Less is More—Achieving SOTA at PolEval 2025 Task 2a: Gender-inclusive LLMs for Polish (Proofreading) with LoRA and Qwen3-8B

Adam Majczyk
Institute of Computer Science
Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warszawa, Poland
adam.majczyk@ipipan.waw.pl

Abstract

In this paper the winning solution of PolEval 2025 Task 2a is presented. The approach utilizes LoRA fine-tuning of the `Qwen3-8B` model. Multiple LoRA matrix ranks are explored. Versions with and without the system prompt in loss calculation are evaluated. New SOTA was established at $F1 = 0.6039$ beating the previously best model at $F1 = 0.5985$. After the task's conclusion the solution was improved upon and $F1 = 0.6283 \pm 0.0056$ was achieved.

The code is available at: https://github.com/amajczyk/2025_PolEval_Task2a_Proofreading.

1 Introduction

In the recent years the interest in Large Language Models (LLMs) has steadily increased worldwide, the adoption of which reaching up to 14% in Eastern Europe in 2024 (Liang et al., 2025), likely to increase further by the end of 2025.

The Polish language is part of a group of languages, which encodes gender in parts of speech, most notably, nouns. Linguist differentiate 5 genders in Polish: 3 masculine, 1 feminine and 1 neuter (both in singular and plural forms) (Przepiórkowski et al., 2012).

A particular challenge, specifically in the professional and formal contexts is the "generic masculine" – one can use a masculine form to refer to groups of people of mixed genders, unknown groups or, in rare cases, groups of females. While grammatically correct, research indicates that such forms are not neutral cognitively – they evoke male mental connotations. It leads to the detriment and devaluation of female applicants in various professional settings (Formanowicz et al., 2013; Formanowicz and Sczesny, 2016).

Commonly available LLMs, both closed and open source, are trained on vast corpora of texts gathered from the Internet. Hence, they inherit the bias and may in turn amplify the pre-existing masculine default. To address and mitigate this issue, the **PolEval 2025 Task 2a: Gender-inclusive LLMs for Polish (Proofreading)** asks the participants to create systems capable of rewriting standard Polish texts into appropriate gender-inclusive versions (Wróblewska, 2025).

In this text the winning solution for the **PolEval 2025 Task 2a: Gender-inclusive LLMs for Polish (Proofreading)** is presented. It is demonstrated that using a smaller model, then the previously available State-of-the-Art (SOTA) (Wróblewska and Żuk, 2025), together with less demanding fine-tuning approaches can yield comparable or better results.

2 Task Description

In this section the task is described and the dataset is presented.

2.1 Task goal – idea

The goal of PolEval 2025 Task 2a is to develop a model/approach for transforming sequences that contain masculine-biased forms (e.g. nouns, verbs, etc.) into sequences that are **gender-inclusive**. For example the Polish sentence: *"Nauczyciel powinien przygotować się do lekcji."* (in English: *"The teacher should prepare for the lesson"*) contains two masculine forms: *"nauczyciel"* (*"teacher"*; feminine: *"nauczycielka"*) and *"powinien"* (*"should"*; feminine: *"powinna"*). For a teacher can be a female or a male, this sentence could be transformed into a gender-inclusive form using one of the schemas provided by the authors of the task (see the Task formulation for more details). One of the proposed forms uses asterisks (*"*"*) to separate them stem of the word from the gendered suffixes (Wróblewska et al., 2025). The transformed sentence using this schema would be:

Proceedings of the PolEval 2025 Workshop, pages 48–53

*"Nauczyciel*ka powin*ien/na przygotować się do lekcji."*.

However, not all such instances of gendered forms should be transformed. For example, if the sentence would refer to the male volleyball team and their achievements (e.g. *"Polscy siatkarze wygrali Mistrzostwa Świata."*, in English: *"The male Polish volleyball team has won the World Championship."*) a transformation of the sentence into a gender-inclusive form would change the original meaning. The authors of the task provide a comprehensive system prompt that defines the rules, when and how (which form to use) to transform such sequences.

It is therefore the goal, to find such gendered instances, decide whether to transform them, based on the wider context of the sequence and, if needed, appropriately apply the correct gender-inclusive schema.

2.2 Task specification

The task belongs to the category of problems named text-rewriting, which may be considered part of the Controlled Text Generation (CTG) domain. Given a `source` sequence written in standard, non-gender-inclusive, Polish, a `system prompt` and a `task prompt` the model should generate a `target` – the `source` rewritten using gender-inclusive forms. The approach shall preserve the original semantic meaning and grammatical structure as close as possible and only transform the parts, where deemed necessary.

