Team Plagiarism-Detector.com at PAN 2026: Strong Detector, Moving Target — Adversarial Augmentation and Out-of-Distribution Findings

Notebook for the PAN Lab at CLEF 2026

Yurii Palkovskii[1,2,*], Alexei Belov[2]

[1]Plagiarism-Detector.com, Ukraine
[2]Zhytomyr State University, Zhytomyr, Ukraine

Abstract

We describe our submission to Subtask 1 of the PAN 2026 "Voight-Kampff" Generative AI Detection task: a supervised human-vs-AI text classifier built by QLoRA-fine-tuning Qwen3-14B-Base on the reused PAN'25 training data, enlarged with a broad, self-generated battery of obfuscation augmentations. Rather than report a single leaderboard number, we organise the paper around three findings that we believe generalise beyond our system. First, broad obfuscation-aware augmentation lets a single supervised detector reach a macro mean of 0.982 on the (blind, still-private) PAN'25 official test set — a close second to the 2025 winner (0.989) and well above the strongest TF-IDF SVM baseline (0.922). Second, the choice of metric aggregation is not a detail: on heterogeneous, single-class test subsets the same predictions score up to 0.10 higher under micro-averaging than under the official macro-average, which can silently flip conclusions. Third, out-of-distribution drift on PAN 2026 is driven not by exotic obfuscation — our detector handles round-trip machine translation at 0.91–0.95 — but overwhelmingly by a single *newest* generator (GPT-5.5) released after our data freeze, on which recall collapses to 39%. We argue that for this task the binding constraint has shifted from obfuscation breadth to newest-model coverage and zero-shot generalisation, and we report all numbers under a strict no-test-fitting integrity boundary.

Keywords

AI-generated text detection, authorship verification, out-of-distribution generalisation, adversarial data augmentation, evaluation methodology, large language models

1. Introduction

The PAN 2026 "Voight-Kampff" Generative AI Detection task [1], one of the shared tasks of PAN 2026 [2], asks a deceptively simple question: given a single, possibly obfuscated text, was it written by a human or generated by a machine? What makes the 2026 edition hard is its construction. The organisers reuse the 2025 *training* data unchanged and evaluate on a *new, blind* test set that deliberately introduces unseen generators and unknown obfuscations. The task is therefore, by design, an out-of-distribution (OOD) generalisation problem rather than a standard supervised one.

Our entry follows the supervised recipe that won PAN 2025 — a QLoRA-fine-tuned Qwen3-14B classifier [3, 4, 5] — but extends it with a much broader, self-generated augmentation corpus spanning eight-plus obfuscation families. All systems were submitted and evaluated on TIRA [6].

Instead of foregrounding a leaderboard placement, we report three findings that we think matter to other participants and to the task design itself:

1. **Augmentation breadth pays off in-distribution but is not the whole story.** Our detector is competitive with the 2025 winner on the blind 2025 test set (Section 4.2).
2. **Micro- vs. macro-averaging can flip the headline.** On the heterogeneous, partly single-class test collection the gap between the two aggregations reaches 0.10 (Section 4.3).

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
EMAIL: palkovskii.yurii@gmail.com (Y. Palkovskii)
URL: https://plagiarism-detector.com (Y. Palkovskii)

3. **Newest-model drift dominates OOD behaviour.** A model released *after* our data freeze, not obfuscation, is what breaks our detector (Section 4.4).

We deliberately treat our post-hoc subset analysis as diagnosis only: because we could reconstruct the blind test's composition from the per-subset evaluation output, we did *not* retrain on that knowledge and re-submit (Section 5).

2. Background and Task Setup

Task and scoring. Subtask 1 is binary detection. For each input text a system emits a score in $[0, 1]$ (<0.5 human, >0.5 machine, exactly 0.5 undecided), and texts must be scored in isolation, with no cross-document features. Systems are scored with five measures — ROC-AUC, Brier, C@1, F_1, $F_{0.5u}$ — and their arithmetic Mean [1]. Crucially, the overall score is a *macro* average over individual test *subsets* (one per text source × obfuscation condition); for single-class subsets, measures that are undefined (e.g. ROC-AUC on an all-machine subset) are dropped from that subset's mean. This macro construction is central to Section 4.3.

Data. The released training material (PAN'25 train+val) is reused unchanged for 2026; there is no separate 2026 training set. It contains 22 LLM generators plus human text across essays, fiction and news, with a class balance near 62% machine. The 2026 test set is withheld (blinded on TIRA) and contains surprise generators and obfuscations from the same broad domain.

