Team LatentAuthors at PAN 2026: DeBERTa with Genre-Aware Fine-Tuning for AI Text Detection

Notebook for the PAN Lab at CLEF 2026

Kholoud Khalil Aldous[1,*], Wajdi Zaghouani[1]

[1]*Northwestern University, Doha, Qatar*

Abstract

This paper presents our approach to the Voight-Kampff Generative AI Detection task at PAN 2026, which requires classifying texts as human-authored or machine-generated, including instances involving previously unseen language models or obfuscation strategies. We propose a detection framework that fine-tunes the `DeBERTa-v3-base` pre-trained language model, augmented with a genre-aware conditioning strategy that prepends a short genre identifier (essays, news, or fiction) to each input, allowing the model to incorporate document type as additional context. Rather than producing hard binary labels, the system outputs calibrated confidence scores, enabling abstention-sensitive metrics to credit the model appropriately under uncertainty. Model selection retains the checkpoint that maximizes the mean of the five official evaluation metrics across training epochs, rather than relying on a single metric or fixed training duration. On the provided validation set, our system substantially outperforms the strongest available baseline, confirming that genre-aware fine-tuning is effective when training and evaluation data are drawn from similar distributions. This advantage, however, does not transfer to the official test set, where performance falls well short of validation-set results. Error analysis reveals a pronounced bias toward predicting human authorship, indicating that the system struggles to generalize to the unseen generators and obfuscation strategies introduced at test time. Our system, LatentAuthors, ranked 31st on the official leaderboard.

Keywords

AI-generated text detection, authorship verification, DeBERTa, transformer fine-tuning, genre-aware classification, PAN 2026

1. Introduction

The proliferation of large language models (LLMs) capable of generating fluent, contextually appropriate text has renewed interest in automatic methods for distinguishing machine-authored content from human-authored writing. Such methods are relevant to academic integrity, journalism, and trust in online information ecosystems. The **Voight-Kampff Generative AI Detection task at PAN** [1] poses this as a binary AI detection task: given a text, a system must decide whether it was machine-authored (class 1) or human-authored (class 0). A general overview of the PAN 2026 lab is given in [2], and the complete results of this year's task are compiled in the official task overview [3]

The 2026 edition introduces additional robustness challenges: LLMs are instructed to mimic specific human authors, and the test set includes texts from novel models and previously unseen obfuscation strategies. These conditions push beyond in-distribution generalization and require systems that respond appropriately to distribution shift.

Our submission, team **LatentAuthors**, takes a supervised fine-tuning approach built on DeBERTa-v3-base [4], a pre-trained encoder model that improves upon standard BERT-style architectures through disentangled attention and a replaced-token-detection pre-training objective. We augment the standard classification objective with lightweight genre conditioning—prepending a genre-specific prefix string to each input—so the model can adapt its feature weighting to the document type. The resulting system produces probability scores that are used directly as submission labels without post-hoc calibration.

CLEF 2026 Working Notes, September 2026, Jena, Germany

*Corresponding author.

kholoud.aldous@northwestern.edu (K. K. Aldous); wajdi.zaghouani@northwestern.edu (W. Zaghouani)

0000-0003-1188-5724 (K. K. Aldous); 0000-0003-1521-5568 (W. Zaghouani)

Table 1
Dataset split statistics.

Split	Human	Machine	Total
Training	9,101	14,606	23,707
Validation	1,277	2,312	3,589

2. Task and Data

2.1. Task Description

The Voight-Kampff Generative AI Detection task at PAN 2026 [1] is a binary classification problem. The dataset is divided into three splits: training, validation, and test. Both the training and validation sets provide a unique identifier, a text passage, a binary label (0 = human, 1 = machine), and a genre field (`essays`, `news`, or `fiction`). The test set provides only the identifier and the text.

Table 1 summarizes the dataset statistics. The test set labels are not yet publicly available and will be released following the official evaluation period.

Systems must output a confidence score $s \in [0, 1]$ for each test instance, where $s < 0.5$ indicates predicted human authorship and $s > 0.5$ indicates predicted machine authorship. A score of exactly 0.5 is interpreted as an abstention and is handled specially by the C@1 and $F_{0.5u}$ metrics to reward systems that acknowledge uncertainty rather than guess.

A key feature of the 2026 edition is that all machine-generated texts were produced with author-mimicry prompts: LLMs were explicitly instructed to write in the style of specific human authors. In addition, the test set may include texts from unseen models or previously unknown obfuscation strategies, directly testing out-of-distribution robustness.

2.2. Evaluation Metrics

Systems are evaluated on five complementary measures, and results are summarized by their arithmetic mean:

- **F1**: Harmonic mean of precision and recall, thresholded at 0.5.
- **ROC-AUC**: Area under the receiver-operating-characteristic curve; measures ranking quality.
- **Brier (complement)**: $1 - \text{BS}$, where BS is the mean squared calibration error; rewards probability calibration.
- **C@1**: Modified accuracy that credits abstentions (score $= 0.5$) with the per-answered accuracy, discouraging overconfident wrong predictions.
- **$F_{0.5u}$**: Precision-weighted F-measure that treats abstentions as false negatives, penalizing systems that abstain excessively.

