POLEVAL 2025

PolEval 2025 Workshop

Proceedings of the PolEval 2025 Workshop

November 21, 2025

The POLEVAL organizers gratefully acknowledge the support from the following sponsors.

Organisational support

PAN

sages

UAM

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN None

Organizing Committee

General Chairs

Łukasz Kobyliński, Institute of Computer Science PAS, Poland
Maciej Ogrodniczuk, Institute of Computer Science PAS, Poland

Publication Chair

Alina Wróblewska, Institute of Computer Science PAS, Poland

Evaluation Platform Administrator

Ryszard Staruch, Adam Mickiewicz University, Poland

Publicity and Social Media Chair

Beata Milewicz, Institute of Computer Science PAS, Poland

Task Organisers

Iwona Christop, Adam Mickiewicz University, Poland
Maciej Czajka, Adam Mickiewicz University, Poland
Łukasz Kobyliński, Institute of Computer Science PAS, Poland
Piotr Przybyła, Universitat Pompeu Fabra, Spain
Jakub Strebeyko, University of Warsaw, Poland
Alina Wróblewska, Institute of Computer Science PAS, Poland
Aleksandra Zwierzchowska, Institute of Computer Science PAS, Poland

Program Committee

Program Chairs

Łukasz Kobyliński, Institute of Computer Science PAS, Poland
Maciej Ogrodniczuk, Institute of Computer Science PAS, Poland
Alina Wróblewska, Institute of Computer Science PAS, Poland

Program Committee

Iwona Christop, Adam Mickiewicz University, Poland
Maciej Czajka, Adam Mickiewicz University, Poland
Łukasz Kobyliński, Institute of Computer Science PAS, Poland
Piotr Przybyła, Universitat Pompeu Fabra, Spain
Alina Wróblewska, Institute of Computer Science PAS, Poland
Aleksandra Zwierzchowska, Institute of Computer Science PAS, Poland

Table of Contents

PolEval 2025

Łukasz Kobyliński[1], Ryszard Staruch[2], Alina Wróblewska[1], and Maciej Ogrodniczuk[1]

[1]Instutute of Computer Science PAS, Warsaw
[2]Adam Mickiewicz University, Poznań
l.kobylinski@ipipan.waw.pl, alina@ipipan.waw.pl
ryszard.staruch@amu.edu.pl, m.ogrodniczuk@ipipan.waw.pl

Abstract

PolEval is an annual shared-task evaluation campaign dedicated to advancing natural language processing for the Polish language. This paper presents an overview of PolEval 2025, the eighth edition of the campaign, which included three completed tasks covering machine-generated text detection, gender-inclusive language generation, and speech emotion recognition. The evaluation was conducted using standardized datasets and metrics via the AmuEval platform. PolEval 2025 attracted 15 teams and over 100 submissions, demonstrating continued engagement from the Polish NLP community. We describe the organization of the campaign, the evaluation setup, and the role of PolEval in fostering reproducible research and community-driven benchmarking.

1 Introduction

PolEval is an annual evaluation campaign, inspired by SemEval, dedicated to benchmarking and advancing natural language processing (NLP) tools for the Polish language. It provides a shared-task framework in which participating teams submit systems to solve a variety of pre-defined tasks using standardised datasets, with their performance evaluated under official protocols. Over the years PolEval has covered multiple NLP domains – from machine translation and named-entity linking to speech recognition, sentiment analysis, and more.

The first edition of PolEval was held in 2017 at the 8th Language & Technology Conference (LTC) in Poznań.[1] It did not have separate proceedings but 10 papers resulting from 2 tasks on part-of-speech tagging (Kobyliński and Ogrodniczuk, 2017) and sentiment analysis (Wawer and Ogrodniczuk, 2017) were included in the LTC proceedings (Vetulani and Paroubek, 2017). In the subsequent years, PolEval results were presented at two other conferences such as AI & NLP Day (Ogrodniczuk and Kobyliński, 2018, 2019, 2020, 2021) and FedCSIS (Kobyliński et al., 2023), and at the Natural Language Processing seminar[2] (Ogrodniczuk and Kobyliński, 2024) and its proceedings were usually published in a separate series.

2 PolEval 2025

In the 8th edition of the campaign – PolEval 2025 – four tasks were selected from the proposals submitted between March and May 2025 and subsequently announced publicly (see Section 2.2). Training data were released in August 2025, followed by the release of the test data in September 2025. An information and dissemination campaign was conducted on social media and at major AI-related events, including FedCSIS 2025 Conference in Cracow and AI & NLP Day jointly organized with Confitura in Warsaw. The evaluation was carried out using the PolEval benchmarking system (see Section 2.3), and the results were announced through the integrated leaderboards on November 17, 2025. The award ceremony and presentation of the winning solution took place at the Data Science Summit Conference 2025.

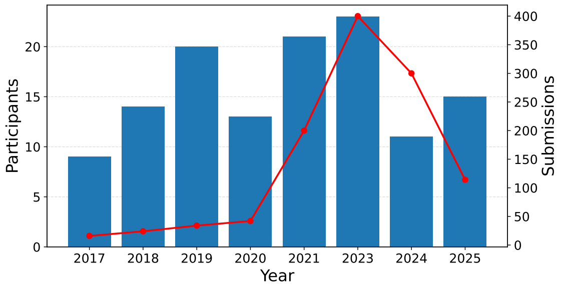

Figure 1: Yearly Numbers of Participants (left axis) and Submissions (right axis).

The 2025 edition of PolEval received over 100

[1] https://ltc.amu.edu.pl/

[2] https://zil.ipipan.waw.pl/seminar

Proceedings of the PolEval 2025 Workshop, pages 1–4
November 21, 2025 ©2025 Association for Computational Linguistics

submissions from 15 participating teams, which is comparable to previous years (see Figure 1).

2.1 Program Committee

General Chairs

Łukasz Kobyliński, ICS PAS,[3] Sages
Maciej Ogrodniczuk, ICS PAS

Evaluation Platform Administrator

Ryszard Staruch, AMU[4]

Publication Chair

Alina Wróblewska, ICS PAS

Publicity and Social Media Chair

Beata Milewicz, ICS PAS

Task Organisers

Iwona Christop, AMU
Maciej Czajka, AMU
Łukasz Kobyliński, ISC PAS
Piotr Przybyła, Universitat Pompeu Fabra, Barcelona
Jakub Strebeyko, University of Warsaw
Alina Wróblewska, ICS PAS
Aleksandra Zwierzchowska, ICS PAS

2.2 Tasks

Task 1: Śmigiel: Spotting Machine-Generated Text from Language Models for Polish
Śmigiel (Przybyła et al., 2025) is a shared evaluation task designed to benchmark the detection of AI-generated vs. human-written text in Polish. The task is formulated as a binary classification problem, and participants can choose among three subtasks: *unsupervised*, *constrained*, and *open*. The accompanying dataset (Strebeyko et al., 2025) includes human-written texts that are drawn from various sources – all distributed under open licenses, and represent diverse domains (reviews, literature, social media posts, Wikipedia articles, parliamentary transcripts, and news). It also includes corresponding machine-generated texts produced using a variety of open-source LLMs.

As LLMs become more widely used in Polish, there is growing need for reliable methods to distinguish between human- vs. LLM-generated texts. Śmigiel seeks to support the development of robust machine-generated text detectors for Polish, with potential applications in journalism, academia, media verification and related fields.

Task 2: Gender-inclusive LLMs for Polish
PolEval 2025 Task 2 (Wróblewska, 2025) aims to support development of LLMs capable of generating grammatically correct and gender-aware Polish across a variety of contexts. The submitted LLMs are evaluated on two subtasks: *gender-inclusive proofreading*, which involves transforming Polish texts written in standard (typically masculine-centric) language into gender-inclusive counterparts, and *gender-sensitive translation*, which consists in translating between Polish and English in both directions while ensuring that the Polish output respects gender-inclusive language norms. Participants are provided with the official dataset, the Inclusive Polish Instruction Set (IPIS, Wróblewska and Żuk, 2025).

By promoting the generation of gender-inclusive Polish, this task seeks to mitigate masculine bias present both in conventional usage and in LLMs trained on standard corpora. In doing so, it contributes to broader social goals related of gender equality, e.g., elimination of sexism from language.

Task 3: Polish Language Document Layout Detection The task has been cancelled due to lack of participants.

Task 4: Polish Speech Emotion Recognition Challenge Task 4 (Christop and Czajka, 2025b) focuses on speech emotion recognition (SER) for the Polish language, where participants are required to develop a system that classifies speech recordings into one of six emotion categories: *anger*, *fear*, *happiness*, *sadness*, *surprise*, and *neutral*. The training data consists of multilingual emotional speech samples from seven non-Polish languages sourced from CAMEO (Christop and Czajka, 2025a). A separate Polish validation set is provided solely for model selection. The final evaluation is conducted on previously unseen Polish speech data. System performance is assessed primarily using the macro-averaged F1 score, with accuracy reported as a secondary metric.

The task addresses the challenge of cross-lingual SER in a low-resource setting, as Polish lacks large, publicly available emotional speech corpora. By restricting access to Polish training data, the shared task promotes the development of language-independent acoustic representations that generalise across languages. This setting reflects realistic deployment conditions and supports research on robust, transferable SER systems, with potential applications in human-computer interaction and

[3] Institute of Compute Science, Polish Academy of Sciences, Warsaw

[4] Adam Mickiewicz University, Poznań

assistive technologies for Polish speakers.

2.3 Evaluation system

PolEval 2025 shared tasks were hosted on a dedicated instance of the AmuEval evaluation platform (Jassem et al., 2024). The platform was chosen for its simplicity, allowing participants to concentrate on solving the tasks rather than dealing with the technical complexities often associated with submitting predictions on evaluation platforms.

Each challenge included one primary evaluation metric that determined the final leaderboard standings, along with up to two additional metrics to provide participants with further insight into the performance of their solutions. Some of these metrics required custom implementation, as they were not available in the platform's default metric set. This integration process was smooth and problem-free, demonstrating that the platform can be effectively customized to the needs of dedicated instances.

Datasets for each task were made available via public repositories on GitHub. The platform itself was used exclusively for the final evaluation of system outputs on held-out test sets. Each task featured two separate test subsets: Test-A and Test-B, organized as distinct challenges on the platform. Metric scores for Test-A were visible to participants during the competition, while scores for Test-B remained hidden until after the submission deadline. This setup was designed to prevent overfitting and ensure a fair comparison of final systems.

The positive experience with AmuEval during PolEval 2025 demonstrates its potential as a reliable foundation for future shared tasks and domain-specific evaluation campaigns.

3 Conclusion and future plans

PolEval has, over multiple editions, established itself as a stable and widely recognized evaluation campaign for Polish natural language processing. The 2025 edition confirmed the continued relevance of shared-task benchmarking for the Polish NLP community, while also highlighting the rapidly evolving landscape of artificial intelligence, in particular the growing dominance of large language models (LLMs).

Looking ahead, one of the main challenges for future editions of PolEval will be to keep pace with these developments while preserving the campaign's core focus on the evaluation of Polish-language competence. As LLM-based systems increasingly achieve strong general performance across languages and tasks, there is a growing need for carefully designed, language-specific evaluations that test phenomena characteristic of Polish. Future PolEval tasks will therefore aim to balance openness to modern, large-scale models with evaluation settings that meaningfully differentiate systems based on their handling of Polish-specific linguistic and cultural properties.

Another important direction for the campaign is strengthening its educational role. We plan to further synchronize PolEval tasks and timelines with NLP- and AI-related university courses conducted in Poland, enabling students to participate as part of their coursework. Such integration has the potential to lower the entry barrier for new participants, foster practical skills in reproducible evaluation, and contribute to the training of the next generation of researchers and practitioners in Polish NLP.

Finally, increasing the visibility and impact of PolEval remains a key objective. Continued cooperation with national and international scientific conferences, workshops, and community events will help broaden the audience for the campaign, attract new participants, and facilitate the dissemination of results. By maintaining strong ties with both the research community and educational initiatives, PolEval aims to remain a relevant and adaptable platform for benchmarking Polish NLP in an era of rapidly advancing AI technologies.

Acknowledgments

We would like to express our sincere gratitude to the organisers of the PolEval 2025 evaluation campaign for their dedication and continuous efforts in supporting the Polish natural language processing community, particularly to the Institute of Computer Science, Polish Academy of Sciences (ICS PAS), for its leading organisational role. We also gratefully acknowledge the generous support of the campaign sponsors, especially Sages, whose contribution made the organisation of the event possible. We extend our sincere thanks to all participating teams and researchers for their enthusiasm, high-quality submissions, and valuable scientific contributions, which are the foundation of the success of the PolEval 2025 campaign. Finally, we thank the organisers of the associated scientific and community events – Data Science Summit, Confitura, the FedCSIS Conference, and the University of Warsaw – for providing the venues and infras-

tructure for dissemination, presentations, and the award ceremony.

References

Iwona Christop and Maciej Czajka. 2025a. Cameo: Collection of multilingual emotional speech corpora. *Preprint*, arXiv:2505.11051.

Iwona Christop and Maciej Czajka. 2025b. Polish Speech Emotion Recognition Challenge. In *Proceedings of the PolEval 2025 Workshop*.

Krzysztof Jassem, Andrzej Gajda, Mateusz Tylka, Ryszard Staruch, Grzegorz Lipiecki, and Szymon Bartanowicz. 2024. Amueval: A user-friendly educational platform for machine learning challenges. In *Proceedings of the Fortieth Information Systems Education Conference (ISECON 2024)*, pages 107–114, Virtual Conference, Chicago, IL, USA. Foundation for Information Technology Education. Virtual conference, October 19, 2024.

Łukasz Kobyliński and Maciej Ogrodniczuk. 2017. Results of the PolEval 2017 competition: Part-of-speech tagging shared task. In (Vetulani and Paroubek, 2017), pages 362–366.

Łukasz Kobyliński, Maciej Ogrodniczuk, Piotr Rybak, Piotr Przybyła, Piotr Pęzik, Agnieszka Mikołajczyk, Wojciech Janowski, Michał Marcińczuk, and Aleksander Smywiński-Pohl. 2023. PolEval 2022/23 challenge tasks and results. In *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, volume 35 of *Annals of Computer Science and Information Systems*, pages 1237–1244.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2018. *Proceedings of the PolEval 2018 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2019. *Proceedings of the PolEval 2019 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2020. *Proceedings of the PolEval 2020 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2021. *Proceedings of the PolEval 2021 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2024. *Proceedings of the PolEval 2024 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Piotr Przybyła, Jakub Strebeyko, and Alina Wróblewska. 2025. PolEval 2025 Task 1 Śmigiel: Spotting Machine-Generated Text from LLMs for Polish. In *Proceedings of the PolEval 2025 Workshop*.

Jakub Strebeyko, Alina Wróblewska, and Piotr Przybyła. 2025. Śmigiel Dataset: Laying Foundations for Investigating Machine-Generated Text Detection in Polish. Unpublished.

Zygmunt Vetulani and Patrick Paroubek, editors. 2017. *Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu, Poznań, Poland.

Aleksander Wawer and Maciej Ogrodniczuk. 2017. Results of the PolEval 2017 competition: Sentiment Analysis shared task. In (Vetulani and Paroubek, 2017), pages 406–409.

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.

PolEval 2025 Task 1 Śmigiel: Spotting Machine-Generated Text from LLMs for Polish

Piotr Przybyła[1,3], Jakub Strebeyko[2], and Alina Wróblewska[3]

[1]Universitat Pompeu Fabra, Barcelona, Spain
[2]University of Warsaw, Warsaw, Poland
[3]Instutute of Computer Science PAS, Warsaw, Poland
piotr.przybyla@upf.edu, j.strebeyko@student.uw.edu.pl, alina@ipipan.waw.pl

Abstract

This paper introduces the first shared task on machine-generated text (MGT) detection for Polish, organised as part of the PolEval 2025 evaluation campaign. The task evaluates participating systems under three scenarios – unsupervised, constrained, and open – designed to reflect different levels of access to training data. In total, seven systems were submitted. The results indicate that MGT detection for Polish is feasible, with the best-performing constrained systems achieving over 90% accuracy on the main evaluation set. However, performance drops when models are tested on unseen domains or generator models, revealing substantial limitations in generalisation. In the most challenging settings, unsupervised approaches beat the supervised ones. This shared task establishes a new benchmark for MGT detection in Polish. The publicly released Śmigiel dataset is intended to support future research on robust and generalisable MGT detection.

1 Introduction

The rapid progress of large language models (LLMs) in recent years has enabled the generation of highly fluent and linguistically correct texts in numerous languages. Although these models demonstrate strong performance in natural language processing (NLP) and natural language generation (NLG) tasks, their reliance on human-authored data and their capacity to emulate human writing styles raise critical concerns regarding authenticity, authorship attribution, and the potential for misuse. Identifying whether a text was written by a human is critical in several contexts:

1. when the act of writing itself is being evaluated, e.g., in education (Frohock, 2025),
2. in preparing high-credibility documents in science (Májovský et al., 2023) or law (Frohock, 2025),
3. in high-risk domains where LLM errors, especially hallucination, may have serious consequences, e.g., in health-related publications (Milmo, 2023),
4. in malicious scenarios enabled by large-scale LLM-generated content, e.g., misinformation (Zhou et al., 2023), disinformation (Vykopal et al., 2024), or fraud (Gressel et al., 2024).

In response to this challenge, research has focused on the development of machine-generated text (MGT) detection systems – tools designed to distinguish between human-authored and AI-generated content (Crothers et al., 2023; Wu et al., 2025). Common authorship indicators include *n*-gram statistics (Gallé et al., 2021; Hamed and Wu, 2024), token probability–based measures such as perplexity (Gehrmann et al., 2019; Wu et al., 2023), embedding-space properties (Tulchinskii et al., 2023), and comparisons with LLM-rewritten variants of the same text (Zhu et al., 2023; Maslo and Gargova, 2025). Supervised methods typically rely on fine-tuned LLMs (Nguyen-Son et al., 2024) or engineered features, including stylistic, discourse-level, and probabilistic cues (Przybyła et al., 2023; Shah et al., 2023; Kim et al., 2024).

At the same time, the need for rigorous evaluation and benchmarking of MGT detection systems has led to several shared tasks. Prominent example include SemEval-2024 Task 8 (Wang et al., 2024) and its successor, the GenAI Content Detection Task 1 (Wang et al., 2025), which covered binary, multi-class, and boundary detection settings across multiple languages. Related challenges addressed cross-domain detection (Dugan et al., 2025), academic essay authenticity (Chowdhury et al., 2025), scientific paper detection (Chamezopoulos et al., 2024), and collaborative human–AI authorship (Bevendorff et al., 2025).

Beyond English, shared tasks have been organised for Dutch (Fivez et al., 2024), Russian

Proceedings of the PolEval 2025 Workshop, pages 5–15
November 21, 2025 ©2025 Association for Computational Linguistics

(Shamardina et al., 2022), and Spanish (Sarvazyan et al., 2023). The AuTexTification dataset (Sarvazyan et al., 2023) and its multilingual extension IberAuTexTification (Sarvazyan et al., 2024) are particularly relevant to our work, as they focus on mixed-authorship and multi-domain settings.

To date, no shared task or large-scale benchmark has been dedicated to MGT detection for Polish. To address this gap, we organised **Śmigiel**[1] – the first shared task dedicated to spotting machine-generated text in Polish (see Section 2), conducted as part of the PolEval 2025 evaluation campaign (Kobyliński et al., 2025). The task is based on a newly created dataset (Strebeyko et al., 2025), developed specifically for this initiative (Section 3). We also provide several strong baseline systems (Section 4), which serves as reference points for the participating submissions (Section 5). The results the Śmigiel shared task are presented in Section 6, followed by an in-depth analysis (Section 7) and a comprehensive discussion (Section 8).

2 Task description

The shared task, together with its corresponding dataset, is called Śmigiel[2] – an extended acronym for **S**potting **M**achine-**G**enerated Text from **LL**Ms for Polish. It was organised at the PolEval 2025[3] (Kobyliński et al., 2025) evaluation campaign.

2.1 Objective

The main objective of this shared task is to benchmark and enhance the state-of-the-art in detecting machine-generated texts in the Polish language across various domains and textual genres. Robust and reliable MGT detection systems will undoubtedly contribute to the broader goals of responsible AI development, supporting critical areas such as media verification, academic and journalistic integrity, and, potentially, digital forensics.

2.2 Procedure

The task is framed as a binary classification problem: distinguishing between human-authored and machine-generated texts. Participating systems are given a collection of text fragments, and must assign each either label 0 (human-written) or label 1 (LLM-generated). To foster the development of models capable of generalising across diverse writing styles, the test dataset includes samples from domains not represented in the training set.

2.3 Śmigiel subtasks

Participants submitted their systems under one of three distinct evaluation subtasks, reflecting different training conditions and levels of methodological constraint.

UNSUPERVISED This subtask is intended for classifiers developed without the use of labelled training data. Systems in this category should rely exclusively on unsupervised methods, heuristic approaches, or pre-trained models used without task-specific fine-tuning.

CONSTRAINED In this subtask, participants are allowed to train their classifiers solely on the Śmigiel dataset provided by the task organizers. The use of any additional external data, pre-trained resources, or synthetic data generation is prohibited, ensuring a fully controlled and comparable experimental setting.

OPEN This subtask imposes no restrictions on training resources. Participants may leverage external datasets, pre-trained models, web-crawled data, or data augmentation techniques.

Submissions within each subtask were evaluated independently and ranked separately, allowing for fair comparison among approaches developed under the same set of constraints.

2.4 Evaluation metric

In the Śmigiel task, `Accuracy` is used as the primary evaluation metric. Accuracy, defined as the proportion of correctly classified instances over the total number of instances, is widely used in text classification (Jurafsky and Martin, 2025). The datasets are balanced, with matching number of human-written and machine-generated fragments.

2.5 Task constraints

The shared task's rules, same for all contestants (except the subtask constraints), are:

1. Publicly available pretrained Polish and multilingual models may be used.
2. Participants may use publicly accessible Polish corpora, lexical resources, knowledge bases, and other structured data resources.

[1] https://github.com/poleval/2025-smigiel

[2] **śmigiel** [’ɕmigjɛl] m I, D. ∼gla gw. «płaski szczebel w drabinach wozu» [Eng. flat rung in a ladder wagon] (Saloni and Bańko, 2012)

[3] https://poleval.pl

3. Participants are expected to prepare a short article, describing their solution with enough details to allow replication of the research.
4. All external models and resources used must be listed in the submission, including bibliographic references or direct links.
5. The use of proprietary or non-public datasets, models, or services is strictly prohibited.
6. Each team is allowed a maximum of three submissions per subtask.

3 Śmigiel Dataset

The Śmigiel dataset (Strebeyko et al., 2025) was developed specifically for the PolEval 2025 shared task on detecting machine-generated text (MGT) in Polish. The dataset pairs human-written text (HWT) passages with machine-generated continuations produced by LLMs. The construction of the Śmigiel dataset involved the following main stages: data collection, generation of MGT counterparts, postprocessing, and final corpus composition. A detailed description of each stage is provided in (Strebeyko et al., 2025).

3.1 HWT data collection

The raw data were compiled from multiple open-access Polish datasets covering diverse textual domains, including literature, reviews, social media, Wikipedia, news, and parliamentary transcripts. The first four domains were used for both training and testing, while news and parliamentary transcripts were reserved for evaluation only.

3.2 MGT data generation

HWT passages served as both stand-alone samples and as prompts (prefixes) for generating MGT counterparts. MGT examples were produced using a variety of Polish-specific and multilingual LLMs of different sizes – ranging from small to large – and across several decoding strategies to enhance output diversity. The following generator models are applied to produce MGT texts:

- small: Bielik-7B-v0.1[4] (Ociepa et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), and Mistral-7B0-v0.3 (Jiang et al., 2023),
- medium: Bielik-11B-v2.3, PLLuM-12B (Kocoń et al., 2025), and Mistral-Nemo (Mistral AI team, 2024),
- large: Gemma-3-27B (Kamath et al., 2025), Llama-3.3-70B.

The resulting raw dataset comprised approximately 460,000 paired HWT–MGT examples.

3.3 Postprocessing

In the post-processing stage, the dataset underwent extensive filtering to remove repeated prefixes, metalinguistic commentary, non-Polish text, and other typical LLM-generated errors. About one-third of MGT and 1% of HWT fragments were discarded. The sampling procedure then ensured a strict balance between the HWT and MGT instances, as well as between the LLM size categories. Texts were length-normalised to control for bias arising from MGT verbosity, and aggregated across domains to form the final dataset.

3.4 Data split

The subsets in the shared task are based on the portions of the Śmigiel dataset (Strebeyko et al., 2025). Namely, the training set, provided for the participants, is equivalent to the $train$ portion. Regarding the test data, in the Śmigiel dataset three subsets are available: $train\_\alpha$ (using the same domains and models as $train$), $train\_\beta$ (including generations in the same domain, but from a model unseen in training data – `llama-lg`) and $train\_\gamma$ (covering domains unseen in training data – parliament and news). In the shared task, we mix them as:

- $\texttt{test}_A$, using 50% of $train\_\alpha$, 33% of $train\_\beta$ and 33% of $train\_\gamma$,
- $\texttt{test}_B$, using 50% of $train\_\alpha$, 67% of $train\_\beta$ and 67% of $train\_\gamma$.

The leaderboard for $\texttt{test}_A$ is a public one, allowing the participants to test their approaches. Results on $\texttt{test}_B$ (more challenging due to higher contribution of data from unseen configurations) are not revealed until the conclusion of the shared task, when they are used to decide the final ranking.

3.5 Final Śmigiel dataset

The final Śmigiel dataset contains 64,000 text samples – 32,000 HWT and 32,000 MGT – balanced across textual domains and model categories. Four domains (literature, reviews, social, wikipedia) are used for training and evaluation. Two domains (news and parliament) unseen in training are reserved exclusively for evaluation.

[4] We use the instruct versions of models available in the HuggingFace repository.

Subtask	System	Method	Model
unsup.	damian96	Binoculars (Śmigiel-calibrated threshold)	Bielik models
	kwrobel	perplexity difference-based approach	Gemma-3-27B & PLLuM-12B
const.	kondziu98	a classifier atop a decoder-only LLM	Qwen3-8B
	grzmot	NA	NA
	tomek	NA	NA
	eevvgg	stylometric feature-based detection	Gaussian Naïve Bayes
open	grzmot	NA	NA

Table 1: Overview of systems used in the PolEval 2025 Task 1: Śmigiel.

In statistical analysis, the Śmigiel dataset achieves close alignment between HWT and MGT in terms of text length, measured in sentences, words and characters. The HWT and MGT instances are comparable in overall length, although they slightly differ in the average number of tokens per sentence. The longest sentences occur in HWT passages from the literature, reviews, social media and Wikipedia, and in MGT passages from the news and parliamentary domains.

The Śmigiel HWT and MGT instances are also compared in terms of perplexity. MGT consistently shows lower perplexity than HWT. This suggests that MGT texts follow more regular and predictable linguistic patterns, likely due to reliance on frequent lexical and syntactic constructions. By contrast, the higher perplexity of HWT texts reflects greater linguistic diversity, structural complexity, and creative variability.

The Śmigiel dataset provides a rigorously curated and domain-balanced resource for training and benchmarking the detection of machine-generated Polish texts. To the best of our knowledge, it is the first publicly available dataset dedicated to MGT detection for Polish.

4 Baseline systems

To provide reference points for the submitted systems' efficacy, three baseline solutions relying on general-purpose text classifiers are presented:

- **BiLSTM** (*Bi-directional Long Short-Term Memory*) neural network, using BERT-tokenised input, embeddings (length 32) and two LSTM layers (Hochreiter and Schmidhuber, 1997) with hidden representation (length 128), and a final dense layer with softmax.
- **BERT** base (Devlin et al., 2018) fine-tuned for text classification.
- **GEMMA2B** (Mesnard et al., 2024), fine-tuned with QLoRa (Dettmers et al., 2023).

We use the implementation of these models from the BODEGA framework (Przybyła et al., 2024).

5 Submitted systems

Below we provide a short summary of the systems, for which description articles were submitted.

5.1 Unsupervised systems

damian96 (Starucha, 2025) The proposed solution adapts the Binoculars method (Hans et al., 2024) to Polish. Binoculars is a zero-shot detector for machine-generated text that relies on the ratio of perplexity to cross-perplexity. The method works by contrasting the outputs of two different LLMs, i.e. comparing how surprising the input text is to one model relative to how surprising another model's predictions of the same text are. In the present implementation, the decision threshold for classifying text as machine-generated or human-written is calibrated on the Śmigiel validation dataset. All experiments are conducted using Bielik models that share the same tokeniser.

kwrobel (Wróbel, 2025) The perplexity difference-based approach detects LLM-generated text by using differences in perplexity behaviour between multilingual and monolingual LLMs. For an input text, character-level normalised perplexities are computed using a pair of multilingual and monolingual models. The difference between these perplexity values serves as the classification signal: if the perplexity difference falls below a predefined threshold, the text is classified as LLM-generated text; if the difference exceeds the threshold, the text is classified as human-written. The core assumption underlying this method is that such model pairs assign relatively low perplexity to LLM-generated texts, resulting in smaller perplexity differences than those observed for human-written texts. Empirical evaluation across 14 tested models indicates that the most effective configuration is the pairing of Gemma-3-27B and PLLuM-12B models.

