Generalizable AI Authership Verification via Adverserial Longformer: Beyond Seen LLMs

Notebook for PAN at CLEF 2026

Mhd Adnan Albani[1], Riad Sonbol[2]

[1]*Independent Researcher*
[2]*Higher Institute for Applied Science and Technology*

Abstract

As large language models (LLM) continue to advance, recognizing AI-generated text from human-authored text became increasingly challenging. A key gap in current works is generalization: detection models tend to latch onto stylistic patterns that are specific to the LLMs they were trained on, rather than learning what fundamentally separates AI-generated text from human writing. As a result, they struggle to generalize when faced with a new LLM they have never seen before. With This paper our team (A_R) addresses this gap with an adversarial based approach. The proposed solution uses Longformer as a backbone with two output heads. The first head classifies whether the text is human or AI-written, while the second head tries to identify which AI model wrote it — but the adversarial signal from the second head is reversed, so the shared backbone is trained to ignore model-specific patterns. We evaluated our approach in the PAN 2026 Voight-Kampff Generative AI Detection task. Our results show that adversarial fine-tuning leads to stronger and more stable ROC-AUC scores compared to the finetuned longformer, supporting the idea that removing model-specific features during training makes the detector more robust across unseen LLMs.

Keywords

PAN 2026, Generative AI Detection, Authorship Verification, Longformer, Adversarial Training, Gradient Reversal Layer, Domain Adaptation

1. Introduction

Large language models have become powerful enough to produce text that is fluent, coherent, and increasingly hard to tell apart from human writing [1, 2]. As these tools become easier to access, the risk of misuse grows — from fake news [3] and academic dishonesty [4] to large-scale automated content on social media [5]. This makes the ability to reliably detect AI-generated text an important and active research problem [6].

Detection methods broadly fall into two camps: supervised classifiers fine-tuned on labelled data [7, 8], and zero-shot approaches that rely on statistical properties of model outputs [9, 10, 11]. While both have shown promise, they share a common weakness: they tend to learn patterns tied to the specific AI systems seen during development, and struggle when faced with new ones [12, 13]. This generalization gap is one of the main challenges in AI-generated text detection.

The PAN shared task at CLEF has tracked this problem across multiple years. The Voight-Kampff Generative AI Authorship Verification task has been running since 2024 [14], with each edition raising the bar on difficulty. The 2025 edition [15] challenged participants to distinguish human-written from AI-generated text across a range of AI systems and genres. The 2026 edition continues and extends this challenge, placing greater emphasis on robustness across AI models — making generalization the defining difficulty of this year's task.

In this paper, we present a system designed to directly address the generalization gap. We fine-tune Longformer [16], a transformer capable of processing documents up to 4,096 tokens, using adversarial training with a Gradient Reversal Layer (GRL) [17]. The GRL pushes the model to ignore which specific AI system wrote a text, forcing it to instead learn features that are common across all AI-generated writing. Our main contributions are:

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

- A Longformer-based Generative AI detection model that handles long documents well beyond the 512-token limit of standard models.
- An adversarial training approach that improves generalization to unseen AI models.
- An empirical evaluation on the PAN 2026 benchmark showing that adversarial fine-tuning leads to stronger and more stable ROC-AUC scores compared to the finetuned LongFormer.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the dataset. Section 4 presents our model. Section 5 covers the evaluation metrics. Section 6 reports our results, and Section 8 concludes.

2. Related Work

The Voight-Kampff Generative AI Authorship Verification task has been part of the PAN lab at CLEF since 2024 [14, 15].In the following subsection we will discuss the main works presented in previous editions of PAN CLEF 2025

2.1. PAN 2024

The 2024 edition [14] was the first instalment of the task, run in a builder-breaker style where PAN participants built detectors and participants developed obfuscation methods to evade them. Systems were given a pair of texts — one human, one machine-generated — and required to output a confidence score. Thirty teams participated, ranked by the mean of ROC-AUC, Brier, C@1, F1, and F0.5u.

The winning system, MarSan [18], scored a perfect 1.000 on the development set but dropped to 0.924 on the official leaderboard ($\Delta = -0.074$), Across nine robustness test variants — including short, Unicode-obfuscated, and paraphrased inputs — the median mean score was 0.990 but the minimum was 0.887, exposing sensitivity to out-of-distribution text conditions despite leading the competition.

2.2. PAN 2025

The 2025 edition [15] introduced a key twist over the previous year: LLMs were instructed to mimic the style of a specific human author rather than generating in their default style, making detection significantly harder. The task retained the builder-breaker format, with the test set containing surprise elements such as new unseen models and unknown obfuscation techniques. Robustness was evaluated across four test variants: clean, short (35 words), Unicode-obfuscated, and paraphrased. Twenty-seven teams participated in Subtask 1, ranked by the arithmetic mean of ROC-AUC, Brier, C@1, F1, and F0.5u.

