Śmigiel Dataset: Laying Foundations for Investigating Machine-Generated Text Detection in Polish

Jakub Strebeyko[1], **Alina Wróblewska**[2], **Piotr Przybyła**[3,2]
[1] University of Warsaw, Warsaw, Poland
[2] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[3] Universitat Pompeu Fabra, Barcelona, Spain
alina@ipipan.waw.pl

Abstract

We present Śmigiel, the first open dataset for training and evaluating machine-generated text (MGT) in Polish. The dataset includes a collection of human-written text fragments from six domains, which are used to prompt text generation by eight language models capable of producing credible Polish text. In addition to the raw corpus of over 462K generated texts, we also release a cleaned source- and domain-balanced dataset suitable for training and evaluating MGT detectors. Finally, we conduct preliminary experiments with text classifiers, showing that task difficulty depends on the text domain, the generating language model, and the availability of similar data in training. The results indicate that MGT detection in Polish can be approached with general-purpose classifiers that generalize well to new LLMs, but struggle to adapt to genres not represented in the training data.

Keywords: machine-generated text detection, corpus, large language models, Polish

1. Introduction

While language models have long been an important tool in the natural language processing (NLP) toolkit, recently it has become possible to use them to generate content that is not easily distinguishable from human-written text. Moreover, these Large Language Models (LLMs) are now widely available to the public through chatbot services, such as ChatGPT, Gemini, or Claude. Because they are easy to access and use, they are increasingly applied in a variety of scenarios of natural language generation (NLG) that were previously performed by humans.

There are many situations where it is important to know if a given piece of text was written by a person. Some examples include:

- when rather than the text itself, we value the act of writing, as a cognitive exercise or evaluation measure, e.g., in education (Frohock, 2025),
- in important document creation, when the credibility of the author plays a crucial role, e.g., in science (Májovský et al., 2023) or law (Frohock, 2025),
- in high-stakes scenarios, where the weaknesses of LLMs (especially hallucination) can lead to serious consequences, e.g., in health-related publications (Milmo, 2023),
- in malicious use-cases, where LLMs can be applied to generate content at scale, e.g., misinformation (Zhou et al., 2023) or fraud (Gressel et al., 2024).

Due to this need, the research on machine-generated text (MGT) detection has been growing recently, in terms of resources (corpora of MGT and human-written text), classifiers (distinguishing the two text types), and evaluation frameworks (especially as shared task events). However, most of this effort has been directed towards English, although there are many more languages, for which LLMs are mature enough to produce convincing output, on some benchmarks surpassing English (Kim et al., 2025).

Here, we present Śmigiel (**S**potting **M**achine-**Ge**nerated Text from **L**LMs for Polish), the first dataset for training and testing MGT detection solutions for Polish. Specifically, we contribute the following:

- a raw collection of 462 360 text fragments in Polish generated by eight state-of-the-art LLMs of various sizes based on human-written prefixes,
- a source- and domain-balanced and cleaned dataset of 64 538 human-written and machine-generated fragments, suitable for training and evaluation of MGT detection tools,
- a preliminary study into the performance of baseline approaches to distinguishing human-written text (HWT) and machine-generated text (MGT).

In order to encourage further research in this area, we openly share the created datasets.[1] It is worth noting that Śmigiel was newly used in Task 1 of the PolEval 2025 (Przybyła et al., 2025).

[1] https://doi.org/10.5281/zenodo.18919631

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10556–10568
11–16 May 2026.

2. Related Work

This section presents previous work in the three research streams that have led to our study: the proliferation and improvement of LLMs capable of generating text in Polish (Section 2.1), shared tasks on MGT detection in various languages (Section 2.2), and the intense development of MGT detection approaches (Section 2.3).

2.1. Large Language Models for Polish

In recent years, several large-scale language models have been developed to support the Polish language and reflect its cultural and linguistic specificity. Earlier research in Polish NLP research mainly focused on transformer-based encoders, e.g., HerBERT (Mroczkowski et al., 2021) or Polish RoBERTa (Dadas et al., 2020). These models were based on the BERT and RoBERTa architectures and were pre-trained and fine-tuned on large Polish corpora. They proved highly effective in handling morphologically rich languages such as Polish and represent a key milestone in Polish NLP.

The next significant step forward was the development of generative large language models. Polish support has increasingly been integrated into most prominent multilingual LLMs, e.g., Llama (Meta Team et al., 2024), Mistral (Jiang et al., 2023; Mistral AI team, 2024), Google Gemma (Gemma Team, 2025), Qwen (Yang et al., 2025), and DeepSeek (DeepSeek-AI, 2024). Their open-weight availability has significantly supported research and downstream adaptation for Polish-based applications.

In addition to these multilingual LLMs, two important families of large-scale Polish-specific LLMs have been developed: Bielik (Ociepa et al., 2025) and PLLuM (Consortium PLLuM, 2025). These initiatives represent an important shift from adapting foreign models to building Polish-native generative systems that better capture linguistic nuance, stylistic variation, and cultural context.

Most of these LLMs can serve a dual function in MGT detection research. They can generate content for training and evaluation, or be used as the underlying architecture for MGT detectors.

