18]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 7: 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
standarised human-annotated IPIS-translation-test												
nllb-200-distilled-1.3b	45.47*	65.59*	66.04*	–	–	–	**38.55***	65.34*	65.14*	–	–	–
bielik-11b-default	50.58	74.12	74.52	47.61	71.11	71.41	34.62	69.46	69.27	32.89	68.67	68.45
gemma-3-27b	48.70	78.41	78.58	**53.81**	**79.16**	**79.39**	37.50	**71.22**	**70.65**	13.95	60.12	58.98

Table 8: Preliminary evaluation of multilingual and Polish-specific LLMs on the **gender-sensitive translation** task using the human-annotated IPIS-translation-test subset. 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.
[Lower section of the result table] Reference Polish–English translation quality is reported for the out-of-the-shelf *NLLB-200-distilled-1.3B*, *Bielik-11B* and Gemma-3-27B models. The models are evaluated on the **standardised** human-annotated IPIS-translation-test comprising standard Polish–standard English translation pairs. Explanations: *user prompts are not used; **highest performance on standard translation**.

H Human vs. LLM evaluation of English-to-Polish translation

H.1 MT-related evaluation

We compare the scoring behaviour of the human evaluator and the LLM judge across the predefined evaluation criteria:

- *completeness*,
- *clarity* (and logical coherence),
- *structure* (i.e., structural fidelity),
- (linguistic) *correctness*,
- *gender inclusivity*.

Both evaluators assign higher scores to the high-BLEU subset of translations generated by IPIS-tuned Bielik-11B than to the low-BLEU one (see Table 9). For the high-BLEU instances, the human evaluator assigns high scores across all MT-related criteria (4.053 – 4.965). For the low-BLEU subset, *completeness* (1.333) and *structure* (3.018) are clearly lower scored, but *clarity* (4.333) and *correctness* (4.509) remain relatively high, i.e., translations with low BLEU scores can still be linguistically well-formed and understandable.

The LLM judge shows more uniform scoring patterns across MT-related criteria. For the high-BLEU translations, the average scores range from 2.404 to 3.053, and for the low-BLEU subset, they range from 1.281 to 1.842. As the LLM judge assigns lower scores to the low-BLEU subset across all criteria than the human evaluator, it may rely more strongly on lexical similarity signals related to BLEU, while humans evaluate quality criteria more independently.

Correlation analysis between human and LLM individual scores shows moderate to strong agreement for *completeness* ($\rho = 0.766$) and *structure* ($\rho = 0.582$). *Clarity* ($\rho = 0.196$) and *correctness* ($\rho = 0.141$), in turn, show weak correlations, which suggest that the LLM judge does not reliably reproduce human judgments for fluency-related aspects of translation.

H.2 Human report on translation errors

The error analysis of Polish translations generated by IPIS-tuned Bielik-11B reveals several reoccurring errors affecting *completeness*, *clarity*, *structure* and linguistic *correctness*.

Completeness: Many translations contain hallucinated content, where Bielik-11B introduces information that is not present in the English source text or in the Polish reference. Furthermore, some passages from the original text are omitted, and the translations are thus incomplete. Another common issue is the repetition or paraphrasing of the same content across multiple items in a list.

Clarity: Problems with clarity and the logical structure are often caused by weak connections between passages in the translated text. In some cases, sentences addressing a similar topic appear detached from their original context, which makes the translation difficult to follow.

Structure: Structural errors occur when Bielik-11B hallucinates new items (e.g., numbered points) or omits existing ones. This results in translations that do not faithfully reproduce the structure of the source text.

Linguistic correctness: Translations also contain grammatical and typographical errors, e.g., incorrect or non-agreeing verb forms (see Ex. 1), missing prepositions in prepositional phrases (Ex. 2), or inconsistent font usage (Ex. 3). Examples:

1. *Członk∗owie/inie Komisji powoływan∗i/e między datą wejścia w życie Traktatu z Lizbony a 31 października 2014 roku wybiera się* [wybieran∗i/e są] *spośród obywatel∗i/ek Państw Członkowskich.*

   [Eng. Members$_{masc/fem}$ of the Commission appointed$_{masc/fem}$ between the date of entry into force of the Treaty of Lisbon and 31 October 2014 are selected$_{masc/fem}$ from among the citizens$_{masc/fem}$ of the Member States.]