Macko [19] fine-tuned Qwen3-14B [20] on the combined training and validation splits. To make the model harder to fool, 10% of AI-generated training texts were modified by swapping characters with visually similar Unicode symbols. Rather than using the official validation set to pick the best checkpoint, they used MIX2k — a separate dataset of 2,000 texts from 75 different AI models across 7 languages — to ensure the selected model generalized beyond the training distribution.

The in-distribution validation mean of 0.9978 collapsed to 0.6960 on MIX2k ($\Delta = -0.302$), revealing severe overfitting to the official split. Selecting checkpoints using MIX2k instead recovered a leaderboard mean of 0.989. The authors explicitly flag near-perfect in-distribution scores as a sign of overfit.

3. Dataset

The PAN 2026 dataset [24] is provided as JSONL files. Each record contains the following fields:

- **text**: the document content, authored either by a human or an AI model.
- **label**: a binary value where 0 indicates human-written and 1 indicates AI-generated.
- **model**: the name of the AI model that produced the text, or `human` for human-authored records.
- **genre**: one of three categories — essays, news, or fiction.

Table 1
PAN 2024–2025: training/validation performance vs. official leaderboard mean.

System	Year	Val/train mean	Leaderboard mean	Gap
BinocularsLLM [18]	2024	0.998	0.924	−0.074
you-shun-you-de	2024	—	0.921	—
baselineavengers [21]	2024	—	0.886	—
mdok, in-dist. val. [19]	2025	0.9978	0.989	−0.009
mdok, MIX2k val. [19]	2025	0.6960	0.989	+0.293
ISG-GNN [22]	2025	0.981	0.929	−0.052
Liu et al. [23]	2025	—	0.928	—

Table 1 gives a quick overview of the top systems across both editions. One pattern stands out clearly: every system that scored near-perfectly on its own validation data dropped noticeably on the official leaderboard. The only exception was when mdok swapped its validation set for an out-of-distribution benchmark, which closed the gap entirely. This tells us that the real challenge is not getting high scores on familiar data — it is building a model that still works when it encounters AI systems or writing styles it has never seen before. Generalisation is the core problem, and it is what this work directly targets.

New to this year, two challenges that make the task harder than previous editions:

- **Style mimicry**: AI-generated texts were produced under the constraint that models mimic the writing style of a specific human author, rather than generating in their default style. This makes stylistic fingerprints less reliable as detection signals.
- **Unseen test conditions**: the test set introduces new AI models not present during training and unknown obfuscation techniques, while remaining within the same domain and genre distribution.

The training set contains 23,710 records (9,101 human, 14,609 AI) and the validation set contains 3,589 records (1,277 human, 2,312 AI), spanning 22 distinct language models from the GPT, Gemini, LLaMA, Mistral, DeepSeek, and Falcon families. Figure 1 shows the record distribution across all 22 AI models and the human class for both the training and validation splits.

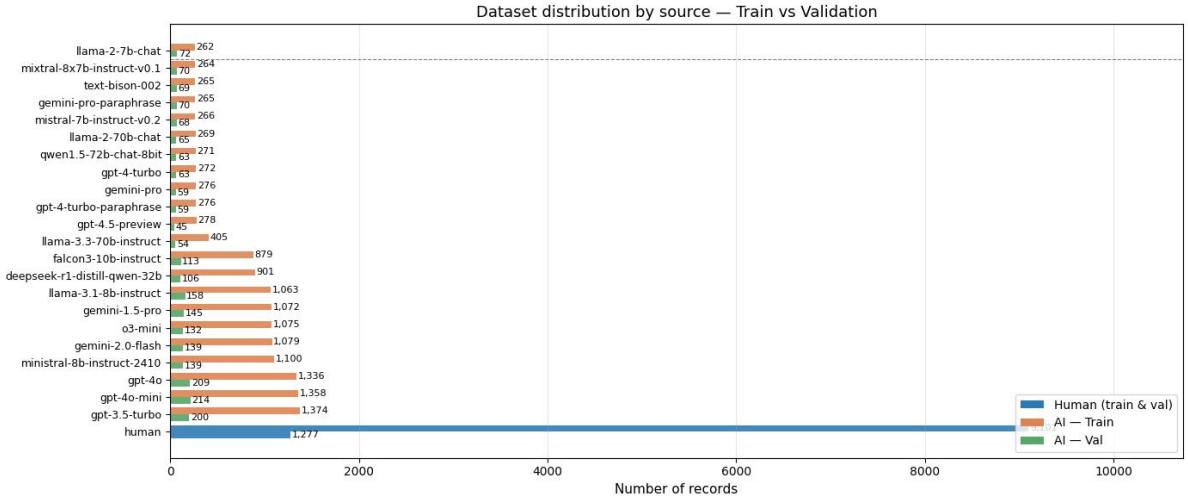

Figure 1: Dataset distribution by source across training and validation splits. Human-authored texts are shown in blue; AI-generated texts in orange (train) and green (val).