5.2 Constrained systems

kondziu98 (Pierzyński, 2025) The winning MGT detection solution is based on the Qwen3–8B model (Yang et al., 2025) fine-tuned for binary classification. To adapt the model to this task, the language modelling head used for token-level generation is replaced with a classification head. This simple upper-layer classifier is fine-tuned on Śmigiel data. The resulting approach is characterised by fast adaptation, resistance to overfitting, and robustness across diverse textual domains and generators.

eevvgg Building on classical computational stylometry, the proposed approach leverages linguistic fingerprints — such as lexical richness, function-word usage, part-of-speech distributions, punctuation patterns, and basic text statistics — extracted using the *pl_core_news_lg* spaCy model. The author hypothesises that these stylistic indicators remain discriminative for modern LLM-generated content and are more robust to domain shift than transformer-based detectors, as they capture general stylistic profiles rather than model-specific artefacts. The resulting handcrafted feature vectors are classified using Gaussian Naïve Bayes, showing that efficient and transparent methods can rival more resource-intensive neural approaches.

The architectures of the remaining solutions submitted to the Śmigiel task were not disclosed by their authors. As a result, a detailed architectural comparison across systems is not possible. The lack of publicly available information limits analysis to empirical performance rather than design choices, training strategies, or model complexity.

6 Results

Table 2 shows the results of the submitted approaches (accuracy on $\texttt{test}_B$) and baseline solutions. In total, we received 7 submissions: two in the unsupervised subtask, four in the constrained subtask and one in open subtask.

We can see that the performance of the unsupervised systems, despite not using any training data, easily exceeds the 50% accuracy of random choice and almost reaches 80%. In the constrained scenario we can see that the submitted solutions exceeded the baseline approaches in all but one case. Results over 90%, even though they are hard to compare with shared tasks in other languages, indicate that MGT detection is fairly manageable for Polish – but far from solved. In the open subtask we received just one approach, which exceeded two of the baselines, but did not reach the levels of the constrained solutions. This is despite the fact that the open subtask allowed for using any resources – those available in the constrained scenario and others. The smaller popularity of this subtask highlights that gathering resources for training MGT detection model is a laborious (and costly) task.

subtask	no.	system	accuracy
unsup.	1.	damian96	**0.7977**
	2.	kwrobel	0.7574
const.	1.	kondziu98	**0.9253**
	2.	grzmot	0.9127
	3.	tomek	0.9103
	B	GEMMA2B	0.8999
	B	BERT	0.8007
	B	BiLSTM	0.7737
	4.	eevvgg	0.4907
open	1.	grzmot	**0.8551**

Table 2: Classification performance for the seven submitted systems (ordered according to accuracy) and three baselines (marked *B*) in the subtasks: unsupervised, constrained and open.

7 Analysis

In the present section, we analyse the results obtained to better understand the task of MGT detection for Polish. Firstly, we perform a quantitative analysis of the detection performance in various scenarios. Secondly, we manually analyse the individual text fragments that are most commonly misclassified (or classified correctly) to understand the difficult and easy aspects of the task.

7.1 Quantitative

Our quantitative analysis consists of comparing performance of the 7 submitted solutions and 3 baselines for various subsets of the test set:

- Firstly, we differentiate *human-written* and *machine-generated* fragments,
- Secondly, we look at the relationship between the train and test data, comparing:
  - *known* data, i.e. produced by the model and the domain seen in training,
  - *new domain*, i.e. belonging to one of two domains not included in training (parliament and news),
  - *new model*, i.e. fragments generated by the model unseen in training (`llama-lg`),

  - *new domain/model*, i.e. the combination of the above,
- Thirdly, we check the accuracy on the four domains available in training data (*social* media messages, *literature* snippets, online *reviews*, *Wikipedia* entries) and two only present in the test data (*parliament* proceedings and *news* articles from Wikinews).
- Finally, we test the recognition performance for different sizes of the model generating text: *small* (under 9 billion parameters), *medium* (above 9, but under 15 billion) and *large* (above 15 billion parameters)

subset		**accuracy**	
	unsup.	**all**	**best**
all	0.7977	0.8124	**0.9253**
human-written	0.8887	0.8154	**0.9436**
machine-generated	0.7063	0.8093	**0.9069**
known	0.8380	0.8685	**0.9739**
new domain*	0.7930	0.7938	**0.9080**
new model*	0.7307	0.7899	**0.8414**
new domain/model*	**0.7126**	0.6928	0.6093
social media	0.7888	0.8439	**0.9643**
literature	0.8255	0.8555	**0.9812**
reviews	0.8291	0.8799	**0.9813**
wikipedia	0.7963	0.8631	**0.9527**
parliament*	0.7876	0.7911	**0.9016**
news*	0.8082	0.8014	**0.9261**
small model	0.6758	0.7943	**0.8945**
medium model	0.7067	0.8157	**0.9443**
large model*	0.7271	0.8151	**0.8886**

Table 3: The accuracy of the evaluated approaches (best unsupervised, all and best overall) for subsets of the test set according to their source, overlap with the training data, domain and model size. See description in text.

Moreover, we compare the performance on the above subsets for predictions coming from:

- the best *unsupervised* model (**damian96**),
- the *best* model overall (**kondziu98**),
- *all* the submissions and baselines.

Table 3 shows the result of the aforementioned analysis, providing performance for subsets divided as explained above. The subsets unseen in training data are marked with the asterisk in the table.

Regarding the sources, we are always getting a better classification performance for human-written fragments (recognised as such), while machine-generated samples are more difficult to recognise. In terms of connections with training data, the test samples coming from the same distribution as training data (*known*) are easy to classify, resulting in the impressive 97% accuracy for the best model. But, the performance worsens when we introduce new domains (to 91%) and even more so with the new model (to 84%). For the new model applied to new domains, the accuracy of the best approach overall falls down to 61% – far lower than the unsupervised approach, standing at 71%. This is a cautionary tale, indicating that MGT detection models are extremely prone to overfitting and might prove unreliable for data unlike what they've seen in training.

Regarding domains, we notice fairly similar detection accuracy, with the exception of parliament and news data, which were not available during training (or pre-training, as they cover current events). Reviews appears to be the easiest domain, which is an encouraging result for the genre which particularly suffers from the deluge of low-credibility machine-generated content online (Martínez Otero, 2021). Among the domains seen in training, social media proves to be most challenging, most likely due to very short length – 36 words on average (opposed e.g. to 196 for wikipedia articles (Strebeyko et al., 2025) provide less clues for predicting provenance.

Regarding model size, we would expect larger models to produce more human-like text, resulting in lower accuracy, but that, interestingly, is not the case. For unsupervised models, it's quite the opposite: the lowest accuracy is observed for output of the smallest generators, while for supervised models, the medium-sized LLMs prove most challenging. This indicates that size is not everything and very credible text can be produced from modest models. We need to acknowledge that our analysis is limited by the fact that for every size bracket we had a different composition of model families, which might be a confounding factor.

Additionally, we checked the performance on text generated with various decoding strategies. The most commonly detected MGTs were the ones created with the `greedy` strategy, with average success rate of 0.83% across the systems in the constrained subtask and baselines. This overrepresentation is most probably due to the fact that the strategy considers only the most probable token at each step, which generally leads to more deterministic outputs and repetitions, especially visible in longer sequences (Wu et al., 2025). In comparison, the strategy with the lowest average detec-

Hard human-written
Z tego powodu odbudowa trybunału jest tak naprawdę jak składanie drobnych fragmentów w całość — wymaga precyzji, wymaga wiedzy i umiejętności, tak aby stworzyć obraz rzeczywistości. Dlatego potrzebujemy nie tylko nowych kawałków, a zatem sędziów, ale także stabilnej ramy, czyli prawnych regulacji i procedur, które pozwolą nam na złożenie tego w spójną całość.
Polska stawia także na rozwój sztucznej inteligencji, AI, jako narzędzia, które powinno służyć człowiekowi, a nie odwrotnie. Wdrażając europejskie regulacje dotyczące AI, musimy zadbać, aby technologie te wspierały rozwój w takich dziedzinach, jak ochrona zdrowia, edukacja czy logistyka, jednocześnie zapewniając ich bezpieczeństwo i odpowiedzialne wykorzystanie. Nasz kraj dąży do tego, aby stać się liderem w produkcji i wdrażaniu innowacji związanych z AI, co w dłuższej perspektywie może przynieść ogromne korzyści
Z tego miejsca pragnę serdecznie podziękować wszystkim obecnym i byłym członkom orkiestry za ich zaangażowanie, pasję oraz trud włożony w rozwój tej wyjątkowej instytucji. Składam również gratulacje i wyrazy uznania mieszkańcom Suchedniowa, którzy od lat wspierają swoją orkiestrę, czyniąc z niej prawdziwy symbol lok
Hard machine-generated
Czy Morawiecki skrytykował już dzisiaj gromkim głosem Niemcy za... cokolwi
Czy pan poseł Adrian Zandberg z Nowej Lewicy mógłby odpowiedzieć publicznie na pytanie, czy Nowa Lewica była finansowana z pieniędzy Gazpromu albo Kremla? Pytałem o to roztargnionego posła Zandberga rok temu w Sejmie.
Oswajanie z seksualną normalnością nie na poziomie tak popularnego shokera ludzkich obrzydliwości, a w bardzo subtelnej, drgającej poezji codzienneg
Easy human-written
@user2618: mniemy taką nadzieje, zresztą już kolejka jest pewnie na jej miejsce by Klau
Przedmiotem artykułu są ekspresywne nazwy osób o wysokim ładunku emocjonalnym, głównie negatywnym. Są to leksemy określające w gwarze polskiej obwodu lwowskiego człowieka próżnego, leniwego, powolnego oraz wolno pracującego. Podstawę materiałową analizy stanowią wyrazy rodzime i pochodzenia obcego, typo
Pan @user4269 wchodząc w tak idiotyczną narrację robi idiotów z własnego elektoratu. Panie Budka przestań Pan robić z Siebie #POśm
Easy machine-generated
Już od samego początku "Sowy mafii" absorbują widza w swoisty świat przestępczości i korupcji, gdzie granice między dobrem a złem są coraz bardziej zacierane. Reżyser z niezwykłym wyczuciem portretuje postacie, które są zarówno fascynujące, jak i przerażające, unosząc się na granicy między realizmem a stylistyczną ekspresją. Każda scena jest starannie skomponowana, a aktorzy dostarczają występy, które są po prostu olśniewające. **(400 more words)**
Wystawa ta prezentować będzie prace artystów młodego pokolenia, którzy w swojej twórczości podejmują tematy związane z kondycją współczesnego człowieka. Artyści ci, poprzez swoje dzieła, starają się odpowiedzieć na pytania dotyczące tożsamości, wolności, relacji międzyludzkich oraz
, którego autorem jest Ludwik Mierosławski, polski generał, pisarz i działacz polityczny, jeden z przywódców powstania styczniowego. Utwór ten stanowi ważne źródło historyczne, pozwalające lepiej zrozumieć okoliczności i przebieg powstania, które było jednym z najwa

Table 4: Random selection of human-written and machine generated fragments that are easy (recognised correctly by all 10 approaches) or hard (incorrectly labelled by all 10 approaches)

tion success rate of **77%** was the (multinominal) `sampling`. Considering a wider array of next token candidates, the strategy increases stochasticity of the outputs, resulting in more exploratory, creative, and human-like generations.

7.2 Qualitative

Table 4 shows a random selection of fragments, either machine-generated or human-written, that were either very hard to recognise (all submitted and baseline solutions providing incorrect answer) or very easy (all correct). In total, we found 10 hard/human cases, 31 hard/machine, 2397 easy/human and 1028 easy/machine fragments. We can see that the difficult cases belong to the domain of parliamentary proceedings, which is understandable, given that that domain was withheld from training. As mentioned, the hard-to-detect instances are generally shorter than average, supporting the role text length (and, in our case, its proxy, the generation genre) play in detectability (Fivez et al., 2024).

During qualitative evaluation of texts, we found that the most easily detectable MGT (i.e. they have been detected by all the evaluation systems) contain repetitions of principal nouns, common phrases, and grammatical patterns, which is known to be an effective indicator of MGT in the literature (Wu et al., 2025) and is characteristic to `greedy` decoding strategy. In extreme cases, the noun gets repeated every other sentence, leading to patterns that are easy to spot, especially in longer passages.

Despite applying rigorous filtering of metalinguisitic artifacts in postprocessing, there were some instances in which they have been included. In the following example, certain "placeholders" for signatures were added at the end of generated text, presumably aiding detection:

Poniżej należy umieścić nazwisko posła lub senatora, który wypowiedź złożył). [Nazwisko posła lub senatora] (np. Jan Kowalski). [Tytuł posła lub senatora] (np. poseł na Sejm). [Partia polityczna] (np. Platforma Obywatelsk

In few generations our prompting method might have unintentionally made the task easier. To avoid gender bias in generations, our parliamentary proceeding prompt mentions that quote is coming from "a polish male parliamentarian or a polish female parliamentarian" (Polish *parlamentarzysta / parmanetarzystka*). This decision resulted in some generations expressing certain variability as per gender of the addresser or the addressed, a feature otherwise unseen among human texts:

W związku z tym, jako ***poseł/posłanka****, uważam, że niezmiernie ważne jest utrzymanie wysokiego poziomu...*

Niech ***Pan/Pani*** *przestanie opowiadać takie rzeczy. To, co* ***Pan/Pani*** *mówi, jest zwyczajną...*

Apart from that, it is not easy to notice many regularities. Most importantly, even the easy machine-generated fragments do not exhibit many visible signs of their provenance, which confirms both the quality of our dataset and the difficulty of the task.

8 Discussion

Generally speaking, the shared task has fulfilled its purpose by systematically evaluating a range of approaches to MGT detection in Polish in the scenarios covering various text genres and generator models. Clearly, the best performance is only achieved for models operating in their 'comfort zone' – classifying text in the same genre and from the same model seen in training. While this framework can have its uses, the lack of generalisability outside of it is the clear challenge in the area.

While the tested approaches – seven submitted by participants and three prepared baselines – represent a variety of solutions, we have to acknowledge that the two non-standard tasks (unsupervised and open) are less popular – just three submissions. These clearly require more effort from the participants, dealing with poorer performance in the former one and obtaining additional resources in the latter. However, the data used in our shared task is openly available (Strebeyko et al., 2025) and we hope it will be used to tackle these demanding scenarios in the future.

Our analysis also gives us some clues on the LLM's quality in Polish, since MGT detection accuracy can be interpreted as a proxy for text apparent credibility. Nevertheless, this analysis is not complete. For truly systematic analysis, we would prefer to independently test models' size (parameter count), novelty (whether it was seen in training), target language (specifically for Polish or multilingual) and family. Unfortunately, for many of these combinations there simply aren't models available, making the analysis biased.

Finally, our effort could be improved by expanding its size and scope: including more genres, more models and more text. The landscape of LLMs and their capabilities evolve constantly, making it necessary to update the benchmarks to make their evaluation results valid. Nevertheless, we hope that Śmigiel will be a valuable starting point for such efforts in the future.

9 Conclusion

This paper reports the results of the first shared task on MGT detection for Polish, organised within the PolEval 2025 evaluation campaign. The task attracted seven submissions across three scenarios – unsupervised, constrained, and open – reflecting different methodological choices. The results demonstrate that MGT detection for Polish is feasible: the top-performing constrained systems exceeded 90% accuracy on the main evaluation set (test B). Nevertheless, the task remains far from solved, particularly in setting that diverge from the training distribution.

Our analysis reveals several challenges for MGT detection. Although supervised approaches achieve strong performance on known domains and generator models, their accuracy drops in the most demanding setting – *new domain/model*. This behaviour highlights the risk of overfitting and the limited generalisation capabilities of supervised solutions. As robustness across domains and generator models is essential for real-world deployment, unsupervised approaches may offer a reliable alternative in such settings.

Overall, the shared task provides a valuable benchmark for MGT detection in a less-resourced language – Polish – and offers insights into both the strengths and limitations of current approaches. By releasing the Śmigiel dataset and an evaluation framework, we aim to foster further research on robust and generalisable methods for MGT detection.

Acknowledgments

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018019.

References

Janek Bevendorff, Daryna Dementieva, Maik Fröbe, Bela Gipp, André Greiner-Petter, Jussi Karlgren, Maximilian Mayerl, Preslav Nakov, Alexander Panchenko, Martin Potthast, Artem Shelmanov, Efstathios Stamatatos, Benno Stein, Yuxia Wang, Matti Wiegmann, and Eva Zangerle. 2025. Overview of PAN 2025: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

Savvas Chamezopoulos, Drahomira Herrmannova, Anita De Waard, Drahomira Herrmannova, Domenic Rosati, and Yury Kashnitsky. 2024. Overview of the DagPap24 shared task on detecting automatically generated scientific paper. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 7–11, Bangkok, Thailand. Association for Computational Linguistics.

Shammur Absar Chowdhury, Hind Almerekhi, Mucahid Kutlu, Kaan Efe Keleş, Fatema Ahmad, Tasnim Mohiuddin, George Mikros, and Firoj Alam. 2025. GenAI content detection task 2: AI vs. human – academic essay authenticity challenge. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 323–333, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Evan N. Crothers, Nathalie Japkowicz, and Herna L. Viktor. 2023. Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. *IEEE Access*, 11:70977–71002.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Fine-tuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.

Liam Dugan, Andrew Zhu, Firoj Alam, Preslav Nakov, Marianna Apidianaki, and Chris Callison-Burch. 2025. GenAI content detection task 3: Cross-domain machine generated text detection challenge. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 377–388, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Pieter Fivez, Walter Daelemans, Tim Van de Cruys, Yury Kashnitsky, Savvas Chamezopoulos, Hadi Mohammadi, Anastasia Giachanou, Ayoub Bagheri, Wessel Poelman, Juraj Vladika, Esther Ploeger, Johannes Bjerva, Florian Matthes, and Hans van Halteren. 2024. The clin33 shared task on the detection of text generated by large language models. *Computational Linguistics in the Netherlands Journal*, 13:233–259.

Christina Frohock. 2025. Ghosts at the Gate: A Call for Vigilance Against AI-Generated Case Hallucinations. *Penn State Law Review*, 130(1).

Matthias Gallé, Jos Rozen, Germán Kruszewski, and Hady Elsahar. 2021. Unsupervised and Distributional Detection of Machine-Generated Text. *Preprint*, arXiv:2111.02878.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

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.

Gilad Gressel, Rahul Pankajakshan, and Yisroel Mirsky. 2024. Discussion Paper: Exploiting LLMs for Scam Automation: A Looming Threat. In *Proceedings of the 3rd ACM Workshop on the Security Implications of Deepfakes and Cheapfakes*, WDC '24, pages 20–24, New York, NY, USA. Association for Computing Machinery.

Ahmed Abdeen Hamed and Xindong Wu. 2024. Detection of ChatGPT fake science with the xFakeSci learning algorithm. *Scientific Reports*, 14(1):16231.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *Preprint*, arXiv:2401.12070.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

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.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition. Online manuscript released August 24, 2025.

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

Zae Myung Kim, Kwang Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. 2024. Threads of Subtlety: Detecting Machine-Generated Texts Through Discourse Motifs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5449–5474, Bangkok, Thailand. 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.

Martin Májovský, Martin Černý, Matěj Kasal, Martin Komarc, and David Netuka. 2023. Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened. *J Med Internet Res*, 25(1):e46924.

Juan María Martínez Otero. 2021. Fake reviews on online platforms: perspectives from the US, UK and EU legislations. *SN Social Sciences*, 1(7):181.

Andrii Maslo and Silvia Gargova. 2025. BuST: A Siamese Transformer Model for AI Text Detection in Bulgarian. In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pages 45–52, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, and 88 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Dan Milmo. 2023. Mushroom pickers urged to avoid foraging books on Amazon that appear to be written by AI. *The Guardian*.

Mistral AI team. 2024. Mistral NeMo.

Hoang-Quoc Nguyen-Son, Minh-Son Dao, and Koji Zettsu. 2024. SimLLM: Detecting Sentences Generated by Large Language Models Using Similarity between the Generation and its Re-generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22340–22352, Miami, Florida, USA. Association for Computational Linguistics.

Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Krzysztof Wróbel, and Adrian Gwoździej. 2025. Bielik v3 small: Technical report. *Preprint*, arXiv:2505.02550.

Konrad Pierzyński. 2025. Detecting Machine-Generated Text in Polish Using Fine-Tuned Qwen. In *Proceedings of the PolEval 2025 Workshop*.

Piotr Przybyła, Nicolau Duran-Silva, and Santiago Egea-Gómez. 2023. I've Seen Things You Machines Wouldn't Believe: Measuring Content Predictability to Identify Automatically-Generated Text. In *Proceedings of the 5th Workshop on Iberian Languages Evaluation Forum (IberLEF 2023)*, Jaén, Spain. CEUR Workshop Proceedings.

Piotr Przybyła, Alexander Shvets, and Horacio Saggion. 2024. Verifying the robustness of automatic credibility assessment. *Natural Language Processing*, 31(5):1134 – 1162.

Zygmunt Saloni and Mirosław Bańko. 2012. Słownik języka polskiego, Warszawa 1958-1969. In Witold Doroszewski, editor, *Poradnik Językowy : organ Towarzystwa Kultury Języka*.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Overview of the AuTexTification 2023 Shared Task: Detection and Attribution of Machine-Generated Text in Multiple Domains. In *Procesamiento del Lenguaje Natural*, Jaén, Spain.

Areg Mikael Sarvazyan, José Ángel González, Francisco Rangel, Paolo Rosso, and Marc Franco-Salvador. 2024. Overview of IberAuTexTification at IberLEF 2024: Detection and Attribution of Machine-Generated Text on Languages of the Iberian Peninsula. *Procesamiento del Lenguaje Natural, Revista*, (73):421–434.

Aditya Shah, Prateek Ranka, Urmi Dedhia, Shruti Prasad, Siddhi Muni, and Kiran Bhowmick. 2023. Detecting and Unmasking AI-Generated Texts through Explainable Artificial Intelligence using Stylistic Features. *International Journal of Advanced Computer Science and Applications*, 14(10).

Tatiana Shamardina, Vladislav Mikhailov, Daniil Cherniankii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the the ruatd shared task 2022 on artificial text detection in russian. In *Computational Linguistics and Intellectual Technologies*, page 497–511. RSUH.

Damian Starucha. 2025. Perplexity-Driven Contrastive Scoring for Unsupervised Detection of AI-Generated Texts in Polish. In *Proceedings of the PolEval 2025 Workshop*.

Jakub Strebeyko, Alina Wróblewska, and Piotr Przybyła. 2025. Śmigiel Dataset: Laying Foundations for Investigating Machine-Generated Text Detection in Polish. Unpublished.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2023. Intrinsic dimension estimation for robust detection of AI-generated texts. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2024. Disinformation capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14830–14847, Bangkok, Thailand. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Ashraf Elozeiri, Saad El Dine Ahmed El Etter, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Nurkhan Laiyk, and 7 others. 2025. GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 244–261, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Krzysztof Wróbel. 2025. Unsupervised Detection of LLM-Generated Polish Text Using Perplexity Difference. In *Proceedings of the PolEval 2025 Workshop*.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. *Computational Linguistics*, 51(1):275–338.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. LLMDet: A Third Party Large Language Models Generated Text Detection Tool. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2113–2133, Singapore. Association for Computational Linguistics.

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.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Biru Zhu, Lifan Yuan, Ganqu Cui, Yangyi Chen, Chong Fu, Bingxiang He, Yangdong Deng, Zhiyuan Liu, Maosong Sun, and Ming Gu. 2023. Beat LLMs at Their Own Game: Zero-Shot LLM-Generated Text Detection via Querying ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7470–7483, Singapore. Association for Computational Linguistics.

Detecting Machine-Generated Text in Polish Using Fine-Tuned Qwen Models

Konrad Pierzyński
Adam Mickiewicz University
ul. Uniwersytetu Poznańskiego 4
61-614 Poznań, Poland
konrad.pierzynski@amu.edu.pl

Abstract

This paper presents a system submitted to the PolEval 2025 ŚMIGIEL shared task on detecting machine-generated Polish text. Within the CONSTRAINED setting, several Qwen3 models were fine-tuned using only the organisers' data and publicly available checkpoints. The study details the dataset, formulates an adaptation of a decoder-only language model for binary classification, and describes the end-to-end training pipeline. The best model, Qwen/Qwen3–8B, attains 93.11% accuracy on Test A and 92.53% on the hidden Test B. These results show that multilingual decoder-based LLMs can be strong discriminators of human- and machine-authored Polish text when appropriately adapted. The discussion also outlines areas in which robustness can be improved and points to avenues that may enable accuracy to surpass the reported results.

1 Introduction

Large language models have transformed modern NLP, enabling fluent text generation across languages, registers, and domains. However, these same capabilities also make it increasingly difficult to distinguish machine-generated text (MGT) from human-written content. Reliable discrimination is important for applications ranging from misinformation mitigation to academic integrity and content authenticity systems.

The ŚMIGIEL shared task at PolEval 2025 (Przybyła et al., 2025) provides a controlled benchmark for MGT detection in Polish—a morphologically rich West Slavic language whose inflectional complexity and syntactic flexibility challenge both generators and discriminators. This paper examines whether a multilingual decoder-only model, trained primarily for next-token prediction, can be effectively repurposed as a binary classifier given only supervised fine-tuning data.

This paper presents a complete pipeline for such adaptation using Qwen3 models. Work is carried out within the *CONSTRAINED* track, where only the organisers' dataset and publicly available pretrained checkpoints may be used. The dataset, architecture modification, training design, evaluation results, and practical observations from model behaviour are described in detail.

All code used in experiments, is publicly available at https://github.com/kpierzynski/poleval-smigiel.

2 Task Description

The ŚMIGIEL task is a binary classification problem. Given an input text, the system must assign:

- **0** – Human (human-written),
- **1** – AI (machine-generated).

The official evaluation metric is accuracy:

$$\text{Accuracy} = \frac{|\text{correct predictions}|}{|\text{all samples}|}.$$

Accuracy reflects the overall proportion of correctly classified instances and serves as the primary ranking criterion.

Although accuracy is intuitive, additional metrics might provide complementary insight into error types. Precision is defined as

$$\text{Precision} = \frac{TP}{TP + FP},$$

indicating how often predicted AI-generated texts are correct. Recall is given by

$$\text{Recall} = \frac{TP}{TP + FN},$$

capturing the fraction of AI-generated texts that are successfully detected. Their harmonic mean,

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

balances the two and is particularly informative under class imbalance or asymmetric error costs.

Evaluation is carried out on two datasets:

- **Test A** — public,
- **Test B** — private (labels withheld).

3 Dataset

3.1 Structure

The dataset consists of two files: `data.tsv` and `labels.tsv`. Each line in the label file corresponds to the same line in the data file, with 35,763 examples in total. The distribution is nearly balanced: 17,890 AI samples and 17,873 Human samples.

3.2 Domains

Human-written data comes from:

- Wikipedia passages,
- literary texts,
- social media posts,
- product reviews.

Machine-generated data was created using a diverse collection of open-source LLMs. According to the organisers' description, the training set includes generations from multiple model families such as LLama, Mistral, Bielik, PLLuM and larger Gemma. This diversity is crucial: it prevents the classifier from overfitting to idiosyncratic traits of a single generator and encourages learning broader stylistic cues indicative of synthetic text.

3.3 Example Data

Below are examples from the training dataset that were used in finetuning.

Label: AI
"Nie do końca tak jest. Program Smart to usługa oferowana przez Allegro, która zapewnia kupującym darmową dostawę i zwrot produktów przy zakupie u danego sprzedawcy. Sprzedawca musi spełnić określone . . . "

Label: Human
"Rozwalone składy owszem są bardzo ważne i na jakiś czas faktycznie zatrzymały strzelanie, ale ilość rozwalonych w ten sposób pocisków nie sądzę aby była duża (w porównaniu do zużycia 50k dziennie). Ukraińcy rozwalali składy przejściowe, które z założenia mają magazynować amunicję dla lokalnego odcinku frontu . . . "

Label: AI
"Ciekawe, że niektórzy senatorowie wyrażali wątpliwości co do jej skuteczności w ochronie prywatności obywateli. Czy uważacie, że ta ustawa jest wystarczająco skuteczna w zapewnieniu bezpieczeństwa . . . "

3.4 Observations

The dataset covers a broad range of domains in which AI-text detection systems are practically relevant. Both human and machine generated samples appear in multiple styles and levels of formality, which makes the task realistic and prevents reliance on superficial cues. The data preparation procedure—in particular the balanced label distribution and sufficient volume—provides a solid basis for training a classifier with good generalisation ability. As a result, the dataset is well created for developing models that can indeed detect AI-generated Polish text across real world scenarios.

