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

(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 **LLM**s 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 meta-linguisitic 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 Finetuning 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 Chernianskii, 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.