AK-YS at PAN 2026: Towards Robust Machine-Generated Text Detection under Domain Shift and Obfuscation

Notebook for the PAN Lab at CLEF 2026

András Kovács[1,†], Yi Shi[1,†]

[1]University of Copenhagen, Denmark

Abstract

Machine-Generated Text (MGT) detection is becoming increasingly challenging, especially with domain shift and obfuscation techniques. To address this challenge, we follow a robustness focused approach: iteratively fine-tuning Qwen3-0.6B and Qwen3-14B models on out-of-distribution and obfuscated datasets. The results show that training on diverse data and using larger models improves performance on popular MGT benchmarks, but model size alone cannot ensure performance in obfuscation settings.

Keywords

Generative AI Detection, Machine Generated Text (MGT) Detection, Large Language Models, PAN 2026

1. Introduction

The advancement of large language models (LLMs) has led to great improvements in the quality of machine generated texts. As AI is increasingly capable of producing human-like content, distinguishing between human-authored and machine-generated text has become an important challenge in natural language processing.

To address this challenge, the *Voight-Kampff AI Detection Task* at PAN 2026 introduces an evaluation setting[1]. The task is formulated as a binary classification problem in which systems must determine whether a given text is human-authored (class 0) or machine-generated (class 1). In order to evaluate the generalisation and robustness of detection systems, part of the generated texts in this task are produced by LLMs instructed to imitate specific human authors and modify their writing style accordingly. In addition, the test set contains several distribution shifts, including previously unseen models and unknown obfuscation strategies.

In this paper, we present our approach to this task. Our system focuses on improving robustness under distribution shifts and stylistic variations through training on diverse data that cover a wide range of domains, styles, and obfuscation strategies.

2. Background

Machine-generated text (MGT) detection has received increasing attention in recent years. PAN has continuously contributed to this research direction by organising shared tasks on generative AI authorship analysis since 2024.[2, 3]

Participants have proposed various methods for the detection of machine-generated text, including stylometric feature-based methods [4] and classifiers based on transformer architectures [5, 6]. Overall, LLM-based approaches generally achieve stronger performance than traditional stylometric systems under standard evaluation settings. However, their robustness remains limited when faced with out-of-distribution data and obfuscation techniques.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

[†]These authors contributed equally.

wch731@alumni.ku.dk (A. Kovács); mrv720@alumni.ku.dk (Y. Shi)

To address these challenges, our approach focuses on improving robustness under distribution shifts and stylistic variation through training on diverse out-of-distribution and obfuscated data. Inspired by recent works that focus on data augmentation [7, 6], we further incorporate multiple obfuscation strategies and heterogeneous data sources, from different domains and writing styles.

3. Method

To build a model that generalises well, our main methodology followed fine-tuning a pre-trained large language model with out-of-distribution and obfuscated data.

3.1. Architecture

We opted to use open-weight large language models that have been fine-tuned and evaluated many times. Namely, we have selected two variants from the Qwen3 series, 0.6B and 14B parameters. Then we fine-tuned these models using LoRA (Low-Rank Adaptation) with a simple binary classification head that uses the LLM embeddings to produce the task specific predictions. The models were trained using the datasets and obfuscation techniques described below for two epochs to ensure good generalisation over the various data sources. To address class imbalance, we assigned different weights to AI and human data in the loss function during training.

3.2. Datasets

We combined the PAN 25/26 training and validation datasets with a series of out-of-distribution datasets covering different domains and writing styles. We used 90% of the data for training and the remaining 10% for validation.

PAN 25/26 training and validation dataset The official training and validation set, covering 3 domains: fiction, essay and news, with fiction constituting the majority of the data.

MultiSocial A multilingual benchmark for machine-generated text detection within the domain of short social media texts, covering 22 languages and five social platforms, containing over 470k texts, including both human-written posts and machine-generated variants produced by seven multilingual LLMs through iterative paraphrasing. The human-written portion was collected from publicly available datasets covering platforms such as Twitter, Telegram, Discord, Gab, and WhatsApp, while the generated part was created through multilingual paraphrasing.