4 Methods

4.1 Base Architecture: Qwen3

Qwen3 is a versatile family of decoder-only Transformer models offered in multiple sizes (from 0.6B to 235B), built on a proven architecture and widely adopted due to strong performance across tasks. The models provide reliable multilingual coverage, including Polish, which makes them a practical choice for downstream applications (Yang et al., 2025).

Although Qwen3 is trained for next-token prediction, its internal representations transfer well to classification. By replacing the generative LM Head with a simple classifier, the pretrained model becomes an effective text encoder suitable for wide range of text classifcation tasks.

4.2 Adapting a Generative LM for Classification

A decoder-only LLM typically ends with a large linear layer projecting hidden states to vocabulary logits. This LM Head is aligned with next-token prediction, not sequence-level classification.

For classification, token-level generation is unnecessary. Instead, the model should produce a single binary label based on the entire input sequence.

The transformation is conceptually simple:

- Generative LM Head is removed.
- A Classification Head is added, a linear projection from the final hidden state to two logits.

This setup parallels how generative models encode semantics: transformer layers still build a contextual representation of the entire sequence, but instead of using it to predict the next token,

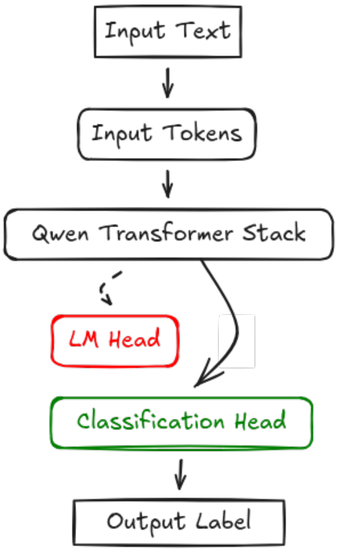

Figure 1: Modified model architecture

the final hidden state is repurposed as the input to a classifier. Last-token pooling was chosen both for its simplicity and because decoder-only models tend to compress sequence-level information into the representation of the final token (Radford et al., 2018).

4.3 Training Pipeline

Before listing the procedural steps, it is useful to describe the overall philosophy guiding finetuning. The goal is to preserve the strong multilingual representations already present in Qwen while allowing a lightweight classifier to specialise in the AI-vs-human distinction. Given the dataset size and the model's multilingual pretraining, training target stable convergence within a single epoch.

The core idea is that almost all parameters belong to the transformer backbone, which already captures syntax, semantics, coherence, and discourse structure, while classification head is comparatively tiny. As a result:

- the model adapts quickly,
- overfitting is unlikely within one epoch,
- most learning consists of shaping decision boundaries in the last-layer representation space.

Fine tuning steps

1. **Tokenizer preparation.** Load the pretrained tokenizer and remap the padding token to the EOS token to ensure consistent padding behaviour.

2. **Model loading.** The original LM Head is replaced by a Classification Head, using `AutoModelForSequenceClassification` with `num_labels=2` (Wolf et al., 2019).

3. **Dataset construction.** Load `data.tsv` and `labels.tsv` using a custom dataset class tailored to the task.

4. **Tokenisation.** Apply truncation to 256 tokens and pad sequences to a fixed length to enable efficient batching.

5. **Dataloaders.** A batch size of 4 provides a good balance between optimisation stability, memory limits, and generalisation behaviour.

6. **Hyperparameters.**

Hyperparameter	Value
Learning rate	5×10^{-5}
Epochs	1
Batch size	4
Optimizer	AdamW

Table 1: Fine tuning hyperparameters.

The settings in Table 1 were chosen based on prior experience to balance convergence speed and memory footprint.

7. **Training loop.** For each batch, perform:

forward pass → compute loss → backward pass → parameter update.

Training one epoch on Qwen3–8B required approximately 1 hour on a single A100 80GB GPU.

8. **Saving the model.** Save the final checkpoint in the standard Hugging Face format for straightforward downstream inference.

5 Results

Table 2 reports the accuracy obtained by the evaluated Qwen3 models of different sizes, showing consistently high performance across all variants. All variants achieve high accuracy on Test A, with results ranging from 89.0% for Qwen3–0.6B to 93.11% for Qwen3–8B. The intermediate model, Qwen3–1.7B, attains 91.55% accuracy, placing it between the smaller and larger variants.

A consistent increase in performance is observed as model size grows. The largest model, Qwen3–8B, achieves the highest accuracy on both evaluation sets, with 93.11% on Test A and 92.53% on Test B. These results indicate a clear scaling trend within the tested Qwen3 family, where larger models benefit from increased parameter capacity.

Model	Params	Test A	Test B
Qwen3–8B	8B	**0.9311**	**0.9253**
Qwen3–1.7B	1.7B	0.9155	–
Qwen3–0.6B	0.6B	0.8900	–

Table 2: Evaluation results.

5.1 Training Behaviour

Even a single epoch was sufficient for convergence, likely due to:

- strong multilingual pretraining of Qwen,
- relatively large dataset,
- straightforward binary classification objective.

Fast adaptation. Within a few thousand steps the model transitioned from near-random predictions to highly stable behaviour. This is characteristic of finetuning decoder-only LLMs on tasks where a linear classifier is sufficient to separate high-level representations.

No overfitting within one epoch. Because only the upper layer is replaced, memorisation is limited; the backbone already encodes broad linguistic patterns, so the classifier mostly learns to interpret them for this specific task.

Generalisation patterns. Qwen-8B performance on Test B closely tracks Test A, suggesting robustness across domains and generator types. It appears to capture subtle stylistic signatures that transcend dataset-specific artefacts.

6 Discussion

6.1 Choice of Qwen3–8B for Test B

Since Qwen3–8B consistently outperformed smaller variants on Test A, it was selected for the final evaluation. Its additional parameter capacity appears to enhance semantic sensitivity and capture soft signals correlated with AI generation.

6.2 Potential Improvements

Although 92.53% accuracy is competitive, crossing 95% or approaching 99% would require substantial refinements. Several avenues merit exploration:

- **More complex classification head.** Currently, only one linear layer is used. Adding more advanced classifier could better capture decision boundaries.
- **Hyperparameter tuning.** Further tuning the hyperparameters could slightly improve model accuracy. Among them, the number of epochs and the batch size will likely have the biggest impact.
- **More epochs with regularisation.** While one epoch prevents overfitting, multiple epochs with dropout or layer-wise learning rate decay might systematically improve model quality.
- **Intermediate model sizes (4B).** This would illuminate whether performance scales smoothly or non-linearly between 1.7B and 8B, helping choose an optimal cost-accuracy trade-off.
- **Very large Qwen3 models (14B–32B).** Testing bigger backbones could provide insights into the upper limit of this architecture for spotting AI generated texts.

Collectively, these steps could improve the classification capabilities of the given architecture and provide insights into the cost-efficiency trade-off, helping to select the best solution.

7 Conclusion

This paper presented a complete system for detecting machine-generated Polish text using finetuned Qwen3 models. By replacing the LM Head with a simple classifier, the models can effectively discriminate between human and machine authorship. Qwen3–8B achieved the strongest results, generalising well to the hidden Test B set. The findings demonstrate that decoder-only multilingual LLMs can serve as accurate detectors even when trained solely on task-provided data. Future work can explore larger models, deeper classification heads, and more advanced training regimes to push accuracy beyond current levels.

References

Piotr Przybyła, Jakub Strebeyko, and Alina Wróblewska. 2025. PolEval 2025 Task 1 Śmigiel: Spotting Machine-Generated Text from LLMs for Polish. In *Proceedings of the PolEval 2025 Workshop*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

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.

Perplexity-Driven Contrastive Scoring for Unsupervised Detection of AI-Generated Texts in Polish

Damian Stachura
Evidence Prime, Kraków, Poland
damian.stachura1@gmail.com

Abstract

The ŚMIGIEL competition (Przybyła et al., 2025) at PolEval 2025 focuses on distinguishing Polish human-written text from AI-generated text. I participated in one of the subtasks that required a zero-shot detection method. My solution adapts the Binoculars detector by pairing language models and using calibrated thresholds. Specifically, I replaced the English language models from the original Binoculars method with models trained on Polish corpora. This approach achieved first place in the chosen competition track. Overall, my findings demonstrate that domain-specific language models and careful thresholding enable state-of-the-art zero-shot AI-text detection performance across new languages and domains. The code is publicly available at https://github.com/damian1996/2025-smigiel.

1 Introduction

Detecting AI-generated text is a pressing task as large language models (LLMs) become pervasive. In educational and content-moderation settings, distinguishing human and machine writing is crucial for ensuring authenticity and preventing misuse. The PolEval 2025 proposed a task caled ŚMIGIEL (Spotting Machine-Generated Text from Language Models for Polish). ŚMIGIEL provides a dataset of Polish texts, both human-authored and AI-generated, challenging participants to build effective detectors for this language. Importantly, the most robust detectors operate in a zero-shot or unsupervised setting, requiring no labeled examples of AI text from the target domain. Zero-shot methods are desirable because they generalize across domains and languages without costly retraining. For example, the Binoculars approach (Hans et al., 2024) contrasts two pretrained LLMs to score a document, achieving over 90% detection at extremely low false positive rates without any model-specific training. Similarly, DetectGPT (Mitchell et al., 2023) uses the log-probability curvature of a single LLM's output to flag generation, also without additional training. These works show that clever use of existing language models can separate human and machine text in a fully unsupervised way.

Perplexity-based methods are a classic example of zero-shot detection. In essence, perplexity measures how surprising or unpredictable a text is according to a language model. Prior studies have found that machine-generated text often has lower perplexity. It means that it is more predictable to an language model than human text. For instance, Liu and Kong (2024) used a sliding window of GPT-2 perplexity to capture patterns of AI text, demonstrating that such a metric can effectively distinguish AI-written samples. A popular AI detector called GPTZero [1] explicitly leverages sentence-level perplexity and a related burstiness metric to score text. The burstiness reflects variability in the writing. Humans tend to vary sentence complexity and word choice more than AI, which writes at a consistently uniform level. Indeed, recent work on stylometric detection shows that machine texts cluster tightly by model, whereas human writing exhibits greater stylistic diversity. My approach builds on these insights. I used Polish LLMs to compute contrastive log-probabilities and note that human Polish writers, like English writers, naturally introduce more variety and idiosyncrasy in their text. By calibrating a threshold on these scores for the ŚMIGIEL data, I effectively adapted the zero-shot Binoculars (Hans et al., 2024) method to Polish. In the end, my unsupervised detector using Polish language models and the Binoculars scoring rule, outperformed all other entries to win first place on the ŚMIGIEL leaderboard.

[1] https://gptzero.me/

Proceedings of the PolEval 2025 Workshop, pages 21–25
November 21, 2025 ©2025 Association for Computational Linguistics

2 Method

2.1 Perplexity Score as AI-Generated Text Detector

Perplexity is classical method for detection of AI-generatex texts. The perplexity of text can be computed under a language model M, defined as the exponentiation of the average negative log-likelihood, reported as log-perplexity, of the tokens in the text. Intuitively, perplexity measures how surprising a string is to a model. In practice, modern LLMs score human-written text as higher-perplexity than text they generate, because the model is trained to assign high probability to its own outputs. Thus, perplexity-based detection flags a text as machine-generated if its perplexity under the model is unusually like presented in Gutiérrez Megías et al. (2024). In formula terms, for a token sequence X of length T and model probability for ith token contitioned on the sequence of tokens $p_\theta(x_i \mid x_{<i})$, the log-perplexity is defined as

$$\log \text{PPL}(M, X) = -\frac{1}{T}\sum_{i=1}^{T} \log p_\theta(x_i|x_{<i}) \quad (1)$$

and the actual perplexity is $\exp(\ell_{\text{ppl}})$. Lower perplexity indicates text closely matches the model's distribution. Since human text tends to have higher perplexity under the model, simple detectors have used a perplexity threshold to distinguish between human and AI text.

2.2 Cross Perplexity Score

Cross-Perplexity, noted as $\text{XPPL}(M_p, M_t; X)$, is a metric derived from cross-entropy and is utilized to assess the difference in the probability distributions predicted by two distinct language models M_p and M_t over a given sequence of text X. This metric is particularly relevant in the field of AI-generated text detection as it can quantify how surprising a text is to the candidate model M_p when its probability distribution is evaluated based on the expected distribution of the reference model M_t. In more details, cross-perplexity is the average per-token cross-entropy when the next-token probabilities of M_t are evaluated under M_p. The final score is computed as a logarithm of XPPL score and stated as log XPPL. Formally,

$$logXPPL = -\frac{1}{T}\sum_{i=1}^{T} M_p(s_i) \cdot \log M_t(s_i) \quad (2)$$

, where log XPPL = log $\text{XPPL}(M_p, M_t, X)$, and $M_p(s_i)$ and $M_t(s_i)$ denote the probability distributions over the vocabulary for the next token at position i. At each step i, the dot product between $M_p(s_i)$ and $logM_t(s_i)$ is computed.

2.3 Binoculars Score

The Binoculars method uses two LLMs to overcome pitfalls of naive perplexity detection (Hans et al., 2024). One of the significant challenges is associated with model prompting due to its potential to influence the linguistic structure of the generated text, resulting in outputs with elevated perplexity scores that may evade detection by perplexity-based AI-generated text classifiers. Furthermore, the typical lack of access to the original prompts makes it impossible to verify the output's adherence to the specified instructions, introducing difficulties in assessing model fidelity.

Let M_p be an observer model and M_t a performer model. Both models are chosen under the assumption that they need to have the same tokenizer. I compute the model's self-perplexity $\text{PPL}(M_t, X)$ and the cross-perplexity $\text{XPPL}(M_p, M_t; X)$, as defined in the earlier sections. The core Binoculars score is then defined as the ratio of computed self-perplexity to cross-perplexity:

$$\text{Binoculars}(X) = \frac{\log \text{PPL}(M_t, X)}{\log \text{XPPL}(M_p, M_t; X)} \quad (3)$$

Intuitively, this ratio re-scales the perplexity by the baseline perplexity that model M_p assigns to the next-token predictions of the model M_t. In typical use, M_p and M_t are two similar models so that their predictions diverge less on machine text than on human text. In the original paper authors used a pair of base model and its instruction-tuned variant. A key advantage is that if a text is high-perplexity only because it is unlikely overall. The cross-perplexity normalizes this: humans tend to choose even more surprising continuations than the model does, so human text yields a larger PPL relative to XPPL. In practice, Binoculars vastly outperforms raw perplexity at low false-positive rates.

2.4 Threshold selection

The Binoculars scores are converted into a binary decision by thresholding. Specifically, the decision threshold are set as the median of binocular score computed over multiple texts. Texts with score below threshold are classified as AI-generated while higher scores are considered human-written. This choice centers the decision boundary on the observed distribution of scores.

2.5 Models

As mentioned earlier, the Binoculars score requires two LLMs that share the same tokenizer. I use three pairs of Polish models proposed by SpeakLeash. Each pair comprising two versions of the same base architecture. I decided to use two instruction tuned models based on the same base model or combination instrunction-tuned and base models. Each of these pairs uses the same tokenizer, satisfying the Binoculars requirement.

- Bielik-11B-v2 (Ociepa et al., 2025b) and Bielik-11B-v2.3-Instruct [2]
- Bielik-11B-v2.5-Instruct [3] and Bielik-11B-v2.6-Instruct [4]
- Bielik-4.5B-v3 (Ociepa et al., 2025a) and Bielik-4.5B-v3.0-Instruct [5]

2.6 My Approach

The final submissions used the proposed Binoculars method with two different thresholding strategies and two backbone models, Bielik-4.5B-v3 and Bielik-4.5B-v3.0-Instruct, for computing the logPPL and logXPPL scores. In the first submission, reported in Table 1, the threshold was determined using the validation dataset. In the second submission, the threshold was set to the median of the Binoculars scores across the entire test set. The resulting performance differences between the two submissions were negligible. The corresponding threshold values were 0.936 and 0.931, respectively.

3 Results

3.1 Task Description

The dataset for ŚMIGIEL competition (Przybyła et al., 2025) consists of Polish texts from the four sources:

- customer reviews
- literature
- social media
- wikipedia

Model 1	Model 2	Score
Bielik-4.5Bv3	Bielik-4.5B-v3.0-Instruct	79.6
Bielik-11B-v2	Bielik-11B-v2.3-Instruct	79.4
Bielik-11B-v2.5-Instruct	Bielik-11Bv2.6-Instruct	72.4

Table 1: Average Binoculars scores for various pairs of Polish LLMs on the validation dataset.

Place	Accuracy
1	**79.77**
2	**79.70**
3	75.74
4	74.79
5	70.23

Table 2: Final ŚMIGIEL Leaderboard. Two first systems were submitted by me.

The human-written and AI-generated texts are evenly distributed across the dataset. Machine-generated texts were produced by multiple open-weight LLMs like LLama 3.1 8B (Grattafiori et al., 2024), Bielik 7B (Ociepa et al., 2024), Mistral 7B (Jiang et al., 2023), Bielik 11B (Ociepa et al., 2025b), PLLuM 12B (Kocoń et al., 2025), and Gemma 3 27B (Team et al., 2025).

3.2 Comparing Model Pairs

I compared three pairs of models, all utilizing the same tokenizer as mentioned in the previous section. The key observation is that a pair consisting of the base model and its respective instruction-tuned model outperforms a pair of two similar instruction-tuned LLMs. This observation is further supported by the pairs of models mentioned in the original Binoculars paper. The results are presented in Table 2.

3.3 Competition Leaderboard

The performance of all participating systems is summarized on the competition leaderboard in Table 1. Among these, my two submissions obtained the highest accuracy scores. The strongest results overall were achieved by using the Binoculars score with Bielik-4.5Bv3 and Bielik-4.5B-v3.0-Instruct models as backbones for it.

3.4 Analysis

I conducted two additional measures to analyze the results achieved by Binoculars. I decided to visualize:

[2] huggingface.co/speakleash/Bielik-11B-v2.3-Instruct
[3] huggingface.co/speakleash/Bielik-11B-v2.5-Instruct
[4] huggingface.co/speakleash/Bielik-11B-v2.6-Instruct
[5] huggingface.co/speakleash/Bielik-4.5B-v3.0-Instruct

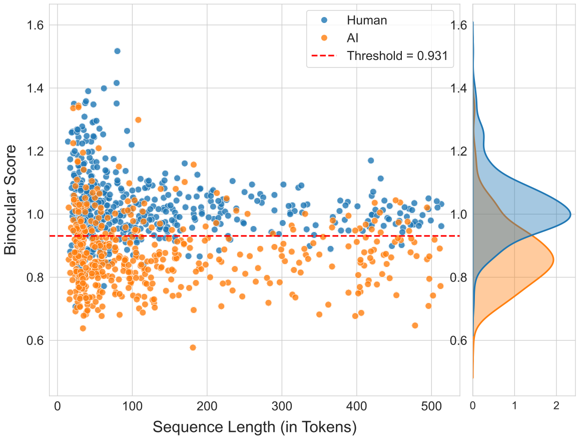

Figure 1: The distribution of the Binoculars scores was visualized as a function of text length (measured in tokens). This analysis used 1000 randomly selected samples from the provided validation dataset, with both classes (AI-generated and human-written) being evenly distributed.

- The distribution of the Binoculars score as a function of text length (in tokens)
- The distribution of the corresponding perplexity and cross-perplexity scores

In the first case, it can be observed that the Binoculars scores scale to similar ranges regardless of the number of tokens in the provided texts. This is a very important property as it demonstrates that the method can be used effectively, even for long texts. I presented it on the Figure 1

I also tested the distribution of the corresponding perplexity and cross-perplexity scores. The key insight from this analysis is that these distributions are very close to one another, suggesting that the specific models chosen as backbones for the Binoculars method may not be critical to its performance. The results for the pair consisting of Bielik-4.5Bv3 and Bielik-4.5B-v3.0-Instruct models are presented in Figure 2. Similarly, the distribution for the pair Bielik-11B-v2 and Bielik-11B-v2.3-Instruct is visible in Figure 3.

4 Discussion

Two key challenges in developing the proposed solution were the selection of models for score computation and the determination of an appropriate decision threshold.

In my experiments, only a narrow set of models was considered due to the requirement that both models share the same tokenizer. One potential direction for improving performance would be to evaluate additional Polish and multilingual large language models.

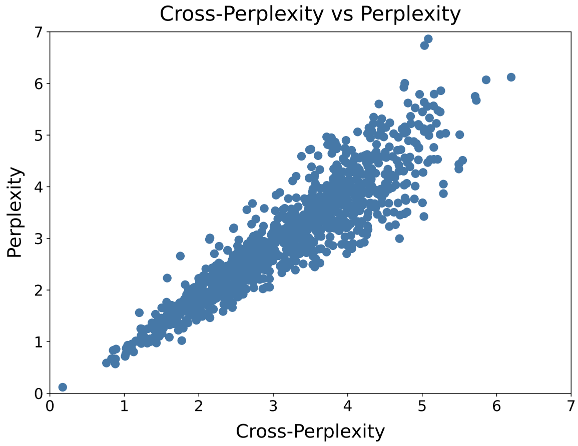

Figure 2: The distribution of the corresponding perplexity and cross-perplexity scores. This analysis used 1000 randomly selected samples from provided validation dataset with both classes being evenly distributed. Models Bielik-4.5Bv3 and Bielik-4.5B-v3.0-Instruct were used for this visualization.

The second major challenge was selecting an optimal threshold. My two thresholds were computed separately for the validation and test datasets, yielding nearly identical values of 0.936 and 0.931, respectively. This similarity indicates that the data distributions of the validation and test sets are highly consistent. For comparison, the original Binoculars paper reported a threshold of 0.896 for an English dataset, suggesting that the distribution of Binoculars scores is broadly similar across languages. Consequently, further refinement of the threshold selection procedure is unlikely to result in substantial performance gains.

An alternative approach could involve the use of a hybrid method for samples with Binoculars scores close to the decision threshold. Such a strategy may help strengthen classification decisions for the most challenging texts.

5 Conclusions

My experiments showed that perplexity-based methods are too naive for AI-generated text detection. The emergence of powerful prompts, which significantly influence model behavior, presents a key pitfall for such detectors. Multiple teams have presented more sophisticated methods in recent years. In my submissions, I decided to measure the performance of Binoculars scores for Polish texts.

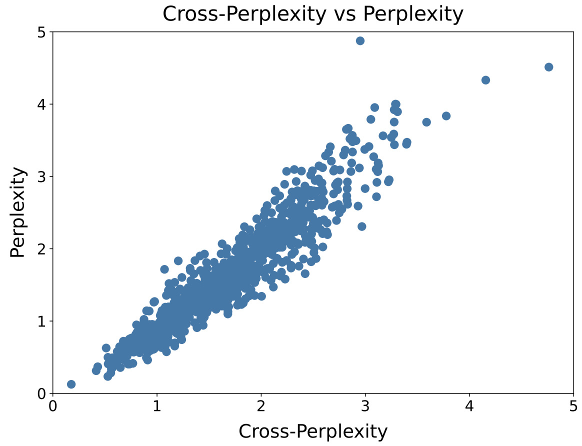

Figure 3: The distribution of the corresponding perplexity and cross-perplexity scores. This analysis used 1000 randomly selected samples from provided validation dataset with both classes being evenly distributed. Models Bielik-11B-v2 and Bielik-11B-v2.3-Instruct ere used for this visualization.

This method performed strongly compared to other submissions and allowed me to achieve first place in the unsupervised subtask of ŚMIGIEL. The best results were achieved using relatively small LLMs with 4.5 B parameters, which is a particularly interesting property as it makes the method relatively efficient and fast.

References

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.

Alberto José Gutiérrez Megías, L. Alfonso Ureña-López, and Eugenio Martínez Cámara. 2024. The influence of the perplexity score in the detection of machine-generated texts. In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 80–85, Lancaster, UK. International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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.

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*.

Xurong Liu and Leilei Kong. 2024. Ai text detection method based on perplexity features with strided sliding window. In *Working Notes of CLEF 2024 — Conference and Labs of the Evaluation Forum (CLEF 2024)*, number 3740 in CEUR Workshop Proceedings, pages 2755–2760, Aachen.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Krzysztof Wróbel, and Adrian Gwoździej. 2025a. Bielik v3 small: Technical report. *Preprint*, arXiv:2505.02550.

Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, and Remigiusz Kinas. 2024. 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. 2025b. Bielik 11b v2 technical report. *Preprint*, arXiv:2505.02410.

Piotr Przybyła, Jakub Strebeyko, and Alina Wróblewska. 2025. PolEval 2025 Task 1 Śmigiel: Spotting Machine-Generated Text from LLMs for Polish. In *Proceedings of the PolEval 2025 Workshop*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Unsupervised Detection of LLM-Generated Polish Text Using Perplexity Difference

Krzysztof Wróbel
Jagiellonian University, SpeakLeash, Enelpol
krzysztof.wrobel@bielik.ai

Abstract

Inspired by zero-shot detection methods that compare perplexity across model pairs, we investigate whether computing perplexity *differences* on whole-text character-level perplexity can effectively detect LLM-generated Polish text. Unlike token-level ratio methods that require compatible tokenizers, our approach enables pairing any models regardless of tokenization. Through systematic evaluation of 91 model pairs on the PolEval 2025 ŚMIGIEL shared task, we identify Gemma-3-27B and PLLuM-12B as optimal, achieving 81.22% accuracy on test data with unseen generators. Our difference-based approach outperforms token-level ratio methods (+5.5pp) and single-model baselines (+8.3pp) without using training labels, capturing asymmetric reactions where human text causes greater perplexity divergence than LLM text. We demonstrate that complementary model pairing (multilingual + monolingual) and architectural quality matter more than raw model size for this task.

1 Introduction

The rapid advancement of large language models (LLMs) has enabled the generation of highly fluent and linguistically correct texts across numerous languages. While these capabilities have transformative applications, they also raise critical concerns regarding authenticity and potential misuse. The need for robust machine-generated text detection systems has become increasingly urgent.

This challenge is particularly acute for non-English languages. The PolEval 2025 ŚMIGIEL shared task (Przybyła et al., 2025)[1] addresses this gap for Polish, offering three subtasks: UNSUPERVISED (no training labels), CONSTRAINED (using only provided data), and OPEN (unrestricted). Our work focuses on the UNSUPERVISED subtask.

Existing approaches face significant limitations. Supervised methods require substantial labeled data and may overfit. Single-model perplexity baselines achieve only 70-72% accuracy. The key challenge is generalizing to texts from LLMs unseen during development.

Inspired by recent zero-shot detection methods like Binoculars (Hans et al., 2024), which uses perplexity *ratios* between model pairs, we investigate whether a simpler mathematical *difference* (subtraction) can achieve competitive performance for Polish text detection. Our hypothesis: model pairs react asymmetrically – both exhibit low perplexity on LLM text (small difference), while human text causes one model to struggle more (large difference). Through systematic evaluation of 91 pairs across 14 models, we identify optimal combinations achieving strong performance without requiring training labels.

Contributions:

1. **Whole-text difference vs. token-level ratio:** We propose computing perplexity differences on whole-text character-level perplexity rather than token-level ratios (Binoculars), achieving superior performance (+5.5pp on Polish) while eliminating tokenizer compatibility constraints and reducing computational cost. This enables systematic evaluation of any model pair.

2. **Systematic Model Selection:** We conduct the first systematic comparison of 91 model pairs across 14 language models with diverse tokenizers for Polish, identifying the Gemma-PLLuM combination as optimal (81.22% accuracy on unseen generators).

3. **Complementarity Principle:** We show that pairing multilingual with monolingual models creates stronger asymmetric reactions than individual model strength, with PLLuM (weak-

[1] ŚMIGIEL: **S**potting **M**achine-**G**enerated Text from **L**LMs for Polish. https://poleval.pl/tasks/task1

Proceedings of the PolEval 2025 Workshop, pages 26–38

est individual discriminator, 19.3%) producing the best pairs.

4. **Generalization:** We demonstrate robust generalization to unseen LLM generators (+1.20pp improvement on test set), showing the method captures fundamental authorship signals rather than generator-specific artifacts.

5. **Efficiency:** We show that architectural quality matters more than size – Gemma-4B achieves near-equivalent performance to Gemma-27B (80.33% vs 81.22%), offering practical deployment advantages.

2 Related Work

2.1 LLM-Generated Text Detection

Recent approaches to detecting machine-generated text include zero-shot methods that compare perplexity across two models. Most notably, **Binoculars** (Hans et al., 2024) computes a *ratio* of log-perplexities between an "Observer" and "Performer" model, achieving state-of-the-art zero-shot detection performance on English text. Crucially, Binoculars computes the ratio *at every token position*, requiring models with compatible tokenizers to ensure token-level alignment. The ratio formulation normalizes for text length and domain scaling. However, most existing work focuses on English, leaving a significant gap for non-English languages like Polish.

In parallel, perplexity differences (rather than ratios) have been explored for data quality assessment (Li et al., 2024), where the perplexity gap between small and large models identifies high-quality training data. However, the use of a simple mathematical *difference* for AI text detection remains underexplored compared to ratio-based methods.

