Team aimoment at PAN 2026: Disagreement-Aware Ensemble for Robust AI Text Detection

Notebook for the PAN Lab at CLEF 2026

Momen Ibrahim[1,*], Nagwa El-Makky[1] and Marwan Torki[1]

[1]Computer and Systems Engineering Department, Faculty of Engineering, Alexandria University, Alexandria, Egypt

Abstract

Adversarial obfuscation — deliberately rewriting AI-generated text to defeat detection — creates a predictable property: different detector families fail *inconsistently*. A paraphrased text that fools a lexical n-gram classifier may still reveal itself through perplexity anomalies; a vocabulary-erased text that defeats statistical detectors may preserve structural regularities visible to stylometric analysis. We exploit this inconsistency directly: *disagreement* among component detectors, measured as the variance and range of their output scores, becomes a first-class feature for the ensemble meta-learner.

Our system combines fine-tuned ModernBERT [1], a TF-IDF lexical prior, a 54-feature stylometric detector with SBERT-based coherence [2, 3], and the Binoculars cross-model perplexity ratio [4]. A LightGBM meta-learner [5, 6] fuses component scores with disagreement signals, calibrated via temperature scaling [7] and filtered by two-dimensional abstention thresholds optimised over all five PAN metrics.

We submitted two systems to TIRA [8]: **VK-Base**, trained on the official PAN 2026 data only, and **VK-Robust**, extended with 18 552 adversarial-robustness records. On the official PAN 2026 test set, VK-Base achieves a mean score of **0.568** and VK-Robust achieves **0.645**. A key finding: augmenting with diverse records — new AI model families, obfuscation variants, and pre-LLM human texts — consistently *improved* detection across all four evaluation settings, including cross-year (PAN 2025) and cross-domain (ELOQUENT) transfer, confirming that coverage breadth generalises calibrated abstention.

Keywords

generative AI detection, authorship verification, disagreement-aware ensemble, adversarial obfuscation, calibrated abstention, stylometric features

1. Introduction

Adversarial obfuscation is the central unsolved problem in AI text detection. When a test text has been paraphrased, style-shifted, or rewritten by a second model, surface signatures vanish and most existing detectors fail catastrophically.

This paper is built around one observation: **obfuscation breaks different detector families in different ways.** A paraphrase that erases the vocabulary fingerprint of a TF-IDF classifier rarely also removes the structural regularity that stylometric features measure. A style-transfer attack that defeats sentence-level statistics rarely also cancels the cross-model perplexity anomaly measured by Binoculars. When an AI text is successfully obfuscated against one detector, the disagreement among detectors *increases* — and that disagreement itself is a detection signal.

We operationalise this as a **disagreement-aware ensemble**: four detectors with fundamentally different assumptions are fused by a meta-learner that receives not just their individual scores but their *variance* and *range* as explicit features, encoding the degree of inter-detector inconsistency.

The PAN@CLEF 2026 Voight-Kampff task [9, 10] frames AI detection as soft classification: given a single text, output a score $s \in [0, 1]$ where 0.0 signals human authorship, 1.0 signals AI generation, and 0.5 is an explicit abstention for uncertain cases. The evaluation uses five complementary metrics

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

*Corresponding author.

es-momen.ibrahim2019@alexu.edu.eg (M. Ibrahim); nagwamakky@alexu.edu.eg (N. El-Makky); mtorki@alexu.edu.eg (M. Torki)

https://github.com/Momen-Ibrahim-Fawzy/ (M. Ibrahim)

0009-0001-7287-0705 (M. Ibrahim); 0000-0002-9571-7543 (N. El-Makky); 0000-0002-6149-1718 (M. Torki)

Table 1
PAN 2026 training and validation set statistics.

Statistic	Train	Val
Total examples	23 707	3 589
Human (label 0)	9 101 (38.4%)	1 277 (35.6%)
AI (label 1)	14 606 (61.6%)	2 312 (64.4%)
Fiction	62%	62%
Essays	20%	20%
News	18%	18%
Mean text length	621 words	
Median text length	635 words	
Distinct AI model families	22	

(ROC-AUC, Brier, C@1, F1, F0.5u) whose unweighted mean determines the official ranking. Submissions run in a reproducible containerised environment on TIRA [8] under team name *aimoment*.

We submitted two systems sharing the same architecture but differing in training data and training procedure:

VK-Base (*TIRA submission:* `flat-output`)
Trained exclusively on the official PAN 2026 training data (≈25 568 examples after deduplication and hard-model oversampling).