2. *Rada składa się z przedstawiciel∗i/ek rządów Państw Członkowskich, szczeblu ministerialnym* [na szczeblu ministerialnym], *któr∗zy/e mogą...*

   [Eng. The Council consists of representatives$_{masc/fem}$ of the governments of the Member States at ministerial level who$_{masc/fem}$ may...]

3. *PRAGNĄc pogłębić...*

   [Eng. DESIRIng to deepen...]

H.3 Evaluation of gender-inclusive modifications

Assessing English-to-Polish translations, the human evaluator counts gender-inclusive modifications (see Table 10). In the high-BLEU subset, 78.7% of candidates (i.e., words requiring inclusive modification) contain gender-inclusive forms,

Test Split	Evaluator	MT-related criteria				Gender inclusivity
		Completeness	Clarity	Structure	Correctness	
high-BLEU	LLM	2.632	2.614	3.053	2.404	3.263
	Human	4.053	4.947	4.632	4.965	4.246
low-BLEU	LLM	1.281	1.491	1.281	1.842	2.947
	Human	1.333	4.333	3.018	4.509	3.579
Pearson and Spearman correlations						
Pearson r (p-value)		0.681 (0.000)	0.145 (0.123)	0.471 (0.000)	0.154 (0.102)	0.517 (0.000)
Spearman ρ (p-value)		0.766 (0.000)	0.196 (0.037)	0.582 (0.000)	0.141 (0.133)	0.539 (0.000)

Table 9: Comparison of LLM-as-a-judge and human evaluations of English-to-Polish test translations generated by IPIS-tuned Bielik-11B (without a system prompt). The generated translations are split into the high-BLEU and the low-BLEU-scored subsets (57 translations each). The LLM-judge is Qwen3-30B-A3B-Thinking-2507. The upper part of the table reports average scores; the bottom part presents the correlation between human and LLM scores across all evaluation criteria and test instances.

compared to 58.6% in the low-BLEU subset. Many of them are correctly introduced gender-inclusive expressions (62.8% of the high-BLEU instances, and 48.3% of the low-BLEU ones).

Category	high-BLEU		low-BLEU	
	#	%	#	%
Candidates for incl. changes	239		273	
Modified candidates:	188	78.7	160	58.6
→ correct incl. forms	150	62.8	132	48.3
→ incorrect incl. forms	38	11.7	28	10.3
Missing incl. changes	51	21.3	113	41.4
Non-personal changes	1	n/a	1	n/a

Table 10: Statistical overview of gender-inclusive expressions in the test translations generated by IPIS-tuned Bielik-11B.

The rate of incorrect inclusive forms is relatively similar across the subsets. Incorrect inclusive forms occur in 11.7% of high-BLEU translations and 10.3% of low-BLEU ones. This suggests that translation quality (as measured by BLEU) does not strongly affect the likelihood of producing incorrect inclusive forms. Examples of incorrect gender-inclusive expressions:

- [incorr.] *wysoki*a przedstawiciel*ka jest wspomagany*a...*[19]
  [corr.] *wysok*i/a przedstawiciel*ka jest wspomagan*y/a...*
  [Eng.] the High$_{masc/fem}$ Representative$_{masc/fem}$ shall be assisted$_{masc/fem}$...
- [incorr.] *mieszkan*com/kom*
  [corr.] *mieszka*ńcom/nkom*
  [Eng.] residents$_{masc/fem}$
- [incorr.] *szefo*wie/wowe*
  [corr.] *szefo*wie/we*
  [Eng.] bosses$_{masc/fem}$

41.4% of low-BLEU candidate translations lack the required gender-inclusive expressions, compared to only 21.3% in the high-BLEU subset. A major reason for this is paraphrasing. IPIS-tuned Bielik-11B sometimes generates several alternative versions of a source text passage, some of which include gender-inclusive forms while others do not. This does not necessarily reflect a weakness of the model; rather, it suggests that the model is uncertain about which variant is most appropriate.

As only one instance in each subset contains an incorrect modification of a non-personal word, such errors appear to be extremely rare in translations generated by IPIS-tuned Bielik-11B, e.g.:

- [incorr.] *Żadn∗emu/ej Państwu Członkowskiemu*
  [corr.] Żadnemu Państwu Członkowskiemu
  [Eng.] No Member State
- [incorr.] *Parlament Europejski wybiera swoj*ego/ą przewodnicząc*ego/ą i sw*oj/e organy...*
  [corr.] Parlament Europejski wybiera swoj*ego/ą przewodnicząc*ego/ą i swoje organy...
  [Eng.] The European Parliament elects its$_{masc/fem}$ President$_{masc/fem}$ and its bodies...