2.2 Perplexity-Based Detection

Traditional single-model perplexity-based detection relies on the observation that LLM-generated text typically has lower perplexity than human text (Zhong et al., 2025; Gutiérrez Megías et al., 2024). However, this conflates authorship with content difficulty – unusual topics yield high perplexity regardless of authorship (the "capybara problem" (Hans et al., 2024)).

Our approach investigates whether a simple perplexity *difference* (subtraction) computed on *whole-text perplexity* can achieve competitive performance with token-level ratio methods while offering computational simplicity and tokenizer independence. Unlike Binoculars, which requires compatible tokenizers to compute ratios at every token position, our character-level normalization enables comparison of any model pair regardless of tokenizer. We demonstrate that for Polish text, character-level normalized differences combined with systematic model pair selection (multilingual + monolingual) achieve strong unsupervised performance.

2.3 Polish Language Models

Bielik family (Ociepa et al., 2024, 2025b,a)[2] (SpeakLeash): Open-source models ranging from 1.5B to 11B parameters, demonstrating strong performance across Polish NLP tasks while maintaining competence in other European languages, making them versatile multilingual models with Polish emphasis.

PLLuM-12B (Kocoń et al., 2025)[3] (HIVE AI Consortium): 12B parameter model continuously trained on ∼150B tokens with primary focus on Polish-language understanding and specialization in Polish administrative contexts.

These models' different training philosophies (Bielik's multilingual versatility vs PLLuM's Polish specialization) provide complementary perspectives when paired with large multilingual models like Gemma, creating powerful asymmetries for detection.

3 Methodology

3.1 Task Definition: PolEval 2025 ŚMIGIEL

The ŚMIGIEL shared task[4] frames machine-generated text detection as binary classification: distinguishing human-written texts (label 0) from LLM-generated texts (label 1). The task includes three subtasks: UNSUPERVISED (no training labels), CONSTRAINED (using only provided training data), and OPEN (unrestricted training). Our solution addresses the UNSUPERVISED subtask.

Dataset characteristics:

Human texts originate from diverse domains under open licenses: Customer reviews, Literature, Social media, and Wikipedia.

[2] https://huggingface.co/speakleash

[3] https://huggingface.co/CYFRAGOVPL/pllum-12b-nc-chat-250715

[4] Task repository: https://github.com/poleval/2025-smigiel

LLM-generated texts are produced by open-source models across three size categories:

- **Small** (7-8B params): Llama 3.1 8B, Bielik 7B, Mistral 7B
- **Medium** (11-12B params): Bielik 11B, Mistral Nemo, PLLuM 12B
- **Large** (27B params): Gemma 3 27B

The dataset is balanced: equal proportions of human/LLM texts, uniform domain representation, and roughly one-third of LLM texts from each size category.

Train set: 35,763 Polish texts (balanced: 50% human, 50% LLM). All texts are proportionally truncated to a character limit for consistency.

Test B set: 18,432 Polish texts (balanced). Crucially, LLM texts are generated by *different models and genres* than the train set, testing generalization to unseen generators.

Unsupervised setting: We do not use training labels. Thresholds are selected as the median of perplexity differences, which is fully unsupervised as it uses only score distributions, not labels (see Methodology for details).

Evaluation metric: The official metric is **Accuracy**, defined as the proportion of correctly classified instances over total instances, following standard practice in text classification research.

3.2 Perplexity Computation

Character-Level Perplexity:

A key technical difference from Binoculars is our use of *whole-text perplexity with character-level normalization* rather than token-level ratios. We compute perplexity using the standard token-based approach and then normalize to character-level by dividing the total negative log-likelihood by the number of characters rather than tokens:

$$\text{PPL}_{\text{char}}(text) = \exp\left(-\frac{1}{|text|_{\text{chars}}} \sum_{i=1}^{|text|_{\text{tokens}}} \log P(t_i|t_{<i})\right)$$

where $|text|_{\text{chars}}$ denotes the number of characters in the text, $|text|_{\text{tokens}}$ is the number of tokens, t_i is the i-th token, $t_{<i}$ represents all preceding tokens, and $P(t_i|t_{<i})$ is the conditional probability of token t_i given the context.

This character-level normalization ensures better comparison across models with different tokenizers, as the same text may be tokenized into different numbers of tokens but always has the same number of characters.

3.3 Perplexity Difference Method

Our method computes the *difference* between character-level perplexities from two language models and uses this difference as a classification signal. While perplexity-based detection has been explored previously (Hans et al., 2024), our contribution lies in the systematic investigation of model pair selection for Polish and the demonstration that character-level normalization combined with complementary model pairings (multilingual + monolingual) achieves superior performance.

Classification Rule:

$$\text{class} = \begin{cases} \text{LLM} & \text{if } \text{PPL}_A - \text{PPL}_B < \theta \\ \text{HUMAN} & \text{otherwise} \end{cases}$$

where PPL_A and PPL_B are character-level perplexities from models A and B, and θ is a threshold.

Threshold Selection:

For unsupervised threshold selection, we use a simple and interpretable approach: the **median** of all perplexity differences in the dataset. This median-based threshold has several advantages:

1. **Truly unsupervised:** Requires no labeled data, only the distribution of difference scores
2. **Robust to outliers:** Median is more stable than mean, resistant to extreme values
3. **Balanced by design:** Automatically balances false positives and false negatives when class distribution is balanced
4. **Simple and interpretable:** No hyperparameter tuning or optimization required

Given a dataset of n texts, we compute:

$$\theta = \text{median}\{\text{PPL}_A(t_i) - \text{PPL}_B(t_i) : i = 1, \ldots, n\}$$

We evaluate two scenarios: (1) *train threshold applied to test* (true generalization), and (2) *test-optimized threshold* (unsupervised upper bound). Both remain fully unsupervised as they use only score distributions, not labels.

Intuition: The method exploits asymmetric model reactions. LLM-generated texts have small perplexity differences (both models find them predictable), while human texts have large differences (models disagree). The median naturally separates these two distributions, as we demonstrate in Section 5.

3.4 Systematic Model Pair Comparison

Models Tested (14 total):

Large multilingual models:

- Gemma-3-27B, Gemma-3-12B, Gemma-3-4B (Team et al., 2025) (Google, multilingual, 4-27B params)
- Llama-3.1-8B (Meta AI, multilingual)
- Mistral-7B-v0.2, Mistral-Nemo (Mistral AI, multilingual)
- Qwen2.5-14B (Alibaba, multilingual)

Polish-specialized models:

- PLLuM-12B-nc-chat-250715 (HIVE AI Consortium/NASK PIB, Polish-Slavic, 12B params)
- Bielik-11B-v2.3, Bielik-11B-v2.6, Bielik-11B-v3.0 (SpeakLeash, Polish+multilingual, 11B params)
- Bielik-7B-v0.1, Bielik-4.5B-v3.0, Bielik-1.5B-v3.0 (SpeakLeash, Polish+multilingual, 1.5-7B params)

Experimental Setup:

All experiments were conducted on NVIDIA A100 40GB GPUs via the PLGrid HPC infrastructure (ACK Cyfronet AGH). We used PyTorch with Hugging Face Transformers for model inference, loading models in bfloat16 precision for numerical stability. Perplexity was computed using a sliding window approach: sequences were truncated at 8,192 tokens maximum, with a window size of 2,048 tokens and stride of 512 tokens.

Computational cost: Processing the full training set (35,763 texts) on A100 40GB requires: Gemma-27B (97 minutes on 2 GPUs), Bielik-11B (51 minutes on 1 GPU), Bielik-4.5B (40 minutes on 1 GPU). The main bottleneck of our implementation is sequential processing of texts and windows, which underutilizes the GPU. Batching would provide the largest speedup.

4 Experiments and Results

4.1 Main Results: Model Pair Performance

Table 1 shows the top-performing model pairs on both train and test B sets.

All top 10 pairs involve Gemma models, demonstrating their critical importance. Remarkably, **Gemma-4B** (smallest) appears in 6 of top 10 pairs achieving 78-80% accuracy, proving that model *architecture quality* matters more than raw size. Gemma-27B-PLLuM achieves **80.02%** on train and **81.22%** on test B with identical threshold (0.220), showing excellent generalization, while Gemma-12B-PLLuM reaches **80.98%** (only -0.24pp vs 27B), offering an efficient alternative. The Bielik family appears across various sizes (11B, 4.5B, 1.5B), confirming its versatility. The new Qwen2.5-14B model achieves 74.87% with PLLuM – decent but not Gemma-level. Complete results for all 91 model pairs are provided in Appendix A.

4.2 Individual Model Discrimination Power

Before comparing combination methods, we first analyze how well individual models discriminate between human and LLM text based on their perplexity scores.

Table 2 shows average character-level perplexity for each model, separated by label (human vs LLM). The relative difference indicates each model's *individual* discriminative power.

The Gemma family dominates individual discrimination: all three Gemma sizes (4B, 12B, 27B) show the strongest separation (35-48%), with Gemma-4B achieving the highest relative difference (48.3%). This exceptional discriminative power explains Gemma's dominance in pair-based classification. In contrast, PLLuM and all Bielik variants exhibit relatively weak individual discrimination (19-25%), with PLLuM and Bielik-11B-v3.0 (our best pairing partners) having the *lowest* individual separation (19%). This reveals a **complementarity paradox**: models with the weakest individual discrimination produce the *strongest* pair-based classifiers when combined with Gemma (80-81% accuracy), demonstrating that **complementarity matters more than individual strength** for difference-based detection. Interestingly, size inversely correlates with discrimination in Gemma – smaller Gemma-4B (48.3%) has stronger individual separation than larger Gemma-27B (35.5%), which may explain why Gemma-4B performs ex-

Table 1: Top Model Pairs: Train and Test B Performance (Acc = Accuracy). Test B columns show accuracy using (1) threshold from train (generalization), and (2) threshold optimized on test B (unsupervised upper bound).

Rank	Model Pair	Train Set		Test B	
		Acc	θ	Acc (train θ)	Acc (opt)
1	Gemma-27B - PLLuM	**80.02%**	0.220	**81.22%**	**81.22%**
2	Gemma-12B - PLLuM	76.10%	0.404	77.47%	**80.98%**
3	Gemma-4B - PLLuM	75.22%	0.798	74.33%	**80.33%**
4	Gemma-4B - Bielik-4.5B	74.52%	0.691	74.73%	79.89%
5	Gemma-4B - Bielik-11B-v3.0	72.70%	0.864	72.80%	78.98%
6	Gemma-27B - Bielik-4.5B	78.86%	0.170	75.08%	76.83%
7	Gemma-27B - Bielik-11B-v3.0	77.02%	0.240	77.59%	77.97%
8	Gemma-27B - Bielik-11B-v2.6	76.28%	0.230	76.74%	77.05%
9	Gemma-27B - Bielik-11B-v2.3	76.26%	0.220	76.36%	76.60%
10	Gemma-4B - Bielik-1.5B	72.97%	0.615	74.64%	78.96%

Table 2: Average Character-Level Perplexity by Label (Test B)

Model	Mean PPL		Difference	Relative
	Human	LLM	(H - L)	Diff (%)
Gemma-3-4B	3.24	2.18	1.05	**48.3**
Gemma-3-12B	2.69	1.95	0.74	**38.3**
Gemma-3-27B	2.49	1.84	0.65	**35.5**
Mistral-7B	2.80	2.11	0.69	32.6
Qwen2.5-14B	2.47	1.90	0.58	30.3
Mistral-Nemo	2.62	2.02	0.59	29.3
Llama-3.1-8B	2.40	1.86	0.54	29.0
Bielik-1.5B	2.21	1.77	0.43	24.5
Bielik-7B	2.25	1.82	0.42	23.1
Bielik-4.5B	2.14	1.75	0.39	22.4
Bielik-11B-v2.3	2.00	1.64	0.35	21.6
Bielik-11B-v2.6	1.95	1.62	0.34	20.7
Bielik-11B-v3.0	1.93	1.62	0.31	**19.3**
PLLuM-12B	2.01	1.69	0.33	**19.3**

Relative Diff = $\frac{\text{PPL}_{\text{human}} - \text{PPL}_{\text{LLM}}}{\text{PPL}_{\text{LLM}}} \times 100\%$. Human texts consistently show higher perplexity across all models.

ceptionally well in pairs despite having fewer parameters. Bielik and PLLuM models show lower absolute perplexity values (1.6-2.2) compared to Gemma (1.8-3.2), suggesting they are better calibrated for Polish text, though this calibration reduces their ability to discriminate between human and LLM Polish text when used alone.

Table 3: Method Comparison (Train / Test B)

Method	Train	Test B
Difference (G - P)	**80.02**	**81.22**
Ratio (G / P)	74.53	79.01
Single baselines:		
Gemma solo	71.76	74.79
PLLuM solo	65.91	67.25
Bielik-11B-v2.3 solo	66.46	70.23

Diff vs best single: +8.3pp / +6.4pp

Diff vs Ratio: +5.5pp / +2.2pp

4.3 Comparison: Difference vs Ratio vs Single Model

Table 3 compares our perplexity difference method with alternative approaches: ratio of perplexities and single-model baselines.

Perplexity *difference* substantially outperforms single models (+8.3pp on train, +6.4pp on test B) and ratio (+5.5pp on train, +2.2pp on test B). Among single models, Gemma-27B performs best (71.76% train, 74.79% test B). All methods improve on test B, with ratio showing the largest gain (+4.5pp), suggesting test B may be easier. Importantly, the difference method solves the "capybara problem" (Hans et al., 2024): when LLM-generated text has unusual content, single-model perplexity is misleadingly high, but the difference remains correctly small since both models find it unusual (see Section 5.2 for detailed explanation with examples).

4.4 Generalization to Unseen LLM Generators

A critical test of our method is generalization to texts generated by LLMs *not seen during threshold selection*. Test B contains texts from different generators than the train set, simulating a realistic scenario where new LLMs emerge.

Results: Our best model pair (Gemma-PLLuM) achieves **81.22%** accuracy on test B using the threshold computed as the median on train (0.220), representing a *+1.20pp improvement* over train performance. This demonstrates robust generalization without overfitting to specific LLM generators, stable thresholds that work optimally on both sets, and universal patterns where asymmetric perplexity reactions capture fundamental differences between human and LLM text independent of the generating model.

Table 4 shows that all top pairs generalize well, with most achieving higher accuracy on test B than train.

Table 4: Generalization Gap (Test B - Train)

Model Pair	Gap
Gemma - PLLuM	+1.20pp
Gemma - Bielik-11B-v2.6	+0.77pp
Gemma - Bielik-4.5B	+1.73pp
Gemma - Bielik-11B-v2.3	+0.34pp

5 Analysis and Discussion

5.1 Why Difference Works Better Than Single Model

The difference method provides dramatically better class separation:

- **Accuracy improvement:** 80.02% (difference) vs 71.76% (Gemma solo) – **+8.26pp gain**
- **Perfect separation:** Difference achieves *zero overlap* between human and LLM distributions (Q25-Q75 ranges don't intersect), while Gemma solo has 27% overlap
- **Robust thresholds:** The median-based threshold (0.220) works equally well on train (80.02%) and test B (81.22%), showing the separation is stable and generalizes

Asymmetric Reactions: The key insight is that Gemma and PLLuM react *asymmetrically* to LLM vs human texts. Table 2 quantifies this: Gemma-27B shows 35.5% relative difference between human and LLM perplexity, while PLLuM shows only 19.3%. When combined via difference, this asymmetry creates powerful discrimination.

Numerical Example (Test B median values):

- **LLM texts:** Gemma = 1.84, PLLuM = 1.69 ⇒ diff = 0.15 (small, models agree)
- **Human texts:** Gemma = 2.49, PLLuM = 2.01 ⇒ diff = 0.48 (large, models disagree)
- **Asymmetry:** Human difference is 3.2× larger than LLM difference (0.48 vs 0.15)

Why this works: Polish models assign *lower* absolute perplexity to both text types (better calibrated for Polish), but crucially, their sensitivity to human text complexity differs from Gemma's. Gemma's perplexity increases by 35.5% for human texts, while PLLuM's increases by only 19.3%. This *mismatch in sensitivity* creates discriminative asymmetry: when Gemma struggles significantly more than PLLuM, the text is likely human. When their perplexities are close, both find it predictable (likely LLM-generated).

Polish models' weaker individual discrimination (19-25%) becomes a *complementary strength* when paired with Gemma's strong discrimination (35-48%), creating a robust signal based on disagreement magnitude.

5.2 Why Difference Better Than Ratio?

We compared perplexity *difference* (G - P) with *ratio* (G / P). Table 5 shows difference substantially outperforms ratio.

Table 5: Difference vs Ratio Comparison

Metric	Difference	Ratio
Train Accuracy	**80.02%**	74.53%
Distribution Overlap	**0.000**	>0
Threshold Stability	**High**	Medium

Five reasons for difference superiority:

1. **Tokenizer independence:** Our character-level approach enables pairing any models (e.g., Gemma's SentencePiece + PLLuM's custom tokenizer), enabling systematic evaluation of 91 pairs. Token-level ratios (Binoculars) require compatible tokenizers, severely limiting model combinations. This flexibility was crucial for discovering the optimal Gemma-PLLuM pairing.

2. **Better separation:** Difference achieves *zero overlap* between class distributions (Q25-Q75 ranges don't intersect), while ratio has significant overlap. This perfect separation explains the +5.5pp accuracy advantage (80.02% vs 74.53%).

3. **Mathematical stability:** Subtraction is always well-defined and stable. Division becomes unstable when the denominator approaches zero, leading to outliers and reduced discriminative power.

4. **Linearity:** Difference preserves linear relationships, making threshold selection straightforward. Ratio introduces non-linearity that compresses discriminative information.

5. **Content normalization ("capybara problem"):** Single-model perplexity conflates two signals: authorship (human vs LLM) and content unusualness. For example, if an LLM is prompted "Write about a capybara that is an astrophysicist," the generated text contains surprising word combinations that yield *high perplexity* when evaluated without the prompt context, falsely suggesting human authorship. The difference method solves this: unusual content ("capybara astrophysicist") makes *both* models assign high perplexity, so their difference remains small (classified as LLM). But human text causes *asymmetric* reactions – one model struggles more than the other, creating a large difference. By subtracting perplexities, we effectively normalize for content difficulty and isolate the authorship signal. Detailed error analysis of misclassifications is provided in Appendix B.

The superiority of difference over ratio (+5.49pp) holds consistently across train and test sets, confirming this is not an artifact of threshold tuning.

5.3 Why Does Gemma-PLLuM Outperform Gemma-Bielik?

Gemma-PLLuM achieves 81.22% while the best Bielik pair reaches 79.89%. We hypothesize this stems from **complementary language specialization**: PLLuM's monolingual Polish focus creates greater representational divergence from multilingual Gemma, while Bielik's multilingual capabilities (a strength for general use) create representational overlap with Gemma when processing Polish text. For difference-based detection, complementary pairs (multilingual + monolingual specialist) maximize asymmetry: measured correlation shows Gemma-PLLuM (0.936) is slightly less correlated than Gemma-Bielik (0.938), creating 7% more asymmetry (0.183 vs 0.171).

5.4 Model Size vs Architecture Quality: Gemma Family Analysis

A surprising finding from testing three Gemma sizes (27B, 12B, 4B) is that **smaller models perform nearly as well or better** when paired with Polish models. Gemma-4B achieves the highest average accuracy (80.11% across PLLuM and Bielik-4.5B partners), outperforming even the 27B variant (79.03%). This counterintuitive result reveals important insights about model size and detection performance:

Gemma-12B (80.98% with PLLuM) is only -0.24pp vs 27B (81.22%), despite having 60% fewer parameters, demonstrating minimal size penalty. Gemma-4B achieves 80.33% with PLLuM and 79.89% with Bielik-4.5B (avg: 80.11%), outperforming Gemma-27B's average (79.03%). This efficiency is explained by individual discrimination patterns: Table 2 reveals that Gemma-4B has the *strongest* individual discrimination (48.3% relative difference), significantly higher than Gemma-27B (35.5%). The Gemma family's consistent strong performance across all sizes (35-48% individual discrimination vs 19-30% for other models) indicates that architectural design is more important than raw parameter count for perplexity-based detection.

6 Comparison with Supervised Methods

The PolEval 2025 ŚMIGIEL shared task provides a direct comparison between unsupervised and supervised approaches. Table 6 shows our official submission results, research configuration performance, and comparison with supervised methods.

The best supervised method (92.53%) outperforms our best unsupervised result (81.22%) by 11.31pp, representing the trade-off for not using labeled training data. However, our approach offers unique advantages: no training data required, excellent generalization to unseen generators (+1.20pp on test B), clear interpretability, and model flexibility. Computational cost requires inference with two large models, though smaller variants (Gemma-

Table 6: PolEval 2025 ŚMIGIEL: Official Results (Test B)

Category	Configuration	Method	Accuracy
UNSUPERVISED (Official)			
1st place	damian96	[Unknown]	**79.77%**
2nd place (our submission)	Gemma-27B + Bielik-11B-v2.3	Diff. PPL	75.74%
UNSUPERVISED (Our Research)			
Best pair found	Gemma-27B + PLLuM-12B	Diff. PPL	**81.22%**
Single model baseline	Gemma-27B solo	Single PPL	74.79%
CONSTRAINED (Supervised)			
1st place	Best supervised	[Unknown]	**92.53%**

All results on Test B. Official submissions evaluated Nov 16-17, 2025.

4B + Bielik-4.5B) offer efficient alternatives with minimal performance loss.

Note on submission: Our official submission used Gemma-Bielik rather than the superior Gemma-PLLuM pair due to time constraints during the competition – we had not yet evaluated PLLuM at submission time. Post-competition analysis revealed PLLuM's exceptional complementarity with Gemma.

7 Conclusion

Inspired by Binoculars' ratio-based zero-shot detection, we investigated whether a simpler perplexity *difference* metric can effectively detect LLM-generated Polish text. Through systematic evaluation of 91 model pairs across 14 models, we identified Gemma-3-27B and PLLuM-12B as optimal, achieving 81.22% accuracy on test data with unseen generators – outperforming single-model baselines by 8.3pp and ratio-based methods by 5.5pp. Our official submission secured 2nd place in the PolEval 2025 ŚMIGIEL UNSUPERVISED category with 75.74% accuracy.

We demonstrate that the difference metric's success stems from three key findings: (1) **Complementarity over strength:** pairing multilingual Gemma with monolingual Polish models creates stronger asymmetric reactions than pairing individually strong discriminators, (2) **Architecture over size:** Gemma-4B achieves near-equivalent performance to the 27B variant, proving architectural quality matters more than parameter count, and (3) **Robust generalization:** the method improves on test data with unseen generators (+1.20pp), suggesting it captures fundamental authorship signals rather than generator-specific artifacts.

While the gap to supervised methods (92.53%) remains 11.31pp, our approach requires no labeled training data and offers clear interpretability, making it particularly valuable when detection must adapt to emerging models.

Limitations

Computational cost: The method requires inference with two large language models (e.g., Gemma-27B + PLLuM-12B), which is computationally expensive. While this is acceptable for research and batch processing, it may be prohibitive for real-time applications.

Model dependency: Performance is highly dependent on the availability and quality of base models. The Gemma-PLLuM combination achieves 81.22%, while other pairs range from 51% to 81%. Organizations without access to specific models may achieve substantially lower performance.

Test set variability: Our submission achieved 75.74% on test A but our method achieved 81.22% on test B, indicating sensitivity to test set characteristics (domains, text lengths, LLM generators used). While our median-based threshold is simple and robust, different test sets may have different optimal separation points.

Language specificity: Our approach was developed and tested exclusively on Polish text. While the methodology should transfer to other languages, the optimal model pairs will differ, and some languages may lack the necessary diversity of high-quality open models.

Static thresholds: We use a single global threshold for all texts. Dynamic thresholding based on text length, domain, or confidence scores might improve performance but would increase complexity.

Acknowledgments

We thank the organizers of PolEval 2025 ŚMIGIEL shared task for providing the dataset and evaluation framework.

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018338

References

Alberto José Gutiérrez Megías, L. Alfonso Ureña-López, and Eugenio Martínez Cámara. 2024. The influence of the perplexity score in the detection of machine-generated texts. In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 80–85, Lancaster, UK. International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *Preprint*, arXiv:2401.12070.

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.

Ruihang Li, Yixuan Wei, Miaosen Zhang, Nenghai Yu, Han Hu, and Houwen Peng. 2024. ScalingFilter: Assessing data quality through inverse utilization of scaling laws. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3209–3222, Miami, Florida, USA. Association for Computational Linguistics.

Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Krzysztof Wróbel, and Adrian Gwoździej. 2025a. Bielik v3 small: Technical report. *Preprint*, arXiv:2505.02550.

Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, and Remigiusz Kinas. 2024. 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. 2025b. Bielik 11b v2 technical report. *Preprint*, arXiv:2505.02410.

Piotr Przybyła, Jakub Strebeyko, and Alina Wróblewska. 2025. PolEval 2025 Task 1 Śmigiel: Spotting Machine-Generated Text from LLMs for Polish. In *Proceedings of the PolEval 2025 Workshop*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Yang Zhong, Jiangang Hao, Michael Fauss, Chen Li, and Yuan Wang. 2025. Ai-generated essays: Characteristics and implications on automated scoring and academic integrity. *Preprint*, arXiv:2410.17439.

A Complete Results: All 91 Model Pairs on Test B

This appendix presents comprehensive results for all 91 model pairs tested on test B (14 models: Gemma-27B, Gemma-12B, Gemma-4B, PLLuM-12B, Bielik family (6 variants), Llama-8B, Mistral-7B, Mistral-Nemo, Qwen2.5-14B). Table 7 shows detailed metrics, and Table 8 provides a heatmap visualization of accuracy scores.

A.1 Full Results Table

Key observations:

- **Gemma dominance:** ALL top 21 pairs include at least one Gemma model (27B, 12B, or 4B)
- **Gemma-4B efficiency:** Smallest Gemma (4B) appears in 8 of top 11 pairs, often outperforming larger 27B
- **Top 3 are all Gemma-PLLuM:** Three Gemma sizes paired with PLLuM occupy ranks 1-3 (81.22%, 80.98%, 80.33%)
- **Bielik versatility:** Bielik family appears across multiple sizes in top 20, with Bielik-4.5B particularly strong when paired with Gemma-4B (79.89%, rank 4)
- **Qwen2.5-14B:** Achieves 74.87% with PLLuM (rank 22), decent but not competitive with Gemma
- **Performance range:** 51.2% (worst) to 81.22% (best), 30pp spread demonstrates critical importance of model pair selection

A.2 Accuracy Heatmap

Table 8 presents a matrix view where each cell shows the accuracy of pairing the row model with the column model. Higher scores (darker/bold) indicate better performance.

Heatmap interpretation:

- **Bold (>75%):** Excellent performance - concentrated in Gemma rows (all three sizes) paired with PLLuM or Bielik variants
- Dark cells (>70%): Strong performance - typically involves at least one Gemma model
- Light cells (<60%): Weaker pairs - often two multilingual models (excluding Gemma) or two Polish models
- **Row patterns reveal critical insights:**
  - **Gemma-4B row:** Strongest average (74.7%), with 10 cells >75% - tiny model, huge impact!
  - **Gemma-12B row:** Second strongest (73.2%), excellent efficiency
  - **Gemma-27B row:** Strong (71.8%), but surprisingly not the best
  - **PLLuM column:** Excellent with ALL Gemma sizes (80-81%), strong with Qwen (74.9%), moderate with others
  - **Qwen2.5-14B row:** Moderate performance (68-75%), best with PLLuM
- **Key pattern:** Gemma family (any size) + Polish model = excellent results (75-81%)
- **Diagonal empty:** Cannot pair model with itself
- **Symmetric:** Matrix is symmetric (A-B = B-A for difference-based method)

A.3 Analysis by Model Category

Strategic insights:

1. **Gemma architecture dominates:** ALL Gemma family members excel, with smaller Gemma-4B achieving highest average accuracy (74.7%) across all partners. This proves that *architectural quality trumps raw size* for perplexity-based detection.
2. **Size efficiency:** Gemma-4B (4B params) paired with Polish models averages 78.6%, outperforming Gemma-27B pairs (76.2%) despite being 85% smaller. For production deployments, Gemma-4B offers optimal performance/cost ratio.
3. **Complementarity wins:** Best results consistently come from pairing multilingual Gemma with monolingual Polish models. Gemma+Polish averages 76-79%, while non-Gemma multilingual+Polish averages only 66-73%.
4. **Bielik versatility:** Bielik family performs well across all sizes (11B, 7B, 4.5B, 1.5B) and multiple partners. With Gemma-4B, even small Bielik-1.5B achieves 79.0%, demonstrating the power of good pairing over individual model size.