Document lengths vary considerably across sources, As shown in Figure 2, Human-authored texts tend to be longer on average than most AI-generated texts, with notable variation across individual AI models. The overall mean length for document in the dataset is more than 600 token, this fact needs to be considered when choosing a suitable encoder to represent the text.

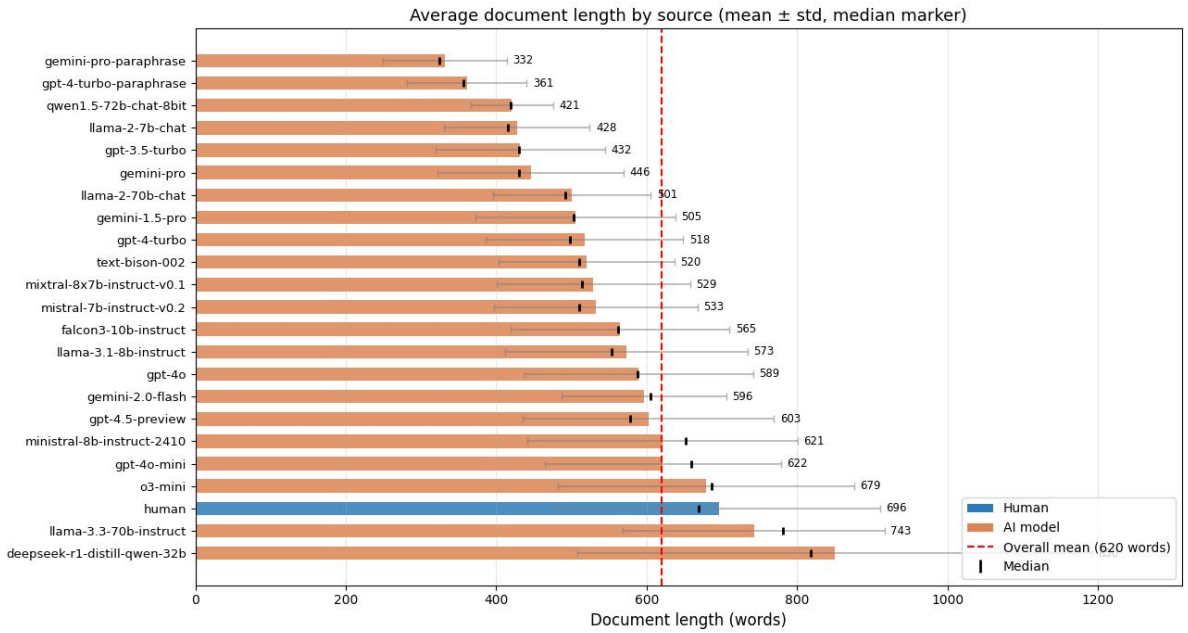

Figure 2: Average document length in words per source, sorted by mean length. Error bars show ±1 standard deviation and vertical ticks mark the median. The red dashed line indicates the overall dataset mean. Human-authored texts are shown in blue; AI-generated texts in orange.

4. Methodology

Figure 3 gives an overview of the system architecture.