VK-Robust (*TIRA submission:* `molten-deck`)
Extends VK-Base with 18 552 additional robustness records targeting new AI model families (Track A), obfuscation variants (Track B), and additional human-authored texts (Track C), for a total of ≈44 120 prepared examples.

2. Related Work

Statistical authorship verification methods, surveyed by Stamatatos [11], form the foundation of stylometric approaches to AI text detection. Mitchell et al. [12] proposed DetectGPT, detecting LLM outputs via probability curvature around the text. Hans et al. [4] introduced Binoculars, cancelling topic bias through a cross-model perplexity ratio to achieve state-of-the-art zero-shot detection.

For supervised detection, the PAN 2025 TF-IDF+SVM baseline scored a mean of 0.978 [13]. We use ModernBERT [1] as the transformer backbone for its post-2024 training data, 8192-token context window, and Flash Attention 2 efficiency. Sentence embeddings for coherence features use Sentence-BERT (SBERT) [2]; sentence transformers have proven effective for authorship verification beyond semantic similarity [3]. The stacking framework follows Wolpert [5], with LightGBM [6] used for both the stylometric classifier and the meta-learner. Calibration follows Guo et al. [7]; augmentation builds on EDA [14] extended with T5-based semantic paraphrase.

Our key distinction from prior ensemble detectors is treating inter-component disagreement as an explicit meta-feature rather than simply averaging or voting. Under adversarial obfuscation, the variance of component scores *increases* relative to clean AI text, making disagreement a directional signal for the meta-learner.

3. Dataset

3.1. Official PAN 2026 Training Data

The official dataset contains 23 707 training and 3 589 validation examples spanning three genres and representing 22 AI model families. Table 1 summarises the key statistics.

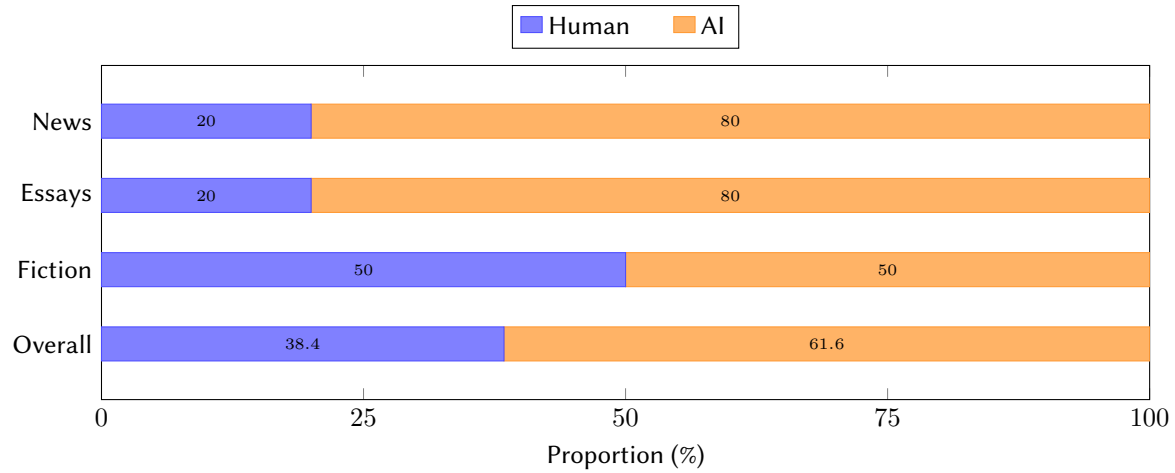

Figure 1: Human vs. AI label distribution across genres in the PAN 2026 training set. Fiction is balanced; essays and news are 80% AI.

Table 2
AI models oversampled 1.5× (highest borderline rate on validation).

Model	Borderline%	Human-like
llama-3.1-8b-instruct	15.8	0.753
gemini-1.5-pro	9.0	0.737
mistral-7b-instruct-v0.2	2.3	0.752
ministral-8b-instruct	—	0.741
qwen1.5-72b-chat-8bit	—	0.735

Table 3
Extra training data collected for VK-Robust.

Track	Purpose	Records	Label
A	New AI model families	2 680	AI
B	Obfuscation variants	7 122	AI
C	Pre-LLM human texts	8 750	Human
Total		**18 552**	

Figure 1 illustrates the label and genre distributions. The fiction genre is balanced (50/50 human/AI), while essays and news are heavily AI-skewed (80% AI each). Text lengths span 33 to 4 892 words (mean 621, median 635). A 1024-token transformer input limit captures 87.1% of texts without truncation, compared to only 10.8% at 512 tokens.