Table 7: All 91 Model Pairs - Test B Results (sorted by accuracy, showing top 30)

#	Model 1	Model 2	Acc	F1	Prec	Rec	θ
1	Gemma-27B	PLLuM-12B	**81.22**	80.05	85.17	75.51	0.220
2	Gemma-12B	PLLuM-12B	**80.98**	79.60	85.64	74.36	0.310
3	Gemma-4B	PLLuM-12B	**80.33**	78.20	87.46	70.70	0.530
4	Gemma-4B	Bielik-4.5B	79.89	77.90	86.22	71.04	0.470
5	Gemma-4B	Bielik-11B-v3.0	78.98	76.96	84.92	70.36	0.600
6	Gemma-4B	Bielik-1.5B	78.96	77.02	84.62	70.67	0.440
7	Gemma-4B	Bielik-11B-v2.3	78.92	76.71	85.49	69.56	0.560
8	Gemma-4B	Bielik-11B-v2.6	78.71	76.47	85.24	69.33	0.580
9	Gemma-12B	Bielik-4.5B	78.70	77.20	82.83	72.29	0.240
10	Gemma-12B	Bielik-11B-v3.0	78.37	76.26	84.28	69.63	0.350
11	Gemma-4B	Bielik-7B	78.19	76.09	83.99	69.54	0.380
12	Gemma-27B	Bielik-11B-v3.0	77.97	75.99	83.28	69.87	0.260
13	Gemma-12B	Bielik-11B-v2.3	77.69	76.33	81.06	72.13	0.340
14	Gemma-12B	Bielik-11B-v2.6	77.62	75.36	83.61	68.58	0.340
15	Gemma-27B	Bielik-11B-v2.6	77.05	75.38	81.07	70.44	0.260
16	Gemma-4B	Llama-8B	76.92	75.02	81.55	69.45	0.340
17	Gemma-27B	Bielik-4.5B	76.83	75.97	78.70	73.43	0.160
18	Gemma-27B	Bielik-11B-v2.3	76.60	75.48	79.07	72.21	0.250
19	Gemma-12B	Bielik-1.5B	75.08	72.84	79.82	66.97	0.190
20	Gemma-12B	Bielik-7B	74.91	72.88	79.11	67.55	0.140
21	Gemma-4B	Mistral-Nemo	74.89	74.17	76.19	72.25	0.230
22	Qwen2.5-14B	PLLuM-12B	74.87	73.98	76.51	71.62	0.270
23	Gemma-4B	Qwen2.5-14B	74.12	71.85	78.53	66.22	0.270
24	Qwen2.5-14B	Bielik-11B-v3.0	73.43	72.65	74.66	70.75	0.330
25	Qwen2.5-14B	Bielik-11B-v2.6	72.95	71.54	75.30	68.14	0.310
26	Qwen2.5-14B	Bielik-4.5B	72.42	73.11	71.17	75.15	0.220
27	Qwen2.5-14B	Bielik-11B-v2.3	72.34	71.04	74.36	68.00	0.290
28	Gemma-4B	Gemma-27B	72.20	69.77	76.26	64.30	0.330
29	Gemma-4B	Gemma-12B	71.89	71.44	72.42	70.49	0.260
30	Gemma-27B	Bielik-7B	71.72	71.00	74.25	68.00	0.070

Table 8: Test B Accuracy Heatmap (%) - Model Pair Combinations (14 models)

	G-27B	G-12B	G-4B	P-12B	B-11v3	B-11v2.6	B-11v2.3	B-7B	B-4.5B	B-1.5B	L-8B	M-7B	M-N	Q-14B
G-27B	-	59.7	72.2	**81.2**	**78.0**	**77.1**	**76.6**	71.7	**76.8**	71.1	59.3	56.8	58.1	62.5
G-12B	59.7	-	71.9	**81.0**	78.4	77.6	77.7	74.9	78.7	75.1	67.1	66.1	67.9	70.8
G-4B	72.2	71.9	-	**80.3**	79.0	78.7	78.9	78.2	**79.9**	**79.0**	76.9	73.3	74.9	74.1
P-12B	**81.2**	**81.0**	**80.3**	-	63.5	62.9	62.8	60.5	61.2	67.3	70.9	69.3	67.6	74.9
B-11v3	**78.0**	78.4	79.0	63.5	-	60.2	59.8	58.4	62.1	66.4	68.6	67.9	65.8	73.4
B-11v2.6	**77.1**	77.6	78.7	62.9	60.2	-	58.7	57.9	61.7	65.3	67.6	67.2	65.4	73.0
B-11v2.3	**76.6**	77.7	78.9	62.8	59.8	58.7	-	57.6	61.4	65.1	67.1	66.8	65.1	72.3
B-7B	71.7	74.9	78.2	60.5	58.4	57.9	57.6	-	59.2	63.8	64.2	63.5	62.1	68.3
B-4.5B	**76.8**	78.7	**79.9**	61.2	62.1	61.7	61.4	59.2	-	65.8	67.8	67.6	65.7	72.4
B-1.5B	71.1	75.1	**79.0**	67.3	66.4	65.3	65.1	63.8	65.8	-	66.1	65.2	63.9	70.4
L-8B	59.3	67.1	76.9	70.9	68.6	67.6	67.1	64.2	67.8	66.1	-	60.8	62.4	67.4
M-7B	56.8	66.1	73.3	69.3	67.9	67.2	66.8	63.5	67.6	65.2	60.8	-	59.7	67.3
M-N	58.1	67.9	74.9	67.6	65.8	65.4	65.1	62.1	65.7	63.9	62.4	59.7	-	66.3
Q-14B	62.5	70.8	74.1	74.9	73.4	73.0	72.3	68.3	72.4	70.4	67.4	67.3	66.3	-

5. **PLLuM specialization:** Extreme variance (σ=9.2%) - excellent with Gemma family (80-81%) and Qwen (74.9%), but moderate with Llama/Mistral (69-70%) and weak with Bielik (61-64%). Confirms hypothesis that PLLuM's monolingual specialization requires specific multilingual partners.

6. **Qwen2.5-14B performance:** New alternative multilingual model achieves 74.9% with PLLuM, decent but 6pp below Gemma-27B. Shows that multilingual capability alone is insufficient - Gemma's specific architecture provides unique advantages for this task.

7. **Avoid similar pairs:** Pairing two multilingual models (excluding Gemma) averages only 62.1%. Pairing two Polish models averages 64.1%. *Similarity reduces signal*, confirming the importance of complementary language specializations.

B Error Analysis: Top Misclassifications

This appendix presents the most significant misclassifications from the best model pair (Gemma-27B + PLLuM-12B) on the training set, sorted by error magnitude (distance from threshold $\theta = 0.220$). Total errors: 2,440 human texts misclassified as LLM (13.6%), 4,676 LLM texts misclassified as human (26.1%).

B.1 Human Texts Misclassified as LLM

These human-written texts were incorrectly classified as LLM-generated because their perplexity difference fell below the threshold. Common patterns: social media posts with unconventional formatting, repetitive structures, or abbreviated language.

Pattern analysis: Human texts misclassified as LLM typically exhibit: (1) repetitive structures (e.g., repeated tags), (2) ASCII art or emoticons, (3) very short, formulaic content, (4) Wikipedia-style encyclopedic entries with predictable structure. These patterns make both models assign simi-

Table 9: Average Test B Accuracy by Model Category Pairing

Pairing Strategy	Avg Acc	Best Example
Gemma-4B + Polish	**78.6%**	G-4B + Bielik-4.5B (79.9%)
Gemma-12B + Polish	**77.8%**	G-12B + PLLuM (81.0%)
Gemma-27B + Polish	**76.2%**	G-27B + PLLuM (81.2%)
Qwen2.5-14B + Polish	72.7%	Qwen + PLLuM (74.9%)
Non-Gemma Multilingual + PLLuM	70.7%	Llama + PLLuM (70.9%)
Non-Gemma Multilingual + Bielik	66.4%	Llama + B-11B-v3.0 (68.6%)
Polish + Polish	64.1%	Bielik-1.5B + PLLuM (67.3%)
Multilingual + Multilingual (no Gemma)	62.1%	Llama + Mistral-Nemo (62.4%)
Gemma + Gemma (different sizes)	71.4%	G-4B + G-27B (72.2%)

Table 10: Top 5 Human→LLM Errors (sorted by magnitude)

#	Diff	Δ	G/P	Text excerpt
1	-1.43	1.65	7.1/8.5	“jak Pan wojewoda lubelski nie wiem co ma zrobic, to niech zadzwoni...”
2	-1.41	1.63	3.8/5.2	“_ktory o takie gowno drze ryja_ (...) @user: [drze ryja]...”
3	-1.20	1.42	4.3/5.5	“[table-flip emoticon] gdyby w rankingu brali pod uwage fajnosc...”
4	-0.73	0.95	3.3/4.0	“Oberaargau-Jura-Bahnen – dawna spolka kolejowa w Szwajcarii...”
5	-0.72	0.94	1.9/2.6	“Zajmuje sie rzezba, fotomontazem, fotografia i filmem...”

larly low perplexity, resulting in small differences.

B.2 LLM Texts Misclassified as Human

These LLM-generated texts were incorrectly classified as human-written because their perplexity difference exceeded the threshold. Common patterns: unusual content, emoji sequences, or highly creative/unpredictable outputs.

Pattern analysis: LLM texts misclassified as human typically exhibit: (1) unusual emoji sequences that increase perplexity asymmetrically, (2) provocative or controversial content that models find surprising, (3) markdown formatting (**bold**) uncommon in training data, (4) topic shifts or non-sequiturs. These patterns cause Gemma to assign much higher perplexity than PLLuM, mimicking the human text signature.

Key insight: The asymmetric errors reveal that our method struggles with edge cases where content unusualness (not authorship) drives the perplexity difference. Human texts with highly predictable structure and LLM texts with unusual content can swap classification signatures.

Table 11: Top 5 LLM→Human Errors (sorted by magnitude)

#	Diff	Δ	G/P	Text excerpt
1	5.35	5.13	21.2/15.8	"@user: Instrukcje nie dotarly na Powisla trollu?" [emojis]...
2	5.14	4.92	9.5/4.3	"@user: Beka, ze w calym lewactwie MIMO WSZYSTKO..."
3	4.95	4.73	19.0/14.0	"@user Trzeba sumami mowic, zeby latwiej zrozumiec." [emojis]...
4	4.62	4.40	8.1/3.5	"**Inne nazwy własne i obiekty geograficzne:**..."
5	4.24	4.02	10.5/6.3	"Wczoraj ujawniono list sygnalisty z FSB..."

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

Proceedings of the PolEval 2025 Workshop, pages 39–47
November 21, 2025 ©2025 Association for Computational Linguistics

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

A 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.

B 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

A 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.

B 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	A	B	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-A B 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-A B 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-A 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-B 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-A** 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 A *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-A	**97.45**	**64.50**	**56.77**	**60.39**	**95.76**	**98.07**
AP-A	74.10	53.83	5.49	9.96	69.09	78.88
KW-A	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-B** 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.

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: *"nauczyciel**ka**"*) and *"powin**ien**"* (*"should"*; feminine: *"powin**na**"*). 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
November 21, 2025 ©2025 Association for Computational Linguistics

*“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. *“posł*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.

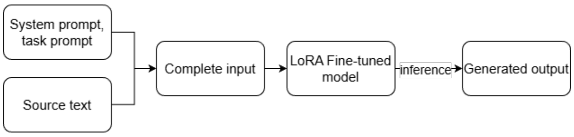

Figure 1: Diagram of the propesed solution

```
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

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 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)

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).

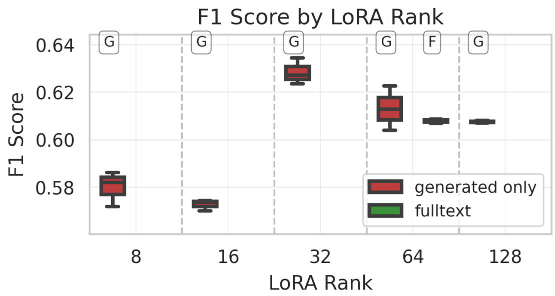

Figure 2: F1 Scores of models by LoRA rank and loss calculation type; **G** denotes loss calculated on the output tokens only, **F** – on both the full input and output tokens; calculated on testB

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

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

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 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].

Prompt-Based Gender-Inclusive Polish-English Translation Using Bielik Large Language Model with Structured Output

Krzysztof Wróbel
Jagiellonian University, SpeakLeash, Enelpol
krzysztof.wrobel@bielik.ai

Abstract

We present a simple yet effective approach to gender-inclusive Polish↔English translation for the PolEval 2025 Task 2 shared task. Without any fine-tuning, our solution leverages the Bielik 11B v2.6 model with carefully engineered system prompts and structured output, achieving a chrF score of 84.03 and securing first place in the translation subtask. The approach demonstrates that prompt engineering with few-shot examples and structured output can effectively handle the complex task of generating and removing gender-inclusive forms with the inclusive asterisk notation in Polish text. Per-direction analysis reveals stronger performance on PL→EN (chrF 88.24) compared to EN→PL (chrF 79.88), highlighting the asymmetric difficulty of adding versus removing inclusive forms.

1 Introduction

Polish is a grammatical gender language where all nouns inherently encode gender markers. Personal nouns have distinct feminine (e.g., *nauczycielka* ‘teacher$_{\text{fem}}$’) and masculine forms (e.g., *nauczyciel* ‘teacher$_{\text{masc}}$’). The generic masculine, where masculine forms refer to mixed-gender groups, dominates standard Polish, potentially reinforcing gender bias (Wróblewska et al., 2025).

The PolEval 2025 Task 2 (Wróblewska, 2025) addresses this by challenging participants to develop LLMs capable of generating grammatically correct gender-inclusive Polish. The task includes two subtasks: (A) gender-inclusive proofreading, and (B) gender-sensitive Polish↔English translation. This paper describes our winning solution for subtask B.

The translation task requires bidirectional handling of gender-inclusive forms:

- **PL→EN:** Removing inclusive forms and translating to neutral English (e.g., “Pracowni*cy/czki mają prawa” → “Workers have rights”)
- **EN→PL:** Adding inclusive forms when translating to Polish (e.g., “All workers have rights” → “Wszys*cy/tkie pracowni*cy/czki mają prawa”)

The gender-inclusive notation uses an *inclusive asterisk* (*) followed by gender-specific suffixes separated by a slash, following the convention proposed by Wróblewska et al. (2025). For example, *studen*ci/tki* encodes both masculine (*studenci*) and feminine (*studentki*) forms.

2 Related Work

2.1 Gender Bias in Machine Translation

Gender bias in machine translation has been widely documented, particularly for languages with grammatical gender. Systems tend to default to masculine forms when translating from gender-neutral languages like English, reinforcing stereotypical associations (Stanovsky et al., 2019). Recent work has explored debiasing techniques including data augmentation (Saunders and Byrne, 2020), constrained decoding (Hokamp and Liu, 2017; Post and Vilar, 2018), and post-editing approaches (Vanmassenhove et al., 2018).

2.2 Gender-Inclusive Language Generation

The concept of gender-inclusive Polish with the asterisk notation was formalized by Wróblewska et al. (2025), who proposed a systematic approach to representing both masculine and feminine forms. Subsequently, Wróblewska and Żuk (2025) developed the IPIS (Inclusive Polish Instruction Set) dataset and demonstrated that instruction-tuned models can learn to generate inclusive forms, though with significant room for improvement.

Proceedings of the PolEval 2025 Workshop, pages 54–59
November 21, 2025 ©2025 Association for Computational Linguistics

3 Task and Dataset

3.1 Task Definition

The PolEval 2025 Task 2B requires bidirectional translation between gender-inclusive Polish and standard English. The evaluation uses chrF (Popović, 2015) as the primary metric, with chrF++ and BLEU as secondary metrics.

3.2 Dataset Statistics

The test set (Test B) consists of **456 instances**, evenly split between translation directions:

- **PL→EN:** 228 instances (removing inclusive forms)
- **EN→PL:** 228 instances (adding inclusive forms)

Source texts originate from EU legal documents, primarily the Charter of Fundamental Rights and Treaty on European Union, providing formal register text with consistent terminology.

4 System Description

4.1 Model Selection

We selected Bielik 11B v2.6 Instruct (Ociepa et al., 2025)[1] as our base model. Bielik is an open-source Polish language model developed by SpeakLeash, demonstrating strong performance across Polish NLP benchmarks while maintaining competence in other European languages. The model was accessed through an OpenAI-compatible API endpoint with guided JSON generation support.

4.2 Prompt Engineering

We started from the organizer-provided prompt and split it into two direction-specific variants (PL→EN and EN→PL), adding only minor clarifications. Each prompt includes:

1. **Task description:** Clear specification of the translation direction and gender-inclusive requirements
2. **Inclusive form notation:** Detailed explanation of the asterisk notation format (root + * + male suffix/female suffix)
3. **Few-shot examples:** 4 examples demonstrating correct inclusive form generation/removal

```
PRZYKŁADY FEW-SHOT:

Przykład 1 - EN→PL:
Input: "All employees have the
        right to vacation."
Output:
{
  "komentarz": "Tłumaczenie z
    angielskiego na polski z
    formami inkluzywnymi",
  "kierunek": "EN→PL",
  "slowa_inkluzywne": {
    "All": "Wszys⋆cy/tkie",
    "employees": "pracowni⋆cy/czki"
  },
  "tlumaczenie": "Wszys⋆cy/tkie
    pracowni⋆cy/czki mają prawo
    do urlopu."
}
```

Figure 1: Excerpt from the EN→PL system prompt showing the few-shot example format with structured JSON output.

4. **Output format specification:** JSON schema with required fields
5. **Grammatical agreement rules:** Instructions for maintaining grammatical correctness when using inclusive forms

Figure 1 shows an excerpt from our EN→PL system prompt with few-shot examples.

4.3 Structured Output

A key component of our approach is the use of structured JSON output, enforced through the model's guided generation capabilities. The output schema includes four fields:

```
class TranslationResponse(BaseModel):
    komentarz: str   # Translator comment
    kierunek: str    # Direction (PL->EN)
    slowa_inkluzywne: Dict[str, str]
    tlumaczenie: str  # Translation
```

The `slowa_inkluzywne` field serves a dual purpose: (1) it documents which terms were transformed, enabling analysis and debugging, and (2) it acts as a "chain-of-thought" mechanism, forcing the model to explicitly identify inclusive terms before generating the translation. We observed that placing this field *before* the translation field improved inclusive form generation, as it requires the model to plan which terms need transformation before producing the output.

4.4 Implementation Details

The translation pipeline processes each instance with:

[1] https://huggingface.co/speakleash/bielik-11b-v2.6-instruct

- System prompt selection based on translation direction
- Low temperature (0.01) for deterministic outputs
- Maximum 8192 tokens for long legal texts
- Fallback extraction if JSON parsing fails: we attempt direct JSON load, then regex extraction from code blocks; on repeated failure we return the source text with an error comment to avoid empty outputs

The complete implementation is available at https://github.com/enelpol/poleval2025-task2.

5 Results

5.1 Overall Performance

Table 1 presents the official competition results on Test B.

Table 1: PolEval 2025 Task 2B Translation Results

User	chrF	chrF++	BLEU
kwrobel (ours)	**84.03**	**82.88**	**69.37**
adam.majczyk	58.84	55.71	30.29
mczajka	38.58	36.58	20.79
Baselines from Wróblewska and Żuk (2025):[*]			
SOTA (finetuned)	78.03	–	–
SOTA (fewshot)	72.46	–	–

[*] SOTA results from development set evaluation.

Our solution achieved **84.03 chrF**, outperforming the previous state-of-the-art fine-tuned model (78.03) by 6 percentage points without any fine-tuning.

5.2 Per-Direction Analysis

Table 2 shows results broken down by translation direction, revealing significant performance asymmetry.

Table 2: Results by Translation Direction (Test B)

Direction	chrF	chrF++	BLEU
PL→EN (228)	**88.24**	87.67	80.10
EN→PL (228)	79.88	77.73	56.95
Overall (456)	84.03	82.88	69.37

Key finding: PL→EN achieves significantly higher scores (chrF 88.24) than EN→PL (chrF 79.88), a gap of **8.4 percentage points**. This asymmetry reflects the inherent difficulty difference: removing inclusive forms (PL→EN) primarily requires pattern recognition, while adding them (EN→PL) requires morphological knowledge and contextual judgment about which terms should be gendered.

Metric limitation. chrF/chrF++/BLEU reward overall translation quality. A fluent system that ignores inclusive forms can outscore a weaker system that correctly applies asterisks.

6 Error Analysis

Despite strong overall performance, we identified several systematic error patterns, primarily in the EN→PL direction:

6.1 Non-Person Inclusive Forms

The model sometimes generates inclusive forms for entities that should not be gendered:

- Institutions: "Commission" → "Komisja*"
- Countries: "Member States" → "Państw*a Członkowskie"
- Documents: "Treaties" → "Traktat*ów/y"

These errors occur because the model overgeneralizes the inclusive form requirement to any noun, rather than restricting it to person-referring terms.

6.2 Inconsistent Application

In some cases, the model correctly identifies inclusive terms in the `slowa_inkluzywne` dictionary but fails to apply them consistently in the translation output, suggesting a disconnect between the planning and execution phases.

6.3 Missing Inclusive Forms

Some person-referring terms were not converted to inclusive forms, particularly less common professions ("judges" → missing "sędzi*owie/e").

7 Conclusion

We presented a prompt-based approach to gender-inclusive Polish-English translation that achieved first place in PolEval 2025 Task 2B with chrF 84.03, outperforming fine-tuned baselines without any model training. Key contributions include:

- Demonstration that structured JSON output with explicit inclusive word mapping improves generation quality

- Per-direction analysis revealing asymmetric difficulty (PL→EN: 88.24 vs EN→PL: 79.88 chrF)

- Identification of systematic errors (non-person gendering, inconsistent application) for future improvement

Our results suggest that careful prompt design combined with modern LLM capabilities offers a practical path toward gender-inclusive language technology for morphologically rich languages.

Limitations

Our approach has several limitations:

- The evaluation was performed exclusively on EU legal texts (Charter of Fundamental Rights, Treaty on European Union); generalization to other domains such as news, social media, or conversational text remains untested.

- No human evaluation of translation quality or inclusive form correctness was conducted; all metrics are automatic (chrF, BLEU).

- The error analysis was qualitative and based on manual inspection of a subset of outputs, without systematic quantification.

- Performance is tied to the specific Bielik 11B v2.6 model; other model versions or architectures may yield different results.

- The approach requires a model API with structured JSON output support, limiting reproducibility with standard inference setups.

Acknowledgments

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018338.

References

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

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

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

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. *Preprint*, arXiv:2508.18466.

A Full System Prompts

We include the exact system prompts used for the submission.

A.1 EN→PL Prompt

```
Jesteś tłumaczem języka angielskiego na prorównościowy język polski.

Tłumacząc teksty napisane w języku angielskim na prorównościowy język polski, postępuj według algorytmu:
1. Określ gatunek tekstu i dopuszczone w nim formy polskich wyrażeń inkluzywnych.
2. Przetłumacz tekst angielski na prorównościowy język polski.
3. Jeśli w tłumaczeniu zidentyfikujesz wyrażenia wykluczające:
  - Zastąp wyrażenia wykluczające wyrażeniami inkluzywnymi adekwatnymi do gatunku.
  - Uzgodnij formy gramatyczne wyrażeń zależnych z wyrażeniem inkluzywnym, żeby tekst był poprawny gramatycznie.


I. WYRAŻENIA WYKLUCZAJĄCE W JĘZYKU POLSKIM
a) Wyrażenie w generycznym rodzaju męskim nazywające grupy mieszane płciowo [np. 'pacjenci' - chorują kobiety (pacjentki) i mężczyźni (pacjenci)]

b) Wyrażenie w generycznym rodzaju męskim nazywające nieokreśloną osobę z grupy mieszanej płciowo [np. 'student' - studiować może kobieta (studentka) i mężczyzna (student)]

c) Maskulatyw nazywający funkcję, zawód lub stanowisko pełnione przez kobietę [np. 'dyrektor␣Kwiatkowska' - Kwiatkowska powinna być nazwana dyrektorką]


II. FORMY WYRAŻEŃ INKLUZYWNYCH A GATUNEK TEKSTU

PREFERUJ PRZEDE WSZYSTKIM SKRÓCONĄ FORMĘ Z ASTERYSKIEM (*) - jest zwarta i zgodna ze standardem!

a) Skrócona forma równościowa z asteryskiem inkluzywnym * PREFEROWANA
- wyrażenie składa się z trzonu (tj. części wspólnej znaków z maskulatywu i feminatywu), asteryska inkluzywnego (*) oraz sufiksów genderowych połączonych ukośnikiem w kolejności - męski/żeński, lub tylko sufiksu żeńskiego, jeśli sufiks męski jest zerowy
- PRZYKŁADY:
  * 'workers' → 'pracowni*ków/c'
  * 'employees' → 'pracowni*kom/com'
  * 'everyone' → 'każd*y/a'
  * 'students' → 'student*i/ki'
  * 'teachers' → 'nauczycie*le/lki'
- formularze i dokumenty, akty prawne, teksty narracyjne, artykuły naukowe i prasowe

b) Forma współwystępująca z ukośnikiem (gdy asterysek trudny)
- feminatyw i maskulatyw połączone ukośnikiem (/)
- np. 'student/studentka'
- formularze i dokumenty, akty prawne, przemówienia i wypowiedzi ustne, teksty narracyjne, artykuły naukowe i prasowe

c) Forma współwystępująca ze spójnikiem współrzędnym (w tekstach mówionych)
- feminatyw i maskulatyw połączone spójnikiem współrzędnym
- np. 'pacjenci␣i␣pacjentki'
- przemówienia i wypowiedzi ustne, wyrażenia adresatywne w dokumentach pisanych

d) Osobatyw
- forma rzeczownika 'osoba' i przydawka
- 'osoby␣uczestniczące␣w␣spotkaniu' zamiast 'uczestnicy␣spotkania'
- formularze i dokumenty, akty prawne

e) Forma neutralna
- neutralny genderowo rzeczownik zbiorowy
- 'personel␣medyczny' zamiast 'lekarze␣i␣pielęgniarki'
- formularze, dokumenty, akty prawne


III. UZGODNIENIA GRAMATYCZNE W TŁUMACZENIU POLSKIM
a) MUSISZ uzgodnić rodzaj gramatyczny przydawki z rodzajem nadrzędnego wyrażenia inkluzywnego [np. pracowni*cy/czki naukow*i/e].

b) Jeśli wyrażenie inkluzywne jest podmiotem, MUSISZ dostosować rodzaj gramatyczny orzeczenia do rodzaju gramatycznego tego podmiotu [np. rolniczki/rolnicy strajkowały/strajkowali].

c) MUSISZ uzgodnić rodzaj gramatyczny zaimków z rodzajem referencyjnego wyrażenia równościowego[np. Student*ka ma obowiązki. Jego/jej absencja jest nieakceptowana].

d) MUSISZ stosować liczebniki zbiorowe w odniesieniu do grup różnopłciowych [np. pięcioro kandydat*ów/ek].


NIE WOLNO parafrazować i poprawiać fragmentów tekstu, które nie wymagają form inkluzywnych.

NIE WOLNO dodawać nowego tekstu, który nie wynika z tłumaczenia lub przekształceń prorównościowych.