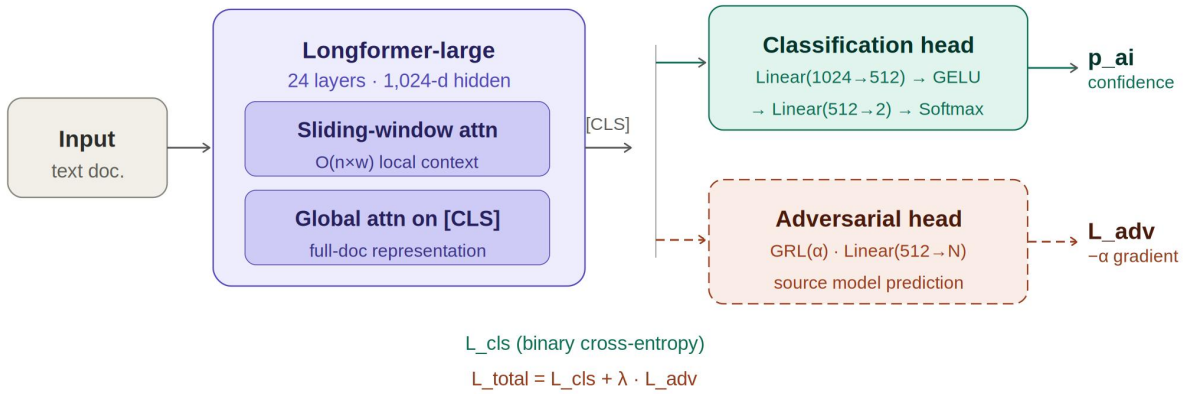

Figure 3: Overview of the adversarial Longformer architecture. The input document is encoded by Longformer-large using sliding-window local attention and global attention on the [CLS] token. The [CLS] representation feeds two heads: a classification head producing the AI confidence score p_{ai}, and an adversarial head with a Gradient Reversal Layer that forces the backbone to discard model-specific features.

4.1. Backbone: Longformer-large

As shown in Figure 2, the majority of documents in the dataset exceed 600 words. Standard BERT-family encoders are limited to 512 tokens, meaning a significant portion of each document would be cut off during tokenisation, discarding stylistic content that may be critical for detection. We therefore use

`allenai/longformer-large-4096` [16] as our backbone, which supports sequences up to 4,096 tokens. The model consists of 24 transformer layers, 16 attention heads, and a hidden dimension of 1,024, initialised from RoBERTa-large and continued with masked language modelling on long documents [16].

Longformer replaces the standard $\mathcal{O}(n^2)$ full self-attention with a sparse attention pattern scaling linearly with sequence length, combining local sliding-window attention with task-motivated global attention [16]. For classification, the `[CLS]` token is assigned global attention, attending to every token in the document and producing a full-document representation suitable for classification.

4.2. Model Heads

The `[CLS]` representation feeds two independent heads, each pushing the shared backbone to different targets during training.

- **Classification head.** This head pushes the backbone toward representations that separate human and AI writing as cleanly as possible. It consists of a two-layer feedforward network that takes the `[CLS]` token as input and maps it to a binary output (human vs. AI). The first layer reduces the 1,024-dimensional representation to 512 dimensions using a linear transformation followed by a GELU activation. A second linear layer then maps this to 2 output logits, one per class. Dropout with a rate of 0.1 is applied before each linear layer to reduce overfitting.

$$\text{Dropout}(0.1) \rightarrow \text{Linear}(1024, 512) \rightarrow \text{GELU} \rightarrow \text{Dropout}(0.1) \rightarrow \text{Linear}(512, 2)$$

- **Adversarial head.** This head pushes the backbone toward learning model-agnostic features. It uses an identical two-layer feedforward architecture but is placed after a Gradient Reversal Layer [17] (GRL). It takes the same `[CLS]` representation, passes it through the GRL, then reduces it from 1,024 to 512 dimensions before mapping to N output logits, one per source model. The GRL acts as an identity function during the forward pass but negates the gradients by $-\alpha$ during backpropagation. This means the adversarial head tries to predict which model wrote the text, while the backbone receives a signal pushing it in the opposite direction — making source model prediction harder rather than easier.

$$\text{GRL}(\alpha) \rightarrow \text{Dropout}(0.1) \rightarrow \text{Linear}(1024, 512) \rightarrow \text{GELU} \rightarrow \text{Dropout}(0.1) \rightarrow \text{Linear}(512, N)$$

The backbone is therefore driven toward representations that are informative for AI detection in general but blind to the identity of the specific generating model — precisely the features needed to generalise to unseen AI models at test time.

4.3. Training Objective

The total training loss combines both heads:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\text{adv}}$$

where $\mathcal{L}_{\text{cls}}$ is the cross-entropy loss for binary AI/human classification, $\mathcal{L}_{\text{adv}}$ is the cross-entropy loss for source model prediction, and λ (adv_weight) controls the adversarial regularisation strength. We experimented with $\lambda \in \{0.1, 0.3, 0.5\}$.

4.4. GRL Alpha Annealing

To stabilise early training, the reversal strength α is annealed from 0 to 1 using a sigmoid schedule:

$$\alpha(t) = \frac{2}{1 + e^{-10 \cdot p}} - 1, \quad p = \frac{t - 1}{T - 1}$$

where t is the current epoch and T is the total number of epochs. Starting at $\alpha = 0$ allows the classifier to learn a meaningful representation before adversarial pressure is applied, with the reversal strength increasing smoothly to $\alpha = 1$ by the final epoch.

4.5. Training Details

Table 2 summarises the hyperparameters used across all experiments. Separate learning rates are applied to the pretrained Longformer backbone and the randomly initialised heads, preventing catastrophic forgetting while allowing the new heads to adapt rapidly.