2.2. MGT Shared Tasks and Corpora

With the rapid development of MGT systems, the need for their benchmarking has become increasingly important. One of the standardized frameworks for evaluating NLP and NLG systems and reporting their performance is a shared task. The main goal of the shared task is to evaluate participating systems on publicly released datasets using a well-defined and consistent evaluation methodology. The MGT detection challenge has already led to several shared tasks organized for multiple languages.

SemEval-2024 Task 8 (Wang et al., 2024) arranged three subtasks: (1) monolingual (English) and multilingual binary classification, (2) multi-way classification aimed at identifying exact text source (human or specific LLMs), and (3) human–machine text boundary detection aimed at identifying the transition point in mixed-authorship texts. Its successor, GenAI Content Detection Task 1 (Wang et al., 2025), included two similar subtasks: English monolingual and multilingual binary classification (covering Arabic, Chinese, Dutch, German, Hebrew, Hindi, Indonesian, Italian, Japanese, Kazakh, Norwegian, Russian, Spanish, Urdu, and Vietnamese). Additionally, two related challenges were organized within GenAI Content Detection: Academic Essay Authenticity Challenge (Chowdhury et al., 2025) and Cross-Domain MGT Detection Challenge (Dugan et al., 2025). The DagPap24 shared task (Chamezopoulos et al., 2024) focused on identifying machine-generated scientific papers, while Generative AI Authorship Verification Task (Bevendorff et al., 2025) introduced two subtasks: (1) binary classification (MGT vs. HWT), and (2) human-AI collaborative text classification.

Since LLMs can generate content in multiple languages, MGT detection shared tasks have also been organized for non-English languages, e.g., the CLIN33 shared task (Fivez et al., 2024) for Dutch, RuATD (Shamardina et al., 2022) for Russian, and AuTexTification (Sarvazyan et al., 2023) for Spanish.

Our approach to creating Śmigiel closely aligns with the latter. The AuTexTification dataset consists of 160K mixed-authorship texts spanning five textual domains. Its follow-up, IberAuTexTification (Sarvazyan et al., 2024), extended the scope of the original challenge. Although the new dataset was only slightly larger (by roughly 8K instances), it was significantly more diverse: a broader range of generation models was employed, and the data covered seven textual domains in six different languages – including Portuguese, Catalan, Basque, and Galician.

2.3. Automatic MGT Detection

The task of distinguishing MGT and HWT has been recognized since capable language models were available (Corston-Oliver et al., 2001; Lavergne et al., 2008; Labbé and Labbé, 2013; Beresneva, 2016). But more recently, the rapid improvement in the capabilities of LLMs has also led to a growing interest in MGT detection (Crothers et al., 2023; Wu et al., 2025).

Researchers have looked into features that differentiate texts from these two sources and have

found some that are distinctive enough to build approaches that are either completely unsupervised or require only slight calibration. This includes *n*-gram frequencies (Gallé et al., 2021; Hamed and Wu, 2024); various measures derived from token probabilities obtained from language models (Gehrmann et al., 2019; Lavergne et al., 2008; Mitchell et al., 2023), especially perplexity (Vasilatos et al., 2025; Wu et al., 2023); text properties in the embedding space (Tulchinskii et al., 2023); and similarities between the investigated text and its LLM-rewritten variant (Zhu et al., 2023; Maslo and Gargova, 2025).

Others have used MGT corpora to train classifiers, either based on fine-tuning general-purpose LLMs (Radford et al., 2018; Rodriguez et al., 2022; Nguyen-Son et al., 2024) or feature engineering, again including measures relying on probabilities from language models (Verma et al., 2024; Przybyła et al., 2023; Venkatraman et al., 2024), but also stylistic (Shah et al., 2023; Corizzo and Leal-Arenas, 2023), discourse-based (Kim et al., 2024) and others (Mindner et al., 2023).

Note that the approaches signaled above have been designed for various languages, but with a dominant position of English. Regarding Polish, only a single preliminary study on MGT email detection (Gryka et al., 2024) has been proposed. To the best of our knowledge, Śmigiel is the first investigation into multi-domain and multi-model MGT detection for Polish.

3. Śmigiel Dataset

In this section, we outline the Śmigiel dataset, describing the procedure for collecting human-written text (HWT) passages and their machine-generated text (MGT) counterparts (see Section 3.1), their postprocessing and sampling (Section 3.2), and the analysis of the resulting dataset (Section 3.3).

3.1. Raw Data

Studies on MGT detection rely on heterogeneous datasets that reflect variability in both underlying text sources and genres, and LLMs used to generate counterparts to HWT passages. In the spirit of open science, it is desirable that such datasets be based on both texts and models not restricted by commercial licenses.