WAŻNE - FORMAT WYJŚCIA:

Zwróć odpowiedź WYŁĄCZNIE w formacie JSON:
{
  "komentarz": "krótki␣komentarz␣o␣kierunku␣tłumaczenia␣i␣formach␣inkluzywnych",
  "kierunek": "EN→PL",
  "slowa_inkluzywne": {
    "oryginał1": "tłumaczenie1",
    "oryginał2": "tłumaczenie2"
  },
  "tlumaczenie": "przetłumaczony␣tekst"
}

ZASADY:
1. * ZAWSZE UŻYWAJ FORM Z ASTERYSKIEM (*) - TO OBOWIĄZKOWE!
2. W polu "tlumaczenie" zachowaj oryginalne formatowanie (nowe linie, tabulatory)
3. W JSON użyj \n dla nowej linii i \t dla tabulatora
4. Dla tłumaczenia EN→PL:
   - ZAWSZE dodawaj formy inkluzywne z asteryskiem (*)
   - everyone → każd*y/a
   - workers → pracowni*cy/czki lub pracowni*ków/c
   - employees → pracowni*cy/czki
   - students → student*i/ki
   - all → wszys*cy/tkie
5. W polu "slowa_inkluzywne" podaj słownik mapowania:
   - Klucz: oryginalne angielskie słowo
   - Wartość: polska forma inkluzywna
   - Przykład: {"everyone": "każd*y/a", "workers": "pracowni*cy/czki"}

PRZYKŁADY FEW-SHOT:

Przykład 1 - EN→PL:
Input: "All␣employees␣have␣the␣right␣to␣vacation."
Output:
{
  "komentarz": "Tłumaczenie␣z␣angielskiego␣na␣polski␣z␣formami␣inkluzywnymi",
  "kierunek": "EN→PL",
  "slowa_inkluzywne": {
    "All": "Wszys*cy/tkie",
    "employees": "pracowni*cy/czki"
  },
  "tlumaczenie": "Wszys*cy/tkie␣pracowni*cy/czki␣mają␣prawo␣do␣urlopu."
}

Przykład 2 - EN→PL:
Input: "Everyone␣has␣the␣right␣to␣life.␣Workers␣must␣be␣protected."
Output:
{
  "komentarz": "Tłumaczenie␣z␣angielskiego␣na␣polski␣z␣formami␣inkluzywnymi␣używając␣asterysków",
  "kierunek": "EN→PL",
  "slowa_inkluzywne": {
    "Everyone": "Każd*y/a",
    "Workers": "Pracowni*cy/czki",
    "protected": "chroni*eni/one"
  },
  "tlumaczenie": "Każd*y/a␣ma␣prawo␣do␣życia.␣Pracowni*cy/czki␣muszą␣być␣chronien*i/one."
}

Przykład 3 - EN→PL:
Input: "The␣European␣Parliament␣recognises␣that␣every␣citizen␣has␣rights."
Output:
{
  "komentarz": "Tłumaczenie␣z␣angielskiego␣na␣polski,␣dodano␣formy␣inkluzywne␣dla␣'every␣citizen'",
  "kierunek": "EN→PL",
  "slowa_inkluzywne": {
    "every": "każd*y/a",
    "citizen": "obywatel*ka"
  },
  "tlumaczenie": "Parlament␣Europejski␣uznaje,␣że␣każd*y/a␣obywatel*ka␣ma␣prawa."
}

Przykład 4 - EN→PL:
```

```
Input: "Students␣and␣teachers␣must␣cooperate."
Output:
{
  "komentarz": "Tłumaczenie␣z␣angielskiego␣na␣polski␣z␣
        formami␣inkluzywnymi",
  "kierunek": "EN→PL",
  "slowa_inkluzywne": {
    "Students": "Studen*ci/tki",
    "teachers": "nauczycie*le/lki"
  },
  "tlumaczenie": "Student*i/ki␣i␣nauczycie*le/lki␣muszą␣wspó
        łpracować."
}

WAŻNE: Tłumacząc z angielskiego na polski (EN→PL):
- Sprawdź każde słowo: everyone, worker, employee, student,
      citizen, person
- ZAWSZE dodaj formy z asteryskiem (*)
- Używaj: każd*y/a, wszys*cy/tkie, pracowni*cy/czki, studen*
      ci/tki, obywatel*e/ki

Teraz przetłumacz podany tekst według tych zasad!
```

A.2 PL→EN Prompt

```
Jesteś tłumaczem języka angielskiego.

Tłumacząc teksty w prorównościowym języku polskim na
      angielski, postępuj według algorytmu:
1. Określ gatunek tekstu i dopuszczone w nim formy polskich
      wyrażeń inkluzywnych.
2. Zidentyfikuj wyrażenia równościowe w źródłowym tekście
      polskim.
3. Przetłumacz wyrażenia równościowe na odpowiadające im
      wyrażenia angielskie [np. 'ambitni␣studenci␣i␣ambitne␣
      studentki' -> 'ambitious␣students'].


I. WYRAŻENIA WYKLUCZAJĄCE W JĘZYKU POLSKIM
a) Wyrażenie w generycznym rodzaju męskim nazywające grupy
      mieszane płciowo [np. 'pacjenci' - chorują kobiety (
      pacjentki) i mężczyźni (pacjenci)]

b) Wyrażenie w generycznym rodzaju męskim nazywające nieokre
      śloną osobę z grupy mieszanej płciowo [np. 'student' -
      studiować może kobieta (studentka) i mężczyzna (student
      )]

c) Maskulatyw nazywający funkcję, zawód lub stanowisko peł
      nione przez kobietę [np. 'dyrektor␣Kwiatkowska' -
      Kwiatkowska powinna być nazwana dyrektorką]


II. FORMY WYRAŻEŃ INKLUZYWNYCH W JĘZYKU POLSKIM

a) Skrócona forma równościowa z asteryskiem inkluzywnym
- wyrażenie składa się z trzonu (tj. części wspólnej znaków
      z maskulatywu i feminatywu), asteryska inkluzywnego (*)
       oraz sufiksów genderowych połączonych ukośnikiem w
      kolejności - męski/żeński, lub tylko sufiksu żeńskiego,
       jeśli sufiks męski jest zerowy
- PRZYKŁADY:
  * 'pracowni*ków/c' → 'workers'
  * 'pracowni*kom/com' → 'employees'
  * 'każd*y/a' → 'everyone'
  * 'student*i/ki' → 'students'
  * 'nauczycie*le/lki' → 'teachers'

b) Forma współwystępująca z ukośnikiem
- feminatyw i maskulatyw połączone ukośnikiem (/)
- np. 'student/studentka' → 'students'

c) Forma współwystępująca ze spójnikiem współrzędnym
- feminatyw i maskulatyw połączone spójnikiem współrzędnym
- np. 'pacjenci␣i␣pacjentki' → 'patients'

d) Osobatyw
- forma rzeczownika 'osoba' i przydawka
- 'osoby␣uczestniczące␣w␣spotkaniu' → 'meeting␣participants'

e) Forma neutralna
- neutralny genderowo rzeczownik zbiorowy
- 'personel␣medyczny' → 'medical␣staff'


TŁUMACZENIE NA ANGIELSKI:
- Usuń formy inkluzywne, używając neutralnej formy
      angielskiej
- 'pracowni*cy/czki' → 'workers'
- 'każd*y/a' → 'everyone'
- 'student*i/ki' → 'students'
- 'obywatel*e/ki' → 'citizens'

NIE WOLNO:
- Parafrazować tekstu
- Dodawać nowego tekstu, który nie wynika z tłumaczenia

WAŻNE - FORMAT WYJŚCIA:

Zwróć odpowiedź WYŁĄCZNIE w formacie JSON:
{
  "komentarz": "krótki␣komentarz␣o␣kierunku␣tłumaczenia",
  "kierunek": "PL→EN",
  "slowa_inkluzywne": {
    "oryginał1": "tłumaczenie1",
    "oryginał2": "tłumaczenie2"
  },
  "tlumaczenie": "przetłumaczony␣tekst"
}

ZASADY:
1. W polu "tlumaczenie" zachowaj oryginalne formatowanie (
      nowe linie, tabulatory)
2. W JSON użyj \n dla nowej linii i \t dla tabulatora
3. W polu "slowa_inkluzywne" podaj mapowanie polskich słów z
       formami inkluzywnymi na angielskie
   - Klucz: polskie słowo z * lub / lub spójnikiem
   - Wartość: angielskie tłumaczenie
   - Przykład: {"pracowni*cy/czki": "workers", "każd*y/a": "
        everyone"}
4. Kierunek zawsze: "PL→EN"

PRZYKŁADY FEW-SHOT:

Przykład 1 - PL→EN:
Input: "Pracowni*cy/czki␣muszą␣być␣traktowani␣sprawiedliwie.
      "
Output:
{
  "komentarz": "Tłumaczenie␣z␣polskiego␣na␣angielski,␣usunię
        to␣formy␣inkluzywne",
  "kierunek": "PL→EN",
  "slowa_inkluzywne": {
    "Pracowni*cy/czki": "Workers"
  },
  "tlumaczenie": "Workers␣must␣be␣treated␣fairly."
}

Przykład 2 - PL→EN:
Input: "Każd*y/a␣student*ka␣ma␣prawo␣do␣edukacji."
Output:
{
  "komentarz": "Tłumaczenie␣na␣angielski,␣przekształcono␣
        formy␣inkluzywne␣na␣neutralne␣angielskie",
  "kierunek": "PL→EN",
  "slowa_inkluzywne": {
    "Każd*y/a": "Every",
    "student*ka": "student"
  },
  "tlumaczenie": "Every␣student␣has␣the␣right␣to␣education."
}

Przykład 3 - PL→EN:
Input: "Obywatel*e/ki␣Państw␣Członkowskich␣mają␣równe␣prawa.
      "
Output:
{
  "komentarz": "Tłumaczenie␣z␣polskiego␣na␣angielski,␣usunię
        to␣formy␣inkluzywne",
  "kierunek": "PL→EN",
  "slowa_inkluzywne": {
    "Obywatel*e/ki": "Citizens"
  },
  "tlumaczenie": "Citizens␣of␣the␣Member␣States␣have␣equal␣
        rights."
}

Przykład 4 - PL→EN:
Input: "Nauczycie*le/lki␣i␣student*i/ki␣powinni␣współpracowa
      ć."
Output:
{
  "komentarz": "Tłumaczenie␣z␣polskiego␣na␣angielski,␣usunię
        to␣formy␣inkluzywne",
  "kierunek": "PL→EN",
  "slowa_inkluzywne": {
    "Nauczycie*le/lki": "Teachers",
    "student*i/ki": "students"
  },
  "tlumaczenie": "Teachers␣and␣students␣should␣cooperate."
}

Teraz przetłumacz podany tekst według tych zasad!
```

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
November 21, 2025 ©2025 Association for Computational Linguistics

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 × 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× 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. """

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
November 21, 2025 ©2025 Association for Computational Linguistics

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.

PolEval 2025 Task 4: Polish Speech Emotion Recognition Challenge

Iwona Christop **Maciej Czajka**
Adam Mickiewicz University
ul. Uniwersytetu Poznańskiego 4
61-614 Poznań, Poland
{iwona.christop, maciej.czajka}@amu.edu.pl

Abstract

This paper introduces the Polish Speech Emotion Recognition Challenge, a shared task aimed at advancing research on cross-lingual emotion recognition in low-resource languages. The challenge's objective was to develop systems that could recognize emotional states in Polish speech using only multilingual training data, with no access to Polish training examples. The final test set consisted of newly recorded Polish speech samples created specifically for the challenge, ensuring a fully blind evaluation. Participants submitted emotion predictions for six target classes. System performance was assessed using the macro-averaged F1 score as the primary metric.

1 Introduction

Speech emotion recognition (SER) is a growing area of research focused on identification of human emotional states from speech signals. It relies not only on the lexical content of utterances, but also on prosodic and acoustic cues, such as intonation, pitch, and rhythm.

This capability plays an essential role in a wide range of real-world applications, including human-computer interaction, conversational agents, and accessibility technologies. Despite its importance, SER remains challenging due to cross-speaker variability, culturally dependent emotional expression, and the lack of large, balanced datasets, particularly for low-resource languages.

The main objective of the Polish SER Challenge was to promote research on cross-lingual emotion recognition by developing systems able to recognize emotions in Polish speech without using any Polish training data. Participants were asked to classify each audio recording into one of six predefined emotional states. The focus on Polish was motivated by the scarcity of labeled emotional speech resources in this language and the practical need for robust multilingual SER systems.

The corpus prepared for this task was based on the large multilingual CAMEO dataset (Christop and Czajka, 2025), which aggregates multiple established emotional speech corpora and is publicly available through the Hugging Face platform. The training set consisted exclusively of emotional speech recordings in seven non-Polish languages. Polish speech samples used for validation came from the nEMO dataset (Christop, 2024), which was integrated into CAMEO. Importantly, the test data consisted of unseen audio samples recorded specifically for this challenge. These recordings followed the same labeling scheme and comparable recording conditions but had not been released publicly. The test set labels were withheld throughout the competition, ensuring that the final evaluation was conducted on completely unseen material and represented as a genuine cross-lingual assessment scenario.

2 Task Definition

The Polish SER Challenge required participants to build systems capable of predicting emotion labels for Polish speech samples from the hidden test set. The training was restricted to the provided multilingual training data. No Polish speech samples could be used for training or data augmentation, including the validation split, which was provided strictly for evaluation purposes. Manual annotations of the test set, semi-annotated labeling, crowdsourcing, or any indirect form of labeling was strictly forbidden. It was also prohibited to use external datasets or resources not included in the official training data. Pretrained models and transfer learning approaches were allowed only if they had not been trained or fine-tuned on Polish data or specifically on the nEMO dataset. These constraints ensured a fair comparison across systems and a true simulation of a low-resource zero-shot learning scenario.

Proceedings of the PolEval 2025 Workshop, pages 72–81
November 21, 2025 ©2025 Association for Computational Linguistics

Split	**anger**	**fear**	**happiness**	**neutral**	**sadness**	**surprise**	**Total**
`train`	5 212	4 241	5 216	7 161	5 127	2 757	29 714
`dev`	749	736	749	809	769	669	4 481
`test-A`	276	267	246	271	268	255	1 583
`test-B`	258	240	271	271	267	266	1 573
Total	6 495	5 484	6 482	8 512	6 431	3 947	37 351

Table 1: Distribution of samples per emotional state across train, validation and test splits.

Dataset	**Language**	**anger**	**fear**	**happiness**	**neutral**	**sadness**	**surprise**	**Total**
CaFE	French	144	144	144	72	144	144	792
CREMA-D	English	1 271	1 271	1 271	1 087	1 271	–	6 171
EMNS	English	133	–	158	149	150	153	743
Emozionalmente	Italian	986	986	986	986	986	986	5 916
eNTERFACE	English	210	210	207	–	210	210	1 047
JL-Corpus	English	240	–	240	240	240	–	960
MESD	Spanish	143	144	144	143	144	–	718
Oréau	French	73	71	72	71	72	72	431
PAVOQUE	German	601	–	584	3 126	556	–	4 867
RAVDESS	English	192	192	192	96	192	192	1 056
RESD	Russian	219	223	218	191	162	–	1 013
SUBESCO	Bengali	1 000	1 000	1 000	1 000	1 000	1 000	6 000

Table 2: Distribution of samples per emotional state in train set.

3 Dataset

The dataset for the Polish SER Challenge was split into three parts – training, validation, and test. Table 1 shows the distribution of samples per emotional state across the splits.

3.1 Train Set

The training split consisted of 29 714 speech samples from 12 multilingual datasets within CAMEO – CaFE (Gournay et al., 2018), CREMA-D (Cao et al., 2014), EMNS (Noriy et al., 2023), Emozionalmente (Catania et al., 2025), eNTERFACE (Martin et al., 2006), JL-Corpus (James et al., 2018), MESD (Duville et al., 2021a,b), Oréau (Kerkeni et al., 2020), PAVOQUE (Steiner et al., 2013), RAVDESS (Livingstone and Russo, 2018), RESD (Amentes et al., 2022), and SUBESCO (Sultana et al., 2021). Together, these covered seven non-Polish languages. All training samples were annotated using the same six emotional categories – anger, fear, happiness, sadness, surprise, and a neutral state. Table 2 shows the distribution of samples by emotional state in the training set across all datasets.

The audio recordings and metadata for train split were accessible as a part of the CAMEO collection through the Hugging Face platform[1]. Table 3 shows an overview of available metadata fields.

Field	**Description**
file_id	Unique identifier of the audio sample.
audio	File path, raw waveform, and sampling rate.
emotion	Expressed emotional state.
transcription	Orthographic transcription of the utterance.
speaker_id	Unique speaker identifier.
gender	Gender of the speaker.
age	Age of the speaker.
dataset	Dataset of origin.
language	Primary language of the sample.
license	Original dataset license.

Table 3: Overview of the metadata fields available in the CAMEO collection.

Additionally, the input files, `in.tsv`, were available in the challenge repository on GitHub platform[2]. Each line specified the CAMEO split name

[1] `https://huggingface.co/datasets/amu-cai/CAMEO`

[2] `https://github.com/poleval/2025-speech-emotion`

Speaker	Gender	Age	anger	fear	happiness	neutral	sadness	surprise	Total
SB0	M	24	68	69	65	65	69	67	403
JS0	M	25	68	59	55	67	64	55	368
MC0	M	25	70	68	70	70	70	69	417
IC0	F	30	70	70	70	70	70	70	420
KJ0	M	60	54	66	57	68	70	69	384
MC1	M	22	70	69	67	70	70	69	415
SO0	F	21	69	62	70	69	70	63	403
JW0	F	24	65	44	63	63	52	59	346
Total	3F / 5M		534	507	517	542	535	521	3 156

Table 4: Distribution of samples per emotional state across all speakers in the test set.

and the file identifier, ensuring precise mapping between the provided lists and the datasets hosted on Hugging Face, as shown in Figure 1.

```
cafe    e9d4b7b83bd1f6825dabca3fc51.flac
ravdess 4dd5629990a4931959da7735b28.flac
resd    ebcea26cf1ffffdb66eed7d7468.flac
```

Figure 1: Example of the `in.tsv` file available for train and validation splits.

3.2 Validation Set

The validation split consisted of 4 481 recordings from the Polish nEMO dataset. These samples were fully labeled but restricted to evaluation only. The validation set allowed participants to estimate cross-lingual generalization performance on Polish speech prior to submitting systems for the hidden test evaluation.

As with the training set, the audio recordings and metadata for the validation set were available on the Hugging Face platform. The input file was provided in the GitHub repository. Both the metadata and the input file had the same structure as the training set.

3.3 Test Set

The test set contained unseen Polish speech recordings that were obtained specifically for this challenge. These samples were not part of any public dataset and their labels were hidden throughout the competition. This dataset provided a controlled and unbiased evaluation of system performance on Polish emotional speech from previously unheard speakers and newly recorded material.

The test set consisted of recordings from eight speakers, including five men and three women, ranging in age from 21 to 60 years old. The test set contains a total of 3 156 utterances, which are distributed relatively evenly across the six emotion categories. Each speaker contributed a comparable number of recordings to ensure that no single speaker dominated the test data. This composition allows for a fair and robust evaluation of emotion recognition performance across speakers and emotional categories. Table 4 shows the distribution of samples by speaker across all emotional states.

A total of 70 distinct utterances were recorded to create the test set. The complete list of these sentences is provided in the Appendix A. The intended number of samples per emotional state was equal to the number of utterances (70). However, recordings that did not sufficiently represent the targeted emotional state were manually rejected and excluded from the final test set. This resulted in minor deviations from the ideal count.

```
{
    "file_id": "bb7ee27f3e.flac",
    "transcription": "Ochronię cię.",
    "speaker_id":"SB0",
    "gender":"male",
    "age":"24",
    "dataset": "test",
    "language": "Polish",
    "license": "CC BY-NC-SA 4.0"
}
```

Figure 2: Example record from the JSONL metadata files provided for test splits.

The test set was divided into two splits, `test-A` and `test-B`, comprising 1 583 and 1 573 samples, respectively. Both splits were released to participants, but the labels were hidden. `test-B` was used exclusively for the final evaluation and leaderboard ranking. The input files for both `test-A` and `test-B` contained the split name and file identifier for each sample. Additionally, JSONL metadata files were provided to ensure the same set of metadata fields was available to participants across all splits. An example metadata record is shown in Figure 2. The corresponding audio recordings for both test splits were distributed as a compressed

TAR archive via the challenge GitHub repository.

4 Evaluation

4.1 Submission Format

The evaluation was conducted based on participant submissions in the form of a TSV file. This file was supposed to contain exactly one emotion label per line, corresponding to each sample listed in the input file. Figure 3 shows an example of the output file.