The main challenge is applying the gender-inclusive forms only, where appropriate. That is, not all masculine forms require transformation. For example, in the case of a specific male individual or group of males (e.g., *"Polscy siatkarze"* referring to the men's national team) no change should be applied. Therefore, the approach must distinguish between generic and specific references.

Evaluation and Normalization

The core metric used for the task is the **F1-score**. Due top the fact, that various gender-inclusive schemas may be correct for any given instance (e.g. *"posłowie/posłanki"* vs. *"poseł*owie/anki"*), the authors provide a normalization pipeline. It expands all shortened forms (slashes, asterisks, underscores) into the full masculine and feminine versions. Then stopwords are removed. Finally, the set of normalized tokens is compared again the ground truth to calculate the appropriate components of the **F1-score**. Secondary metrics such as `BLEU`, `chrF`, and `chrF++` are calculated by the authors to assess general text quality. They are, however, omitted in this paper.

2.3 Dataset

The data provided for the task, based on the IPIS-proofreading dataset (Wróblewska and Żuk, 2025), contains 23,532 training examples, 2,732 validation examples, 2,639 testA (participants could upload predictions on this sample before the final submission date and receive appropriate metrics) and 2,639 testB (the actual submission sample) examples for Polish gender-inclusive language transformation.

Each example consists of a `source` - the sequence to be transformed and a `task prompt` - a short instruction for the model provided by the authors. The training and dev samples also contain a `target` - the expected output of the model given the `source` and `task prompt`.

An obvious observation one can make (see Table 1), it the fact that the texts to be transformed (`source`) are vastly shorter (average length of 466.37 characters) than the system prompt (3167 characters). The total model input is 3726.83 characters long on average. This makes the inputs roughly 7.74 times longer than the expected outputs (on average 481.77 characters). This imbalance is the motivation behind calculating training/-validation loss on output tokens only. For more details regarding the training please refer to Section 3.3.

3 System description

In this section the devised solution is presented. The model selection, training, inference and hardware are described.

3.1 Core methodology

The core idea behind the solution is fine-tuning an LLM using a sequence-to-sequence/instruct approach. The system prompt provided by the Task's authors was used without any modifications. It is important to note, that only the Polish version of the system prompt was used (the authors provide both an English and Polish version). It defines the proper usage of gender-inclusive schemas and rules of their application. The main idea may be visualized as seen in Figure 1.

During training the final prompt follows a three-turn structure (see Listing 1). where

Table 1: Dataset statistics: character-level sequence lengths

Metric	Train	Val/Dev	Test A	Test B
Source text				
Mean	466.37	457.40	586.57	593.15
Std Dev	333.89	431.22	416.05	406.75
Median	378.00	376.00	430.00	437.00
Min	6.00	7.00	8.00	4.00
Max	3387.00	4797.00	2757.00	2757.00
95th %	1292.00	1178.35	1356.10	1329.10
99th %	1802.00	2457.94	1627.62	1610.10
Target text[2]				
Mean	481.77	474.50	—	614.18
Std Dev	346.76	452.46	—	425.82
Median	390.00	387.00	—	445.00
Min	6.00	7.00	—	4.00
Max	3446.00	5078.00	—	2902.00
95th %	1340.00	1244.45	—	1378.30
99th %	1862.69	2627.93	—	1669.62
Task prompt				
Mean	93.45	93.80	93.87	93.69
Std Dev	24.90	24.94	25.08	24.78
Median	94.00	96.00	94.00	94.00
Min	37.00	37.00	37.00	37.00
Max	163.00	163.00	163.00	163.00
95th %	131.00	131.45	133.00	131.00
99th %	153.00	153.00	162.00	153.00
Total input[1]				
Mean	3726.83	3718.20	3847.45	3853.84
Std Dev	334.80	431.83	415.86	406.82
Median	3639.00	3636.00	3695.00	3700.00
Min	3222.00	3224.00	3220.00	3231.00
Max	6642.00	8057.00	6001.00	6021.00
95th %	4549.45	4433.00	4608.00	4586.60
99th %	5064.00	5706.59	4883.96	4869.24

[1] System (3167 char.) + Task prompt + Source
[2] Blank fields denote the fact that test sets were shared with no target sequences available to the participants at the time of the competition. Test set B was shared with the targets after the announcement of the results.

