Lightweight IPIS Instruction Tuning of Bielik-7B for Gender-Inclusive Polish↔English Translation: System Description for PolEval 2025 Task 2 (IPIS-translation)

Mateusz Czajka
Adam Mickiewicz University in Poznań, Poland
matcza11@st.amu.edu.pl

Abstract

We describe a compact but fully open-source system submitted to PolEval 2025 Task 2 (Gender-inclusive LLMs for Polish), subtask B: IPIS-translation. The goal of this subtask is gender-sensitive Polish↔English translation, including the production of gender-inclusive Polish outputs that follow specific orthographic conventions such as gender stars and slash forms. Our method performs instruction tuning of the Polish LLM Bielik-7B-Instruct using parameter-efficient LoRA adapters, with optional 4-bit NF4 quantization for single-GPU training. Samples from the Inclusive Polish Instruction Set (IPIS) are converted into a chat-style format with a task-provided gender-inclusive system prompt. Despite a deliberately lightweight tuning budget and greedy decoding, our submission placed 3rd on the hidden test B split, achieving bleu_pe = 20.7871. We detail the training and inference pipeline, discuss design choices and limitations, and outline directions for improving inclusive translation quality in Polish.

1 Introduction

Polish is a grammatical gender language in which virtually all nouns and many other parts of speech encode gender. In addition to distinct masculine and feminine forms of personal nouns, masculine forms routinely act as generics for mixed or unspecified groups, a phenomenon known as the *generic masculine*. This generic dominance is reflected in media, institutional communication, and web text, and is increasingly viewed as a form of linguistic sexism that may reinforce exclusion and bias (Wróblewska and Żuk, 2025).

As large language models (LLMs) trained on Polish become a standard component in translation and content-generation pipelines, they also inherit patterns of masculine-centric usage. This motivates the development of *gender-inclusive LLMs for Polish* that can produce outputs aligning with inclusive language guidelines and emerging community standards (Wróblewska and Żuk, 2025; Wróblewska et al., 2025).

PolEval 2025 Task 2 directly targets this problem by introducing the Inclusive Polish Instruction Set (IPIS) and two evaluation subtasks: (A) gender-inclusive proofreading and (B) gender-sensitive Polish↔English translation (PolEval Organizers, 2025; PolEval Task 2 Organizers, 2025). In this paper we address **only subtask B** (IPIS-translation). The shared task is designed around clear constraints: systems must be based on publicly available, open-source LLMs; the IPIS train and dev splits can be used for tuning and augmentation; and proprietary models cannot see any part of the IPIS data (PolEval Organizers, 2025).

Our goals were: (1) to build a fully reproducible pipeline based only on public tools and models; (2) to design a simple, robust prompt and training setup compatible with IPIS; and (3) to explore how far we can go with a relatively small Polish LLM (Bielik-7B-Instruct) and parameter-efficient fine-tuning on a single GPU.

1.1 Related Work

Our approach builds directly on the IPIS framework introduced by Wróblewska and Żuk (2025), who propose instruction tuning as a principled way to inject gender-inclusive behavior into Polish LLMs. Their work defines the IPIS dataset, system prompts and normalization procedures used in the shared task.

For modeling, we rely on the open-source Bielik family of Polish instruction-tuned LLMs (SpeakLeash Team, 2025). The task description reports strong results for larger Bielik and PLLuM models (11–12B parameters) tuned on IPIS in both proofreading and translation subtasks (PolEval Organizers, 2025; PolEval Task 2 Organizers, 2025). Our system can be seen as a lighter-weight variant that focuses on Bielik-7B and LoRA-based instruction

Proceedings of the PolEval 2025 Workshop, pages 66–71

tuning.

On the evaluation side, we follow established MT metrics, in particular BLEU (Papineni et al., 2002) and chrF (Popović, 2015), combined with task-specific normalization of inclusive forms (Wróblewska and Żuk, 2025; PolEval Task 2 Organizers, 2025).

2 Task Description

2.1 Subtasks and Objectives

PolEval 2025 Task 2 comprises two subtasks (PolEval Organizers, 2025; PolEval Task 2 Organizers, 2025).

Subtask A: gender-inclusive proofreading. The system receives standard Polish text and is asked to rewrite it into a gender-inclusive version. The evaluation normalizes inclusive forms (e.g. star and slash variants) and focuses on content-level F_1 of expanded masculine and feminine realizations.