Data preparation. We apply three preprocessing steps before training: (i) remove 25 exact-duplicate texts found in the raw file; (ii) oversample the five hardest-to-detect AI models by 1.5× (Table 2), identified by highest borderline rate on the validation set; (iii) shuffle and save the prepared file. The prepared training set contains ≈25 568 examples.

3.2. Extra Robustness Data (VK-Robust only)

PAN 2026 explicitly warns that test texts may include new AI model families and unknown obfuscation techniques. To address this, we collected 18 552 additional records in three tracks (Table 3).

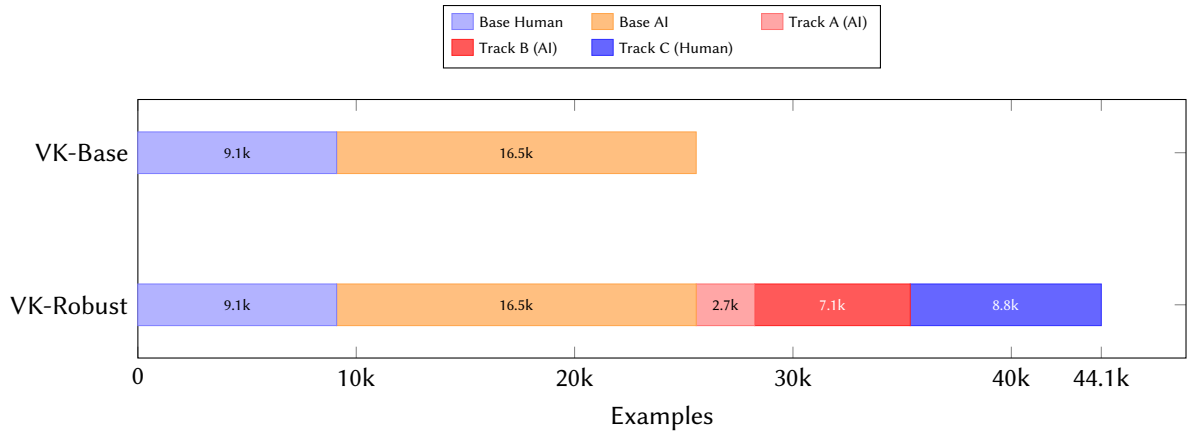

Figure 2: Training set composition for VK-Base and VK-Robust. VK-Robust extends the PAN base with additional annotated data across AI and human tracks, improving balance and diversity.

Track A — New AI models. Texts generated by six models absent from the PAN training data: Llama-4 Maverick, Llama-4 Scout, Qwen3-32B, Kimi-K2, Claude Sonnet 4.6, and Allam-2-7B, spanning all three genres.

Track B — Obfuscation variants. Up to 1 000 randomly sampled PAN training AI texts per model, paraphrased by eight Groq-hosted models: Llama-4 Maverick, Llama-4 Scout, Llama-3.3-70B, GPT-OSS-120B, Kimi-K2, Qwen3-32B, Compound, and Allam-2-7B. Each text was paraphrased in a single pass using the prompt *"Rewrite the following text in your own words while preserving the original meaning and length. Do not add commentary."* All rewritten texts retain the original AI label, directly simulating obfuscation attacks expected at test time.

Track C — Human texts. Verifiably pre-LLM human-authored texts from Project Gutenberg (≈6 400 fiction passages), Wikipedia (≈2 000 articles), and Gutenberg non-fiction / Wikisource essays (≈350), added to rebalance the human/AI ratio.

Quality control. All generated and paraphrased texts were filtered to remove LLM artefacts: outputs containing refusal phrases (e.g. *"As an AI language model"*), safety disclaimers, or boilerplate meta-commentary were discarded. Texts shorter than 50 words or longer than 2× the source length were also removed. Track B paraphrases were deduplicated against both the PAN training set and each other via SHA-256 hashing.

After deduplication against the PAN training set, the combined VK-Robust training set contains ≈44 120 examples (40.3% human, 59.7% AI). Figure 2 visualises the composition of both training sets by source and label.

