PolEval 2025 Task 2: Gender-inclusive LLMs for Polish

Alina Wróblewska
Institute of Computer Science
Polish Academy of Sciences, Warsaw, Poland
alina@ipipan.waw.pl

Abstract

This paper presents the results of the PolEval 2025 shared task on gender-inclusive large language models for Polish. The primary goal of this task is to encourage the development of models capable of generating grammatically well-formed, contextually appropriate, and gender-inclusive output – a property of increasing importance in both human-centred NLP and NLG applications. To support this objective, we employed the newly developed Inclusive Polish Instruction Set (IPIS), a high-quality, human-annotated resource designed to guide models toward gender-inclusive behaviour. The shared task comprised two subtasks: *gender-inclusive proofreading*, which evaluates the ability of a model to transform masculine-generic Polish text into an inclusive equivalent, and *gender-sensitive Polish-English translation*, which investigates gender marking across languages. A total of six system submissions were received – three for each subtask. The evaluation demonstrates that the top-performing gender-inclusive systems outperform both the baseline and state-of-the-art models. Together, these results, along with the high-quality IPIS dataset and evaluation methodology, establish strong benchmarks for future research on gender inclusivity in Polish NLP.

1 Introduction

Polish is a grammatical gender language in which all nouns inherently encode grammatical gender markers as an integral part of the grammatical system, i.e., grammatical gender is a grammatical category that determines how nouns are classified into declensions groups. For example, *śliwka* [a plum] is feminine, *jabłko* [an apple] is neutral, while *pomidor* [a tomato] is masculine. All adjective, numeral, pronoun and verb forms associated with a noun must match the noun's grammatical gender.

Additionally, personal nouns are paired into mixed-gender dyads (e.g., *nauczycielka* [a female teacher] – *nauczyciel* [a male teacher]) to emphasize natural gender diversity across functions, professions, or roles. Although feminine personal nouns typically denote female individuals or groups of women, masculine personal nouns can refer not only to male individuals or male groups but also to mixed-gender groups and even to women, a phenomenon known as the *generic masculine*, e.g., *niemiecka polityk Ursula von der Leyen* [German$_{\text{fem}}$ politician$_{\text{masc}}$ Ursula$_{\text{fem}}$ von der Leyen].

Although the grammatical system of Polish allows for naming individuals according to their natural gender (i.e., female or male), standard Polish remains heavily androcentric. This is reflected in a strong dominance of masculine expressions over feminine ones, which may be interpreted as reinforcing gender bias and exclusion.

The dominance of masculine expressions over feminine ones in a language constitutes a form of gender discrimination (Gender Equality Commission, Council of Europe, 2016; European Parliament, 2018). Recognising the harmful effects of sexist language, the Council of Europe encourages its member states to eliminate sexism from linguistic practices and to promote forms that support gender equality. In recent years, Polish media, public administration, and government communications have increasingly used a gender-inclusive language (e.g., the term *ministra* [female minister] is now used to refer to a woman holding this position, whereas only a few years ago, *minister* [male minister] was applied to both women or men).

However, current large language models (LLMs) trained on standard Polish corpora struggle to generate gender-inclusive forms, often defaulting to generic masculine (Wróblewska et al., 2025). As a result, instead of supporting linguistic progress, they can hinder or slow the ongoing shift towards more gender-inclusive language. As LLMs become increasingly integrated into communication, translation, and content generation systems, it is crucial to ensure that their large-scale outputs align with the principles of gender inclusivity, particularly in grammatical gender-marking languages such as Polish.

In line with EU recommendations and with the aim of simulating positive change at the national level, a shared task dedicated to developing gender-inclusive LLMs for Polish was organised at the PolEval 2025 Workshop (Kobyliński et al., 2025).

2 Task description

2.1 Task objective

The PolEval 2025 Task 2: *Gender-inclusive LLMs for Polish*[1] aims to raise the community awareness of gender inequalities embedded in Polish and to foster development of LLMs capable of generating grammatically correct, contextually appropriate, and gender-inclusive language. Participants are encouraged to treat gender inclusivity as an essential, build-in property of LLMs rather than an optional add-on. This requires addressing a range of challenges, including the proper generation of feminine and masculine forms, the handling of mixed-gender references, and the avoidance of defaulting to the generic masculine.