Baselines and prior art. The official baselines are a TF-IDF SVM, PPMd compression, and Binoculars [7]. On the blind 2025 test the TF-IDF SVM is unusually strong (macro mean 0.922), reflecting a large surface-feature gap between human and machine text (Section 4.1). The 2025 winner, *mdok* [3], fine-tuned Qwen3-14B-Base with QLoRA and a single homoglyph augmentation, reaching macro mean 0.989 [8]. Our system shares that backbone but differs in its augmentation strategy and calibration.

3. System: An Adversarially-Augmented Qwen3-14B Detector

3.1. Backbone and fine-tuning

We fine-tune `Qwen3-14B-Base` [4] for sequence classification with QLoRA [5, 9]: 4-bit NF4 weights with double quantisation, LoRA adapters (r=64, α=16, dropout 0.1) on all linear projections plus a trained classification head. Following mdok [3] we use `paged_adamw_32bit` with a constant learning rate and 3% warmup. We depart from mdok in three ways: (i) a focal loss [10] (γ=2.0, α=[0.85, 0.15]) to counter the class imbalance; (ii) genre-balanced sampling so the model does not over-fit the fiction-heavy corpus; and (iii) checkpoint selection by the *minimum* ROC-AUC across four held-out OOD validation distributions (`ood_min_auc`), a worst-case criterion that favours robustness over average-case fit. The 14B adapter trains in roughly 17 h on a single 23 GB GPU and runs inference on the TIRA A100 slot.

A practical observation: a second training epoch (warm-started from the first, halved learning rate) *reduced* OOD performance on every axis (e.g. −0.025 ROC-AUC on the perturbed validation set), with training loss collapsing while validation loss rose. The augmented corpus is exhausted of novel signal within one pass, so we ship the single-epoch checkpoint.

3.2. Training corpus and augmentation

The core contribution of our system is corpus construction. On top of the PAN'25 base we add a large, self-generated augmentation set (≈46k documents total, ≈79% machine after augmentation) produced with frontier models (Claude and Gemini families) and dedicated paraphrasers, spanning eight-plus obfuscation families:

- **Style/voice transforms**: "humanizer" rewrites (including child-voice and ESL-learner registers), style mimicry of named author voices, and multi-step relay rewriting.
- **Heavy paraphrase**: DIPPER-class rewriting that leaves no sentence intact [11].
- **Format/surface attacks**: ELOQUENT-style reformatting, fingerprint-vocabulary swaps, and homoglyph substitution following Macko et al. [12].
- **Newest-model coverage available at freeze time**: text from GPT-5.4 and 2.5-generation Gemini / Claude-4 models, to push the training distribution toward the most recent generators we could access.

The design intent is to attack precisely the surface markers that the TF-IDF baseline exploits, so the supervised model is forced to learn deeper discriminative signal rather than re-implementing TF-IDF with more parameters.

3.3. Calibration

Raw classifier probabilities are mapped to the submission scale with a linear remap around a single decision threshold T, followed by a ± 0.05 "undecided" snap toward 0.5 (which the C@1 measure rewards). T is chosen by a sweep that maximises the worst-case per-distribution mean over our OOD validation sets, yielding T=0.99.

4. Findings

4.1. Why TF-IDF is a brutal baseline: the fingerprint-vocabulary gap

A dissection of the training corpus explains the unusually strong TF-IDF baseline. A small set of "LLM-fingerprint" words (*delve, tapestry, leverage, nuanced, robust*, …) appears about 18× more often per thousand words in machine text than in human text. By contrast, several commonly-cited stylometric markers are near-useless on this corpus: contraction rate (ratio 1.04) and mean sentence length (ratio 0.98) barely separate the classes. A leave-one-generator-out TF-IDF sweep scores ROC-AUC > 0.97 on every fold, because the corpus pairs generators with genres such that a held-out generator always has stylistic neighbours in training. The implication for an OOD test is sharp: "new generator" alone cannot break a surface-feature detector — only obfuscation that removes the fingerprint, or a generator deliberately tuned away from it, can. This motivated our augmentation-first strategy.

