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
November 21, 2025 ©2025 Association for Computational Linguistics

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...”