By advancing gender-inclusive LLMs, i.e., language models that can recognise and reproduce inclusive linguistic patterns, the shared task contributes to practical solutions for mitigating gender bias in Polish language generation. The initiative also aligns with broader societal and institutional efforts to promote gender equality through language. In doing so, it underscores the potential of AI technologies to support more inclusive, equitable, and socially responsible communication practices.

2.2 Task definition

The PolEval 2025 Task 2 consists in developing gender-inclusive LLMs for Polish and evaluating them on two subtasks:

Ⓐ **Gender-inclusive Proofreading** consists in transforming a text passage written in standard Polish into its gender-inclusive version.
Gender-inclusive proofreading is not merely a stylistic transformation, as in a paraphrasing task. Rather, it requires the model to revise underlying internal representations and patterns of reasoning related to natural gender, and to draw on relevant world knowledge. Consider Example (1):

(1) *Polscy olimpijczycy, w tym siatkarze, wrócili do kraju.*[2]
[gloss.] Polish olympians$_{\text{masc}}$, including volleyball players$_{\text{masc}}$, have returned$_{\text{masc}}$ to the country.

(2) *Polskie olimpijki i polscy olimpijczycy – siatkarze – wróciły/wrócili do kraju.*
[gloss.] Polish olympians$_{\text{fem}}$ and Polish olympians$_{\text{masc}}$ – volleyball players$_{\text{masc}}$ – have returned$_{\text{fem}}$/returned$_{\text{masc}}$ to the country.

Given that at Paris 2024 only female athletes and the men's volleyball team from Poland won medals, the sentence (1) should be reformulated as in Example (2). Performing this proofreading requires substantially more than surface-level paraphrasing.

Gender-inclusive proofreading is of practical significance, especially in light of the current demand to revise government and administrative documents into gender-inclusive forms.

Ⓑ **Gender-sensitive Polish–English Translation** consists in translating a text passage written in gender-inclusive Polish into English or an English text passage into gender-inclusive Polish.
Gender-sensitive translation is not a trivial task. Since English↔gender-inclusive Polish datasets hardly exist, LLMs are typically trained on standard parallel corpora, reducing their chances to acquire differences in gender encoding.

In Polish, gender-inclusive expressions are often realised as mixed-gender dyads of nouns, pronouns, adjectives, and verbs. In English translations, these forms are typically rendered as gender-neutral expressions, resulting in a significantly lower token count, see (3) and (3-a). Reproducing these dyadic forms through duplication in English translation is generally incorrect, cf. (3-c). Moreover, given its unnatural sounding, it is questionable whether the gender distinctions present in the source text should be explicitly retained in translation, see (3-b), or whether a more neutral translation is preferable, see (3-a).

(3) *Szczyt technologiczny zgromadził liczne uczestniczki i licznych uczestników, które/którzy uczestniczyły/uczestniczyli w różnorodnych prelekcjach.*

[1] https://github.com/poleval/2025-gender-inclusive-llms

[2] The following colour scheme is used throughout the article: masculine noun phrase, feminine noun phrase, masculine predicate, feminine predicate.

[gloss.] Tech Summit attracted numerous attendees$_{fem}$ and numerous attendees$_{masc}$[3] who$_{fem}$/who$_{masc}$ participate$_{fem}$/participate$_{masc}$ in a variety of lectures

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

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

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

Depending on the context, English gender-neutral expressions should be translated into Polish either as mixed-gender dyads, see (4-a), or as a single gendered form, see (5-a). Translations relying on generic masculine forms are not acceptable, see (4-b), and (5-b). Likewise, translations that contradict world knowledge are unacceptable, see (4-c) and (5-c).

(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$_{fem}$[4] 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$_{masc}$[5] and patients$_{fem}$ positively rated$_{masc}$/rated$_{fem}$ Medifem

2.3 Task specification

Data: Participants are provided with the Inclusive Polish Instruction Set (IPIS), see Section 3.

Working phase: Using the training and development subsets of the IPIS dataset, participants are expected to adapt and improve an open-source LLM to ensure gender inclusivity.