Table 2
Training hyperparameters.

Hyperparameter	Value
Backbone	allenai/longformer-large-4096
Max sequence length	2,048 tokens
Batch size	8
Learning rate	1×10^{-7} (backbone), 5×10^{-7} (heads)
Optimiser	AdamW (weight decay 0.01)
Training epochs	Ranged from 10 to 13
Adversarial weight λ	{0.1, 0.2, 0.3, 0.4, 0.5}

5. Evaluation Metrics

Following the PAN 2026 evaluation protocol, we report results on six complementary metrics. Together they are designed to penalise different failure modes, making it difficult to game any single one without hurting the others.

- **ROC-AUC:** Measures discrimination ability across all thresholds, independent of calibration. A score of 1.0 is perfect; 0.5 is random.
- **Brier Score (complement):** Computed as $1 - \text{MSE}(p, y)$, it rewards well-calibrated probabilities and heavily penalises overconfident wrong predictions.
- **C@1:** Modified accuracy that rewards abstaining (outputting 0.5) over making a wrong confident prediction, computed as $\frac{1}{n}(n_c + n_u \cdot \frac{n_c}{n})$.
- **F1:** Macro-averaged harmonic mean of precision and recall across both classes, ensuring neither class can be ignored.
- **F0.5u:** Precision-weighted F-measure ($\beta = 0.5$) that treats abstentions as false negatives, preventing systems from gaming C@1 by over-abstaining.
- **Mean:** Arithmetic mean of all five metrics above, used as the primary ranking score. A high mean requires consistent performance across all failure modes simultaneously.

Systems are ranked according to the **Mean** score, as it requires consistently strong performance across all five metrics and cannot be gamed by optimising for any single one in isolation.

6. Results and Discussion

6.1. Official Results

Table 3: Official results on PAN 2026 and PAN 2025 test datasets. Best result per column in **bold**.

Model	Split	ROC-AUC	Brier	C@1	F1	F0.5u	Mean
Longformer (Finetuned)	2026 Dataset	**0.885**	0.558	0.550	0.574	**0.765**	0.666
	2025 Dataset	**0.966**	0.907	**0.904**	**0.927**	**0.960**	**0.933**
+ GRL ($\lambda = 0.1$)	2026 Dataset	0.786	0.570	0.554	0.568	0.759	0.645
	2025 Dataset	0.924	0.904	0.898	0.923	0.954	0.921
+ GRL ($\lambda = 0.2$)	2026 Dataset	0.788	0.583	0.547	0.572	0.761	0.650
	2025 Dataset	0.925	0.904	0.895	0.921	0.951	0.919
+ GRL ($\lambda = 0.3$)	2026 Dataset	0.751	0.588	0.542	0.566	0.757	0.641
	2025 Dataset	0.922	0.904	0.895	0.921	0.950	0.918
+ GRL ($\lambda = 0.4$)	2026 Dataset	0.763	0.611	0.551	0.578	**0.765**	0.654
	2025 Dataset	0.928	0.906	0.894	0.810	0.949	0.920
+ GRL ($\lambda = 0.5$)	2026 Dataset	0.856	**0.658**	**0.551**	**0.579**	0.764	**0.682**
	2025 Dataset	0.944	**0.911**	0.885	0.914	0.938	0.918

Table 3 reports our results on the official PAN test sets for both 2025 and 2026. On the 2025 test set, the non-adversarial Longformer achieved a mean of 0.933, which would place it 2nd on the 2025 leaderboard — a strong result given that the 2025 test distribution closely matches our training data.

On the 2026 test set, all models show lower scores, which is expected: the 2026 edition explicitly tests robustness by introducing new AI models and unknown obfuscation techniques not seen during training. The non-adversarial baseline drops to a mean of 0.666 on the 2026 test set. The adversarial variants, while slightly weaker on the 2025 data, hold up better on the harder 2026 conditions — with $\lambda = 0.5$ reaching a mean of 0.682, a gain of 0.016 over the baseline. This gap, though modest, is consistent across all λ values and points in the right direction: by discarding model-specific patterns during training, the adversarial objective gives the model a small but meaningful advantage when faced with AI systems it has never seen before.

6.2. Train & Val Results

Table 4 reports the performance of our adversarial model variants and the finetuned Longformer on the held-out validation set.

Table 4: Performance comparison across adversarial weight values on the PAN 2026 training and held-out validation sets. Best result per column in **bold**.