MULTITuDE A benchmarking dataset for multilingual machine-generated text detection, containing 74,081 human-authored and machine-generated texts in 11 languages (ar, ca, cs, de, en, es, nl, pt, ru, uk, and zh) generated by 8 multilingual LLMs. The dataset primarily consists of news texts: human-written articles sourced from multilingual news dataset *MassiveSumm*, and machine-generated texts produced by prompting with news headlines.

OpenGen by Krishna et al. [8] uses Wikipedia articles as its sourcce. It consists of 3k two-sentence chunks sampled from the validation split of *WikiText-103*. This was used in a prompt, and from the generated reply, the subsequent 300 tokens were concatenated to serve as the machine-generated text continuation.

LFQA by Krishna et al. [8] is a long-form question-answering dataset, consisting of questions scraped from Reddit. 500 questions from six domains were selected and paired with their corresponding longest human-written answers, resulting in 3k QA pairs.

AuthorMix by Fisher et al. [9] was created primarily for authorship obfuscation, comprising 30k long-form texts from 14 authors and 4 domains, spanning presidential speeches, novels, and diary-style blogs.

Kaggle Essays The human essays are sourced from publicly available essay repositories, academic writings, and manually written samples. The AI-generated part is generated via prompting.

Kaggle Academic A balanced dataset with a total of 6,069 text samples. The human-written part was collected from academic abstracts, and the LLMs were prompted to generate the AI part representing academic, informational, and general domains.

3.3. Obfuscation

We use an obfuscated version of MULTITuDE [10], which includes a diverse set of obfuscation techniques.

Among them, the rule-based techniques:

- **GPTZzzs** A dictionary-based AI detection bypasser, replacing a given ratio of words with their synonyms. [11]
- **GPTZero-Bypasser** Performs Unicode-based obfuscation through homoglyph substitution and random zero-width joiner insertion. [12]
- **Homoglyph Attack** Replaces characters with visually confusable Unicode alternatives using the confusables library.

LLM-based techniques:

- **Backtranslation** Translates texts through an intermediate language and back to the original language using M2M100 models.
- **Paraphrase with different Models** Prompts GPT-3.5-Turbo, PEGASUS and DIPPER to generate paraphrased versions of the texts, preserving semantic content while altering lexical and syntactic patterns.
- **ALISON** Focuses on modifying stylistic features to evade authorship attribution systems while preserving the semantic meaning of the original text. [13]
- **DFTFooler** A lightweight adversarial obfuscation approach designed to make machine-generated text appear human-written to detection systems by generating perturbations using only a pre-trained language model. [14]

As publicly available datasets for obfuscated machine-generated text remain scarce, we further applied the rule-based obfuscation techniques to 5% of our total data. Unlike LLM-based paraphrasing methods, rule-based obfuscation introduce limited noise while better preserving the stylistic properties of human-written texts. In addition, we swapped neighbouring characters to mimic human typographical errors in 1% of the obfuscation part.

4. Results

It was known in advance that the PAN challenge will evaluate submissions on a set of metrics, which we also used during our own validation and testing. Specifically:

- **ROC-AUC**: The area under the ROC (Receiver Operating Characteristic) curve.
- **Brier**: The complement of the Brier score (mean squared loss).
- **C@1**: A modified accuracy score that assigns non-answers (score = 0.5) the average accuracy of the remaining cases.

- **F1**: The harmonic mean of precision and recall.
- **F0.5u**: A modified F0.5 measure (precision-weighted F measure) that treats non-answers (score = 0.5) as false negatives.
- The arithmetic **mean** of all the metrics above.

4.1. Our Test Results

Our methodology relied on iteratively fine-tuning a Qwen3-0.6B and a Qwen3-14B parameter models on out of distribution and obfuscated data. We conducted three training cycles. In the first cycle (0.6B-pan and 14B-pan), the models were fine-tuned only on the official train and validation datasets provided by the challenge organisers. In the second cycle (0.6B-ood and 14B-ood), we used several public datasets (PAN, Kaggle Essays, Kaggle Academic, OpenGPTText, MULTITuDE for training, and LFQA and OpenGen for parameter selection) to expose the models to more diverse genres and writing styles. In the third and last cycle (0.6B-obf and 14B-obf), we further expanded the training data (by the datasets listed in the 3.2. Datasets subsection), and applied various obfuscation techniques (see 3.3. Obfuscation subsection) for data augmentation.