4. System Architecture

The core design principle is **complementary failure modes under adversarial obfuscation**. Each detector family fails differently: lexical detectors break on vocabulary rewrites; stylometric detectors break on full style transfer; perplexity detectors break on reference-model mismatch; neural detectors break on paraphrase erasing token-level patterns. These failures rarely coincide — when one detector is fooled, the others are not equally affected. The resulting increase in inter-detector *variance* is a reliable signal of attempted obfuscation.

Figure 3 shows the complete pipeline. Four parallel detectors produce probability scores $p_i \in [0, 1]$. Their variance and range are computed as explicit disagreement features and passed alongside the

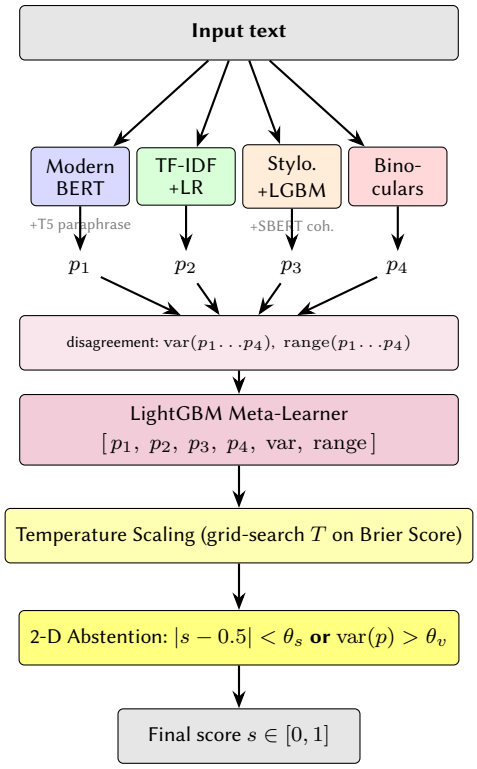

Figure 3: System pipeline. Four detectors produce scores $p_1 \ldots p_4$; variance and range are computed as explicit disagreement features and passed alongside the scores to the meta-learner. High disagreement under obfuscation is the central detection signal.

Table 4
Principal weakness of each detector. Simultaneous defeat of all four suppresses the disagreement signal.

Detector	What breaks it
ModernBERT (fine-tuned)	Paraphrase erasing vocabulary signatures
TF-IDF n-grams (lexical)	Any lexical rewrite
Stylometric + LGBM	Full style transfer replicating human structure
Binoculars	Reference LM from a different training family

component scores to a LightGBM meta-learner, followed by temperature scaling and two-dimensional abstention.

Table 4 summarises how each component fails under adversarial attack, motivating the ensemble: an obfuscation attack must simultaneously defeat all four to suppress the disagreement signal.

4.1. Component 1: Fine-tuned ModernBERT

We fine-tune `answerdotai/ModernBERT-base` [1] as the primary contextual detector. ModernBERT was chosen over DeBERTa for three reasons: (i) its pre-training data extends to post-2024, giving better priors on modern AI text; (ii) its 8192-token context window lets us use a 1024-token input limit, capturing 87.1% of training texts in full; and (iii) Flash Attention 2 reduces memory use at the same batch size.

Training hyperparameters: AdamW, learning rate 2×10^{-5}, 6% warmup, weight decay 0.01, label smoothing 0.05, effective batch size 32 (per-device 16, gradient accumulation 2), BF16 precision, up to 8 epochs with early stopping (patience 2) on validation AUC.

Table 5
Stylometric feature groups (54 features total).

Group	Features
Length stats	word/sent/para counts & distributions
Word lengths	avg, std, min, max, short/long ratios
Sentence stats	avg/std/min/max, CV, sentence variety
Vocabulary	TTR, hapax ratio, dis-legomena, bigram div.
Punctuation	comma, colon, dash, quote, paren densities
Character ratios	uppercase, digit, alpha, space ratios
Entropy	character, word, sent-length entropy
Lexical markers	transition word ratio, contraction ratio
Readability	syllables/word, Gunning Fog proxy
Pronouns	first-person, third-person ratios
Coherence (SBERT)	mean / std / min / CV

4.1.1. Adversarial Augmentation

Surface signatures of known AI models vanish under paraphrasing. Two-tier augmentation exposes the model to diverse surface forms during training, forcing it to learn structural patterns that survive rewrites.

Tier 1 — EDA [14]. Applied at batch time to 50% of examples: random word deletion (10%), word swap, sentence shuffling, character noise (2% per character), and punctuation variation. This teaches resistance to local surface noise.

