Instruction fine-tuning using pragmatic layer

Adrianna Paszkowska
Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen, Germany
pasz.adrianna@gmail.com

Abstract

The Polish language, like some Slavic and Romance languages, has a masculine-centric bias in its generic forms, leading to frequent use of masculine nouns when referring to women or mixed-gender groups. This presents a linguistic challenge for the development of gender-inclusive technologies, addressed in PolEval task 2.

This paper presents a pragmatic instruction fine-tuning approach, using Low-Rank Adaptation (LoRA) on the pre-trained Polish PLT5 sequence-to-sequence model.

1 Introduction

Task 2 of PolEval 2025 covers gender-inclusivity in Polish. Subtask A specifically addresses proofreading, transforming standard text into inclusive versions. The approach introduced by (Wróblewska and Żuk, 2025) focuses on semantic and morphological transformations through instruction fine-tuning—for example, converting masculine "student" to feminine "studentka" by teaching substitution rules to an LLM. The PolEval 2025 task (Wróblewska, 2025) builds directly on this work.

However, there is a possibility to enhance this understanding using pragmatics—interpreting meaning based on context, speaker intent, and social knowledge. We propose adding a pragmatic layer to the instructions to guide the model's decision-making, significantly improving contextual accuracy and resolving ambiguity that morphological approaches cannot.

2 Background

Gender-inclusive language has gained increased attention in recent years, both in linguistic and NLP research. This is evidenced by international workshops like GeBNLP (Faleńska et al., 2024) and GITT (Savoldi et al., 2024), as well as the PolEval 2025 shared task (Kobyliński et al., 2025) for Polish specifically. According to (Wróblewska and Żuk, 2025), such language is especially relevant in professional and social contexts.

The linguistic challenge for gender inclusivity in Polish stems from the language's rich inflectional morphology with gender encoded across multiple grammatical categories: nouns, adjectives, numerals, and past-tense verbs. Here, the masculine form is traditionally unmarked and used generically, creating ambiguity when referring to women or mixed groups. This poses a significant challenge for automatic text processing, as models often default to masculine interpretations even when feminine or neutral forms would be more appropriate (Wróblewska and Żuk, 2025).

Models often fail in contexts requiring pragmatic inference (Gan et al., 2025), such as coreference resolution, speaker intent, social context, and presuppositions embedded in group nouns. This is particularly relevant for inclusive language generation, where the correct target form often depends on cues.

Instruction tuning and in-context learning have proven effective for making models behave in more context-sensitive ways (Sato et al., 2025), but they require prompt engineering and substantial model capacity. In Polish, very few studies have explored combining pragmatically-informed instructions with fine-tuning.

Proceedings of the PolEval 2025 Workshop, pages 60–65

3 Model

Due to hardware constraints, our design prioritized balancing model size with task-specific usability. Section 3.1 introduces the base encoder-decoder model, while Section 3.2 describes the parameter-efficient fine-tuning technique we employed. The complete set of training hyperparameters is provided in Table 2 and discussed in Section 3.3.

3.1 Base Architecture and Setup

We used the PLT5-base model (Chrabrowa et al., 2022), an encoder-decoder T5 architecture pre-trained on a large corpus of Polish text. This architecture is optimal for the sequence-to-sequence proofreading task. All experiments were conducted on a GPU A100 instance provided by Google Colab. The system was implemented in PyTorch using the HuggingFace Transformers and PEFT libraries. Due to resource constraints, the experimental setup focused on a single fine-tuning run selected through prompt design and preliminary inspection.

3.2 LoRA Implementation Details

To enable efficient adaptation while preserving the model's pre-trained knowledge, we applied Low-Rank Adaptation (LoRA) (Hu et al., 2021). This parameter-efficient fine-tuning technique was crucial to prevent catastrophic forgetting given our long, detailed prompts and to ensure memory-efficient training. The configuration, detailed in Table 1, targeted the attention mechanisms and feed-forward layers to minimize trainable parameters while allowing adaptation to the pragmatic reasoning task.