3.1.1. Human-written Text

To compile a corpus of Polish HWT passages, several datasets are surveyed according to two main criteria: permissive licensing and recency. The latter is crucial to minimize overlap with data potentially used in LLM training, reducing the bias in MGT detection tasks.

domain	dataset	genre
literature	•PLSC (Poświata et al., 2024)	articles
	•open-coursebooks-pl Poświata (2024)	course books
	•Wikiźródła (2025)	books
reviews	•PolEmo (Kocoń et al., 2019)	business, physician, courses and product review
	•Allegro Reviews (Rybak et al., 2020)	product review
	•Filmweb (Przybyła, 2024)	movie review
	•Filmweb+ (Narolski, 2020)	movie review
social	•TwitterEmo (Bogdanowicz et al., 2023)	tweets
	•BAN-PL (Kolos et al., 2024)	posts
wikipedia	•Polish Wikipedia (2025)	encyclopedic entry
news	•Polish Wikinews (2025)	news article
parlamint	•ParlaMint (Erjavec et al., 2025)	parliamentary transcripts

Table 1: Sources of HWT data used in Śmigiel.

Balancing the recency requirement with the relative scarcity of Polish resources, texts from multiple relatively new and publicly available datasets are collected (see Table 1). The source datasets are grouped into six general domains. Four of them – *literature*, *reviews*, *social*, and *wikipedia* – are used for both training and testing, while *news* articles and *parliament* transcripts of Polish Parliament hearings are reserved for testing only.

The texts from the source datasets serve a dual purpose: after postprocessing, they constitute the HWT portion of the Śmigiel dataset and act as seeds or prefixes for LLMs to generate corresponding MGT equivalents.

3.1.2. Machine-generated Text

To construct a set of MGT examples, we prompt LLMs to produce texts of a specific type, each initialized with a prefix derived from an HWT source.

The prefix length varies across genres, with social media texts using the shortest prefixes (minimum 26 characters) and encyclopedic articles requiring the longest ones (at least 50 characters). Prefixes are formed by sequentially extracting complete sentences from HWT passages until the predefined character length for a given domain is

reached, using the sentence segmentation functionality provided by the LAMBO tokeniser (Przybyła, 2022).

```
Preserving its style, continue the
newspaper article that begins with
"As reported by the Silesian Cycling
Coalition in a message dated April 25,
Silesian Railways is abandoning its plan
to introduce bicycle seat reservations."
Do not add comments. Do not insert blank
lines between paragraphs.
```

Figure 1: Translation of an example prompt for news generation.

The HWT prefixes are incorporated into user prompts specifically designed for a given domain. A prompt instructs an LLM to continue writing while keeping the original text style (see Figure 1). To enhance the consistency between MGT and HWT examples, additional metadata details are sometimes included in the prompt, e.g., in the case of reviews, it proved beneficial to specify the type of entity being reviewed – whether it is a product, a hotel, or a movie. Similarly, the congruency of generated coursebook entries improves once the chapter or section title is included in the prompt. In addition, prompts for some genres are enhanced to include instructions to steer the output away from certain formal shortcomings, such as the excessive use of newline characters at the end of Figure 1.

Our research focuses on the Polish language; it is thus essential to employ Polish-specific LLMs to generate Polish texts. Additionally, we also use some multilingual LLMs. The size of a model has an impact on the quality of generated texts: smaller models often produce less coherent outputs with noticeable linguistic errors, while larger models tend to generate more fluent and errorless texts (cf. Gemma Team et al., 2024). We use models of various sizes:

- small: **Bielik-sm** (Ociepa et al., 2025, speakleash/Bielik-7B-Instruct-v0.1[2]), **Llama-sm** (Meta Team et al., 2024, meta-llama/Llama-3.1-8B-Instruct), and **Mistral-sm** (Jiang et al., 2023, mistralai/Mistral-7B-Instruct-v0.3),
- medium: **Bielik-md** (speakleash/Bielik-11B-v2.3-Instruct), **Pllum-md** (Consortium PLLuM, 2025, CYFRAGOVPL/PLLuM-12B-nc-chat), and **Mistral-md** (Mistral AI team, 2024, mistralai/Mistral-Nemo-Instruct-2407),
- large: **Gemma-lg** (Gemma Team, 2025, google/gemma-3-27b-it), **Llama-lg** (meta-llama/Llama-3.3-70B-Instruct).

[2]We use models available in the HuggingFace repository.

To increase the diversity of generated outputs, we apply one of the eight decoding strategies for each generation: *greedy* decoding (no sampling, used as the default), *beam search* (with two beams, e.g., alternative continuations), *contrastive search* (with a repetition penalty), *diverse beam search* (six beams divided into three groups, also with a repetition penalty), and the Llama/PLLuM variant of sampling.

The generations produced by each LLM are then combined into a single CSV table along with their corresponding metadata. In cases where an LLM reproduced the seeding prefix in its generation, this prefix gets removed from the output to avoid mixed authorship.

3.1.3. Summary of Raw Data

A collection of raw data examples includes 460K HWT samples paired with an equal number of MGT counterparts. A detailed statistical summary is presented in Table 2. The selected textual domains vary significantly in sample length. On average, *wikipedia* articles are the longest texts (in terms of the number of characters, words, and sentences per sample), followed by *reviews*, *news*, *literature* excerpts, *parlamint* transcripts, and *social* media posts.