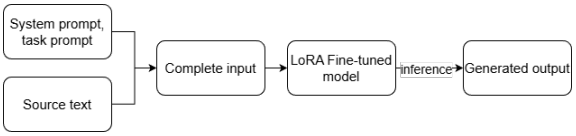

Figure 1: Diagram of the propesed solution

```
messages = [
    {"role": "system", "content": SYSTEM_PROMPT},
    {"role": "user", "content": f"{prompt} {source}"},
    {"role": "assistant", "content": target}
]
```

Listing 1: Training prompt schema

`SYSTEM_PROMPT` contains the system prompt, `prompt` is the task instruction provided in the dataset (e.g., *"Jeśli w tekście pojawiają się treści wykluczające lub krzywdzące, przeredaguj go na wersję inkluzywną. Tekst:"* which translates to: *"If the text contains exclusionary or harmful content, rephrase it into an inclusive version. Text:"*), `source` is the input text in standard (non-gender-inclusive) Polish, and `target` is the expected gender-inclusive output. This structure is then formatted in accordance with the Qwen3-Instruct chat template.

3.2 Model and framework selection

The presented solution is based on the `Qwen3-8B` model (Yang et al., 2025). Despite being a multilingual model `Qwen3-8B` was chosen due to its recent publication (April 2025) and generally high position in LLM rankings among its parameter size class (artificialanalysis.ai, 2025). The specific version (i.e. unsloth/Qwen3-8B-unsloth-bnb-4bit) uses 4-bit quantization to reduce the GPU RAM requirements to meet the limited specification of the used hardware (Nvidia RTX 4080 Desktop and Nvidia RTX 4090 Laptop GPUs were used, both being 16GB video cards).

For the same reason LoRA fine-tuning was utilized (Hu et al., 2021). The specific optimization framework, i.e. `Unsloth`, was chosen for its vast optimizations for low RAM GPU fine-tuning and inference (Han and Han, 2025).

Due to time constrains only this model with a single specific parameter setting was tested for the competition. As the authors allowed a post-competition expansion upon the submitted model, in the text, a study with more training parameters is conducted (see Section 4.2).

3.3 Training and optimization details

The training objective, contrary to the previously achieved SOTA (Wróblewska and Żuk, 2025), is calculated only on the output tokens, rather than also memorizing the instructions. The justification of the approach is to prevent the model of overfitting to the prompts (both system and task specific), given the relatively small sample size and the fact, that the lengths of the texts undergoing transformation are on average noticeably shorter, than the length of the system prompt (see Table 1).

The detailed configuration of the submitted solution's training parameters is summarized in Table 2.

Table 2: LoRA Adapter and Training Configuration

Parameter	Value/Setting
LoRA Adapter Configuration	
Rank (r)	64
Applied Layers	QKV, Output, Gate, Up, Down
Optimization and Training	
Epochs	2
Optimizer	AdamW-8bit
Learning Rate (LR)	2×10^{-4}
LR Scheduler	Cosine
Warmup Steps	10
Weight Decay	0.01
Batching and Sequence	
Effective Batch Size	2 (Accumulation steps: 2)
Max Sequence Length	4096 tokens
Checkpointing	
Save Frequency	Every 500 steps
Selection Criterion	Lowest validation loss

Table 3: Inference Hyperparameters and Configuration for Gender-Inclusive Text Generation

Parameter	Value/Setting
Model Configuration	
Base Model	Qwen3-8B-Instruct
LoRA Checkpoint	checkpoint-23000 (lowest val. loss)
Quantization	4-bit (bitsandbytes)
Max Seq. Length	4096 tokens
Generation Parameters	
Sampling Strategy	Nucleus sampling (top-p)
Temperature (T)	0.3
Top-p (p)	0.9
Top-k (k)	50
Max New Tokens	4096
Processing	
Batch Size	1 (sequential)
Checkpoint Interval	Every 25 examples
Post-processing	Remove chat template artifacts
Output Format	JSONL/TSV (IPIS format)

3.4 Inference details

The fine-tuned model transforms the sequences for previously unseen texts written in standard Polish. The inputs are formatted identically to the training structure (see Listing 1).

Inference is performed using a low, however non-zero, temperature of $T = 0.3$. This was chosen to ensure that the model follows the gender-inclusive rules closely, while still maintaining some *creativity*. The maximum sequence length was set to 4096 to ensure that even long texts generate correctly. For more details about inference parameters please refer to Table 3.