Subtask B: gender-sensitive translation. In IPIS-translation, the system performs translation either from inclusive Polish to standard English or from standard English to gender-inclusive Polish. Inputs include a task-specific prompt and an input passage, while targets either follow inclusive Polish conventions or standard English norms, depending on the direction. **Our system participated exclusively in subtask B.**

2.2 IPIS-translation Format

Each IPIS-translation instance is a JSON object with the following fields (PolEval Task 2 Organizers, 2025; IPIPAN / PolEval Task 2 Team, 2025):

- `prompt`: task description (e.g. "Translate into inclusive Polish. Text to translate:"),
- `source`: passage to translate,
- `target`: reference translation (standard EN or inclusive PL),
- `prompt_language`: language of `prompt` (EN or PL),
- `source_language`: language of `source` (EN or PL),
- `target_language`: language of `target` (EN or PL).

To support instruction-style models, the dataset also provides a `messages` field containing an explicit chat-style dialogue between user and assistant.

2.3 Evaluation Metrics

For subtask B, the primary ranking metric is chrF, with chrF++ and BLEU as additional indicators of translation quality (Popović, 2015; Papineni et al., 2002; PolEval Task 2 Organizers, 2025). All metrics are computed directly on the system outputs and the corresponding gold-standard references, without any token-level normalization or expansion of gender-inclusive notation.

The public leaderboard additionally reports bleu_pe, which applies BLEU to normalized outputs and is particularly sensitive to exact lexical choices and orthographic variants.

2.4 Task Constraints

The shared task imposes several important constraints (PolEval Organizers, 2025; PolEval Task 2 Organizers, 2025):

- Systems must rely on publicly available pretrained models.
- Proprietary and closed-source LLMs are prohibited during training and evaluation.
- IPIS train and dev data may be used freely for fine-tuning and augmentation, but may not be fed into proprietary models.
- All external resources must be documented and cited.

Our system strictly adheres to these constraints.

3 Data

3.1 Dataset Size

The IPIS-translation train split contains 1,728 instances, and the test split contains 760 instances (PolEval Task 2 Organizers, 2025). The organizers also provide a public test A file which we use as a development set for manual inspection and debugging. Final scoring is performed on the hidden test B split.

3.2 Inclusive Conventions

Gender-inclusive Polish in IPIS follows guidelines developed in previous work on the inclusive asterisk ("asterysk inkluzywny") (Wróblewska et al., 2025), including:

- star-internal forms (e.g. *pracownic*y/e*),
- slash forms (e.g. *Polaków/Polek*),
- ordering of feminine and masculine alternants.

Models are expected to learn these patterns and preserve inclusiveness while remaining grammatically correct.

3.3 Preprocessing

We treat the IPIS JSONL files as the single source of truth and perform only minimal preprocessing:

- we read each JSON line as a Python dictionary,
- we keep the raw `prompt` and `source` texts,
- we use the original `target` for supervised training,
- we do not modify punctuation, casing or line breaks.

This deliberately conservative choice avoids accidentally normalizing away inclusive markers that are important for evaluation.

4 Method

4.1 Base Model: Bielik-7B-Instruct

We base our system on `speakleash/Bielik-7B-Instruct-v0.1`, a Polish instruction-tuned LLM released on HuggingFace (SpeakLeash Team, 2025). Bielik-7B is designed for Polish and bilingual tasks and comes with a chat interface compatible with the HuggingFace `transformers` library. We use the causal language modeling head for both training and generation.

4.2 Prompt and Chat Template

For each instance we construct a three-message chat:

- **System:** a gender-inclusive translation guideline prompt from the shared-task repository.
- **User:** the concatenation of `prompt` and `source`.
- **Assistant:** the reference `target` during training; left empty at inference time.

When the tokenizer includes an `apply_chat_template` method with a defined template, we call it directly. Otherwise we fall back to a simple textual format:

```
<|system|>
SYSTEM_PROMPT
<|user|>
PROMPT + SOURCE
<|assistant|>
TARGET (training) / empty (inference)
```

This approach allows our pipeline to adapt to different LLMs if needed while remaining compatible with Bielik's chat conventions.

4.3 System Prompt

We follow the task recommendation and use the official Polish system prompt for translation from the task repository (PolEval Task 2 Organizers, 2025; Wróblewska and Żuk, 2025).