4.2. Augmentation breadth is competitive in-distribution

Table 1 reports the macro mean on the blind, still-private 2025 official test set (re-run under the unchanged 2025 evaluator; identity of the dataset confirmed by the organisers). Our Qwen3 system (TIRA handle `plagiarism-detector-com-reg2`) reaches 0.982, a close second to the 2025 winner (0.989) and +0.060 over the strongest baseline. An earlier ModernBERT version of our system [13], trained on a narrower augmentation set, reaches 0.921 — essentially the TF-IDF baseline — which isolates the contribution of both the stronger backbone and the broader augmentation. On the separate ELOQUENT "breaker" adversarial set our Qwen3 system scores 0.936, the best of our submitted versions. We make no claim of beating the 2025 winner: on the identical blind set we are 0.007 below it.

4.3. The micro–macro trap

The official ranking metric is a macro average over test subsets, but tooling (and intuition) often reports a micro average pooled over all documents. These can differ sharply, because the hardest subsets are small in document count but each counts equally under macro. Table 2 shows the same predictions scored both ways. In-distribution the gap is mild (0.006); on the 2026 test it widens to **0.104**. A team reading only the micro number would conclude its 2026 system is nearly as strong as in 2025 (0.946); the macro number it is actually ranked on tells a very different story (0.842).

Table 1
Macro mean on the blind, private PAN'25 official test set (the 2026 evaluation reuses this evaluator). Winner and baseline figures are from the official 2025 task overview [8]; our figures are recomputed with the same macro / drop-undefined rule. We do not claim to beat the winner.

System	Macro Mean
mdok (2025 winner) [3]	0.989
Ours, Qwen3-14B + broad aug. (v0.6)	**0.982**
Baseline TF-IDF SVM	0.922
Ours, ModernBERT + narrow aug. (earlier)	0.921
Baseline Binoculars [7]	0.790

Table 2
Same model, same predictions, two aggregations. Macro is the official ranking metric; micro is the pooled average commonly shown by tooling. The gap widens out-of-distribution.

System	Test set	Micro	Macro
Qwen3 v0.6	PAN'25 (blind)	0.988	0.982
ModernBERT	PAN'25 (blind)	0.930	0.921
Qwen3 v0.6	PAN'26 (self-computed)	0.946	0.842

A subtle hazard compounds this: on single-class subsets (all-human or all-machine), measures such as F_1 or ROC-AUC are undefined. A naive implementation that scores an undefined measure as 0 understates those subsets severely — in one of our intermediate analyses it turned a *perfect* all-human subset (0 false positives) into a reported 0.40. The official rule (drop undefined measures from that subset's mean) is essential, and we recommend participants verify their local evaluation matches it before trusting any subset breakdown.

4.4. Newest-model drift, not obfuscation, dominates OOD

The same Qwen3 detector that scores 0.982 macro on the 2025 test drifts to 0.842 on our self-computed 2026 evaluation (−0.14).[1] Decomposing the drift (Table 3) overturns the obvious hypothesis. *Translation* obfuscation — including a seven-language round-trip chain — is handled well (0.91–0.95). The drift is concentrated almost entirely in one place: **GPT-5.5**, on which recall collapses to 39%. GPT-5.4, which *is* in our training data, transfers cleanly (0.956, 97.5% recall); watermarked LLaMA-3.3 is moderate (0.88).

The explanation is a timeline, not a bug. GPT-5.5 was released on 2026-04-23, *after* our corpus was generated. It did not exist when we trained, so it was untrainable — the canonical OOD situation the task is built to expose. Notably, our one-version-older coverage (GPT-5.4) did transfer, suggesting the failure is the *one-version* jump to an unseen newest model rather than a general inability to detect strong models.

5. Limitations and a Research-Integrity Boundary

Provisional 2026 numbers. Section 4.4's figures are self-computed; the official 2026 macro may differ, and the split we scored may not be the final ranking set.

No test-fitting. Because TIRA returns a per-subset breakdown, we were able to reconstruct much of the blind 2026 test's composition (its generators and obfuscation types). We treat this strictly as post-hoc *diagnosis*. We did *not* generate GPT-5.5-targeted training data and re-submit against the same

[1]These 2026 numbers are self-computed on a public challenge split using the official evaluator; the official 2026 leaderboard was not yet available at writing, so they are provisional.

Table 3
Per-generator macro mean on the self-computed PAN'26 split. The drift is concentrated in GPT-5.5 (released after our data freeze), not in translation obfuscation.