Across all domains, MGT instances tend to be longer than their HWT counterparts. Word lengths are generally consistent – averaging nearly six characters per word, except in *news*, *literature*, and *parlamint*, where words are slightly longer. In contrast, average sentence lengths show clearer distinctions: HWT sentences are longer in *literature* passages, while the longest MGT sentences are in *parlamint* transcripts and encyclopedic articles (*wikipedia*). *Social* media posts, *news* articles, and *reviews* have comparable overall sentence lengths.

3.2. Postprocessing

In this stage, the raw corpus of HWT passages and MGT generations is trimmed, filtered, sampled, and aggregated to make it fit for training and evaluating MGT detection models.

3.2.1. Filtering

In the filtering stage, we load all HWT and MGT fragments and apply a process that consists of dropping some of them and trimming undesired content from others. The process has been designed iteratively based on manual inspection of generated outputs. Specifically:

1. We trim the non-letter characters from the beginning, which are often added by LLMs – especially quotations.

domain	#docs	#characters mean	#characters median	#words mean	#words median	#sentences mean	#sentences median	avg word length	avg sent length
human-written texts (HWT)									
literature	84000	830.87	506.0	132.80	79.0	6.27	4.0	6.17	20.98
reviews	84000	1131.90	556.0	197.59	100.0	10.85	6.0	5.70	18.98
social	84000	252.77	191.0	45.70	34.0	3.15	3.0	5.63	16.31
wikipedia	84000	1623.11	804.5	271.88	135.0	16.96	10.0	5.95	14.49
news	42120	959.61	756.0	155.13	123.0	8.36	7.0	6.20	19.55
parlamint	84240	614.67	530.0	102.75	88.0	5.32	5.0	6.01	20.54
machine-generated texts (MGT)									
literature	84000	886.75	596.0	141.93	91.0	8.07	5.0	6.24	17.91
reviews	84000	1199.50	643.0	208.72	107.0	12.42	6.0	5.97	19.13
social	84000	462.89	382.0	85.90	69.0	5.78	5.0	5.58	16.28
wikipedia	84000	1942.81	1016.0	465.82	170.0	19.54	10.0	5.94	20.45
news	42120	1072.57	899.0	174.30	140.0	9.01	7.0	6.39	20.36
parlamint	84240	720.51	647.0	116.17	104.0	5.58	5.0	6.23	30.48

Table 2: Statistical summary of **raw data** across different domains.

2. If an MGT fragment begins with a repetition of the HWT prefix, as is often the case, we trim it from both fragments.[3]
3. We remove repeated newlines and whitespace characters.
4. We strip the introductions provided by the model and ending with a colon, e.g., "Below is the rest of the text:".
5. When generations created by different LLMs for the same HWT prefix have an overlapping beginning, we remove it, cutting on sentence boundaries (detected using a regular expression).
6. We detect the language for the whole fragment (using `langdetect`[4]) and drop those that are not in Polish.
7. We split text into lines and remove those that:
   - start with one of the standard LLM openings, e.g., "Oczywiście" [*Of course*], "Przepraszam" [*I'm sorry*], "Nie mogę" [*I can't*],
   - are in parentheses (often added by LLMs as comments),
   - are in English,
   - are standard Sejm greetings (for parliamentary domain).
8. In case of generations including looped repetitions, we trim the repeated fragment from the end.
9. Drop fragments that express LLM's refusal to answer the prompt, e.g. "Przepraszam, ale jako model językowy…" [*I'm sorry, but as a language model …*].

After the process above, we drop any fragments that have fewer than 70 characters left. Depending on the LLM and text domain, between 30% and 40% of MGT fragments (and 1% of the HWT ones) get discarded through this process.

3.2.2. Sampling

We group all text fragments into *tuples*, each containing one HWT fragment and up to eight LLM-generated continuations prompted with a prefix extracted from that fragment. Each tuple is then converted into a data *instance* by selecting a single HWT or MGT fragment and discarding the remaining ones. The process is designed so that the final dataset has the following properties:

- No two fragments coming from the same tuple can be included in the dataset.
- Half of the instances are human-written and half are machine-generated.
- One-third of the generated instances are coming from small models (*Bielik-sm*, *Mistral-sm*, and *Llama-sm*), one-third from the medium-sized ones (*Bielik-md*, *Pllum-md*, and *Mistral-md*), and one-third from the large ones (*Gemma-lg* and *Llama-lg*).
- The text lengths of HWT and MGT instances have the same distribution.

The above criteria are met by randomly selecting the fragment to be used in the given instance and trimming it to 95% of the length of the shorter of the two: the HWT fragment and its MGT equivalent.

[3] This operation is not applied to the HWT fragment, if it would take away more than 50% of the fragment.