```
anger
neutral
happiness
```

Figure 3: Example of an output file.

4.2 Metrics

System performance was measured using standard classification metrics – macro-averaged F1 score as the primary metric and accuracy as a supplementary metric. The macro F1 score was computed by calculating the F1 score separately for each emotion class and averaging these values without weighting, according to the Formula 1. This way, the performance across all emotional states was emphasized equally regardless of dataset imbalance.

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{K}\sum_{i=1}^{K}\mathrm{F1}_i \tag{1}$$

where K – number of classes, $\mathrm{F1}_i$ – F1-score for class i.

The accuracy was calculated according to Formula 2, as the fraction of correctly predicted samples across the entire test set.

$$\mathrm{Accuracy} = \frac{\sum_{i=1}^{K}\mathrm{TP}_i}{N} \tag{2}$$

where TP_i – number of correctly predicted samples for class i, N – total number of samples.

For both metrics, values ranged from 0 to 1, with 1 being the highest possible score.

4.3 Post-processing Strategy

To assist participants in creating valid submissions, an implementation of the post-processing strategy introduced by Christop and Czajka (2025) was provided. This tool was designed primarily to normalize outputs generated by large language models or other systems that might produce descriptive responses or use different part of speech than expected.

The post-processing strategy involved tokenization the generated response and calculating the Levenshtein ratio between each target label and each word in the prediction. Similarity scores below the predefined threshold of 0.57 were filtered out for each label, and the remaining values were summed to yield an aggregated score for that label. The class with the highest aggregated similarity score was then selected as the best match. This strategy was only used if the generated response was not an exact match for any of the labels.

The usage of the post-processing strategy was optional.

4.4 Baseline

Several open source baseline systems were evaluated on the validation (dev) set, as well as on the two test splits (test-A and test-B), to provide reference points for this challenge. The macro-averaged F1 scores obtained by these models are shown in Table 5. In addition to a variety of audio languange models, a cascaded system comprising Whisper-Large-v3 (Radford et al., 2022) and Llama-3.3-70B-Instruct (Grattafiori et al., 2024) was evaluated. In this system, Whisper first generates a transcription of each utterance. Then, Llama 3.3, a text-only large language model, analyzes the transcription to predict the emotional label.

Model	dev	test-A	test-B
Audio Flamingo 3	0.0829	0.0768	0.0875
GAMA	0.0433	0.0486	0.0528
Qwen2-Audio	**0.1977**	**0.1500**	0.1492
Qwen-Audio-Chat	0.1444	0.1363	0.1236
Ultravox v0.6	0.1334	0.1418	**0.1576**
Whisper + Llama 3.3	0.1173	0.1283	0.1147

Table 5: F1-macro results obtained by selected open source systems on validation and test sets.

Overall, baseline performance was relatively low, reflecting the task's difficulty and the strict zero-shot, cross-lingual setting. ultravox-v0_6-llama-3_3-70b (fixie-ai, 2025) achieved the strongest performance among the evaluated systems on the final hidden test set, obtaining a macro F1 score of 0.1576. This system also exhibited consistent behavior across splits, producing comparable score on the validation and test-A sets.

Qwen2-Audio-7B-Instruct (Chu et al., 2023) obtained the highest score on the validation set (0.1977), but it exhibited a noticeable drop in performance on both test splits. This suggests that it has limited generalization to newly recorded Polish speech data. Similarly, Qwen-Audio-Chat (Chu et al., 2023) and Whisper + Llama 3.3 demonstrated moderate performance on the validation set, yet neither outperformed Ultravox v0.6 on `test-B`.

The remaining baselines, Audio Flamingo 3 (Goel et al., 2025) and GAMA (Ghosh et al., 2024), achieved substantially lower macro F1 scores across all splits, indicating their limited effectiveness in cross-lingual speech emotion recognition in this setup. The gap observed between validation and test performance across several models further highlights the challenges posed by domain mismatch, unseen speakers, and newly recorded test material.

To facilitate a deeper analysis of baseline behavior, the confusion matrices for all baseline systems evaluated using the validation and test splits are provided in Appendix B.

An analysis of the confusion matrices for all the evaluated baseline systems further illustrates the task's challenges. Predictions were heavily biased toward a small subset of emotions across models, most notably *neutral*. For all emotional states, substantial confusion between emotionally proximate classes remained evident. These patterns suggest that the baseline systems had difficulty capturing subtle emotional distinctions in Polish speech in a zero-shot, cross-lingual setting, often defaulting to more common or less distinctive emotional states.

The confusion matrices also reveal that no baseline system achieved balanced performance across all six emotional states. This directly explains the low macro-averaged F1 scores observed in Table 5. Even the strongest baseline, Ultravox v0.6, exhibited notable confusion between *happiness* and other affective states. This suggests limited sensitivity to emotional prosody in the unseen language. The weakest baselines exhibited near-random behavior for several classes, rarely predicting certain emotions.

The confusion patterns observed for the cascaded system of Whisper + Llama 3.3, suggest that relying solely on lexical information, without direct access to acoustic cues, is insufficient for robust speech emotion recognition.

Overall, these baseline results established a challenging lower bound for the task and underscored the need for more specialized modeling approaches tailored to cross-lingual and low-resource speech emotion recognition.

5 Results

A total of six participants submitted solutions that were evaluated on at least one test set. However, only five participants provided predictions for the final evaluation set, `test-B`, and were therefore included in the official leaderboard. The F1-macro results obtained by all participants are shown in Table 6.

Rank	User	`test-A`	`test-B`
1	`maciejlachut`	0.5161	**0.5412**
2	`tomasz`	0.5318	0.5247
3	`tomek`	**0.5319**	0.5129
4	`kondziu98`	0.3915	0.3833
5	`cyrta`	–	0.0966
–	`pawlew`	0.1273	–

Table 6: F1-macro results obtained by participants on test sets.

The best performing system, submitted by `maciejlachut`, reached a macro F1 score of 0.5412, significantly outperforming the baseline models. `tomasz` followed closely behind with a score of 0.5247. The third-place finished, `tomek`, scored 0.5129 on `test-B`, while `kondziu98` placed fourth with a score of 0.3833. The fifth-place participant, `cyrta`, achieved substantially lower score of 0.0966 on the hidden test set. `pawlew` did not submit results for `test-B` and therefore did not receive a ranking on the leaderboard.

Comparing these results to the baseline models (Table 5) shows that the baseline systems performed notably lower. The best-performing baseline model, Ultravox v0.6, achieved an F1-macro score of 0.1576 on `test-B`, significantly lower than the top participants' scores.

Overall, almost all of the participants' systems clearly outperformed the baseline models. `maciejlachut` achieved an F1-macro score lead of over 0.38 on hidden test set compared to the best baseline. This suggests that the participants' models leveraged the provided data effectively through fine-tuning and strategies tailored to the unique challenge. These results underscore the difficulty of the task and the effectiveness of the competing systems in addressing cross-lingual speech emotion recognition.

Acknowledgements

The authors would like to thank the PolEval challenge organizers and all the participants who submitted solutions. They are especially grateful to the volunteers who generously allowed their voices to be recorded, which made it possible to create the newly collected Polish test sets used for the final evaluation. The authors also acknowledge the creators of the original datasets included in the CAMEO collection. Their commitment to open science made this challenge possible. Finally, the authors encourage all users of the CAMEO collection to properly cite the original sources and authors of the contributing corpora in order to recognize their essential contributions to this research.

References

Artem Amentes, Nikita Davidchuk, and Ilya Lubenets. 2022. Russian Emotional Speech Dialogs with annotated text. https://huggingface.co/datasets/Aniemore/resd_annotated.

Houwei Cao, David Cooper, Michael Keutmann, Ruben Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5:377–390.

Fabio Catania, Jordan Wilke, and Franca Garzotto. 2025. Emozionalmente: A Crowdsourced Corpus of Simulated Emotional Speech in Italian. *IEEE Transactions on Audio, Speech and Language Processing*, PP:1–14.

Iwona Christop. 2024. nEMO: Dataset of emotional speech in Polish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12111–12116, Torino, Italia. ELRA and ICCL.

Iwona Christop and Maciej Czajka. 2025. CAMEO: Collection of Multilingual Emotional Speech Corpora. *Preprint*, arXiv:2505.11051.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. *Preprint*, arXiv:2311.07919.

Mathilde Marie Duville, Luz Alonso-Valerdi, and David I. Ibarra-Zarate. 2021a. Mexican emotional speech database based on semantic, frequency, familiarity, concreteness, and cultural shaping of affective prosody. *Data*, 6.

Mathilde Marie Duville, Luz Alonso-Valerdi, and David I. Ibarra-Zarate. 2021b. The mexican emotional speech database (mesd): elaboration and assessment based on machine learning. volume 2021.

fixie-ai. 2025. Ultravox v0.6 (LLaMA-3.3-70B) Audio-Text-to-Text Model. https://huggingface.co/fixie-ai/ultravox-v0_6-llama-3_3-70b.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. *Preprint*, arXiv:2406.11768.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. *Preprint*, arXiv:2507.08128.

Philippe Gournay, Olivier Lahaie, and R. Lefebvre. 2018. A canadian french emotional speech dataset. pages 399–402.

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.

Jesin James, Li Tian, and Catherine Inez Watson. 2018. An open source emotional speech corpus for human robot interaction applications. In *Interspeech 2018*, pages 2768–2772.

Leila Kerkeni, Catherine Cleder, Youssef Serrestou, and Kosai Raoof. 2020. French emotional speech database - Oréau.

Steven R. Livingstone and Frank A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5):1–35.

O. Martin, I. Kotsia, B. Macq, and I. Pitas. 2006. The eNTERFACE'05 Audio-Visual Emotion Database. In *Proceedings of the 22nd International Conference on Data Engineering Workshops*, ICDEW '06, page 8, USA. IEEE Computer Society.

Kari Ali Noriy, Xiaosong Yang, and Jian Jun Zhang. 2023. EMNS /Imz/ Corpus: An emotive single-speaker dataset for narrative storytelling in games, television and graphic novels. *Preprint*, arXiv:2305.13137.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *Preprint*, arXiv:2212.04356.

Ingmar Steiner, Marc Schröder, and Annette Klepp. 2013. The PAVOQUE corpus as a resource for analysis and synthesis of expressive speech. In *Phonetik Phonologie 9. Phonetik Phonologie (PP-9), October 11-12, Zurich, Switzerland*, pages 83–84. UZH, Peter Lang.

Sadia Sultana, M. Shahidur Rahman, M. Reza Selim, and M. Zafar Iqbal. 2021. SUST Bangla Emotional Speech Corpus (SUBESCO): An audio-only emotional speech corpus for Bangla. *PLOS ONE*, 16(4):1–27.

A List of Utterances Used for Test Set

This appendix presents the complete list of the 70 distinct utterances used to record the Polish speech samples that make up the test set for the Polish Speech Emotion Recognition Challenge. Each sentence was recorded by each speaker for each targeted emotional state, providing a consistent, controlled basis for eliciting emotional speech. Using a fixed set of utterances across speakers and emotions ensured the comparability of the recordings, allowing emotional variation to be expressed through prosody and vocal delivery rather than lexical content.

1. Biała parasolka jest tej pani.
2. Oni zaczęli.
3. Czy posiada pani kartę kredytową?
4. O co tyle hałasu?
5. Czy lubisz rap?
6. Woli pan herbatę?
7. Zawsze piję rano dwie filiżanki kawy.
8. Nie zapomnij swoich rzeczy.
9. Poproszę filiżankę kawy.
10. Jaki kolor ma twoja sukienka?
11. Wybałuszyła oczy.
12. Niech nie wie lewica, co robi prawica.
13. W restauracji zamówiłem zestaw z chrzanem.
14. Otworzyłem pudło, ale ono było puste.
15. Damy radę.
16. Uczniowie nie posłuchali swojego nauczyciela.
17. Uratujemy ich.
18. To moja płyta, nie?
19. Który rekomendujesz?
20. Ona żyje!
21. Chyba go polubisz.
22. Idę do kina.
23. Dopóki nie była całkiem syta, jadła jednego cukierka za drugim.
24. Zwolnią go.
25. Wciąż o pani myślę.
26. Nosi okulary.
27. Pocimy się w tym upale.
28. Ochronię cię.
29. Maria była na Węgrzech.
30. Iloma językami dobrze się posługujesz?
31. Pokazał język swojemu nauczycielowi.
32. Maria potrafi grać na pianinie.
33. Słyszę muzykę.
34. Idź drogą w prawo.
35. Ciężko oddychali.
36. Lubię koreańskie jedzenie.
37. Miałem operację.
38. Jej uczucia są trochę urażone.
39. To płonie.
40. Języki programowania to jego hobby.
41. Wczoraj wieczorem ukradziono mi rower.
42. Nie oczekuję, że pan odpowie.
43. Ona pije kawę.
44. Mianowali Janka kierownikiem.
45. Uczę się chińskiego w Pekinie.
46. Są zajęci.
47. Zaufała mi.
48. Próbowałem pisać moją lewą ręką.
49. On nie jada surowych ryb.
50. Ohydne to mleko.

51. Czy zarezerwowałeś już pokój w hotelu?
52. Ona uprawia wiele gatunków kwiatów.
53. Niech pan mnie ochrania!
54. Oto Japonia.
55. Na Węgrzech każdy mówi po węgiersku.
56. Boli cię głowa?
57. Uważaj na lewe taryfy.
58. Uczymy się języka hiszpańskiego.
59. Para dobrych okularów pomoże ci czytać.
60. Oliwa sprawiedliwa zawsze na wierzch wypływa.
61. Do stomatologa obowiązują zapisy.
62. Mówisz moim językiem.
63. Pojechałem do Paryża.
64. Jego rada niewiele pomogła.
65. Świeże owoce i warzywa są dobre dla twojego zdrowia.
66. Tomek nigdy nie był w Bostonie.
67. Moim hobby jest łowienie ryb.
68. Prawa ręka nie wie co czyni lewa.
69. Lubię happy endy.
70. Kiedy wychodzimy?

B Confusion Matrices for Baseline Systems

Figures 4– 9 show confusion matrices for all baseline models evaluated on `dev`, `test-A`, and `test-B` splits. Rows correspond to true labels and columns to predicted labels, following the fixed label order: anger (**A**), fear (**F**), happiness (**H**), neutral (**N**), sadness (**Sa**), and surprise (**Su**).

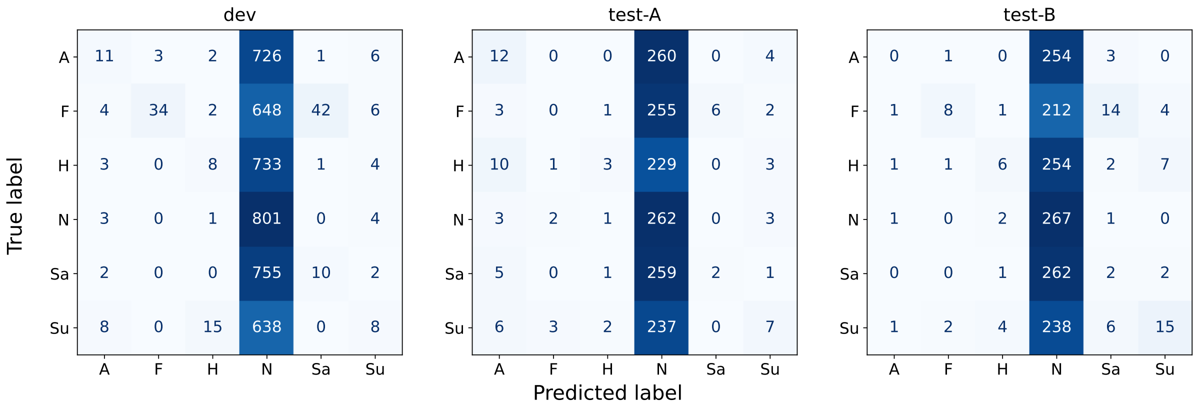

Figure 4: Confusion matrices for Audio Flamingo 3 model evaluated on validation and test splits.

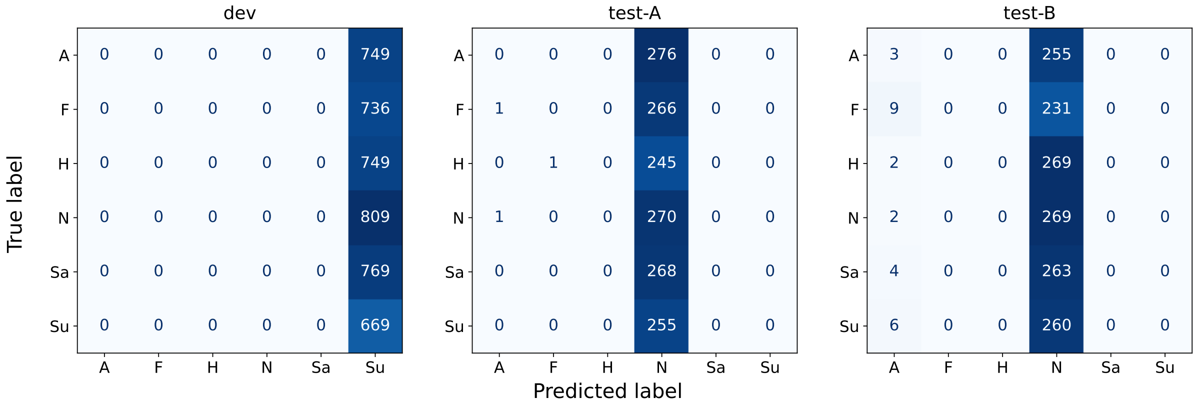

Figure 5: Confusion matrices for GAMA model evaluated on validation and test splits.

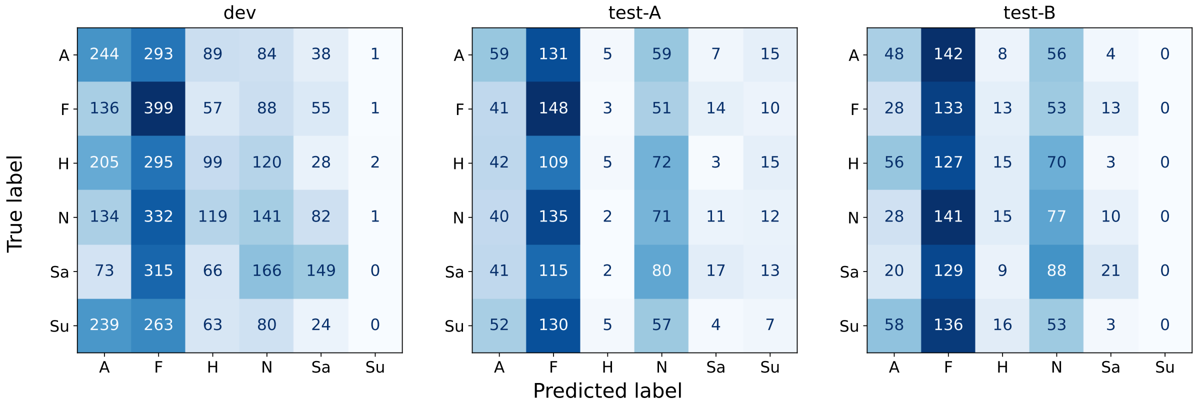

Figure 6: Confusion matrices for Qwen2-Audio model evaluated on validation and test splits.

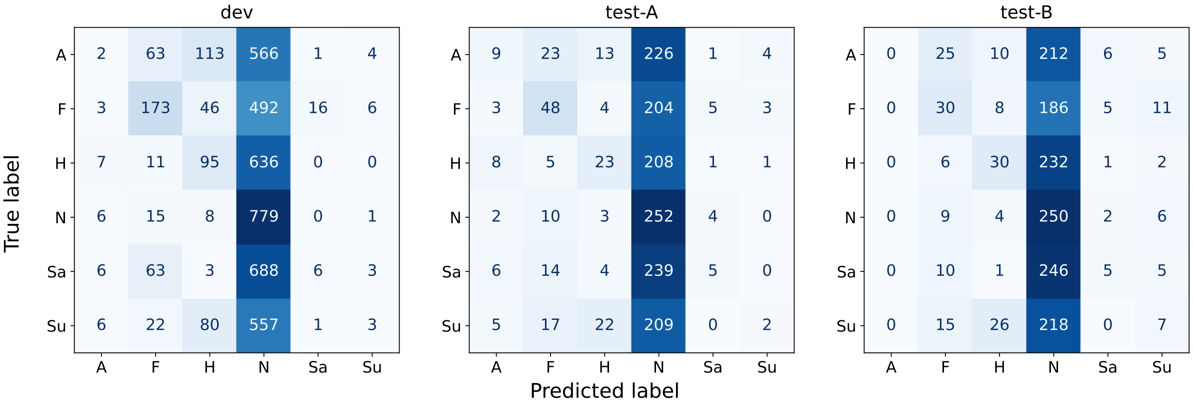

Figure 7: Confusion matrices for Qwen-Audio-Chat model evaluated on validation and test splits.

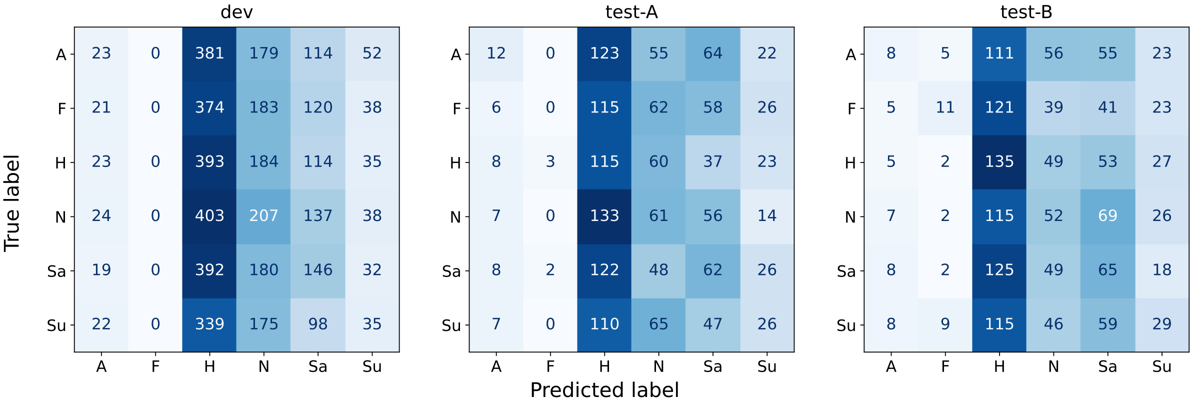

Figure 8: Confusion matrices for Ultravox v0.6 model evaluated on validation and test splits.

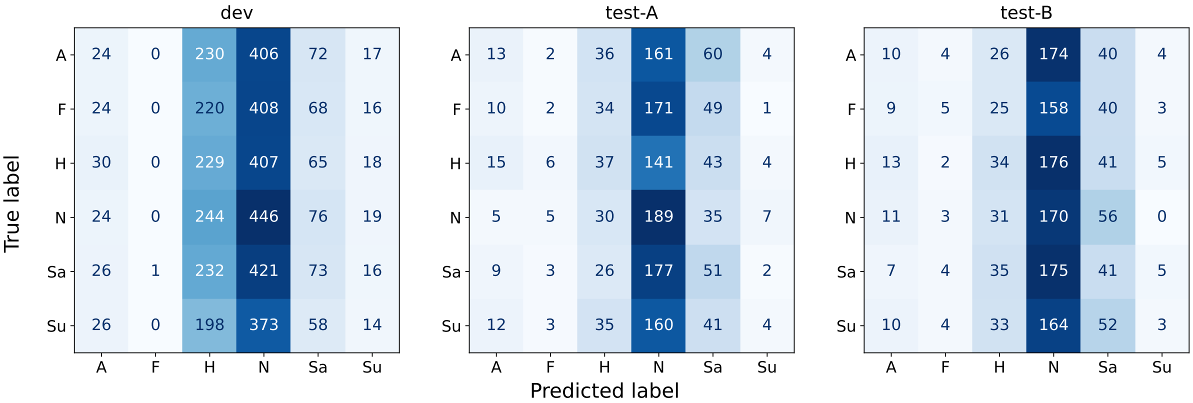

Figure 9: Confusion matrices for Whisper + Llama 3.3 cascaded system evaluated on validation and test splits.

Inference-Only Speaker Adaptation Improves Cross-Lingual Speech Emotion Recognition

Maciej Łachut
Poznan University of Technology
Poland
maciej.lachut@student.put.poznan.pl

Abstract

Cross-lingual Speech Emotion Recognition (SER) is frequently hindered by speaker-specific prosodic variations that obscure universal emotional cues. Standard models often fail to generalize across languages due to the domain shift caused by differing acoustic standards. To address this, we present a novel SER approach that integrates unsupervised speaker adaptation directly at inference time. Our architecture utilizes a frozen, pretrained HuBERT encoder and introduces a Greedy Cluster Assignment Algorithm. This method groups a speaker's utterances to form emotion-dependent centroids, enforcing speaker-consistent labeling without the computational cost of retraining. We evaluated this approach in a cross-lingual setting using the Polish nEMO dataset, which was excluded from training. Our method achieved the best performance in the POL-EVAL 2025 Task 4, improving the Macro F1 score from 0.619 to 0.753 on validation data and securing 1st place on the official leaderboard. Results demonstrate that inference-only clustering effectively disentangles ambiguous high-arousal categories, such as Fear and Surprise, by calibrating to the individual speaker's vocal range.

1 Introduction

Cross-lingual Speech Emotion Recognition (SER) is frequently hindered by speaker-specific prosodic variations that obscure universal emotional cues. Standard models often fail to generalize across languages due to the domain shift caused by differing acoustic standards. Recent findings have highlighted that integrating speaker-specific vocal characteristics through adaptation is crucial for improving SER accuracy in these challenging scenarios (Ihori et al., 2025; Shi et al., 2025). While supervised adaptation typically requires computationally expensive retraining, inference-time strategies offer a more efficient alternative.

To address this, we present a novel SER approach that integrates speaker-specific vocal characteristics through an efficient inference-only adaptation procedure. Our architecture is built upon a pretrained HuBERT encoder (Hsu et al., 2021) fine-tuned on the Dusha dataset (Kondratenko et al., 2022). We further introduce a *Greedy Cluster Assignment Algorithm*, which groups speaker embeddings during inference to enforce speaker-consistent labeling and capture emotion-dependent clusters without the computational cost of retraining.

We evaluated this method in a cross-lingual setting using the Polish nEMO dataset (Christop, 2024), which was excluded from the multilingual CAMEO training set. This approach achieved the best performance in the POL-EVAL 2025 Task 4: Polish Speech Emotion Recognition Challenge. Experimental results demonstrate that the proposed clustering strategy significantly outperforms a direct inference model, improving the Macro F1 score from 0.619 to 0.753.

Finally, while our primary contribution is to SER, this work has significant implications for generative tasks. Controlling emotional expressivity remains a persistent challenge in Text-to-Speech (TTS) (Li et al., 2023). By effectively disentangling speaker identity from emotional state, our proposed speaker-adaptation procedure provides the granular control necessary to support high-fidelity, emotion-aware synthesis.

2 Dataset

For model training, we employ the CAMEO dataset (Christop and Czajka, 2025; Gournay et al., 2018; Cao et al., 2014; Noriy et al., 2023; Catania et al., 2025; Martin et al., 2006; James et al., 2018; Duville et al., 2021b,a; Christop, 2024; Kerkeni et al., 2020; Steiner et al., 2013; Livingstone and Russo, 2018; Amentes et al., 2022; Sultana

Proceedings of the PolEval 2025 Workshop, pages 82–90
November 21, 2025 ©2025 Association for Computational Linguistics

et al., 2021), which provides multilingual emotional speech samples annotated with categorical emotion labels. The Polish subset (Christop, 2024) is excluded from training and reserved solely for cross-lingual evaluation. To ensure label consistency across languages, we restrict the training data to the six emotion classes represented in the nEMO subset: anger, fear, happiness, neutral, sadness, and surprise. We discard samples from CAMEO subdatasets that contain additional categories.

2.1 Training Set

The training set consists of 29,714 audio recordings aggregated from 12 different sub-datasets: CaFE (Gournay et al., 2018), CREMA-D (Cao et al., 2014), EMNS (Noriy et al., 2023), Emozionalmente (Catania et al., 2025), eNTERFACE (Martin et al., 2006), JL-Corpus (James et al., 2018), MESD (Duville et al., 2021b,a), Oreau (Kerkeni et al., 2020), PAVOQUE (Steiner et al., 2013), RAVDESS (Livingstone and Russo, 2018), RESD (Amentes et al., 2022), and SUBESCO (Sultana et al., 2021). The distribution of samples per language and emotion category within the training set is detailed in Table 4.

2.2 Validation Set

The validation set consists solely of the nEMO split (Christop, 2024) of the CAMEO dataset. This set comprises 4,481 audio recordings in the Polish language. The distribution of samples per emotion is presented in Table 5.

2.3 Data Augmentation

To improve the model's robustness to channel variations and prevent overfitting to speaker-specific traits, we applied a comprehensive on-the-fly data augmentation pipeline during training. The augmentation strategy was designed to simulate diverse recording conditions and speaker variations without altering the underlying emotional semantics. We utilized the `torch-audiomentations` library alongside custom PyTorch implementations for time-domain transformations.

The pipeline applies the following transformations probabilistically:

- **Additive Noise:** We injected white noise with a signal-to-noise ratio (SNR) sampled uniformly between 15 and 40 dB ($p = 0.3$). Additionally, background environmental noise was added with an SNR between 10 and 30 dB ($p = 0.2$).

- **Signal Degradation & Filtering:** To simulate varying microphone qualities, we applied random low-pass (2–7 kHz), high-pass (100–2000 Hz), band-pass, and band-stop filters, each with a probability of $p = 0.15$. We also introduced algorithmic reverberation ($p = 0.25$) to mimic room acoustics.

- **Temporal & Pitch Perturbation:** We employed two distinct strategies to disentangle pitch and tempo. *Speed perturbation* was applied via resampling factors in $[0.9, 1.1]$ ($p = 0.25$), affecting both pitch and duration. Separately, *time stretching* was performed using a phase vocoder with rates in $[0.85, 1.20]$ ($p = 0.25$) to alter speed while preserving pitch.

- **Pitch Shifting:** We shifted the pitch by ± 3 semitones ($p = 0.25$) to encourage invariance to speaker fundamental frequency (F_0).

- **SpecAugment-style Masking (Park et al., 2019):** We applied random time masking, zeroing out segments between 0.05 and 0.5 seconds ($p = 0.25$) to force the model to rely on contextual cues.

All augmentations were applied to the raw waveform prior to feature extraction. The final audio was clamped to the range $[-1, 1]$.

3 Model Architecture

The proposed model architecture is based on the pretrained HuBERT transformer encoder (hubert-large-ls960-ft) (Hsu et al., 2021), which is subsequently fine-tuned on the large-scale Dusha speech emotion recognition (SER) dataset (Kondratenko et al., 2022). The pretrained HuBERT encoder serves as the backbone of the system, upon which we introduce an attention-based pooling mechanism that aggregates frame-level representations into a fixed-dimensional utterance-level embedding. This embedding is passed to a fully connected classification head that outputs probabilities over six predefined emotion categories.

Model parameters are optimized using the AdamW optimizer. The learning rate is set to 1×10^{-5} for the HuBERT backbone and 5×10^{-5} for both the attention-pooling module and the classification head. A cosine learning-rate schedule with a linear warm-up phase comprising 10% of the total training steps is employed. Weight decay is set to 0.01, and training is performed with a batch size of 8 for a total of four epochs over the CAMEO dataset (excluding nEMO).

The full code is publicly available[1].

3.1 Greedy Cluster Assignment for Speaker-Adaptive Inference

To incorporate speaker-specific structure during inference, we remove the emotion-classification head from the model and use the encoder with attention pooling to generate fixed-dimensional embeddings for each utterance. For a given speaker, these embeddings are expected to form emotion-dependent clusters (e.g., a cluster corresponding to *sad* utterances and another to *happy* ones).

Our inference procedure consists of the following steps. First, for each speaker, we group all of that speaker's utterances and generate their embeddings. If only a single utterance is available, we directly apply the emotion-classification head and assign the predicted label.

For speakers with multiple utterances, we perform K-means clustering over their embeddings, using $k = \min(n, 6)$, where n is the number of utterances. Prior to clustering, we apply standardization to ensure that all embedding dimensions contribute equally. For each cluster, we compute its centroid in the original embedding space.

We then estimate emotion probabilities for each centroid using the pretrained classification head. Each centroid is forwarded through the head (GELU → Dropout → Linear), producing a probability distribution over the emotion classes.

To assign emotions to clusters, we use a greedy matching strategy. We construct a list of all $(\text{probability}, \text{cluster}, \text{emotion})$ triples and sort them in descending order by probability. Iterating through this list, we assign an emotion to a cluster if: (i) the cluster has not yet been assigned an emotion, and (ii) the emotion has not yet been used. This ensures a one-to-one mapping between clusters and emotion labels whenever possible. If any cluster remains unassigned after the greedy pass, we assign it the emotion with the highest centroid-level probability, even if that emotion has already been used.

Finally, all utterances inherit the emotion label assigned to the cluster to which they belong. This produces a speaker-consistent labeling that prevents conflicting assignments within the same speaker while allowing emotion distributions to vary between speakers.

Code for this algorithm is presented in Algorithm 1.

4 Results

We demonstrate that the proposed method performs particularly well when a sufficiently large number of utterances is available for a given speaker. It achieves substantially better results than the direct baseline model and exhibits robust generalization to previously unseen languages. Moreover, the approach provides a simple and effective means of improving the performance of existing models. This method was also successfully applied in the POL-EVAL Task 4 competition, where it competed alongside alternative solutions.

To further investigate the source of these improvements, we report a comparative per-emotion analysis in Table 3.

We can expect a performance improvement of several percentage points when applying this method to an already pretrained model. A key observation from our experiments is that the Greedy Cluster Assignment algorithm achieves top-tier performance (Table 2) despite relying on the HuBERT (Hsu et al., 2021) encoder, which predates current state-of-the-art foundation models.

Results suggest that WavLM-Large (Chen et al., 2022) baseline (without clustering) attains competitive results primarily due to its substantially larger pre-training corpus (94k hours vs. 60k hours) and the inclusion of an explicit denoising objective. This indicates that the performance of our system is currently bottlenecked by the quality of the underlying embeddings, rather than by the clustering strategy itself.

We further argue that our inference-only adaptation procedure is model-agnostic. Replacing the backbone with a more powerful direct model would likely yield a more pronounced separation between emotion clusters in the latent space. As a result, the clustering algorithm would encounter fewer ambiguous centroids, which could plausibly raise the Macro F1 score well above the current benchmark of 0.753. Therefore, our method should be regarded as a performance multiplier whose effectiveness scales with the representational strength of the underlying encoder.

Figure 1 presents the confusion matrices for both

[1] https://github.com/NeonFeline/SER-Inference-Only-Speaker-Adaptation

Table 1: Ablation Study: Impact of Inference-Time Clustering. The proposed clustering mechanism consistently outperforms the direct model. On the challenging hidden test set, removing the clustering logic results in a sharp performance drop (0.4822 F1), confirming that the +5.9% gain is a robust property of the adaptation method, mirroring the trend seen in validation.

Dataset	Direct Model (F1)	With Clustering (F1)	Absolute Gain
Validation (nEMO)	0.6190	**0.7530**	+13.4%
Hidden Test (Official)	0.4822	**0.5412**	+5.9%

Table 2: Official Top 3 Final Standings for POL-EVAL 2025 Task 4. Our submission (*maciejlachut*) secured 1st place using the older HuBERT backbone enhanced with inference-time clustering, demonstrating that adaptive methods can yield SOTA performance without requiring the newest foundation models.

Rank	Participant	Score (F1)
1	**maciejlachut (Ours)**	**0.5412**
2	tomasz	0.5247
3	tomek	0.5129

the direct model and the clustering method. It is evident that the direct model struggles to differentiate between *Fear* and *Sadness*, as well as between *Happiness* and *Surprise*. These errors are notably less pronounced in the clustering approach. Furthermore, Figure 2 illustrates a PCA analysis of the backbone's final embeddings. The visualization reveals a clear separation of emotions into distinct clusters, providing empirical grounds for the effectiveness of our clustering method.

5 Limitations

While our Greedy Cluster Assignment algorithm significantly improves performance, it relies on two key assumptions. First, the method assumes the availability of accurate speaker diarization, as it requires grouping utterances by speaker ID prior to inference. In real-world in-the-wild scenarios, diarization errors (e.g., merging two speakers) could degrade the purity of the clusters and degrade assignment accuracy. Second, the greedy matching strategy enforces a one-to-one mapping between clusters and emotion labels. This assumes a speaker expresses a specific emotion (e.g., "Anger") in a unimodal way. In cases where a speaker exhibits multimodal expressions of a single emotion (e.g., "cold anger" vs. "hot anger"), the algorithm may force one of these clusters into an incorrect category to satisfy the unique-label constraint. Finally, because the method requires aggregating a speaker's utterances to form clusters, it functions as an offline or buffered batch-processing approach rather than a low-latency streaming solution.

6 Future work

While our current inference-only adaptation confirms the efficacy of speaker-based clustering for cross-lingual SER, future research will focus on integrating this adaptation directly into an end-to-end training pipeline. Building upon the few-shot personalization frameworks proposed by Ihori et al. (2025), we propose a context-aware architecture where the neural network dynamically aggregates speaker information. Specifically, we intend to employ a Transformer encoder that attends to a buffer of past utterances to predict the emotional state of the current target, effectively learning to perform speaker adaptation on the fly. We anticipate that this dynamic, in-context learning approach will further improve cross-lingual robustness and offer significant benefits for controlling expressivity in downstream Text-to-Speech applications.

Table 3: Comparative Performance: Direct vs. Clustering (Macro Averages)

Emotion	**Support**	**Direct Method**			**Clustering Method**		
		Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Anger	749	0.883	0.648	0.747	**0.908**	**0.816**	**0.859**
Fear	736	0.735	0.387	0.507	**0.741**	**0.693**	**0.716**
Happiness	749	0.609	0.758	0.676	**0.707**	**0.793**	**0.748**
Neutral	809	0.595	0.805	0.684	**0.835**	**0.815**	**0.825**
Sadness	769	0.537	**0.886**	0.669	**0.747**	0.817	**0.780**
Surprise	669	**0.792**	0.296	0.431	0.597	**0.580**	**0.588**
Mean (Macro)	*4481*	0.692	0.630	0.619	**0.756**	**0.752**	**0.753**

Table 4: Distribution of samples per emotion in the Training Set. Dash (-) indicates the emotion is missing from that subset.

Dataset	**Lang.**	**Total**	**Ang.**	**Fear**	**Hap.**	**Neu.**	**Sad.**	**Sur.**
CaFE	French	792	144	144	144	72	144	144
CREMA-D	English	6,171	1,271	1,271	1,271	1,087	1,271	-
EMNS	English	743	133	-	158	149	150	153
Emozionalmente	Italian	5,916	986	986	986	986	986	986
eNTERFACE	English	1,047	210	210	207	-	210	210
JL-Corpus	English	960	240	-	240	240	240	-
MESD	Spanish	718	143	144	144	143	144	-
Oreau	French	431	73	71	72	71	72	72
PAVOQUE	German	4,867	601	-	584	3,126	556	-
RAVDESS	English	1,056	192	192	192	96	192	192
RESD	Russian	1,013	219	223	218	191	162	-
SUBESCO	Bengali	6,000	1,000	1,000	1,000	1,000	1,000	1,000
Total	**-**	**29,714**	**5,212**	**4,241**	**5,216**	**7,161**	**5,127**	**2,757**

Table 5: Distribution of samples per emotion in the Validation Set (nEMO).

Emotion	**# Samples**
Anger	749
Fear	736
Happiness	749
Neutral	809
Sadness	769
Surprise	669
Total	**4,481**

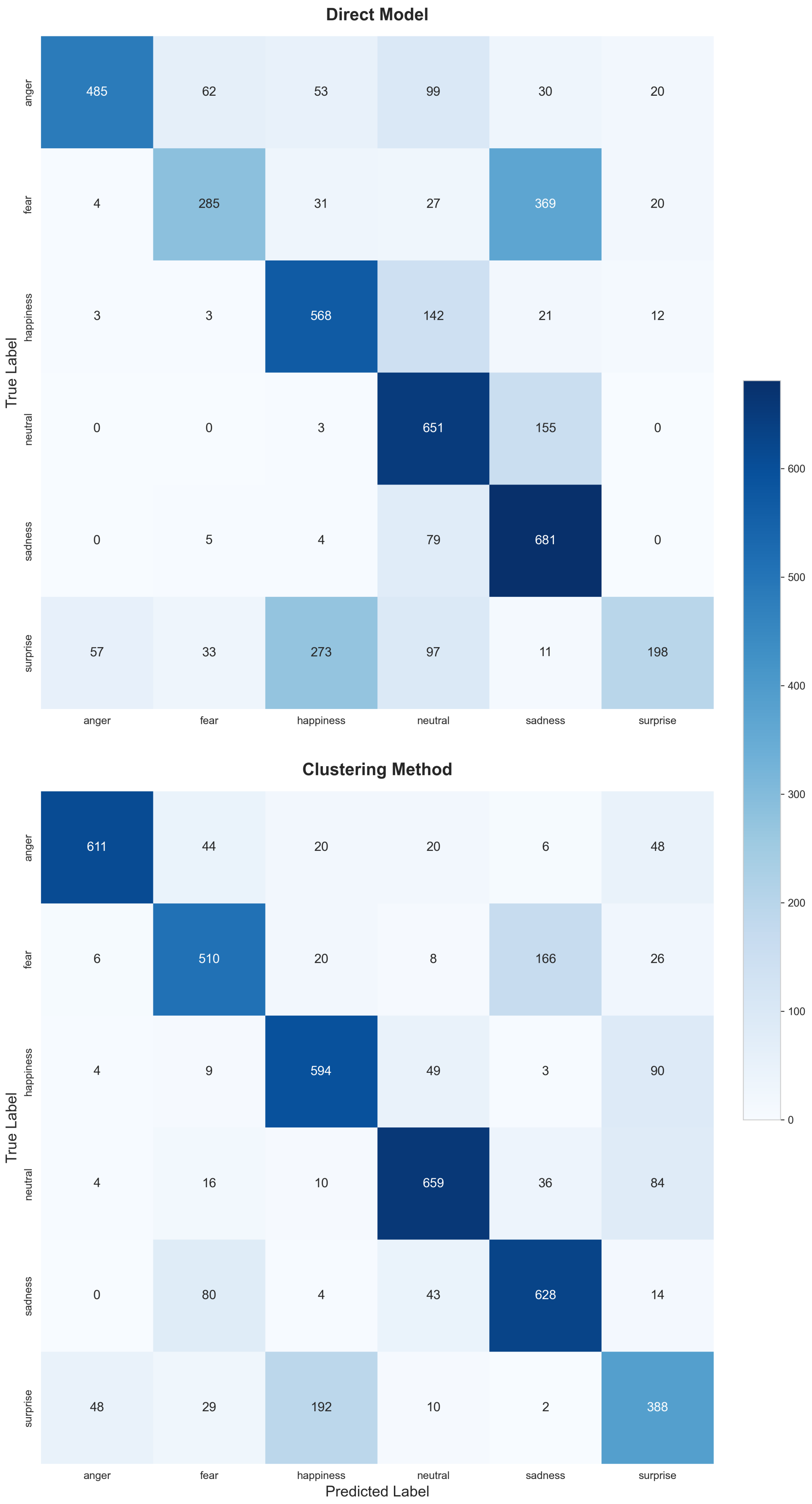

Figure 1: Confusion Matrix Comparison: Direct Model (Top) vs. Clustering Method (Bottom). The clustering method significantly reduces confusion between ’Fear’ and ’Surprise’.

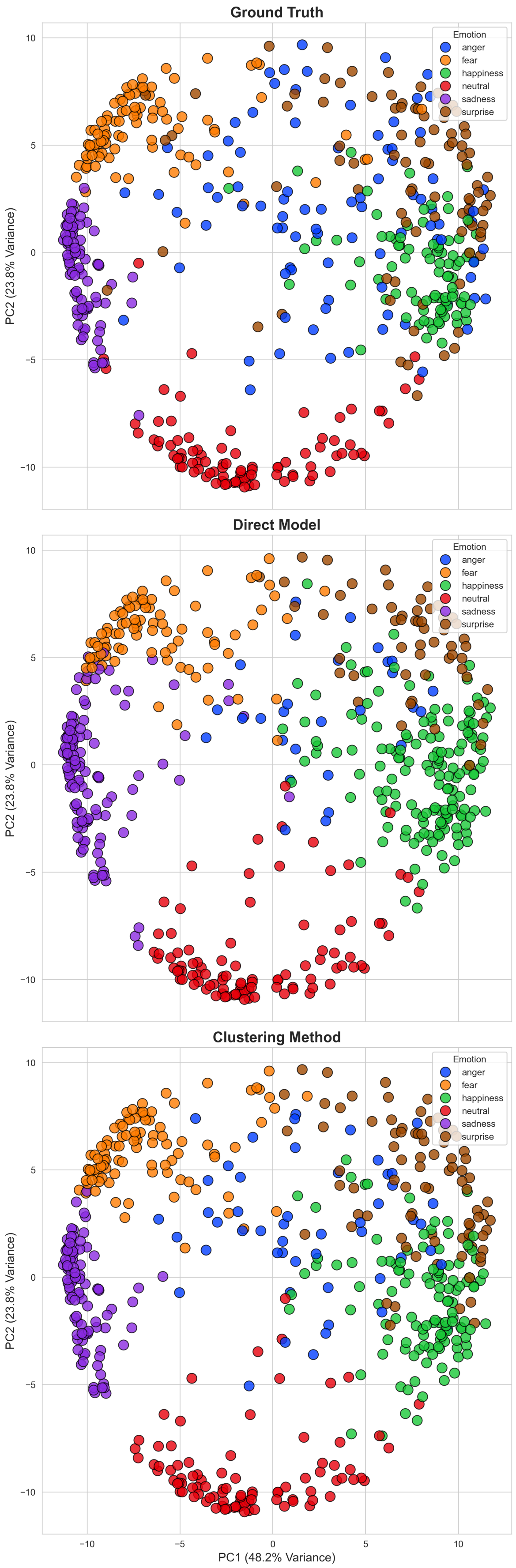

Figure 2: PCA Visualization of embeddings for one person: The clustering method (Bottom) shows better separation of emotion classes compared to the direct model (Center) and approaches Ground Truth (Top).

Algorithm 1: Greedy Cluster Assignment for Speaker-Adaptive Inference