Model	Split	ROC-AUC	Brier	C@1	F1	F0.5u	Mean
Longformer (Finetuned)	Train	0.994	**0.989**	**0.988**	**0.987**	0.989	**0.988**
	Val	0.995	**0.990**	**0.989**	**0.988**	0.988	**0.990**
+ GRL ($\lambda = 0.1$)	Train	0.996	0.986	0.983	0.982	0.984	0.986
	Val	0.997	0.987	0.986	0.985	0.986	0.988
+ GRL ($\lambda = 0.2$)	Train	0.997	0.986	0.982	0.981	0.983	0.986
	Val	**0.998**	0.986	0.983	0.981	0.983	0.986
+ GRL ($\lambda = 0.3$)	Train	0.997	0.985	0.982	0.980	**0.987**	0.985
	Val	0.997	0.985	0.982	0.980	0.982	0.985
+ GRL ($\lambda = 0.4$)	Train	0.997	0.984	0.980	0.978	0.981	0.984
	Val	0.997	0.984	0.981	0.979	0.982	0.984
+ GRL ($\lambda = 0.5$)	Train	0.997	0.978	0.972	0.969	0.973	0.978
	Val	0.997	0.978	0.972	0.969	0.973	0.978

6.3. Effect of Adversarial Weight λ

As λ increases, the mean score decreases — from 0.990 at Longformer(Finetuned) down to 0.978 at $\lambda = 0.5$. However, ROC-AUC improved consistently across all adversarial variants (0.995 Longformer (Finetuned) vs. up to 0.998 at $\lambda = 0.2$), showing that adversarial training produces more separable representations even when the overall mean drops slightly. The train–val gap is negligible across all variants, confirming that none of the models overfit to the training set. Overall, $\lambda = 0.1$ offers the best balance: a small ROC-AUC gain with minimal cost to the other metrics.

7. Error Analysis

Figure 4 shows the error rate per source model and detector. Of the 23 sources tested, 19 produce zero errors under every variant. Errors concentrate on two sources: **gemini-1.5-pro** (up to 7.6 %, all false negatives) and **human** text (up to 6.7 %, all false positives), suggesting the model occasionally struggles to distinguish highly fluent text in either direction. No adversarial variant consistently outperforms the the finetuned Longformer on both sources; Adv_0.5 degrades on both while Adv_0.4 and Adv_0.1 each approach Longformer(finetuned) performance on one of them.

False positives (human text predicted as AI) outnumber false negatives (AI text predicted as human) across all variants, with Adv_0.5 showing the largest imbalance (85 FP vs. 14 FN). As shown in Figure 5, errors tend to occur at low confidence values rather than with high certainty, which is a desirable property: a simple confidence threshold could allow the model to abstain on most erroneous predictions while retaining the majority of correct ones — directly benefiting the C@1 metric.