[4] https://pypi.org/project/langdetect/

source	#docs	#characters		#words		#sentences		avg word length	avg sent length
		mean	median	mean	median	mean	median		
bielik-sm	3342	656.16	364.0	108.42	59.0	6.25	4.0	6.09	17.00
llama-sm	3661	644.66	353.0	108.91	59.0	5.98	3.0	5.84	18.25
mistral-sm	3064	637.44	367.0	104.02	59.0	5.47	3.0	6.17	18.64
bielik-md	3672	707.66	375.0	114.54	60.0	6.4	4.0	6.23	16.75
pllum-md	3410	677.83	387.0	111.01	62.0	6.47	4.0	6.12	16.83
mistral-md	3300	571.63	337.0	98.49	56.0	5.68	3.0	5.94	17.70
gemma-lg	8452	614.41	342.0	100.66	58.0	5.98	4.0	5.97	16.85
llama-lg	3354	708.52	397.5	115.28	65.0	5.11	3.0	6.13	28.49
human	32283	657.98	369.0	110.12	62.0	5.96	4.0	5.96	19.37

Table 3: Statistical summary of Śmigiel samples categorized by source type (MGT vs. HWT).

domain	#docs	#characters		#words		#sentences		avg word length	avg sent length
		mean	median	mean	median	mean	median		
human-written texts (HWT)									
literature	5798	630.28	432.0	100.44	67.0	4.97	4.0	6.21	20.48
reviews	5401	869.23	511.0	151.24	91.0	8.47	6.0	5.69	19.35
social	5377	199.41	152.0	35.92	27.0	2.45	2.0	5.66	16.68
wikipedia	5738	1176.37	650.0	195.97	106.0	10.25	6.0	6.10	19.48
news	2593	749.46	603.0	121.3	98.0	6.88	6.0	6.19	18.76
parlamint	7376	423.95	355.0	71.02	59.0	3.77	3.0	6.02	20.62
machine-generated texts (MGT)									
literature	5948	633.85	437.0	97.13	67.0	5.3	4.0	6.32	17.56
reviews	5205	863.31	503.0	146.6	82.0	7.85	5.0	5.98	18.39
social	5382	196.25	151.0	36.29	28.0	2.88	2.0	5.50	13.69
wikipedia	5775	1144.89	670.0	193.57	112.0	10.87	6.0	5.96	17.84
news	2622	720.26	573.0	112.47	89.0	5.94	5.0	6.39	19.10
parlamint	7323	416.72	346.0	67.25	56.0	3.46	3.0	6.21	23.10

Table 4: Statistical summary of Śmigiel dataset across different domains.

3.2.3. Aggregation

The process described above is executed independently for each domain. The resulting instances are then concatenated and shuffled to produce the corpus described in Section 3.3.

3.3. Śmigiel Final Composition

3.3.1. Statistical Overview of Śmigiel

The final Śmigiel dataset is balanced not only with respect to the main categories (MGT and HWT), but also to the textual domains and generations produced by LLMs of different sizes. The dataset contains 32K HWT and 32K MGT examples, including: (1) 10K MGT samples generated by small LLMs, 10K by medium LLMs, and 12K by large LLMs (see Table 3); and (2) approximately 5.5K HWT and MGT samples drawn from each of four textual domains – *literature*, *reviews*, *social*, *wikipedia* (see Table 4). Additionally, the Śmigiel dataset includes two unseen domains reserved exclusively for testing: *news* with 2.6K HWT and MGT samples each and *parlamint* with 7K samples per category.

To mitigate potential detection of MGT instances based on textual length – given that MGT raw samples are, on average, longer than their HWT counterparts across all domains (Section 3.1.3) – MGT passages are trimmed to match the length of their corresponding HWT originals. As a result, the Śmigiel MGT and HWT instances are closely aligned in length (Table 3): MGT passages average 572–708 characters, 98–115 words, and 5–6 sentences, compared to 658 characters, 110 words, and 6 sentences in HWT fragments. The median values show even closer correspondence between the two categories.

Word length does not differ between HWT and MGT samples. Sentence lengths, however, vary across the two groups: HWT sentences average

19 tokens, while MGT sentences are slightly shorter (17–18 tokens), except for those generated by Llama-lg, which average 28 tokens per sentence.

Similarly, the Śmigiel HWT and MGT instances are comparable in overall length across domains, although they slightly differ in the average number of tokens per sentence.

3.3.2. Analysis of Dataset Quality

Perplexity is a widely used metric for evaluating performance of language models. Lower perplexity values indicate greater confidence and accuracy in predicting the next word in a sequence, while higher values suggest weaker predictive ability. In our study, perplexity is used to compare HWT and MGT instances, offering insight into how natural or predictable each text type appears.

LLM size	LLM	count	mean	
			gpt2-sm	gpt2-xl
small	bielik	1000	42.30	27.49
	llama	1000	31.46	22.75
	mistral	1000	59.39	46.19
medium	bielik	1000	36.55	25.23
	pllum	1000	33.72	22.91
	mistral	1000	49.40	37.12
large	gemma	1000	45.15	27.59
	llama	1000	36.22	23.68
human		1000	84.17	52.56

Table 5: Mean perplexity of Śmigiel MGT and HWT instances.

Perplexity values, based on the log-likelihoods of text tokens conditioned on preceding tokens, are estimated using the *sdadas/polish-gpt2-small* and *sdadas/polish-gpt2-xl* models (Dadas, 2025). Mean perplexity values are calculated over 1000 text samples for each tested LLM and for HWT texts (see Table 5). The perplexity values estimated using the smaller *polish-gpt2-small* model are much higher than those obtained with *polish-gpt2-xl*; however, the Pearson correlation between the two sets of scores remains very strong ($\rho = 0.96$).