Testing phase: Using the test subset of the IPIS dataset, the submitted outputs of the gender-inclusive LLMs are evaluated in the PolEval benchmarking system[6] (Kobyliński et al., 2025).

System prompt: Gender-inclusive system prompts based on Wróblewska et al. (2025) are available in the task repository. Participants are encouraged to employ these system prompts during both training and inference.

Modifications to the system prompt, as well as alternative uses of the IPIS dataset (e.g., for data augmentation), are permitted, provided they remain consistent with the task requirements and uphold principles of fair competition.

2.4 Task constraints

1. Participants may use publicly available pre-trained language models, both Polish-specific and multilingual.
2. The use of proprietary or closed-source LLMs is prohibited.
3. The training and development subsets of the IPIS dataset may be used freely for any task-related purpose, including (but not limited) to LLM instruction-tuning, fine-tuning, and data augmentation.
4. Participants may use publicly accessible linguistic resources, such as Polish corpora, lexical databases, knowledge graphs, and other structured data resources.

[3] Both women and men participate in Tech Summits.

[4] 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.

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

[6] https://poleval.amueval.pl

5. All external resources and models used for developing a gender-inclusive LLM must be clearly documented in the final description, including appropriate bibliographic references and/or direct URLs.
6. The use of non-public datasets, tools, or models is strictly forbidden.
7. It is prohibited to input any portion of the IPIS dataset – whether training or development instances – into proprietary LLMs (e.g., ChatGPT, Claude) for any reason, including data augmentation.
8. Each team is allowed to submit a maximum of three runs per task.
9. Participants are expected to prepare an article describing their solution in sufficient detail to allow replication of the research.

3 IPIS dataset

Inclusive Polish Instruction Set (Wróblewska and Żuk, 2025) is a collection of instructions designed to improve the gender sensitivity and inclusiveness of LLMs in the Polish language scenario. The IPIS dataset is built on a gender-inclusive text corpus manually annotated in the PLLuM project (Kocoń et al., 2025).

3.1 IPIS format

Ⓐ Gender-inclusive Proofreading Each IPIS-proofreading sample consists of three components:

1. **user prompt** (`prompt`) – a specification of the given task,
2. **input text passage** (`source`) – a text passage requiring a gender-inclusive proofreading,
3. **desired output** (`target`) – the expected response corresponding to the user instruction and an input text passage. This serves as the ground truth for evaluating and optimising LLM's predictions.

Ⓑ Gender-sensitive Polish–English Translation Each IPIS-translation sample consists of three main components and language specifications (see Figure 1):

1. **user prompt** (`prompt`) – a specification of the given task,
2. **input text passage** (`source`) – a text passage to translate,
3. **desired output** (`target`) – an expected translation in standard English or gender-inclusive Polish. This serves as the ground truth for evaluating and optimising LLM's predictions,
4. `prompt_language` – the language of prompt (either EN or PL)
5. `source_language` – the language of a passage to translate, either inclusive Polish (PL) or standard English (EN)
6. `target_language` – the language of a reference translation, either standard English (EN) or gender-inclusive Polish (PL).

```
{"source_resource_id": "EU_Karta_Praw_Podstawowych",
"ipis_id": "IPIS_translation_dev_117",
"prompt": "Translate into inclusive Polish. Text to
    translate: ",
"source": "Article 28\nRight of collective bargaining
   and action\nWorkers and employers, or their
   respective organisations, have, in accordance
   with Union law and national laws and practices,
   the right to negotiate and conclude collective
   agreements at the appropriate levels and, in
   cases of conflicts of interest, to take
   collective action to defend their interests,
   including strike action.",
"target": "Artykuł 28\nPrawo do rokowań i działań
   zbiorowych\nPracownic*y/e i pracodaw*cy/czynie,
   lub ich odpowiednie organizacje, mają, zgodnie
   z prawem Unii oraz ustawodawstwami i praktykami
   krajowymi, prawo do negocjowania i zawierania
   układów zbiorowych pracy na odpowiednich
   poziomach oraz do podejmowania, w przypadkach
   konfliktu interesów, działań zbiorowych, w tym
   strajku, w obronie swoich interesów.",
"prompt_language": "EN",
"source_language": "EN",
"target_language": "PL"}
```

Figure 1: The instance of the IPIS-translation subset.