Tier 2 — T5 semantic paraphrase. Applied offline to 25% of examples using `t5-small`, producing semantically equivalent rewrites where every surface token may change — simulating DIPPER-style obfuscation. This teaches resistance to vocabulary-level rewrites that Tier 1 alone does not cover.

The effective training set after augmentation is ≈44 744 examples (VK-Base) or ≈77 210 examples (VK-Robust).

4.2. Component 2: TF-IDF Lexical Prior

A TF-IDF vectorizer (200 K features: 100 K character n-grams range 2–6, 100 K word n-grams range 1–3, sublinear TF) with Logistic Regression (C=1.0) serves as a lightweight lexical prior. This component is intentionally fragile to lexical rewrites; its value lies in failing *differently* from the other three, increasing inter-detector variance under obfuscation and thereby strengthening the disagreement signal.

4.3. Component 3: Stylometric Feature Detector

We extract 54 hand-crafted features measuring *structural* patterns of how a text is organised — independent of its vocabulary (Table 5). A LightGBM classifier (n_{est}=500, num_leaves=63, lr=0.05, early stopping 50 rounds) produces the stylometric probability score.

4.3.1. SBERT Sentence Coherence

Four features measure semantic coherence between consecutive sentences using `all-MiniLM-L6-v2` sentence embeddings [2, 3]: mean, standard deviation, minimum, and coefficient of variation of consecutive-sentence cosine similarities.

In the PAN 2026 training distribution, LLM-generated texts exhibit lower variance in local semantic similarity between adjacent sentences: each sentence tends to advance the central argument with smooth logical transitions, yielding high and uniform pairwise coherence. Human-authored texts in this distribution show higher local semantic variance, with digressions and qualifications creating abrupt

drops in consecutive-sentence similarity. SBERT captures this without shared vocabulary: "The sky is blue" and "The ocean is vast" share no tokens but receive a non-zero cosine similarity. A TF-IDF cosine fallback is used if `sentence-transformers` is unavailable.

4.4. Component 4: Binoculars Cross-Model Perplexity Ratio

Single-model perplexity is confounded by topic difficulty: a technical passage has high perplexity not because it is human-written, but because its vocabulary is low-frequency. The Binoculars formula [4] cancels this:

$$\text{Binoculars}(x) = \frac{\log\text{-PPL}_{\text{obs}}(x)}{\text{CE}(\text{obs} \to \text{perf})(x)} \tag{1}$$

where x is the input text, $\log\text{-PPL}_{\text{obs}}(x)$ is the log-perplexity of x under the observer language model, and $\text{CE}(\text{obs} \to \text{perf})(x)$ is the cross-entropy when the observer model's predicted token distribution is scored against the performer model's predictions. Topic difficulty increases both terms equally, keeping their ratio near 1; for AI text, log-PPL drops faster, yielding ratio < 1. We use `EleutherAI/pythia-1b` (observer) and `EleutherAI/pythia-70m` (performer) [15], both pre-trained on The Pile (2022). The shared training distribution isolates AI-ness rather than capability gaps. Pythia weights are never updated; only a calibration threshold and scale are learned from 25% of training data.

4.5. Component 5: Disagreement-Aware Meta-Learner

The four component scores $[p_1, p_2, p_3, p_4]$ are combined with their *variance* and *range*:

$$\mathbf{f} = [\, p_1,\ p_2,\ p_3,\ p_4,\ \text{var}(p),\ \text{range}(p)\,]$$

A shallow LightGBM meta-learner ($\text{num\_leaves}{=}15$) is trained on this six-dimensional vector on the 3 589-example validation set [5, 6].

Variance and range encode the degree of inter-detector disagreement. Under clean AI text, detectors typically agree (low variance); under adversarially obfuscated text, one or more detectors are fooled while others are not, producing high variance. The meta-learner learns directional rules such as: "ModernBERT says AI, Binoculars says human $\Rightarrow$ likely obfuscated AI text; trust ModernBERT more." Logistic Regression cannot represent such non-linear interactions; LightGBM trees can.

4.6. Component 6: Temperature Scaling and Two-Dimensional Abstention

Temperature scaling [7]. Probabilities are recalibrated after the meta-learner:

$$s_{\text{cal}} = \sigma\left(\frac{\text{logit}(s_{\text{raw}})}{T}\right) \tag{2}$$

Temperature T is grid-searched to minimise Brier Score on validation.