Parameter	Value
Rank (r)	16
Alpha (α)	32
Attention Mechanisms	q, v, k, o
Feed-Forward Networks	wi, wo, lm_head
Trainable Parameters	5,433,344
Total Parameters	280,536,320
Percentage Trainable	1.94%

Table 1: LoRA Configuration Details

3.3 Training Hyperparameters

The training setup utilized the AdamW optimizer (Loshchilov and Hutter, 2019). We used a low learning rate of 1×10^{-4} to ensure stable convergence and prevent catastrophic forgetting of the pre-trained Polish knowledge. The complete hyperparameters are listed in Table 2.

Parameter	Value
Learning Rate	1×10^{-4}
Batch Size	8
Epochs	3
Max Sequence Length	256
Optimizer	AdamW
Beam Search Settings	4

Table 2: Training Hyperparameters

4 Data Collection and Preparation

The dataset used for this study was provided as part of PolEval 2025 Task 2 (Wróblewska, 2025), which originates from the IPIS dataset (Wróblewska and Żuk, 2025). It consists of sentence-level and short-paragraph examples annotated with inclusive rewritings. The organizers supplied two files: `train.jsonl` and `dev.jsonl`, each containing prompt-target pairs of original and inclusive text. The dataset covers a wide variety of linguistic constructions.

Before training, all examples were normalized using the following procedure:

- Removal of stray whitespace and non-standard punctuation.
- Tokenization using the PLT5 sentencepiece model without additional preprocessing, to maintain compatibility with the pre-trained vocabulary.
- Truncation or padding to the maximum length of 256 tokens.
- Removal of encoding artifacts, including Unicode replacement characters (U+FFFD), mojibake (e.g., CJK characters like U+898F resulting from encoding mismatches), and empty or malformed strings.

No aggressive augmentation was applied due to the nature of the task; modifications risked altering gender cues in unintended ways. Instead, the few-shot prompt examples were designed to supplement the dataset with controlled boundary cases.

5 Prompt

The complete instruction prompt, including all few-shot examples and defined rules, is provided in Appendix A for full reproducibility.

5.1 Prompt Design

Our prompt design builds upon the instruction template used in previous work on gender-inclusive transformation (Wróblewska and Żuk, 2025). The training data contained variations of these queries, which we analyzed to define specific transformation cases. The main goal was to teach the model to distinguish when a speaker refers specifically to a man or a male group versus when they are using common Polish generic masculine forms.

We framed the model as a pragmatic editor, matching the proofreading task. Its specific role was to identify pragmatic failures: instances where the language might be grammatically correct but contextually exclusionary. To define this intention, we designed several basic rules.

5.1.1 Coreference Cues (Case A)

This rule instructs the model to use gender signals outside the noun phrase to determine the target gender, forcing it to resolve long-distance dependencies. For example, spotting specific entities ("Anna") or pronouns ("jej" 'her') to override the default implicit gender of words like "lekarz" (doctor).

5.1.2 Role Cues and Presuppositions (Case B)

Generic masculine terms in plural contexts (e.g., "studenci" 'students [masculine/mixed]') pragmatically presuppose mixed groups. When such a group noun is detected, the probability of the "Dual-Inclusive" class (e.g., "studentki i studenci" 'female and male students') is maximized over the generic masculine.

5.1.3 Default and Markedness (Case C)

In Polish, the masculine is "unmarked" (default). Inclusive language attempts to "unmark" the feminine or introduce a neutral form. This rule defines the fallback behavior. In the absence of strong positive signals (like a female name), the model is instructed to default to the "Osoba-form" (neutral, e.g., "osoba kierująca" 'person driving') rather than hallucinating a specific gender.

5.1.4 Cost

Following the Principle of Least Effort (Sperber and Wilson, 1995), speakers often use the short, unmarked masculine form as a low-cost default option. The model, acting as a pragmatic editor, must counter this tendency. The cost of inclusive language (e.g., using "osoba kierująca" or the dual-inclusive "studentka/student") is higher but achieves a crucial social goal (maximizing inclusivity). Therefore, when a rule requires transformation, the model must accept this higher communicative cost to override the speaker's pragmatic preference for brevity.

5.2 Few-shot Examples

We provided eight few-shot examples (Cases A-1 through C-2) representing boundary cases for the rules defined above. All examples are listed in full in Appendix A.

- Case A-2 demonstrates the override of a generic term by a specific named entity.
  Example: "Pani Anna jest lekarzem." ('Ms Anna is a doctor') → The feminine context of "Anna" triggers the transformation to "lekarką" ('Ms Anna is a doctor [feminine]').