MGT texts consistently show lower perplexity than HWT texts, suggesting that LLM-generated texts tend to follow more predictable linguistic patterns. This lower perplexity may reflect higher internal consistency in token sequences, and more frequent use of common lexical items and syntactic structures learned from large-scale training data. In contrast, the higher perplexity observed in HWT texts implies greater intra-language diversity and unpredictability, which can stem from richer lexical variety, more complex sentence structures, and context-dependent creativity typical of human authorship. These findings highlight that, although MGT texts may appear fluent and grammatically coherent, they often lack irregularities and nuanced variations typical in natural human language use.

4. MGT Detection

In order to understand how credible the MGT content appears to ML models, we perform preliminary experiments on the MGT detection task. We split the dataset into one training and three test portions (Section 4.1) and then run general-purpose text classifiers of several sizes (Section 4.2), measuring the classification accuracy for each combination (Section 4.3).

4.1. Experimental Setup

We divide our final Śmigiel dataset into the following portions:

- $train$: 80% of the fragments in the *literature*, *reviews*, *social* and *wikipedia* domains, including generations by all models except *Llama-lg*,
- $test\_\alpha$: 10% of the fragments in the same domain as $train$, including generations by the same models,
- $test\_\beta$: 10% of the fragments in the same domain as $train$, including *Llama-lg* generations,
- $test\_\gamma$: all of the fragments in the *news* and *parlamint* domains.

Each of the portions includes a matching number of MGT and HWT fragments.

This split allows us to train classifiers on the $train$ instances and check their performance when tested on similar text ($test\_\alpha$), generations coming from an unseen model ($test\_\beta$), and text data from an unseen domain ($test\_\gamma$).

4.2. Baseline Classifiers

We check the performance of the following baseline classifiers:

- BiLSTM neural network, using BERT-tokenised input, 32-long embeddings and two LSTM layers (Hochreiter and Schmidhuber, 1997) with 128-long hidden representation, and a final dense layer followed by softmax.
- BERT base (Devlin et al., 2018) fine-tuned for text classification.
- *google/gemma-2b* (Gemma Team and Google DeepMind, 2024), fine-tuned with QLoRa (Dettmers et al., 2023).

classifier	accuracy on test set		
	$test\_\alpha$	$test\_\beta$	$test\_\gamma$
BiLSTM	0.8619	0.8669	0.7357
BERT	0.8977	0.8988	0.7588
GEMMA	0.9685	0.9683	0.8684

Table 6: Classification performance for three types of models, measured as accuracy on the test sets: matching the distribution of training set ($test\_\alpha$), using unseen LLM as generator ($test\_\beta$) or texts from unseen domains ($test\_\gamma$).

We used the code of BODEGA, a framework for testing text classification in the misinformation domain, and all computational details are described in the associated article (Przybyła et al., 2024).

4.3. Results

Table 6 shows the performance of the classification on the three selected test sets. BiLSTM and BERT behave similarly, obtaining good performance on the test set matching the training data and comparable for text coming from the new model. However, when challenged with texts from new domains, they reach much poorer results, i.e., over 10% lower accuracy. This indicates the general-purpose solutions do not generalize well to new genres. Interestingly, this also applies to the largest model *Gemma-2b*, which achieves impressive performance on the first two datasets, but much worse on new data.

5. Discussion

In general, the goals that motivated the creation of Śmigiel are satisfied, as the resulting resource represents a broad picture of Polish texts generated by LLMs. It accounts for a variety of domains and source models of different families and sizes.

The MGT detection effort is a preliminary one, but it already shows interesting insights. Most noticeably, the performance achieved by the baseline methods indicates that the classification, while far from trivial, is noticeably easier than in studies for other languages. For example, in AuTexTification (Sarvazyan et al., 2023) the Transformer-based baseline led to F-score of 57% for English and 69% for Spanish, while our experiments yielded 76-90% accuracy of BERT. A likely explanation for this is that MGT text in Polish is easier to recognize because less web content is available for pre-training in this language (Wenzek et al., 2020), resulting in weaker support for LLMs.

Indeed, our manual analysis has revealed clearly visible MGT features in the generated text. This is most obvious in social media domain, where most LLMs use a formal style of language, clearly discernible from human-written content. In this challenge, prompts play a crucial role. Our approach was based on iterative improvements based on manual inspection of results. However, a more systematic solution can yield better results.

One obvious limitation that a lot of LLM output is discarded in post-processing: 462K of raw generations lead to 32K final training examples. Some of it (between 30% and 40%) is filtered out due to low quality (Section 3.2.1), and this step is undoubtedly necessary. However, the rest is discarded to achieve domain balance and avoid repetitions during sampling (Section 3.2.2). Depending on the downstream MGT classifier or other solutions that use the data, this step might be skipped, allowing access to a much larger dataset.