3.2 IPIS size

Ⓐ Gender-inclusive Proofreading The gender-inclusive proofreading test, development and training subsets contain 5278, 2732 and 23,532 instances, respectively. All IPIS-proofreading partitions are balanced with respect to the proportion of gender-inclusive transformations.

Ⓑ Gender-sensitive Polish–English Translation The gender-sensitive translation test and training subsets contain 760 and 1728 instances, respectively.

4 Evaluation

4.1 Methodology

Ⓐ Gender-inclusive Proofreading To evaluate the ability of the gender-inclusive LLM to generate gender-inclusive language, its outputs are compared against gold standard test instances. The normalised LLM-generated texts (see Section 4.2 for details how to normalise LLMs' outputs) are assessed using the primary evaluation metric: ***F1-measure***. The textual quality of LLM-proofread passages is further evaluated using the secondary metrics: ***chrF*** and ***chrF++*** (Popović, 2015) and ***BLEU*** (Papineni et al., 2002).

Ⓑ Gender-sensitive Polish–English Translation To evaluate the ability of the gender-inclusive LLM to process and generate gender-inclusive Polish in the Polish↔English translation setting, model outputs are compared against gold standard test instances and ranked using the primary evaluation metric – ***chrF*** (Popović, 2015). Translation quality is further assessed using two secondary metrics: ***chrF++*** and ***BLEU***.

4.2 Normalisation procedure

Various gender-inclusive alternatives are possible, e.g., for *posłowie* 'deputies':

- posłanki i posłowie
- posłowie i posłanki
- posłowie/posłanki
- posłanki/posłowie
- posł*owie/anki

For the evaluation of gender-inclusive proofreading, the gender-inclusive `generated_target` samples must be normalised. The normalisation process consists in expanding all gender-inclusive expressions, especially those containing slashes or gender stars (asterisks), into a pair of masculine and feminine forms, followed by filtering out predefined stop words (i.e., punctuation marks, subordinating and coordinating conjunctions). Accordingly, the notation variants listed above for 'deputies' are normalised as `[posłowie posłanki]`.

In the normalisation steps, tokenisation is performed with Lambo (Przybyła, 2022), and part-of-speech tagging – with Combo (Klimaszewski and Wróblewska, 2021).

Table 1: Performance of **gender-inclusive proofreading** with Llama-PLLuM-8B (a small Polish-specific LLM) and Bielik-11B (the best LLM overall) in their baseline versions (*default* and *few-shot*) and the SOTA versions (*tuned*).

LLM	Acc	Prec	Rec	F_1	BLEU	chrF
Baseline						
Llama-PLLuM-8B						
default	34.88	0.18	0.22	0.20	32.04	57.13
default-pl	32.36	0.46	0.44	0.45	41.94	51.25
default-en	41.29	0.37	1.01	0.54	40.69	67.13
fewshot	31.34	0.49	0.58	0.53	37.51	52.38
fewshot-pl	38.05	0.56	0.66	0.60	46.65	58.55
fewshot-en	37.36	0.47	0.62	0.53	43.77	58.28
Bielik-11B						
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**	83.94
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
SOTA						
Llama-PLLuM-8B						
tuned	97.08	61.91	46.40	53.04	94.28	97.64
tuned-pl	95.86	50.87	36.68	42.63	93.40	96.85
tuned-en	96.19	51.93	44.25	47.79	93.78	97.25
Bielik-11B						
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

4.3 Baseline and SOTA

In PolEval 2025 Task 2: *Gender-inclusive LLMs for Polish*, the baseline and state-of-the-art (SOTA) are defined with reference to Wróblewska and Żuk (2025), which is the first systematic demonstration that instruction tuning can yield gender-inclusive LLMs for Polish.

Baseline The baseline includes several configurations of off-the-shelf LLMs (pre-trained, not-tuned) under zero-shot (denoted *default*) or few-shot (*fewshot*) evaluation settings, possibly with a gender-inclusive system prompt in Polish (*-pl*) or English (*-en*) added at inference time.