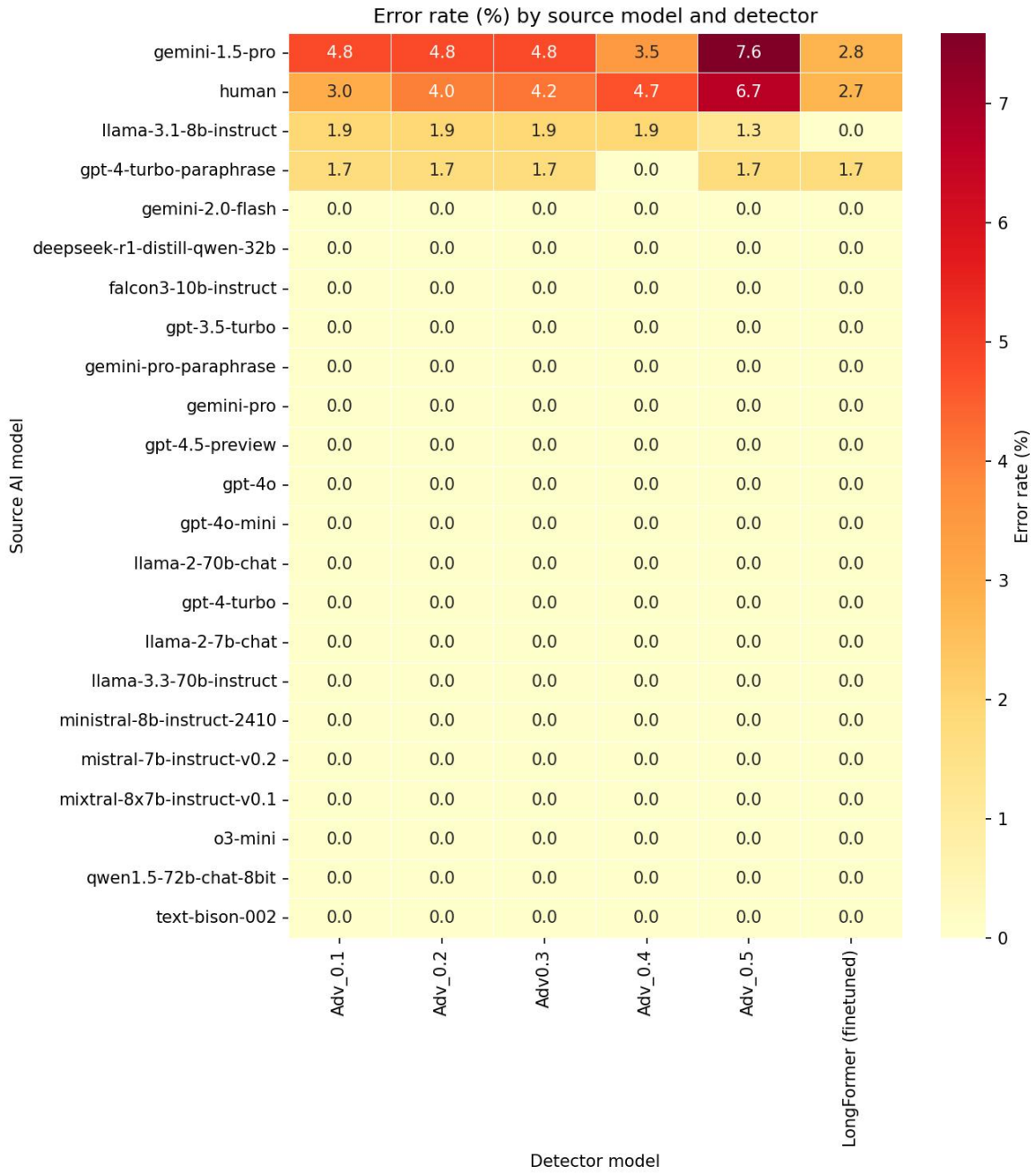

Figure 4: Error rate (%) by source model and detector. Darker cells indicate higher error rates. Errors are confined to four sources; all other sources are classified perfectly.

8. Conclusion and Future Work

We presented an adversarial Longformer-based approach for the PAN 2026 Generative AI Authorship Verification task. The three central contributions are: (1) the use of Longformer [16] to overcome the 512-token limitation of standard classifiers, enabling full-document stylistic analysis; (2) an adversarial training objective via Gradient Reversal Layer [17] that forces model-agnostic AI detection; and (3) calibrated probability output aligned with the task's C@1, Brier, and F0.5u metrics.

Future work will explore more sophisticated targets, genre-aware calibration, and extending sequence length to the full 4,096-token capacity of Longformer for documents longer than 2,048 tokens.

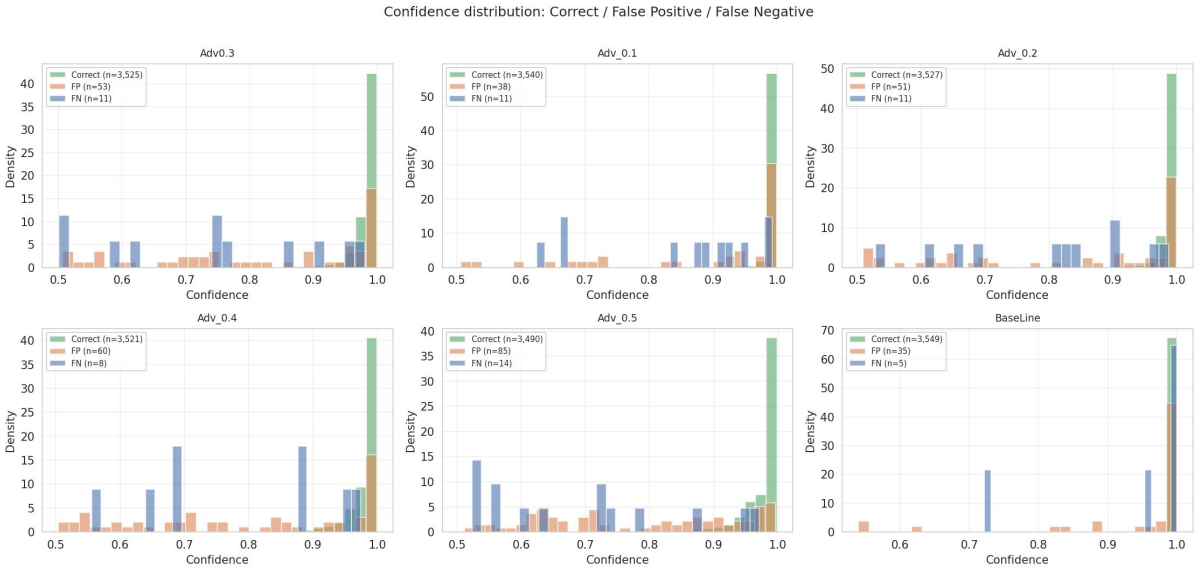

Figure 5: Confidence distributions for correct predictions, false positives, and false negatives across all detector variants. Correct predictions cluster at confidence $= 1.0$; errors spread across the lower range.