We envisage the following further use of our dataset:

- Development of MGT detection solutions for Polish. The baselines presented here are very simple and could be improved upon, drawing from research in other languages (Crothers et al., 2023; Wu et al., 2025).
- Deep linguistic analysis of machine-generated language properties. Such studies have been performed for other languages (Zhu et al., 2023; Berber Sardinha, 2024; Amirjalili et al., 2024) and the breadth of our corpus allows to investigate various morphological, lexical and syntactic phenomena.

To make this future work possible, we share the created datasets.[5] The Śmigiel dataset has already been used to conduct a shared task on MGT detection within the PolEval evaluation campaign described elsewhere (Przybyła et al., 2025), including a manual analysis of the differences between human-written and machine-generated fragments.

6. Conclusion

The presented study provides a foundation for investigating MGT detection in Polish. We outlined the broader context of MGT detection research and highlighted the authors' specific contributions to this area within the Polish setting. We described the collection and postprocessing of the Śmigiel dataset, along with the preliminary statistical analysis of its content and underlying raw data. The experimental framework establishes a reliable benchmarking environment for testing MGT detection systems – the first initiative of this kind for Polish.

Our experiments revealed that the effectiveness of MGT detection models for Polish remains high,

[5] https://doi.org/10.5281/zenodo.18919631

as long as they are trained and tested within the same domain. They also generalize well to texts generated by new LLMs unseen during training, suggesting that effective systems can be developed for many important applications in academia, medicine, and law. However, the observed performance drop on out-of-domain data indicates substantial room for improvement – particularly in scenarios where universal detection systems must operate across diverse styles and content, such as in social media contexts.

7. Acknowledgements

This work was supported by the *Ramón y Cajal* grant RYC2024-050327-I, funded by the Spanish State Research Agency (MICIU/AEI/10.13039/501100011033) and by the European Social Fund Plus (ESF+) of the European Union. We also 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.

8. Bibliographical References

Forough Amirjalili, Masoud Neysani, and Ahmadreza Nikbakht. 2024. Exploring the boundaries of authorship: a comparative analysis of AI-generated text and human academic writing in English literature. *Frontiers in Education*, Volume 9 -.

Tony Berber Sardinha. 2024. AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1):100083.

Daria Beresneva. 2016. Computer-Generated Text Detection Using Machine Learning: A Systematic Review. In *Natural Language Processing and Information Systems*, pages 421–426, Cham. Springer International Publishing.

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.

Stanisław Bogdanowicz, Hanna Cwynar, Aleksandra Zwierzchowska, Cezary Klamra, Witold Kieraś, and Łukasz Kobyliński. 2023. Twitteremo: Annotating emotions and sentiment in polish twitter. In *Computational Science – ICCS 2023*, pages 212–220, Cham. Springer Nature Switzerland.

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.

Consortium PLLuM. 2025. Pllum: A family of polish large language models.

Roberto Corizzo and Sebastian Leal-Arenas. 2023. A Deep Fusion Model for Human vs. Machine-Generated Essay Classification. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10.

Simon Corston-Oliver, Michael Gamon, and Chris Brockett. 2001. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 148–155, Toulouse, France. Association for 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.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020. Pre-training polish transformer-based language models at scale. In *Artificial Intelligence and Soft Computing*, pages 301–314. Springer International Publishing.

Sławomir Dadas. 2025. Polish gpt-2 models.

DeepSeek-AI. 2024. Deepseek-v3 technical report.

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.

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.

Gemma Team. 2025. Gemma 3.

Gemma Team and Google DeepMind. 2024. Gemma: Open Models Based on Gemini Research and Technology. Technical report, Google DeepMind.

Gemma Team, Morgane Riviere, Shreya Pathak, and Pier Giuseppe Sessa et al. 2024. Gemma 2: Improving open language models at a practical size.

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.

Paweł Gryka, Kacper Gradoń, Marek Kozłowski, Miłosz Kutyła, and Artur Janicki. 2024. Detection of AI-Generated Emails - A Case Study. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, ARES '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.

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.

Yekyung Kim, Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. One ruler to measure them all: Benchmarking multilingual long-context language models.

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.

Jan Kocoń, Piotr Miłkowski, and Monika Zaśko-Zielińska. 2019. Multi-level sentiment analysis of PolEmo 2.0: Extended corpus of multi-domain consumer reviews. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 980–991, Hong Kong, China. Association for Computational Linguistics.

Anna Kolos, Inez Okulska, Kinga Głąbińska, Agnieszka Karlinska, Emilia Wisnios, Paweł Ellerik, and Andrzej Prałat. 2024. BAN-PL: A Polish dataset of banned harmful and offensive content from wykop.pl web service. In *Proceedings*

of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 2107–2118, Torino, Italia. ELRA and ICCL.

Cyril Labbé and Dominique Labbé. 2013. Duplicate and fake publications in the scientific literature: How many SCIgen papers in computer science? *Scientometrics*, 94(1):379–396.

Thomas Lavergne, Tanguy Urvoy, and François Yvon. 2008. Detecting fake content with relative entropy scoring. In *Proceedings of the 2008 International Conference on Uncovering Plagiarism, Authorship and Social Software Misuse - Volume 377*, PAN'08, pages 27–31, Aachen, DEU. CEUR-WS.org.