SOTA The SOTA of the shared task is represented by LLMs that were instruction-tuned on the human-crafted IPIS-train (*tuned*), using parameter-efficient adaptation (LoRA, Hu et al., 2021). The

Table 2: Performance of **gender-sensitive translation** with Mistral-Nemo (a multilingual LLM) and Bielik-11B (the best LLM) in their baseline versions (*default* and *fewshot*) and the SOTA versions (*tuned*). Explanations: baseline results are underlined and SOTA is in **bold**.

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++*
Mistral-Nemo												
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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**
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
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
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

IPIS-tuned models optionally receive the gender-inclusive system prompt at inference (*tuned-pl/en*).

Tested LLMs A range of small- and medium-sized models is tested:

- multilingual LLMs:
  - **Llama-8B** (Grattafiori et al., 2024),
  - **Mistral-7B** (Jiang et al., 2023),
  - **Mistral-Nemo** (Mistral AI team, 2024),
- Polish-specific LLMs:
  - **Bielik-7B** (Ociepa et al., 2024b),
  - **Llama-PLLuM-8B** (Kocoń et al., 2025),
  - **Bielik-11B** (Ociepa et al., 2024a),
  - **PLLuM-12B** (Kocoń et al., 2025).

Discussion On the gender-inclusive proofreading task, the gap between the baselines and the IPIS-tuned models is substantial. For example, the Polish-specific model, Bielik-11B, in its default configuration achieves an F1 score close to zero (see Table 1). In contrast, the IPIS-tuned variant of the same model reaches an F1 score of 60, accompanied by very high BLEU and chrF values. These results show that IPIS-based instruction tuning yields consistent, correct, and fluent gender-inclusive rewritings. It is also noteworthy that system prompts improve performance only for baseline models, but not for IPIS-tuned SOTA models.

For the gender-sensitive translation task, performance differs substantially between the two translation directions. In the Polish-to-English direction, instruction tuning yields only a modest improvement (see the Bielik-11B results in Table 2), but the baseline Mistral-Nemo models perform nearly as well as the SOTA Bielik-11B model. In English-to-Polish translation, however, instruction-tuned models are either outperformed by few-shot prompting (as in the case of Bielik-11B) or they surpass the baselines but only marginally (as observed for Mistral-Nemo).

These findings confirm that instruction tuning on a carefully crafted, human-annotated, gender-inclusive dataset is markedly more effective at steering LLMs toward gender-inclusive Polish than

Table 3: Overview of the systems participating in PolEval 2025 Task 2

System	Ⓐ	Ⓑ	LLM	Method	System prompt
AM	1	2	Qwen3-8B	LoRA adapter	PL
KW	3	1	Bielik-11B	few-shot/chain-of-thought prompting	modified PL
AP	2	–	plt5-base	LoRA adapter	pragmatic instruction
MC	–	3	Bielik-7B	LoRA adapter	PL

prompting or few-shot learning alone, provided that the training data are sufficiently large. These results establish a clear benchmark for PolEval 2025 submissions.

5 Submitted systems

This section outlines the systems participating in PolEval 2025 Task 2 (see Table 3 for a summary).

Majczyk (2025) – AM-ⒶⒷ The author proposes a parameter-efficient adaptation of the Qwen3-8B model (Yang et al., 2025) using LoRA (Hu et al., 2021) trained on the provided IPIS dataset. The fine-tuning process uses the official Polish system prompt supplied with the shared task, which contains guidelines for gender-inclusive proofreading. The adapted model wins the gender-inclusive proofreading subtask and gains the second place in gender-sensitive translation.

Wróbel (2025) – KW-ⒶⒷ The proposed prompt-based approach builds on the Bielik-11B v2.6 model (Ociepa et al., 2025), and employs carefully engineered system prompts with translation or proofreading examples and a structured JSON output format. The translation problem is formulated as the addition of gender-inclusive forms in EN→PL translation and the removal of such forms in PL→EN translation. The identification of terms requiring modification (adding or removing feminine forms) and the generation of the final translation are two steps of chain-of-thought reasoning. The proofreading task resembles the translation into gender-inclusive Polish. With this setup, the system achieved first place in the gender-sensitive translation subtask and third place in the gender-inclusive proofreading subtask.