```
Input  : Embeddings $E = \{e_1, \ldots, e_N\}$, speaker IDs $S = \{s_1, \ldots, s_N\}$, pretrained emotion head
         $H$, maximum clusters $K = 6$
Output : Predicted emotion labels $L = \{l_1, \ldots, l_N\}$
Initialize $L \leftarrow$ array of default label (e.g., neutral);
Group indices by speaker: $G \leftarrow$ map from speaker $s$ to indices $i$ where $s_i = s$;
foreach speaker $s$ in $G$ do
    $I \leftarrow G[s]$ ;                                   // Global indices for speaker $s$
    $E_s \leftarrow \{e_i \mid i \in I\}$ ;                          // Speaker embeddings
    $n \leftarrow |I|$;
    if $n = 1$ then
        Compute $p = H(E_s)$ ;
        $L[I[0]] \leftarrow \arg\max_e p$ ;
        continue ;
    $k \leftarrow \min(n, K)$ ;
    Standardize $E_s$ ;
    Perform K-means clustering on $E_s$ with $k$ clusters $\rightarrow$ cluster labels $C \in \{0, \ldots, k-1\}^n$ and
      centroids $M$;
    Compute emotion probabilities for centroids: $P \leftarrow H(M)$;
    Initialize $cluster\_to\_emotion \leftarrow \emptyset$, $used\_emotions \leftarrow \emptyset$;
    Create list of tuples $(P[c, e], c, e)$ for all clusters $c$ and emotions $e$;
    Sort list in descending order by probability;
    foreach tuple $(prob, c, e)$ in sorted list do
        if $c \notin cluster\_to\_emotion$ and $e \notin used\_emotions$ then
            $cluster\_to\_emotion[c] \leftarrow e$;
            Add $e$ to $used\_emotions$;
            if $|cluster\_to\_emotion| = k$ then
                break
    // Fallback: Assign remaining clusters to their highest probability emotion
    for $c \leftarrow 0$ to $k - 1$ do
        if $c \notin cluster\_to\_emotion$ then
            $cluster\_to\_emotion[c] \leftarrow \arg\max_e P[c, e]$;
    // Map local cluster labels back to global utterance indices
    for $j \leftarrow 0$ to $n - 1$ do
        $i \leftarrow I[j]$ ;                                          // Get global index
        $c_{\text{label}} \leftarrow C[j]$ ;                                // Get local cluster label
        $L[i] \leftarrow cluster\_to\_emotion[c_{\text{label}}]$;
return $L$;
```

References

Artem Amentes, Nikita Davidchuk, and Ilya Lubenets. 2022. Russian Emotional Speech Dialogs with annotated text.

Houwei Cao, David Cooper, Michael Keutmann, Ruben Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5:377–390.

Fabio Catania, Jordan Wilke, and Franca Garzotto. 2025. Emozionalmente: A Crowdsourced Corpus of Simulated Emotional Speech in Italian. *IEEE Transactions on Audio, Speech and Language Processing*, PP:1–14.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*.

Iwona Christop. 2024. nEMO: Dataset of emotional speech in Polish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12111–12116, Torino, Italia. ELRA and ICCL.

Iwona Christop and Maciej Czajka. 2025. CAMEO: Collection of multilingual emotional speech corpora. *Preprint*, arXiv:2505.11051.

Mathilde Marie Duville, Luz Alonso-Valerdi, and David I. Ibarra-Zarate. 2021a. Mexican Emotional Speech Database Based on Semantic, Frequency, Familiarity, Concreteness, and Cultural Shaping of Affective Prosody. *Data*, 6.

Mathilde Marie Duville, Luz Alonso-Valerdi, and David I. Ibarra-Zarate. 2021b. The Mexican Emotional Speech Database (MESD): elaboration and assessment based on machine learning. volume 2021.

Philippe Gournay, Olivier Lahaie, and Roch Lefebvre. 2018. A Canadian French Emotional Speech Dataset. In *Proceedings of the 9th ACM Multimedia Systems Conference*, MMSys '18, pages 399–402, New York, NY, USA. Association for Computing Machinery.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Mana Ihori, Taiga Yamane, Naotaka Kawata, Naoki Makishima, Tomohiro Tanaka, Satoshi Suzuki, Shota Orihashi, and Ryo Masumura. 2025. Few-shot personalization via in-context learning for speech emotion recognition based on speech-language model. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Accepted.

Jesin James, Li Tian, and Catherine Watson. 2018. An Open Source Emotional Speech Corpus for Human Robot Interaction Applications. pages 2768–2772.

Leila Kerkeni, Catherine Cleder, Youssef Serrestou, and Kosai Raoof. 2020. French emotional speech database - Oréau.

Vladimir Kondratenko, Artem Sokolov, Nikolay Karpov, Oleg Kutuzov, Nikita Savushkin, and Fyodor Minkin. 2022. Large raw emotional dataset with aggregation mechanism. *arXiv preprint arXiv:2212.12266*.

Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *arXiv preprint arXiv:2306.07691*.

Steven R. Livingstone and Frank A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5):1–35.

O. Martin, I. Kotsia, B. Macq, and I. Pitas. 2006. The eNTERFACE' 05 Audio-Visual Emotion Database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 8–8.

Kari Ali Noriy, Xiaosong Yang, and Jian Jun Zhang. 2023. EMNS /Imz/ Corpus: An emotive single-speaker dataset for narrative storytelling in games, television and graphic novels. *Preprint*, arXiv:2305.13137.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of Interspeech*, pages 2613–2617.

Jiacheng Shi, Hongfei Du, Y. Alicia Hong, and Ye Gao. 2025. EMO-TTA: Improving test-time adaptation of audio-language models for speech emotion recognition. *arXiv preprint arXiv:2509.25495*.

Ingmar Steiner, Marc Schröder, and Annette Klepp. 2013. The PAVOQUE corpus as a resource for analysis and synthesis of expressive speech. In *Phonetik & Phonologie 9 (P&P-9)*, pages 83–84, Zurich, Switzerland. UZH, Peter Lang.

Sadia Sultana, M. Shahidur Rahman, M. Reza Selim, and M. Zafar Iqbal. 2021. SUST Bangla Emotional Speech Corpus (SUBESCO): An audio-only emotional speech corpus for Bangla. *PLOS ONE*, 16(4):1–27.

Zero-Shot Transfer of Pretrained Speech Representations for Multilingual Emotion Recognition

Tomasz Kuczyński
Adam Mickiewicz University ul. Uniwersytetu Poznańskiego 4 61-614 Poznań, Poland
tomkuc2@st.amu.edu.pl

Abstract

Speech emotion recognition remains a challenging task, particularly in low-resource language settings. In this work, we explore the development of a system capable of identifying emotional states in Polish speech using training data exclusively from other languages. Our approach relies on a pretrained speech representation model and follows a strict zero-shot training paradigm, enabling cross-lingual knowledge transfer without access to any Polish data. The system was developed in the context of the Polish Speech Emotion Recognition Challenge (PolEval 2025), which required participants to train models solely on multilingual resources and evaluate them on Polish speech in a zero-shot setup. We present a complete solution encompassing model selection, audio preprocessing, and fine-tuning strategy, and discuss the potential of large-scale language models for cross-lingual emotion recognition.

1 Introduction

Speech Emotion Recognition (SER) is a critical subfield of affective computing, focusing on the automatic identification of human emotional states from vocal signals (Schuller et al., 2011). Speech is a rich communication channel that conveys not only linguistic but also paralinguistic information, making SER increasingly relevant in human-computer interaction, voice service analytics, assistive technologies, and remote psychological assessment (El Ayadi et al., 2011; George and Ilyas, 2024). Accurate detection of emotion in speech enables more natural and empathetic interactions in dialog systems and facilitates better understanding of users' intentions and mental states.

In recent years, advances in deep learning and self-supervised representation learning have significantly improved the performance of SER systems, especially for high-resource languages such as English, German, or Mandarin (Latif et al., 2021). Models such as wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022) allow for the extraction of effective speech representations with minimal task-specific supervision. However, most of these developments have been concentrated on resource-rich languages, which limits their applicability in low-resource or underrepresented linguistic contexts.

The challenge of multilingual SER lies in the language dependence of emotional prosody. While certain acoustic correlates of emotion, such as pitch variation, speech rate, and energy, are considered relatively universal, the way emotions are expressed and perceived is strongly influenced by cultural, contextual, and phonological factors (Shochi et al., 2009; Pell et al., 2009). This makes cross-lingual transfer under zero-shot conditions particularly difficult. Models trained on emotions in one language may fail to recognize or may incorrectly classify equivalent expressions in another.

In recent years, numerous studies have attempted to address the challenges of cross-lingual knowledge transfer in speech emotion recognition, particularly in scenarios where no supervised data is available for the target language. Among the strategies explored to improve cross-lingual generalization are multilingual fine-tuning, techniques for aligning emotional representations across languages (for example, through contrastive learning or adversarial training), and the use of shared feature spaces derived from self-supervised speech models. Despite encouraging results in experimental settings, zero-shot performance, where models are trained without any access to data from the target language, remains substantially lower compared to supervised approaches based on direct fine-tuning.

The Polish Speech Emotion Recognition Challenge (PolEval 2025) directly addresses this problem by proposing a strict zero-shot setup in which participants are required to train their models exclusively on multilingual data and evaluate them solely on Polish speech. This setup reflects real-world deployment conditions for SER systems in low-resource languages. While Polish does not entirely lack emotional speech corpora, high-quality, large-scale, publicly available datasets have only recently become accessible, such as the nEMO corpus (Christop, 2024). As a result, Polish remains underrepresented in the international SER research landscape, making it a compelling target language for empirical evaluation of zero-shot transfer.

In this work, we investigate whether pre-trained self-supervised speech models, specifically WavLM, can serve as a robust foundation for multilingual SER under a zero-shot paradigm. Our contributions are threefold:

- We design and implement a zero-shot SER system based on WavLM, fine-tuned exclusively on speech data from non-Polish corpora.
- We evaluate the system on the official PolEval 2025 benchmark, strictly adhering to the zero-shot evaluation constraints.
- We analyze the challenges of cross-lingual generalization in emotional representation and discuss future directions, including the use of larger models, multitask learning, and data augmentation techniques.

2 Task Description

The Polish Speech Emotion Recognition Challenge (PolEval 2025) is designed as a strict zero-shot classification task. The objective is to recognize one of six discrete emotional categories from spoken utterances: *anger*, *fear*, *happiness*, *neutral*, *sadness*, and *surprise*. No Polish data may be used during training, and all models are evaluated exclusively on Polish speech during testing.

2.1 Training Data

The training set consists of 29,714 utterances derived from 12 publicly available speech emotion recognition corpora. These corpora cover seven different languages and are unified within the multilingual CAMEO dataset. All samples are annotated using a consistent set of six emotion labels.

Table 1 summarizes the total number of training examples per emotion aggregated across all corpora. Although the dataset is relatively large, class imbalance remains a potential challenge.

Emotion	# Samples
Anger	5,212
Fear	4,241
Happiness	5,216
Neutral	7,161
Sadness	5,127
Surprise	2,757

Table 1: Total number of training samples per emotion (aggregated across all corpora).

2.2 Validation and Evaluation Data

The validation set used in this task is based on the existing nEMO corpus, which contains Polish emotional speech recordings labeled with six emotion categories: *anger*, *fear*, *happiness*, *neutral*, *sadness*, and *surprise*. For the purposes of the PolEval 2025 challenge, the entire corpus was adopted as the *dev* set and made available to participants for model tuning and selection. The class distribution within the validation set is shown in Table 2. While the dataset is relatively balanced, minor variations in class frequency are present.

In addition to the validation set, the organizers prepared two separate test sets for final evaluation:

- **test-A**, containing 1,583 recordings, was released during the competition to allow for intermediate evaluation under test-like conditions;
- **test-B**, containing 1,573 recordings, remained unreleased until the end of the challenge and was used to determine the final leaderboard.

For both test sets, emotion labels and class distributions were withheld. This ensured an unbiased zero-shot evaluation scenario, preventing any data leakage or task-specific tuning.

Emotion	# Samples
Anger	749
Fear	736
Happiness	749
Neutral	809
Sadness	769
Surprise	669

Table 2: Number of validation samples per emotion in the Polish nEMO corpus.

3 System Architecture

The architecture of the proposed system consists of three main components: an acoustic representation extraction layer, a classification layer, and a data preparation procedure. The system is based on the WavLM-Base+ model, which was further fine-tuned for the emotion classification task using a multilingual training set.

3.1 Representation Extraction

Raw audio signals were processed using the WavLM-Base+ model together with the feature extractor from the `transformers` library. All input samples were resampled to 16 kHz and standardized to a fixed length of 3.5 seconds, allowing for consistent batch processing. This duration corresponded to the typical utterance length in the test sets and ensured stable training in a resource-constrained environment.

Model parameters were fine-tuned for the classification task without modifying the internal architecture or layers. The weights were initialized using the official version released by the model's authors.

3.2 Classification Layer

The classification head was implemented as a single-layer neural classifier placed on top of the final hidden state of the base encoder. It consists of a dropout layer (with a dropout rate of 0.1) followed by a fully connected linear transformation projecting the encoder output into a 6-dimensional logits vector, corresponding to the six emotion categories defined in the dataset. The input to the classifier was the contextualized embedding of the [CLS] token, representing the aggregate sentence representation.

The output logits were passed to a softmax function during evaluation to produce class probabilities, while during training the raw logits were used directly in the cross-entropy loss function. No class weighting or label smoothing was applied.

This simple classification structure was chosen to avoid overfitting, preserve the generalization capacity of the base encoder, and ensure comparability with related works using similar low-parameter output layers. The design aligns with the standard setup in transformer-based text classification tasks.

3.3 Training Configuration

The model was trained using the AdamW optimizer with a learning rate of 2×10^{-5} and a weight decay of 0.01. A batch size of 16 was used, and gradient accumulation was set to 2 steps, effectively simulating a batch size of 32, which allowed efficient training on the available hardware.

We trained the model for 10 epochs. A linear warm-up was applied during the first 10% of total training steps, followed by a linear learning rate decay schedule. No early stopping or learning rate restarts were used. Dropout with a rate of 0.1 was applied before the final classification layer to reduce overfitting.

After each epoch, the model was evaluated on a held-out validation set, and the checkpoint achieving the highest macro-averaged F1 score was saved as the final model used for test evaluation.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	3×10^{-5}
Batch size (per device)	4
Gradient accumulation	4 steps
Effective batch size	16
Number of epochs	5
Warm-up ratio	10% of total steps
Learning rate schedule	Linear decay
Dropout rate	0.1
Early stopping	No
Evaluation metric	Macro-averaged F1
Mixed precision (FP16)	Yes

Table 3: Summary of training hyperparameters used for fine-tuning the WavLM model.

3.4 Classification Layer

A single-layer classification head was appended to the model's output, consisting of a dropout layer followed by a fully connected linear layer. Its purpose was to assign each representation to one of six emotion classes. The model was trained using a standard cross-entropy loss without class weighting. Optimization was performed using the AdamW algorithm.

Training hyperparameters (number of epochs, learning rate, batch size, gradient accumulation) were selected empirically, considering available computational resources and training stability. The model was evaluated after each epoch on the validation set, and the checkpoint with the highest macro-averaged F1 score was selected as the final model.

3.5 Data Preparation

The training set was constructed based on metadata provided by the PolEval organizers, which included references to specific utterances from the CAMEO corpus. Only samples with clear assignments to one of the six target emotions were retained.

No data augmentation or additional language filtering was applied. The structure of the dataset was preserved in accordance with the competition rules and the zero-shot constraints, without using any Polish data during training.

4 Evaluation and Results

The system was evaluated on three datasets: a validation set (*dev*) and two independent test sets (*test-A* and *test-B*), all containing exclusively Polish speech. The final ranking in the PolEval 2025 competition was based on the macro-averaged F1 score obtained on the hidden *test-B* set, which was not accessible during model development.

Macro-averaged F1 was adopted as the primary evaluation metric, as it accounts for class imbalance and better reflects overall model performance across all emotion categories. For comparison, we also report accuracy, although its interpretative value is limited in the presence of imbalanced labels (Sokolova and Lapalme, 2009).

Dataset	F1-score	Accuracy
Dev	0.7045	0.7120
Test-A	0.5318	0.5338
Test-B	0.5233	0.5429

Table 4: Evaluation results on the validation and test sets.

As shown in Table 4, the system achieved an F1 score of 0.7045 on the validation set, indicating good generalization to previously unseen Polish data. Performance on the test sets was notably lower, highlighting the difficulty of the zero-shot emotion recognition task.

The gap in performance between the validation and test sets is a well-known phenomenon in cross-lingual transfer settings, where the model is evaluated on entirely new speakers, acoustic conditions, or lexical material. This illustrates the inherent challenge of transferring emotional representations across languages without any adaptation to the target language.

Despite these limitations, the system achieved competitive results under strict zero-shot constraints, confirming the potential of multilingual pretrained models for emotion recognition in under-resourced languages.

5 Discussion

The results obtained in this study confirm that multilingual pretrained models can serve as a reliable foundation for cross-lingual speech emotion recognition under zero-shot conditions. The use of the WavLM model enabled the extraction of speech representations that generalized successfully to Polish, despite the language being entirely absent from the training data. This suggests that the model was able to capture acoustic patterns relevant to emotion recognition that are not strictly language-dependent (Chen et al., 2022).

However, a detailed analysis of the results reveals certain limitations. The significant drop in performance between the validation set and the test sets indicates that full generalization to previously unseen data remains difficult. This is especially evident in the case of the test sets, which were not accessible during model selection and better reflect realistic deployment conditions. The observed discrepancy may have several causes.

First, emotional expression varies across languages due to phonological, cultural, and contextual factors (Pell et al., 2009; Shochi et al., 2009), which may have affected the model's ability to transfer knowledge from the training data to Polish. Second, the emotion categories used in the training set may not have fully matched the way those emotions are realized in Polish speech. While label definitions were harmonized, their acoustic realizations could still differ significantly across languages and corpora.

In addition, domain shift between the training and test data may have contributed to the performance degradation. Differences in speaker characteristics, recording quality, speaking style, or class distribution could have introduced inconsistencies that the model was unable to resolve. These factors are particularly important in speech emotion recognition, where prosodic cues are subtle and highly sensitive to contextual and acoustic variation.

It is also worth noting that the relatively higher score achieved on the validation set compared to the test sets may indicate a degree of implicit adaptation during model selection — even though no Polish data was used for training. The validation set was used as a criterion for selecting the best checkpoint, which, while in line with the competition rules, may have introduced a mild form of adaptation to that specific dataset.

Despite these limitations, the system achieved competitive results under strict zero-shot constraints. This confirms that models pretrained on large and diverse multilingual corpora can successfully transfer emotion-related features across languages. These findings support the validity of the zero-shot approach as a feasible strategy for emotion recognition in low-resource languages and highlight the need for further research into the mechanisms and limitations of cross-lingual transfer.

6 Conclusion

The primary objective of this study was to evaluate the feasibility of cross-lingual speech emotion recognition in a strict zero-shot scenario, where the target language (Polish) is completely absent from the training data. To this end, we developed a system based on the pretrained WavLM model, which was fine-tuned for emotion classification using the multilingual CAMEO corpus. A key constraint of the task was the strict separation between training and evaluation languages, reflecting realistic conditions for deploying systems in low-resource settings.

The results obtained on the official test sets of the PolEval 2025 challenge demonstrate that self-supervised models pretrained on large-scale, diverse audio data can effectively learn emotion-related representations that generalize across languages. Despite a noticeable performance drop compared to the validation set, the system maintained stable performance on unseen Polish data, suggesting that certain prosodic features associated with emotion are sufficiently language-independent to support zero-shot transfer.

The proposed approach highlights the practical potential of leveraging multilingual pretrained models for building SER systems in languages lacking annotated emotional speech corpora. This opens up opportunities for applications in voice-based interaction systems, affective computing in social media analytics, and mental health monitoring in conversational technologies.

Future work could explore several strategies to further improve cross-lingual generalization. These include domain adaptation methods, contrastive learning for more discriminative emotion representations, multitask learning that combines emotion recognition with auxiliary tasks such as speaker or language identification, and the use of larger or domain-specific pretrained models. It would also be valuable to investigate the impact of class imbalance, stylistic variation, and sociolinguistic factors on the quality of zero-shot transfer.

The findings presented in this paper represent a step toward building more robust and language-independent emotion recognition systems capable of operating in truly low-resource conditions.

Resources

To facilitate reproducibility and further research, we provide public access to the source code implementing our system at github.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12449–12460.

Sanyuan Chen, Chengyi Wang, Yu Wu, Shujie Wu, Yanmin Qian, and Dong Yu. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Iwona Christop. 2024. nemo: Dataset of emotional speech in polish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12111–12116, Torino, Italia. ELRA and ICCL.

M. El Ayadi, M. S. Kamel, and F. Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Swapna Mol George and P. Muhamed Ilyas. 2024. A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. *Neurocomputing*, 568:127015.

Wei-Ning Hsu, Bastian Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*, pages 1474–1484.

Siddique Latif, Junaid Qadir, Adnan Qayyum, Muhammad Usama, and Shahzad Younis. 2021. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14:342–356.

Marc D. Pell, Laura Monetta, Silke Paulmann, and Sonja A. Kotz. 2009. Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, 33(2):107–120.

B. Schuller, S. Steidl, and A. Batliner. 2011. Recognizing realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9–10):1062–1087.

T. Shochi, V. Aubergé, and A. Rilliard. 2009. Cross-cultural perception of prosodic functions in expressive speech: some japanese/french/chinese examples. *Intercultural Pragmatics*, 6(2):237–266.

M. Sokolova and G. Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Author Index