If this file is missing, we fall back to a concise default: *"Translate the text into gender-inclusive Polish according to the inclusive language guidelines, preserving stars and slash forms."*

The same prompt is used during training and inference to encourage the model to internalize the inclusive style.

4.4 Parameter-Efficient Fine-Tuning

We employ LoRA via the `peft` library. Before attaching adapters, the model is prepared for k-bit training (using the standard recipe for 4-bit quantization). We target attention and MLP projections: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`. Our LoRA configuration exactly matches the implementation:

- rank $r = 64$,
- scaling $\alpha = 16$,
- dropout 0.05,
- bias: none,
- task type: causal language modeling.

This configuration strikes a balance between expressivity and memory footprint; it is sufficient to adapt the model to inclusive translation while keeping training feasible on a single RTX 5090 GPU.

4.5 Training Objective

We treat the task as standard causal language modeling on the combined prompt and target text. Each example is tokenized up to a maximum length of 2,048 tokens; shorter inputs are padded to this length. For training we set the labels to be equal to `input_ids` (i.e. full causal loss over the sequence). We do not mask out the prompt region; in practice, the model learns to copy the system and user turns and to focus its generative capacity on the assistant segment.

5 Experimental Setup

5.1 Hardware and Software

All experiments are run on a single NVIDIA RTX 5090 GPU (Blackwell architecture, sm_120) with CUDA 12.8. We use:

- PyTorch with cu128 builds,
- `transformers` for model and tokenizer,
- `bitsandbytes` for 4-bit quantization,
- `peft` for LoRA,
- `accelerate` for efficient device placement.

5.2 Hyperparameters

Training uses the HuggingFace `Trainer` with:

- epochs: 2,
- per-device batch size: 1,
- gradient accumulation: 8,
- effective batch size: 8,
- learning rate: 1×10^{-4},
- precision: bf16,
- max sequence length: 2,048.

For 4-bit training we enable NF4 with double quantization and use paged AdamW with 32-bit optimizer states; otherwise, we use standard AdamW.

5.3 Inference

At inference time (test B), we load: (1) the base Bielik-7B-Instruct model, and (2) the trained LoRA adapter. We then construct inputs using the same system and user messages as during training, with an empty assistant turn.

Generation is purely greedy: `do_sample=False`, `temperature=0.0`, `max_new_tokens=256`. We trim the prompt portion of the decoded sequence to obtain `generated_target`. The outputs are serialized as JSONL with the fields `ipis_id` and `generated_target` and written into the TSV-compatible format required by the PolEval platform (PolEval Task 2 Organizers, 2025).

6 Results

Our submission achieved **3rd place** in the IPIS-translation subtask on hidden test B, with:

- bleu_pe = 20.7871.

The best system in the leaderboard reached bleu_pe of approximately 69.3687, indicating a substantial performance gap to the state of the art.[1]

Due to the competition setting, we do not have access to detailed per-system scores on chrF or chrF++ for test B. However, based on the task description and baseline results reported by the organizers, larger tuned Polish models (e.g. 11B or 12B) achieve stronger metrics than smaller baselines (Wróblewska and Żuk, 2025; PolEval Organizers, 2025).

Training script:
`train_translation_gpu.py`
Inference script:
`predict_translation_gpu_B.py`
Training command:

```
python train_translation_gpu.py --load-in-4bit
```

Code and configuration files are released at:
https://github.com/Matcza11/PolEval_Gender_Inclusive_LLMsTranslation

7 Qualitative Analysis

7.1 Preservation of Inclusive Markers

The system generally respects the star and slash conventions present in the training data. When the prompt explicitly indicates inclusive output, generated text often mirrors the reference pattern, such as: *pracownic*y/e*, *Pol*aków/ek*, *uczniów/uczennic*. In some cases, however, the model chooses only a masculine form or reverts to standard Polish, which is likely penalized under the normalization procedure.

[1] Value reported here as indicative context from the PolEval challenge page.

7.2 Literal vs. Inclusive Tension

For EN→PL direction, we occasionally observe a tension between literal semantic translation and strict adherence to inclusive guidelines. For example, when the English source lacks explicit gender marking, the model must decide whether to add inclusive forms or remain neutral. Our system sometimes defaults to a non-inclusive but grammatically correct sentence, which harms inclusive evaluation while remaining acceptable for general translation.