- Case B-2 demonstrates complex agreement where the numeral must also change.
  Example: "Pięciu kandydatów otrzymało nagrody"('Five candidates received an award.') → "Pięcioro kandydat*ów/ek" ('Five candidates [neutral/collective]')

6 Results and Comparative Analysis

Our model was evaluated on the held-out `TestB.jsonl` test set from the PolEval Task 2 data. Following the official task guidelines, we report the standard classification metrics, focusing on the F1-score as the primary measure of balance between precision and recall. We also include the best-reported result from the official task for contextual comparison, which used PLLuM-12B (Kocoń et al., 2025) as baseline and Bielik-11B (Ociepa et al., 2024) for SOTA.

- F1-score: The harmonic mean of precision and recall.

- Precision: The fraction of relevant instances among the retrieved instances.

- Recall: The fraction of relevant instances that were retrieved.

Model	Precision	Recall	F1
PLLuM-12B	2.56	6.28	3.65
Our Model	**53.83**	**15.49**	**9.96**
Bielik-11B	63.93	56.26	59.85

Table 3: Comparative Token-Level Results on the PolEval Test B Set

The PI-LoRA model achieved an F1-score of 9.96. While this value appears low, it represents a 2.7 $\times$ improvement over the official PLLuM-12B (Kocoń et al., 2025) baseline F1 3.65 for the proofreading task, validating the effectiveness of the pragmatic instruction layer. The model's performance is characterized by a high Precision (53.83%) coupled with a low Recall 15.49%. This indicates that when the model decides to make a gender-inclusive transformation, it is highly accurate (few false positives), but it misses the majority of required transformations (many false negatives).

7 Discussion and Limitations

The results suggest that incorporating pragmatic instructions effectively guides the model toward more contextually appropriate inclusive forms. Unlike purely morphological systems, the proposed approach reasons about speaker intent and coreference, which are essential for handling ambiguous or discourse-dependent phenomena.

However, several limitations remain. First, the model depends heavily on the quality of the instruction: poorly defined pragmatic rules may lead to inconsistent behavior. Second, the PLT5 architecture, while strong for Polish, is relatively small compared to modern multilingual LLMs, limiting its capacity for complex discourse reasoning.

7.1 Future Work

Future work can focus on addressing the observed low recall:

- Scaling and Capacity: Applying the pragmatic instruction set to a larger foundational model to better internalize the complex morphological and syntactic agreement rules necessary for high recall.
- Rule Refinement: Adjusting the Default Rule to be less conservative, potentially by instructing the model to default to the dual-inclusive form (studentki/studenci) instead of the more complex Osoba-form in ambiguous contexts, thereby increasing the rate of necessary transformations.
- Rule Conjunction: Adapting the merger of the full (Wróblewska and Żuk, 2025) prompt with pragmatic rules.
- Training Curriculum: Implementing a curriculum learning approach where simple transformations (Rule A) are taught before complex agreement patterns (Rule B), aiming to stabilize learning and mitigate the current high conservatism.

8 Conclusion

This paper presented a pragmatic instruction fine-tuning approach using LoRA on the PLT5 model for gender-inclusive language transformation in Polish. By defining explicit rules for pragmatic reasoning (Coreference, Role Cues, and Cost), we achieved a 2.7$\times$ F1-score improvement over the official baseline, demonstrating that targeted instruction can successfully align LLMs for context-sensitive social tasks. The high precision validates our strategy of prioritizing contextual accuracy.

9 Acknowledgments

The model output results were normalized by Alina Wróblewska, using the script for normalization included in the task repository (Wróblewska, 2025).

This article is not peer-reviewed, but reviewed by the organizing committee.

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. *arXiv preprint arXiv:2205.08808*.

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

Yujian Gan, Yuan Liang, Yanni Lin, Juntao Yu, and Massimo Poesio. 2025. Improving llms' learning for coreference resolution. *Preprint*, arXiv:2509.11466.

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.

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. *arXiv preprint arXiv:2511.03823*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, SpeakLeash Team, and Cyfronet Team. 2024. Bielik-11b-v2 model card. Accessed: 2024-08-28.

Takuma Sato, Seiya Kawano, and Koichiro Yoshino. 2025. Pragmatic theories enhance understanding of implied meanings in llms.