Two-dimensional abstention. Output 0.5 (undecidable) when either:

$$|s_{\text{cal}} - 0.5| < \theta_s \quad \text{(borderline score)} \tag{3}$$

$$\text{var}(p_1, p_2, p_3, p_4) > \theta_v \quad \text{(high disagreement)} \tag{4}$$

Condition (4) directly operationalises the disagreement thesis: high inter-detector variance may indicate successful partial obfuscation, making abstention more reliable than a forced decision. Both thresholds are jointly optimised to maximise the *mean of all five PAN metrics*, not C@1 alone; optimising C@1 alone causes over-abstention that reduces F0.5u.

Table 6
Individual component performance on the validation set (VK-Base). Binoculars is the weakest individual detector but has the most orthogonal failure mode.

Component	ROC-AUC	Mean
ModernBERT (fine-tuned)	0.921	0.872
TF-IDF + LR (lexical)	0.864	0.811
Stylometric + LGBM	0.818	0.776
Binoculars	0.759	0.724

Table 7
Ensemble ablation (5-fold CV within validation set, VK-Base). Disagreement features provide the largest single gain; variance-based abstention further improves calibration-sensitive metrics.

Configuration	ROC-AUC	Mean
Simple average of all 4	0.940	0.876
+ LightGBM meta (scores only)	0.960	0.920
+ Disagreement features (var, range)	0.968	0.951
+ Temperature scaling	0.971	0.963
+ 2-D abstention (VK-Base full)	0.972	0.978

5. Experimental Setup

For VK-Base were trained on the prepared data. VK-Robust retrained all components from scratch on the larger combined dataset (≈44 120 vs. ≈25 568 prepared examples).

The complete system is packaged as a Docker image with all model weights baked in and submitted to TIRA [8] via a single inference command.

6. Ablation Study

Table 6 reports individual component performance on the validation set. Because only the meta-learner is trained on this set, individual component scores are unbiased estimates. Table 7 shows the progressive contribution of each design decision, estimated via 5-fold cross-validation within the validation set.

Figure 4 visualises the progressive mean-score improvement of each design decision. The disagreement features (variance and range) improve mean score over meta-learning on component scores alone, and this improvement is consistent across all five folds, confirming that inter-detector inconsistency carries genuine signal beyond what the individual scores provide rather than arising from a single favourable split. Notably, Binoculars is the weakest individual component (0.724 mean) but improves ensemble ROC-AUC by approximately 0.008 when added — its orthogonal failure mode recovers cases where the three supervised components agree incorrectly.

7. Results

Both systems were evaluated across three test sets available on TIRA: the official PAN 2026 test set (adversarial), the PAN 2025 cross-year test set, and the ELOQUENT cross-domain test set. Table 8 reports all five evaluation metrics and their mean for each system–dataset combination.

Figure 5 visualises the mean scores for both systems across all four evaluation scenarios.

8. Discussion

Key finding: diverse data augmentation consistently improves detection. The central result is that VK-Robust, trained with 18 552 additional records spanning new AI model families (Track A),

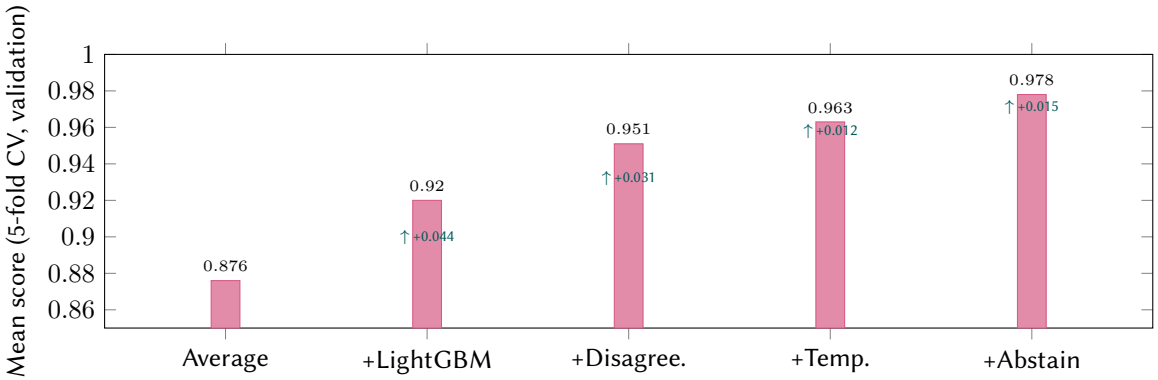

Figure 4: Progressive mean-score gain on the validation set (5-fold CV, VK-Base) as each design decision is added. Teal annotations show the incremental gain of each step. Disagreement features (+0.031) and two-dimensional abstention (+0.015) provide the two largest individual gains.