7.3 Formatting and Line Breaks

Some reference texts contain structured elements (e.g. article titles, section headers, numbered items). Our model reproduces these reasonably well but occasionally alters line breaks or spacing, which BLEU-style metrics may punish despite end-to-end readability.

7.4 Error Types

Common error types include:

- missing one side of an inclusive pair (e.g. *ochrona konsumentów* instead of *ochrona konsument*ów/ek*),
- incorrect star placement inside a morpheme (e.g. *pracowni*cy/ce* instead of *pracownic*y/e*),
- overuse of inclusive markers in contexts where neutral forms would suffice (e.g. *Wymogi ochrony konsument*ów/ek są uwzględniane...* instead of *Wymogi ochrony konsumenckiej są uwzględniane...*),
- small semantic omissions at sentence end (e.g. EN: *Specific provisions shall apply to those Member States whose currency is the euro.* PL (wrong): *Do Państw Członkowskich stosuje się postanowienia szczególne.* (ommited: *„których walutą jest euro”*)).

These errors suggest that longer training and explicit regularization of inclusive patterns could improve performance.

8 Discussion

Our system demonstrates that even a relatively small Polish instruction-tuned LLM can be effectively adapted to IPIS-translation using simple LoRA tuning on a single GPU. However, the gap to top systems highlights several avenues for improvement.

Model capacity and recency. The task report and related work suggest that larger Bielik and PLLuM models, especially when tuned on IPIS, achieve higher chrF and BLEU scores for both proofreading and translation (Wróblewska and Żuk, 2025; SpeakLeash Team, 2025; PolEval Organizers, 2025). Exploring Bielik-11B or 12B variants with the same pipeline is a natural next step.

Decoding strategy. Greedy decoding is robust but conservative. Beam search or nucleus sampling with reranking by an auxiliary inclusive-language classifier might yield outputs closer to reference forms without sacrificing grammaticality.

Data augmentation. The shared task allows various forms of data augmentation based on IPIS. For example, one could systematically swap feminine/masculine order in inclusive pairs, or generate paraphrastic prompts, to improve robustness to small orthographic variations.

Multitask learning. Joint tuning on both IPIS-proofreading and IPIS-translation may help the model internalize inclusive patterns more strongly and transfer them across tasks. We leave such extensions for future work.

9 Conclusion

We presented a lightweight PolEval 2025 IPIS-translation system based on Bielik-7B-Instruct with LoRA adapters and optional 4-bit quantization. Using a simple chat-style prompt format and a short training schedule, we achieved 3rd place on the hidden test B split, with bleu_pe=20.7871.

Beyond the competition, the pipeline serves as a reproducible reference for training gender-inclusive Polish translation models under strict open-source constraints. Future work will explore larger models, richer decoding, and explicit regularization of inclusive patterns.

Limitations

Our approach has several limitations: (1) only a single base model was explored; (2) hyperparameters were chosen heuristically without a thorough search; (3) we did not experiment with multilingual backbones or additional corpora beyond IPIS. Moreover, the qualitative analysis is restricted by the lack of access to official gold labels for test B.

Ethics Statement

The system is explicitly designed to promote gender-inclusive language in Polish, which aligns with recommendations of European institutions to avoid sexist language and support gender equality (Wróblewska et al., 2025; Wróblewska and Żuk, 2025). At the same time, inclusive forms are socially debated and evolving. Any deployment of such systems should respect user preferences and provide transparent control over the degree and style of inclusivity.

We use only publicly available datasets and open-source models. No proprietary LLMs see the IPIS data, and no personal information beyond what is contained in the shared task resources is processed.

References

IPIPAN / PolEval Task 2 Team. 2025. Inclusive polish instruction set (ipis). Dataset description accessed during PolEval 2025.

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.

PolEval Organizers. 2025. Poleval 2025 task 2: Gender-inclusive llms for polish. Accessed: 2025-12-05.

PolEval Task 2 Organizers. 2025. 2025-gender-inclusive-llms: Task repository and ipis dataset description. Accessed: 2025-12-05.

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.

SpeakLeash Team. 2025. Bielik-7b-instruct-v0.1. Polish instruction-tuned causal LLM.

Alina Wróblewska, Martyna Lewandowska, Aleksandra Tomaszewska, Karol Saputa, and Maciej Ogrodniczuk. 2025. Koncepcja form równościowych z asteryskiem inkluzywnym. *Język Polski*, CV(2):97–118.

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