Paszkowska (2025) – AP-Ⓐ The author proposes a pragmatically motivated approach to gender-inclusive proofreading. The plt5-base model (Chrabrowa et al., 2022) is adapted using the LoRA technique. The prompt design contains explicit pragmatic cues, i.e., *coreference cues*, *role cues*, *presuppositions*, *markedness*, and *cost*, which guide the model toward contextually appropriate gender-inclusive forms. The resulting system achieves relatively high precision but low recall.

Czajka (2025) – MC-Ⓑ The lightweight translation system is based on Bielik-7B-Instruct (Ociepa et al., 2024b) enhanced with LoRA adapters and optional 4-bit quantisation. The author reformatted training samples from the IPIS-translation subset into a chat-style dialogue, accompanied by a task-specific gender-inclusive prompt. Despite its intentionally modest budget and the use of greedy decoding, the system achieved 3rd place in the PolEval 2025 Task 2 (translation subtask).

6 Results

Table 4 reports the overall evaluation results for the gender-inclusive proofreading subtask. Among the three submitted systems, **AM-Ⓐ** achieved the highest scores and slightly surpassed the SOTA model. All participating systems outperformed the baseline across all metrics except recall: the systems ranked second and third achieved lower recall values (with the second system failing to generate outputs for several input texts).

Table 4: Results for Ⓐ *Gender-inclusive Proofreading*

LLM	Acc	Prec	Rec	F_1	BLEU	chrF
baseline	60.55	1.60	13.62	2.86	56.56	84.72
AM-Ⓐ	**97.45**	**64.50**	**56.77**	**60.39**	**95.76**	**98.07**
AP-Ⓐ	74.10	53.83	5.49	9.96	69.09	78.88
KW-Ⓐ	90.63	8.43	7.51	7.94	88.34	94.21
SOTA	97.37	63.93	56.26	59.85	95.22	97.99

The results for the gender-sensitive translation subtask are presented in Table 5. The **KW-Ⓑ** system substantially outperformed all other systems, including the current SOTA once. Due to the very small size of the IPIS-translation dataset, the author of the top-performing system chose not to instruction-tune the underlying Bielik-11B model; instead, the system relies on carefully designed sys-

Table 5: Results for B *Gender-sensitive Translation*. Baseline values that also represent SOTA results are marked with ★.

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++*
Mistral-Nemo												
baseline	54.52	75.89	74.15	53.42	75.78	73.97	43.02★	72.46★	69.61★	43.17★	72.82★	70.00★
KW-B	**80.32**	**88.18**	**87.60**	**79.89**	**88.31**	**87.74**	**55.45**	**79.52**	**77.25**	**58.40**	**80.25**	**78.19**
AM-B	36.21	63.00	60.47	36.83	64.83	62.09	23.31	54.88	50.65	22.76	52.66	48.71
MC-B	26.51	44.61	42.90	28.29	48.08	46.32	13.65	31.11	28.45	12.19	30.29	27.44
SOTA	57.55	78.03	76.66	57.45	77.93	76.52	43.02	72.46	69.61	43.17	72.82	70.00

tem prompts enriched with translation examples.

Across all systems, the English→Polish translation direction proves considerably more challenging than the reverse. Furthermore, the language of the user prompt does not measurably affect translation quality, indicating that model behaviour is dominated by the system-prompt configuration and training regime rather than input-prompt language.

7 Conclusion

This paper presents the results of Task 2: *Gender-inclusive LLMs for Polish* organised within the PolEval 2025 workshop. The evaluation demonstrate that the top-performing gender-inclusive systems outperform both the baseline and state-of-the-art models. These findings highlight the effectiveness of IPIS-based approaches and establish strong benchmarks for future research on gender inclusivity in Polish NLP.