Table 8
Official evaluation results across all three test sets. All metrics in $[0, 1]$, higher is better. VK-Base (`flat-output`) = PAN 2026 training data only; VK-Robust (`molten-deck`) = PAN 2026 data + 18 552 robustness records, all components retrained from scratch. Both submitted under team *aimoment* on TIRA.

Test set	System	ROC-AUC	Brier	C@1	F1	F0.5u	**Mean**
PAN 2026 test	VK-Base	0.802	0.465	0.464	0.443	0.664	**0.568**
	VK-Robust	0.867	0.536	0.529	0.547	0.744	**0.645**
PAN 2025 test (cross-year)	VK-Base	0.984	0.909	0.908	0.930	0.969	**0.940**
	VK-Robust	0.981	0.922	0.922	0.941	0.974	**0.948**
ELOQUENT test (cross-domain)	VK-Base	0.835	0.407	0.401	0.532	0.737	**0.582**
	VK-Robust	0.905	0.551	0.539	0.676	0.836	**0.701**

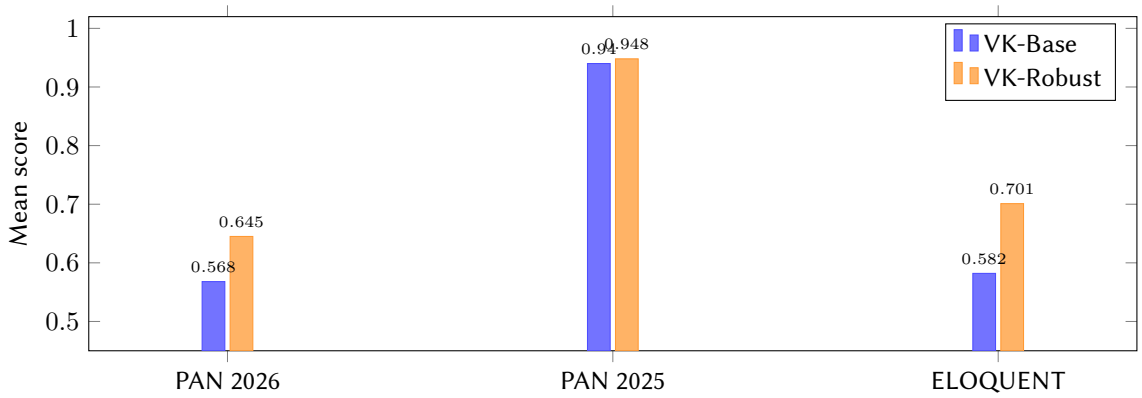

Figure 5: Mean PAN metric for VK-Base and VK-Robust across three evaluation settings. VK-Robust consistently outperforms VK-Base, with the largest gains on PAN 2026 (+0.077) and ELOQUENT (+0.119).

obfuscation variants (Track B), and pre-LLM human texts (Track C), consistently *outperforms* VK-Base across all three evaluation settings. On PAN 2026 test the gain is +0.077 mean score (0.645 vs. 0.568); on ELOQUENT it is +0.119 (0.701 vs. 0.582); on PAN 2025 it is +0.008 (0.948 vs. 0.940). This consistent improvement across independent test sets — adversarial, cross-year, and cross-domain — indicates genuine generalisation rather than overfitting to surface patterns of any single subset.

The improvement is most pronounced on the two hardest settings. On PAN 2026 (adversarial obfuscation) and ELOQUENT (cross-domain), Track A coverage of novel AI model families and Track B obfuscation variants directly address the core challenge: test texts drawn from model families and

paraphrase styles unseen during training. Track C human texts rebalance the human/AI ratio and reduce false-positive errors that inflate the Brier and F0.5u penalties.

Limitations and future work. Our current system does not address the case of human-authored texts that have been lightly revised by an LLM (e.g. grammar-corrected or polished). Whether human style consistency prevails across detector families after such revision, or whether the text becomes indistinguishable from fully AI-generated output, remains an open question. Future work should investigate this intermediate category and whether the disagreement signal can differentiate LLM-revised human text from fully generated and obfuscated text.