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.

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.

Meta Team, Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri et al. 2024. The Llama 3 Herd of Models.

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

Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. *Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT*, pages 152–170. Springer Nature Singapore.

Mistral AI team. 2024. Mistral NeMo.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.

Robert Mroczkowski, Piotr Rybak, Alina Wr 'oblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

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 Regeneration. 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.

Piotr Przybyła. 2022. LAMBO: Layered Approach to Multi-level BOundary identification.

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.

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*, pages 5–15, Warsaw. Institute of Computer Science PAS and Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Juan Diego Rodriguez, Todd Hay, David Gros, Zain Shamsi, and Ravi Srinivasan. 2022. Cross-Domain Detection of GPT-2-Generated Technical Text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1213–1233, Seattle, United States. Association for Computational Linguistics.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: Comprehensive benchmark for polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

pages 1191–1201, Online. Association for Computational Linguistics.

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.

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.

Christoforos Vasilatos, Manaar Alam, Talal Rahwan, Yasir Zaki, and Michail Maniatakos. 2025. HowkGPT: Investigating the Detection of ChatGPT-generated University Student Homework through Context-Aware Perplexity Analysis.

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2024. GPT-who: An Information Density-based Machine-Generated Text Detector. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 103–115, Mexico City, Mexico. Association for Computational Linguistics.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1702–1717, Mexico City, Mexico. 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, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 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.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv 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.

9. Language Resource References

Erjavec, Tomaž and Kopp, Matyáš and Kuzman Pungeršek, Taja and Ljubešić, Nikola and Ogrodniczuk, Maciej and Osenova, Petya and Agirrezabal, Manex and Agnoloni, Tommaso and Aires, José and Albini, Monica and Alkorta, Jon and Antiba-Cartazo, Iván and Arrieta, Ekain and Barcala, Mario and Bardanca, Daniel and Barkarson, Starkaður and Bartolini, Roberto and Battistoni, Roberto and Bel, Nuria and Bonet Ramos, Maria del Mar and Calzada Pérez, María and Cardoso, Aida and Çöltekin, Çağrı and Coole, Matthew and Darǵis, Roberts and de Libano, Ruben and Depoorter, Griet and Diwersy, Sascha and Dodé, Réka and Fernandez, Kike and Fernández Rei, Elisa and Frontini, Francesca and Garcia, Marcos and García Díaz, Noelia and García Louzao, Pedro and Gavriilidou, Maria and Gkoumas, Dimitris and Grigorov, Ilko and Grigorova, Vladislava and Haltrup Hansen, Dorte and Iruskieta, Mikel and Jarlbrink, Johan and Jelencsik-Mátyus, Kinga and Jongejan, Bart and Kahusk, Neeme and Kirnbauer, Martin and Kryvenko, Anna and Ligeti-Nagy, Noémi and Luxardo, Giancarlo and Magariños, Carmen and Magnusson, Måns and Marchetti, Carlo and Marx, Maarten and Meden, Katja and Mendes, Amália and Mochtak, Michal and Mölder, Martin and Montemagni, Simonetta and Navarretta, Costanza and Nitoń, Bartłomiej and Norén, Fredrik Mohammadi and Nwadukwe, Amanda and Ojsteršek, Mihael and Pančur, Andrej and Papavassiliou, Vassilis and Pereira, Rui and Pérez Lago, María and Piperidis, Stelios and Pirker, Hannes and Pisani, Marilina and Pol, Henk van der and Prokopidis, Prokopis and Quochi, Valeria and Rayson, Paul and Regueira, Xosé Luís and Rii, Andriana and Rudolf, Michał and Ruisi, Manuela and Rupnik, Peter and Schopper, Daniel and Simov, Kiril and Sinikallio, Laura and Skubic, Jure and Tungland, Lars Magne and Tuominen, Jouni and van Heusden, Ruben and Varga, Zsófia and Vázquez Abuín, Marta and Venturi, Giulia and Vidal Miguéns, Adrián and Vider, Kadri and Vivel Couso, Ainhoa and Vladu, Adina Ioana and Wissik, Tanja and Yrjänäinen, Väinö and Zevallos, Rodolfo and Fišer, Darja. 2025. *Multilingual comparable corpora of parliamentary debates ParlaMint 5.0*. Slovenian language resource repository CLARIN.SI.

Narolski, Paweł. 2020. *Filmweb+*.

Poświata, Rafał. 2024. *Open Coursebooks PL*.

Rafał Poświata, Sławomir Dadas, and Michał Perełkiewicz. 2024. Pl-mteb: Polish massive text embedding benchmark. *arXiv preprint arXiv:2405.10138*.

Przybyła, Kamil. 2024. *Polish movie reviews dataset*.

Wikinews, contributors. 2025. *Wikinews*.

Wikipedia, contributors. 2025. *Wikipedia, The Free Encyclopedia*.

Wikiźródła, contributors. 2025. *Wikiźródła — wolna biblioteka*.