Acknowledgments

The authors would like to thank the PAN 2026 organizers for providing the dataset and evaluation infrastructure.

Declaration on Generative AI

The authors have employed Generative AI tools for grammar correction and language enhancement only. No AI tool was used to generate scientific content, results, analysis, or conclusions.

References

[1] T. Brown, B. Mann, N. Ryder, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020) 1877–1901.

[2] J. Achiam, S. Adler, S. Agarwal, et al., GPT-4 technical report (2023). ArXiv:2303.08774.

[3] R. Zellers, A. Holtzman, H. Rashkin, et al., Defending against neural fake news, in: Advances in Neural Information Processing Systems, volume 32, 2019.

[4] J. P. Wahle, T. Ruas, F. Kirstein, B. Gipp, How large language models are transforming machine-paraphrase plagiarism, in: Proceedings of EMNLP 2022, 2022, pp. 952–963.

[5] I. Vykopal, M. Pikuliak, I. Srba, R. Moro, D. Macko, M. Bielikova, Disinformation capabilities of large language models, in: Proceedings of the 62nd Annual Meeting of the ACL, 2024, pp. 14830–14847.

[6] G. Jawahar, M. Abdul-Mageed, L. V. Lakshmanan, Automatic detection of machine generated text: A critical survey, arXiv preprint arXiv:2011.01314 (2020).

[7] I. Solaiman, M. Brundage, J. Clark, et al., Release strategies and the social impacts of language models (2019). ArXiv:1908.09203.

[8] B. Guo, X. Zhang, Z. Wang, et al., How close is ChatGPT to human experts? comparison corpus, evaluation, and detection (2023). ArXiv:2301.07597.

[9] 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), volume 202 of *PMLR*, 2023, pp. 24950–24962.

[10] J. Su, T. Zhuo, D. Wang, P. Nakov, Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 12395–12412.
[11] 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), volume 235 of *PMLR*, 2024.
[12] A. Bakhtin, S. Gross, M. Ott, et al., Real or fake? learning to discriminate machine from human generated text (2019). ArXiv:1906.03351.
[13] A. Uchendu, T. Le, K. Shu, D. Lee, Authorship attribution for neural text generation, in: Proceedings of EMNLP 2020, 2020, pp. 8384–8395.
[14] J. Bevendorff, M. Wiegmann, J. Karlgren, et al., Overview of the Voight-Kampff generative AI authorship verification task at PAN and ELOQUENT 2024, in: Working Notes of CLEF 2024, CEUR-WS.org, 2024.
[15] 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.
[16] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer (????).
[17] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: Proceedings of the 32nd International Conference on Machine Learning (ICML), 2015, pp. 1180–1189.
[18] E. Tavan, M. Najafi, MarSan at PAN: BinocularsLLM, fusing Binoculars' insight with the proficiency of large language models for machine-generated text detection, in: Working Notes of CLEF 2024, CEUR-WS.org, 2024.
[19] D. Macko, mdok of KInIT: Robustly fine-tuned LLM for binary and multiclass AI-generated text detection, in: Working Notes of CLEF 2025, CEUR-WS.org, 2025.
[20] Qwen Team, Qwen3 Technical Report, Technical Report, 2025. ArXiv:2505.09388.
[21] L. Lorenz, F. Z. Aygüler, F. Schlatt, N. Mirzakhmedova, BaselineAvengers at PAN 2024: Often-forgotten baselines for LLM-generated text detection, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[22] A. Valdez-Valenzuela, H. Gómez-Adorno, M. Montes-y Gómez, AI-generated text detection using ISGraphs and graph neural networks, in: Workshop on New Perspectives in Advancing Graph Machine Learning, NeurIPS 2025, 2025.
[23] J. Liu, L. Kong, Z. Peng, F. Chen, Generative AI authorship verification based on contrastive-enhanced dual-model decision system, in: Working Notes of CLEF 2025, CEUR-WS.org, 2025.
[24] J. Bevendorff, M. Wiegmann, M. Potthast, B. Stein, PAN'25/26 Generative AI Detection: Voight-Kampff AI Detection Sensitivity, 2025. URL: https://zenodo.org/records/14962653. doi:10.5281/zenodo.14962653.