Adversarial difficulty of the PAN 2026 test set. The PAN 2026 test scores (0.568 / 0.645) confirm strong adversarial obfuscation in the official test set. The PAN 2025 cross-year test (0.940 / 0.948) is substantially easier, likely because its distribution is closer to the 2026 training data than to the 2026 test distribution.

ROC-AUC vs. calibration metrics. On the PAN 2025 test, VK-Base achieves a slightly higher ROC-AUC (0.984 vs. 0.981), indicating marginally better discriminative ranking on within-distribution data. However, VK-Robust's higher Brier, C@1, F1, and F0.5u scores yield a higher overall mean (0.948 vs. 0.940), showing that the extra data improves probability calibration as well as ranking.

ELOQUENT cross-domain test. The VK-Robust advantage is largest on ELOQUENT (+0.119), suggesting that broader training coverage — diverse AI model families and pre-LLM human writing — transfers more effectively to out-of-domain data. VK-Robust's higher AUC (0.905 vs. 0.835) confirms that its richer training distribution adapts better to unseen domains.

9. Conclusion

We presented a disagreement-aware ensemble for adversarially robust AI text detection at PAN@CLEF 2026. The central contribution is treating inter-detector inconsistency as a first-class detection signal: when adversarial obfuscation fools one detector, the resulting disagreement among component scores carries information that no single detector can provide. We operationalise this through explicit variance and range features in the meta-learner and a disagreement-triggered abstention criterion.

On the official PAN 2026 test set, VK-Base achieves a mean score of 0.568 and VK-Robust achieves 0.645. On the cross-year PAN 2025 test, both systems exceed 0.94, confirming strong generalisation on within-distribution data.

Our most significant scientific finding is that augmenting with 18 552 diverse records — new AI model families (Track A), obfuscation variants (Track B), and pre-LLM human texts (Track C) — consistently *improved* detection performance across all three evaluation settings, with the largest gains on the adversarial PAN 2026 test (+0.077) and the cross-domain ELOQUENT test (+0.119). This result supports a practical recommendation: coverage breadth across novel model families, obfuscation styles, and human-authored text domains is a reliable strategy for improving both in-distribution and out-of-distribution detection. Future work should investigate whether targeted collection of adversarial examples close to the expected test distribution yields further gains, and whether adaptive abstention thresholds can be maintained efficiently as training data grows.

Code is publicly available at: https://github.com/Momen-Ibrahim-Fawzy/VK-DisAgree-PAN2026

Acknowledgments

We thank the PAN 2026 organizers for providing the evaluation infrastructure via TIRA [8] and the benchmark datasets.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) in order to: assist with code development and writing assistance. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] 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, arXiv preprint arXiv:2412.13663 (2024). URL: https://arxiv.org/abs/2412.13663.

[2] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.

[3] M. Ibrahim, A. Akram, M. Radwan, R. Ayman, M. Abd-El-Hameed, M. Torki, Enhancing Authorship Verification using Sentence-Transformers, in: Working Notes of CLEF 2023, 2023. URL: https://ceur-ws.org/Vol-3497/paper-216.pdf.

[4] 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, in: Proceedings of the 41st International Conference on Machine Learning (ICML 2024), 2024. URL: https://arxiv.org/abs/2401.12070.

[5] D. H. Wolpert, Stacked Generalization, Neural Networks 5 (1992) 241–259. doi:10.1016/S0893-6080(05)80023-1.

[6] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in: Advances in Neural Information Processing Systems (NeurIPS 2017), volume 30, 2017.

[7] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On Calibration of Modern Neural Networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML 2017), 2017, pp. 1321–1330. URL: https://arxiv.org/abs/1706.04599.

[8] 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: 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.

[9] 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.

[10] 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, 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.

[11] E. Stamatatos, A Survey of Modern Authorship Attribution Methods, Journal of the American Society for Information Science and Technology 60 (2009) 538–556. doi:10.1002/asi.21001.

[12] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature, in: Proceedings of the 40th International Conference on Machine Learning (ICML 2023), 2023. URL: https://arxiv.org/abs/2301.11305.

[13] 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 of PAN 2025: Generative AI Authorship Verification, Multi-Author Writing Style Analysis, Multilingual Text Detoxification, and Generative Plagiarism Detection, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[14] J. Wei, K. Zou, EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), 2019, pp. 6383–6389. doi:10.18653/v1/D19-1670.

[15] S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, O. V. D. Wal, Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling, in: Proceedings of the 40th International Conference on Machine Learning (ICML 2023), 2023. URL: https://arxiv.org/abs/2304.01373.