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

Dan Sperber and Deirdre Wilson. 1995. *Relevance: Communication and Cognition*, 2nd edition. Blackwell Publishing.

Alina Wróblewska. 2025. PolEval 2025 Task 2: Gender-inclusive LLMs for Polish. In *Proceedings of the PolEval 2025 Workshop*.

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

A Pragmatic Instruction

PRAGMATIC_INSTRUCTION =
""" You are a pragmatic editor. Take the speaker's intentions into account.
Detect pragmatic failure: technically correct language that is contextually exclusionary or socially problematic.
Consider speaker intent and social context, not just dictionary meaning (e.g., generic masculine terms when the context implies mixed groups, juxtaposition of male/female titles, or professional titles that should be neutral). Include coreference cues.

Rules for pragmatic contextual reasoning:
A. COREFERENCE CUES: Search for explicit gender signals (named subjects like Anna, Pan, or pronouns like jej, jego) outside the immediately gendered term. If a female cue is present, the generic masculine term MUST be transformed into female or dual-inclusive form.
B. ROLE CUES: If the sentence mentions roles that suggest mixed groups (e.g., studenci, lekarze) or words like 'wszyscy', 'każda', transform generic masculine into dual-inclusive form (Masculine nouns/feminine nouns or Masculine and feminine nouns with a coordinating conjunction).
C. Presuppositions: A generic masculine term in a broad context presupposes a mixed group, making the masculine form exclusionary by default. That automatically triggers the transformation for the mixed-group.
D. DEFAULT: If NO cues indicate a mixed or specific female group, default to the least gender-specific form appropriate for the genre (Osoba-form or Neutral form) for unspecified persons.
E. COST: Speakers aim for maximum contextual effects with minimum effort. Inclusive language (like Osoba-form) requires more effort but achieves a desired social effect.
F. MARKEDNESS: Your goal is to unmark the female form. When a clear female cue is present, the transformation must always go to the feminized noun.

Do not change anything else. Do not add any new text beyond inclusive transformations.
"""

"""— FEW-SHOT EXAMPLES —

Case A-1 (Male Cue)
Input Text: Pan Kowalski jest lekarzem.
Target Text: Pan Kowalski jest lekarzem.
Rule Reinforced: No Correction. Masculine cue present; generic masculine is pragmatically correct.

Case A-2 (Female Cue)
Input Text: Pani Anna jest lekarzem.
Target Text: Pani Anna jest lekarką.
Rule Reinforced: Correction Required. Female cue overrides the generic masculine term.

Case A-3 (Ambiguous Cue)
Input Text: Ktoś został zwycięzcą.
Target Text: Ktoś został osobą zwycięską.
Rule Reinforced: Rule C Default. No explicit gender cue; use neutral form when allowed by genre.

Case B-1 (Mixed Group, Explicit Role)
Input Text: Studenci napisali raport.
Target Text: Studenci/studentki napisali/napisały raport.
Rule Reinforced: Dual-inclusive form. Mixed group indicated by role; generic masculine must be transformed.

Case B-2 (Mixed Group, Collective Numeral)
Input Text: Pięciu kandydatów otrzymało nagrody.
Target Text: Pięcioro kandydat*ów/ek otrzymało nagrody.
Rule Reinforced: Collective numerals + dual-inclusive. Indicates mixed-gender group; enforce grammatical agreement.

Case B-3 (Profession, Female Present)
Input Text: Dyrektor Kowalska podpisała dokument.
Target Text: Dyrektorka Kowalska podpisała dokument.
Rule Reinforced: Feminine form override. Profession expressed in masculine must match female subject.

Case C-1 (Generic Masculine, No Cue, Genre Allows Neutral)
Input Text: Kierownicy spotkali się na konferencji.
Target Text: Osoby kierujące spotkały się na konferencji.
Rule Reinforced: Neutral/default form. No gender cues; default to least gender-specific form when appropriate.

Case C-2 (Coreference with Pronoun)
Input Text: Student dostał nagrodę. Jego praca była doskonała.
Target Text: Student*ka/student dostał/dostała nagrodę. Jego/jej praca była doskonała.
Rule Reinforced: Coreference + agreement. Pronouns and verbs must match the inclusive expression. """