After inference a simple post-processing step is performed to remove any special tokens and trim the text appropriately to cut off the prompts and any other template artifacts. The predictions are then saved to the required `.tsv` format and the provided normalization script is applied.

4 Results

In this section the results of the submitted model and the expanded analysis are presented.

4.1 Submitted model

Having applied the fine-tuned `Qwen3-8B` model with LoRA matrixes of rank 64, the generated outputs yielded F1-Score of 60.39. The result is a slight improvement over the previously available SOTA F1-Score of 59.85), that used a larger 11B-parameter Bielik model (Ociepa et al., 2025). One can therefore draw a conclusion that a smaller model can yield comparable or better results. For more detailed metrics, see Table 4.

Table 4: Performance comparison of existing solutions and the submitted model; calculated on testB

Model	Prec	Rec	F1
PLLuM-12B (Baseline)	2.56	6.28	3.64
Bielik-11B-tuned (SOTA)	63.93	56.26	59.85
Qwen3-8B-LoRA (Ours)	**64.50**	**56.77**	**60.39**

4.2 Post competition improvements

The solution was improved upon following the conclusion of the competition. The following sections focus on the influence of the adapter matrix ranks and approaches to calculating the loss during training. The other parameters in training and inference remain unchanged, as in the submitted model (see Table 2).

Influence of LoRA rank

Since the inference uses non-zero temperature ($T = 0.3$), inference for the analyses was performed 3 times for each rank (except for rank 64, for which 2 additional runs, except the submission run were performed, to equal 3 in total). For each additional setting explored, inference was performed using the best performing checkpoint (based on validation loss).

Based on the results presented in Figure 2 one can draw a conclusion, that larger (32 and up) LoRA ranks are generally better. There is, however, a diminishing returns effect starting from the rank of 32, i.e. ranks 64 and 128 do not provide

additional benefits and may introduce over-fitting. We observe, that rank 32 appears to be the sweet spot for this particular task (average F1-Score of 0.6283 ± 0.0056), outperforming the rank 64 chosen for the submitted model (average F1-Score of 0.6131 ± 0.0094).

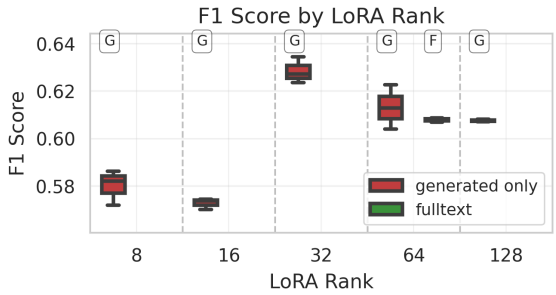

Figure 2: F1 Scores of models by LoRA rank and loss calculation type; **G** denotes loss calculated on the output tokens only, **F** – on both the full input and output tokens; calculated on testB

As a non-zero temperature ($T = 0.3$) was used during inference determinism was also explored. In Table 5 the results of the analysis are presented. On average 72.72%-74.16% of the generated texts were identical given a set LoRA rank across 3 inference runs. Interestingly, larger ranks did not yield smaller differences across the texts, with rank 64 exhibiting an average edit distance of 35.46 – highest among the chosen LoRA rank set.

Table 5: Inference consistency across LoRA ranks (3 passes per rank); calculated on testB

LoRA Rank	Determinism (%)[1]	Mean Edit Distance[2]
r=8	73.40	34.40
r=16	72.72	19.98
r=32	73.78	19.72
r=64[3]	74.16	35.46
r=128	73.32	18.66

[1] Percentage of examples with identical outputs across 3 inference passes (temperature=0.3)
[2] Levenshtein distance (characters) for examples with differences
[3] Includes the submitted inference run

Influence of approach to loss calculation

As stated priorly (see Sections 2.3 and 3.3) the proposed solution calculates loss only using the generated tokens. To verify if that approach yields better results, a model was trained using the same parameters as the submitted solution (see Tables 2 and 3), with the only difference being the loss calculation. For this test, the training and validation loss calculations take all tokens, i.e. input and output tokens, into consideration. The performance of this model is slightly worse on average ($F1 = 0.6078 \pm 0.0009$), than the one trained on output tokens only ($F1 = 0.6131 \pm 0.0094$). For more details see Table 7.