The models were evaluated using a 10% split from the unseen M4GT benchmark made by Wang, et al.[15], which contains an English machine vs human text classification subtask. Table 1 shows the metrics across all six model variants.

Table 1
Performance comparison across model iterations on the M4GT dataset. Best results per metric are shown in bold.

Model	ROC-AUC	Brier	C@1	F1	F0.5u	Mean
0.6B-pan	0.852	0.695	0.669	0.505	0.717	0.688
14B-pan	0.803	0.531	0.516	0.060	0.138	0.409
0.6B-ood	0.895	0.775	0.747	0.675	0.814	0.781
14B-ood	**0.978**	**0.939**	**0.925**	**0.922**	**0.948**	**0.942**
0.6B-obf	0.902	0.823	0.795	0.762	0.843	0.825
14B-obf	0.839	0.785	0.720	0.404	0.573	0.664

We observe that, in general, increasing the amount and diversity of training data resulted in better performing models. The best-performing model on the M4GT benchmark is the 14B-ood, in line with our expectations, as the benchmark does not contain any obfuscated samples, so training on such samples may reduce performance on clean text.

However, since the official test set was expected to contain obfuscated data, the 0.6B-obf and 14B-obf models were selected for submission. Our hypothesis was that larger models should be able to capture more complex patterns inherent to diverse writing styles and obfuscated text.

4.2. Official Results

Table 2 shows the performance of the final fine-tuned models (0.6B-obf and 14B-obf) on the PAN25, PAN26 and ELOQUENT LAB test sets.

The official test results partially confirm our initial assumption and training methodology. The most significant results, the PAN26 test dataset, shows the larger 14B model outperforming the smaller 0.6B variant, achieving a mean score of 0.811 compared to 0.773. This proves our hypothesis, that the increased parameter count allows the model to better capture complex patterns in the obfuscated data.

However, the results on the ELOQUENT dataset contradict the above. This benchmark contains text generated by language models specifically designed to evade MGT detectors. On this dataset, the smaller model performed better, suggesting that model size alone does not necessarily improve robustness, especially against adversarial attacks. To fully understand the reasoning behind this, an analysis of the individual submissions to the ELOQUENT challenge would be required.

Table 2
Performance of the final fine-tuned Qwen3 models across test datasets. Best results per metric per dataset are shown in bold.

Dataset	Model	ROC-AUC	Brier	C@1	F1	F0.5u	Mean
PAN25	Qwen3-0.6B	**0.991**	**0.957**	**0.949**	**0.962**	**0.982**	**0.968**
	Qwen3-14B	0.950	0.867	0.851	0.882	0.942	0.898
PAN26	Qwen3-0.6B	**0.968**	0.700	0.652	0.696	0.851	0.773
	Qwen3-14B	0.911	**0.754**	**0.723**	**0.774**	**0.891**	**0.811**
ELOQUENT	Qwen3-0.6B	**0.854**	**0.605**	**0.593**	**0.724**	**0.865**	**0.728**
	Qwen3-14B	0.849	0.405	0.366	0.489	0.705	0.563

Lastly, both models achieved strong results on the PAN25 test set, though this is not a good representation of general performance, as last year's task's setup was known and incorporated into our training methodology.

5. Conclusion

We investigated the robustness of machine-generated text (MGT) detection by fine-tuning open-weight large language models from the Qwen3 series on increasingly diverse and obfuscated datasets. Using a LoRA-based binary classification head we trained 0.6B and 14B parameter variants in three training cycles; baseline training on the PAN dataset, out-of-distribution (OOD) training, and obfuscation augmented. We included multilingual datasets across diverse topics and writing styles, using both rule-based and LLM-based obfuscation techniques. In summary, our training methodology aimed to expose the models to MGT evasion techniques and improve generalisation across domains.

The results show that training on diverse data improves performance on popular MGT benchmarks. Additionally, larger models were able to capture complex patterns inherent to obfuscated text. However, model size alone cannot ensure performance, as carefully designed adversarial text generation methods were able to deceive models in certain configurations.