Generator / subset	Macro Mean	Machine recall
GPT-4o	1.000	199/199
GPT-5.4 *(in training)*	0.956	776/796
Watermarked LLaMA-3.3	0.881	183/199
GPT-5.5 *(unseen, post-freeze)*	**0.447**	**45/114**
Round-trip translation (5 subsets)	0.91–0.95	—
Human	0.999	0 false pos.

test: doing so would convert a blind OOD measurement into a fitted one, defeating the purpose of the evaluation. Our reported results are from the blind submission; GPT-5.5 coverage is framed only as future work for a genuinely unseen evaluation.

Worst-case selection. Our `ood_min_auc` selection and worst-case threshold sweep optimise the minimum over distributions rather than the macro average the leaderboard uses; a macro-aligned calibration would likely recover a small amount of additional macro mean without retraining.

6. Conclusion

A strong supervised backbone plus broad obfuscation-aware augmentation produces a detector that is competitive with the prior state of the art in-distribution and robust to a wide range of obfuscations, including aggressive round-trip translation. Yet two factors — not model capacity — govern the out-of-distribution outcome that the task actually measures. The first is mundane but decisive: reporting and optimising the correct *macro* metric. The second is structural: robustness to the *newest* generator, which by construction did not exist at training time. For this task we conclude that the binding constraint has shifted from obfuscation breadth, which is now largely solved by augmentation, to newest-model coverage and zero-shot generalisation to models released after the data freeze. Future work that does not require peeking at the test set: macro-aligned threshold calibration, hard-negative mining for better zero-shot transfer, and keeping the training distribution continuously refreshed with the newest available generators.

Acknowledgments

We thank the PAN and ELOQUENT organisers for the task and the TIRA team for the evaluation platform.

Declaration on Generative AI

During the preparation of this work, the authors used Anthropic's Claude (Claude Code / Opus) and Google Gemini for: (i) implementing and debugging the training, inference, and packaging code; (ii) generating the adversarial augmentation corpus described in Section 3; (iii) data analysis and the production of tables; and (iv) drafting and editing the manuscript text. All experimental design, scientific claims, and reported numbers were verified by the authors against logged experimental outputs; the authors reviewed and edited all content and take full responsibility for the publication's content.

References

[1] J. Bevendorff, R. R. Gunti, J. Karlgren, M. Fröbe, M. Potthast, B. Stein, Overview of the third "Voight-Kampff" Generative AI / LLM Detection Task at PAN and ELOQUENT 2026, in: A. Barrón-Cedeño, A. G. S. de Herrera, E. S. Salido, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[2] 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, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, A. Barrón-Cedeño, A. G. S. de Herrera, E. S. Salido, S. MacAvaney, J. M. Struß (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2026.

[3] D. Macko, mdok of KInIT: Robustly Fine-tuned LLM for Binary and Multiclass AI-Generated Text Detection, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-4038/paper_307.pdf.

[4] Qwen Team, Qwen3 Technical Report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[5] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient Finetuning of Quantized LLMs, in: Advances in Neural Information Processing Systems 36 (NeurIPS 2023), 2023. URL: https://arxiv.org/abs/2305.14314. arXiv:2305.14314.

[6] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

[7] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, T. Goldstein, Spotting LLMs with Binoculars: Zero-Shot Detection of Machine-Generated Text, 2024. URL: https://arxiv.org/abs/2401.12070. arXiv:2401.12070.

[8] J. Bevendorff, Y. Wang, J. Karlgren, M. Wiegmann, M. Fröbe, A. Tsivgun, J. Su, Z. Xie, M. Abassy, J. Mansurov, R. Xing, M. N. Ta, K. A. Elozeiri, T. Gu, R. V. Tomar, J. Geng, E. Artemova, A. Shelmanov, N. Habash, E. Stamatatos, I. Gurevych, P. Nakov, M. Potthast, B. Stein, Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 3504–3534.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, in: The Tenth International Conference on Learning Representations (ICLR 2022), 2022. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal Loss for Dense Object Detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), 2017, pp. 2999–3007. arXiv:1708.02002.

[11] K. Krishna, Y. Song, M. Karpinska, J. Wieting, M. Iyyer, Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval is an Effective Defense, in: Advances in Neural Information Processing Systems 36 (NeurIPS 2023), 2023. URL: https://arxiv.org/abs/2303.13408. arXiv:2303.13408.

[12] 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, 2024, pp. 6348–6368.

URL: https://aclanthology.org/2024.findings-emnlp.369/.

[13] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, I. Poli, Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference, 2024. URL: https://arxiv.org/abs/2412.13663. arXiv:2412.13663.