Table 6: Average metrics by LoRA rank (mean ± std dev) across 3 runs; calculated on testB

Rank/Type	Acc.	Prec.	Rec.	F1
8	0.9740 (±0.0007)	0.6424 (±0.0136)	0.5287 (±0.0042)	0.5800 (±0.0074)
16	0.9740 (±0.0001)	0.6460 (±0.0003)	0.5143 (±0.0038)	0.5727 (±0.0024)
32	**0.9764** (±0.0004)	**0.6670** (±0.0050)	**0.5940** (±0.0061)	**0.6283** (±0.0056)
64[1]	0.9753 (±0.0008)	0.6587 (±0.0121)	0.5735 (±0.0083)	0.6131 (±0.0094)
128	0.9754 (±0.0001)	0.6540 (±0.0019)	0.5671 (±0.0009)	0.6074 (±0.0005)

[1] Includes the submitted inference run

Table 7: Average metrics by loss calculation type (Rank 64)

Model Type	Acc.	Prec.	Rec.	F1
Full-text	0.9751 (±0.0001)	**0.6596** (±0.0034)	0.5635 (±0.0020)	0.6078 (±0.0009)
Generated Only[1]	**0.9753** (±0.0008)	0.6587 (±0.0121)	**0.5735** (±0.0083)	**0.6131** (±0.0094)

[1] Includes the submitted inference run

An independent (no repeatable seeds were set during inference) samples t-test has been performed to verify, if the training using generated tokens only for loss calculation yields significantly better values of $F1$, compared to calculating loss on all tokens. At p-value of 0.191, the difference is deemed statistically insignificant. A Mann-Whitney U-test yields the same conclusion (p-value of 0.35). However, it has to be emphasised, that given the low sample sizes of 3, the power of such tests is low.

5 Limitations

Due to many factors, including underpowered hardware and lack of time, the proposed solution exhibits several limitations. Most notably, only a small subset of parameters is explored for a single model. Additionally, as inference is performed

with a non-zero temperature ($T = 0.3$) and no seed applied during generation, the results are not repeatable. Therefore, for the post-competition improvements, 3 inference runs are performed for each parameter configuration. This is, however, also deemed a limiting factor, as statistical power is not high enough to reach a satisfactory conclusion.

6 Future works

In the future, the solution may be improved upon by exploring larger models or a broader training and inference parameter set. Aside from that, semi- or fully automated prompt engineering approaches could potentially yield better results. It would also be advisable, to verify what results would the proposed fine-tuning approach yield with the Bielik-11B model used in the previously available SOTA.

7 Conclusions

We conclude, that using a smaller model (8B vs 11B) can yield comparable, if not better results, when paired with an efficient fine-tuning approach. For the gender-inclusive proofreading task (**PolEval 2025 Task 2a: Gender-inclusive LLMs for Polish (Proofreading)**) a new SOTA was established during the competition, achieving an F1-Score of 0.6039. It was improved upon post-competition and an F1-Score of 0.6283 ± 0.0056 was reached. No statistically significant conclusion was drawn, whether calculating training loss on generated tokens only is better than using both the input and generated tokens. Further research in this domain is required.

References

artificialanalysis.ai. 2025. LLM Leaderboard - Comparison of over 100 AI models from OpenAI, Google, DeepSeek & others.

Magdalena Formanowicz, Sylwia Bedynska, Aleksandra Cisłak, Friederike Braun, and Sabine Sczesny. 2013. Side effects of gender-fair language: How feminine job titles influence the evaluation of female applicants. *European Journal of Social Psychology*, 43(1):62–71. Place: US Publisher: John Wiley & Sons.

Magdalena Formanowicz and Sabine Sczesny. 2016. Gender-Fair Language and Professional Self-Reference: The Case of Female Psychologists in Polish. *Journal of Mixed Methods Research*, 10(1):64–81. Publisher: SAGE Publications.

Daniel Han and Michael Han. 2025. Unsloth Docs | Unsloth Documentation.

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. *arXiv preprint*. ArXiv:2106.09685 [cs].

Weixin Liang, Yaohui Zhang, Mihai Codreanu, Jiayu Wang, Hancheng Cao, and James Zou. 2025. The Widespread Adoption of Large Language Model-Assisted Writing Across Society. *arXiv preprint*. ArXiv:2502.09747 [cs] version: 2.

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

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego: praca zbiorowa*. Wydawnictwo Naukowe PWN, Warszawa.

Alina Wróblewska. 2025. PolEval 2025 Task 2: Gender-inclusive LLMs for Polish. 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. *arXiv preprint*. ArXiv:2508.18466 [cs].

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. *arXiv preprint*. ArXiv:2505.09388 [cs].