References

[1] J. Bevendorff, M. Fröbe, A. Greiner-Petter, A. Jakoby, M. Mayerl, P. Nakov, H. Plutz, M. Potthast, B. Stein, M. N. Ta, Y. Wang, E. Zangerle, Overview of pan 2026: Voight-kampff generative ai detection, text watermarking, multi-author writing style analysis, generative plagiarism detection, and reasoning trajectory detection, 2026. arXiv:arXiv:2602.09147.

[2] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, E. Stakovskii, B. Stein, M. Taulé, D. Ustalov, X. Wang, M. Wiegmann, S. M. Yimam, E. Zangerle, Overview ofnbsp;pan 2024: Multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, andnbsp;generative ai authorship verification condensed lab overview, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2024, p. 231–259. URL: https://doi.org/10.1007/978-3-031-71908-0_11. doi:10.1007/978-3-031-71908-0_11.

[3] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview ofnbsp;pan 2025: Voight-kampff generative ai detection, multilingual text detoxification, multi-author writing style analysis, andnbsp;generative plagiarism detection, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 16th International Conference

of the CLEF Association, CLEF 2025, Madrid, Spain, September 9–12, 2025, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2025, p. 388–411. URL: https://doi.org/10.1007/978-3-032-04354-2_21. doi:10.1007/978-3-032-04354-2_21.

[4] J. K. Ochab, M. Matias, T. Boba, T. Walkowiak, Styloch at pan: Gradient-boosted trees with frequency-based stylometric features, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[5] D. Macko, mdok of kinit: Robustly fine-tuned llm for binary and multiclass ai-generated text detection, 2025. URL: https://arxiv.org/abs/2506.01702. arXiv:2506.01702.

[6] J. Liu, L. Kong, Z. Peng, F. Chen, Generative ai authorship verification based on contrastive-enhanced dual-model decision system, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[7] D. Macko, R. Moro, I. Srba, Increasing the robustness of the fine-tuned multilingual machine-generated text detectors, 2025. URL: https://arxiv.org/abs/2503.15128. arXiv:2503.15128.

[8] K. Krishna, Y. Song, M. Karpinska, J. Wieting, M. Iyyer, Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense, 2023. URL: https://arxiv.org/abs/2303.13408. arXiv:2303.13408.

[9] J. Fisher, S. Hallinan, X. Lu, M. Gordon, Z. Harchaoui, Y. Choi, Styleremix: Interpretable authorship obfuscation via distillation and perturbation of style elements, 2024. URL: https://arxiv.org/abs/2408.15666. arXiv:2408.15666.

[10] D. Macko, R. Moro, A. Uchendu, I. Srba, J. S. Lucas, M. Yamashita, N. I. Tripto, D. Lee, J. Simko, M. Bielikova, Authorship obfuscation in multilingual machine-generated text detection, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, 2024, p. 6348–6368. URL: http://dx.doi.org/10.18653/v1/2024.findings-emnlp.369. doi:10.18653/v1/2024.findings-emnlp.369.

[11] J. Hickey, Gptzzzs: Large language model detection evasion through grammar and vocabulary modification, https://github.com/Declipsonator/gptzzzs, 2024. GitHub repository, accessed 2026-05-26.

[12] Jayyt12161, Gptzero-bypasser, https://github.com/jayyt12161/GPTZero-Bypasser, 2024. GitHub repository, accessed 2026-05-26.

[13] E. Xing, S. Venkatraman, T. Le, D. Lee, Alison: Fast and effective stylometric authorship obfuscation, in: AAAI, 2024.

[14] J. Pu, Z. Sarwar, S. M. Abdullah, A. Rehman, Y. Kim, P. Bhattacharya, M. Javed, B. Viswanath, Deepfake text detection: Limitations and opportunities, in: Proc. of IEEE S&P, 2023.

[15] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, O. M. Afzal, T. Mahmoud, G. Puccetti, T. Arnold, et al., M4gt-bench: Evaluation benchmark for black-box machine-generated text detection, to appear in ACL 2024 (2024).