The PolEval 2025 shared task on gender-inclusive LLMs for Polish introduces the first systematic evaluation benchmark dedicated to assessing gender inclusivity in Polish language generation. To the best of our knowledge, this is the first scientific effort of this kind worldwide. The task is built around the Inclusive Polish Instruction Set (IPIS) and comprises two complementary subtasks – gender-inclusive proofreading and gender-sensitive Polish–English translation – together with a dedicated evaluation methodology. This design enables evaluation of not only the grammatical and semantic correctness of LLM outputs, but also their ability to explicitly encode inclusive gender marking.

The results of the gender-inclusive proofreading subtask demonstrate that high-quality instruction tuning is a highly effective strategy for guiding LLMs towards inclusive language use. The winning system, **AM-A**, achieved the strongest overall performance, even slightly surpassing the SOTA model across all metrics, including F1= 60.39 and chrF= 98.07. Notably, all submitted systems outperformed the baseline, confirming that targeted modelling approaches – whether instruction-tuned or prompt-engineered – can substantially improve inclusive rewriting in Polish.

Taken together with the outcomes of the translation subtask, these findings underscore three broader conclusions. First, gender inclusivity in LLM output can be significantly advanced through carefully designed, human-curated instruction datasets such as IPIS. Second, while instruction tuning is highly effective for the proofreading task, its benefits for translation depend more strongly on the size and representativeness of the available training data. Third, the shared task establishes clear, data-driven performance baselines that will support consistent evaluation and encourage further methodological innovation.

Overall, the PolEval 2025 results highlight both the feasibility and the importance of developing LLMs capable of generating contextually appropriate, grammatically correct, and gender-inclusive Polish. We hope that the resources and benchmarks introduced here will stimulate continued research on inclusivity-aware language technologies and contribute to more equitable and user-aligned NLP systems.

Acknowledgments

We gratefully acknowledge Poland's high-performance computing infrastructure PLGrid (HPC Centres: ACK Cyfronet AGH) for providing computer facilities and support within the computational grant no. PLG/2022/015872.

References

Aleksandra Chrabrowa, Łukasz Dragan, Karol Grzegorczyk, Dariusz Kajtoch, Mikołaj Koszowski, Robert Mroczkowski, and Piotr Rybak. 2022. Evaluation of transfer learning for Polish with a text-to-text model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4374–4394. European Language Resources Association.

Mateusz Czajka. 2025. Lightweight IPIS Instruction Tuning of Bielik-7B for Gender-Inclusive Polish↔English Translation: System Description for PolEval 2025 Task 2 (IPIS-translation). In *Proceedings of the PolEval 2025 Workshop*.

European Parliament. 2018. Gender-Neutral Language in the European Parliament. Accessed on November 7, 2025.

Gender Equality Commission, Council of Europe. 2016. Gender Equality Glossary. Accessed on November 7, 2025.

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.

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

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

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. Association for Computational Linguistics.

Łukasz Kobyliński, Ryszard Staruch, Alina Wróblewska, and Maciej Ogrodniczuk. 2025. PolEval 2025. In *Proceedings of the PolEval 2025 Workshop*.

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.

Adam Majczyk. 2025. Less is More: Achieving SOTA at PolEval 2025 Task 2a (Gender-inclusive Proofreading for Polish) with LoRA and Qwen3-8B. In *Proceedings of the PolEval 2025 Workshop*.

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

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

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

Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, and Remigiusz Kinas. 2025. Bielik 11b v2 technical report. *Preprint*, arXiv:2505.02410.

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. Association for Computational Linguistics.

Adrianna Paszkowska. 2025. Instruction fine-tuning using pragmatic layer. In *Proceedings of the PolEval 2025 Workshop*.

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. ACL.

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

Krzysztof Wróbel. 2025. Prompt-Based Gender-Inclusive Polish-English Translation Using Bielik Large Language Model with Structured Output. In *Proceedings of the PolEval 2025 Workshop*.

Alina Wróblewska, Martyna Lewandowska, Aleksandra Tomaszewska, Karol Saputa, and Maciej Ogrodniczuk. 2025. Koncepcja form równościowych z asteryskiem inkluzywnym. *Język Polski*, 105(2):97–117.

Alina Wróblewska and Bartosz Żuk. 2025. Integrating gender inclusivity into large language models via instruction tuning. *Preprint*, arXiv:2508.18466.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.