The five-metric mean is the primary ranking criterion on the Tira leaderboard.

3. System Description

3.1. Model Architecture

Our system is built on `DeBERTa-v3-base`, available through the Hugging Face transformers library [4]. DeBERTa-v3 combines two improvements over the original DeBERTa architecture: (1) *disentangled attention*, which encodes content and relative position using separate embedding vectors, and (2) an ELECTRA-style replaced-token-detection (RTD) pre-training objective with gradient-disentangled embedding sharing, which is more sample-efficient than masked-language modeling and results in representations with stronger discriminative power for fine-grained classification.

Table 2
Training hyperparameters.

Hyperparameter	Value
Base model	`DeBERTa-v3-base`
Maximum sequence length	512
Batch size	16
Gradient accumulation steps	2
Peak learning rate	2×10^{-5}
Optimizer	AdamW
Weight decay	0.01
LR schedule	Linear with warm-up
Warm-up ratio	0.10
Training epochs	3
Gradient clipping (ℓ_2)	1.0
Random seed	42

We add a two-class linear head on top of the `[CLS]` representation and fine-tune the entire model end-to-end. The classification layer maps the 768-dimensional hidden state to logits for the `human` and `machine` classes; during inference, a softmax is applied and the `machine`-class probability is output as the confidence score.

3.2. Genre-Aware Prefix

A key design decision is the insertion of a short genre-identifying prefix to each input sequence before tokenization. Specifically, texts labeled `essays`, `news`, and `fiction` are inserted with "`Essay text:`", "`News text:`", and "`Fiction text:`" respectively, while any instance with an unrecognized or missing genre field falls back to the plain prefix "`Text:`". This prefix occupies only a few tokens but provides the model with an explicit, reliable signal about the domain, freeing the attention mechanism from having to infer genre from content alone. Since the test set omits the genre field, the default "`Text:`" prefix ensures the system can still process any such instance without modification.

3.3. Tokenization and Input Preparation

All inputs are tokenized using the DeBERTa-v3-base SentencePiece tokenizer with a maximum sequence length of 512 tokens. Sequences shorter than 512 tokens are padded to the maximum length; longer sequences are truncated. The effective vocabulary size is approximately 128,000 tokens, enabling the model to handle diverse writing styles without excessive out-of-vocabulary fragmentation.

3.4. Training Procedure

Training hyperparameters are summarized in Table 2. These values were not tuned via a dedicated search: learning rate (2×10^{-5}), batch size (16), weight decay (0.01), and maximum sequence length (512) follow the configuration reported by a high-performing PAN 2025 system fine-tuning the same base model on an analogous classification task [5], which placed second on the official leaderboard. The remaining hyperparameters—number of training epochs and warm-up ratio—were fixed following standard practice for transformer fine-tuning rather than drawn from that system. We use the AdamW optimizer with a linear learning-rate schedule, with warm-up covering the first 10% of training steps. Gradient accumulation over 2 mini-batches yields the reported batch size of 16, and gradients are clipped to a maximum ℓ_2-norm of 1.0 to prevent instability during the early warm-up phase. After each epoch, the model is evaluated on the validation set and all five official metrics are computed. The checkpoint achieving the highest arithmetic mean score is saved as the final model.

Table 3
Validation set results. Baseline scores are from the official task description [1]. Best score per column is in **bold**.

System	F1	ROC-AUC	Brier	C@1	$F_{0.5u}$	Mean
PPMd Compression [7, 8]	0.812	0.786	0.799	0.757	0.778	0.786
Binoculars [9]	0.872	0.918	0.867	0.844	0.882	0.877
TF-IDF SVM (baseline)	0.980	0.996	0.951	0.984	0.981	0.978
LatentAuthors (ours)	**0.996**	**1.000**	**0.996**	**0.994**	**0.996**	**0.996**

3.5. Deployment

The system is packaged as a Docker image for submission to the Tira platform [6], which evaluates runs in a sandboxed, network-isolated environment.

4. Experiments

4.1. Experimental Setup

All experiments use the official PAN 2026 training and validation splits. Training is conducted on a single GPU. We report results on the validation set using the five official metrics and their arithmetic mean. Baseline scores are taken from the official task description.

4.2. Baseline Comparison

The PAN 2026 organizers provide three baselines:

- **TF-IDF SVM**: A linear support vector machine trained on character and word n-gram TF-IDF features.
- **PPMd Compression**: A compression-based cosine-similarity method [7, 8].
- **Binoculars** [9]: A zero-shot detector that contrasts perplexity scores from two pre-trained LLMs.

Validation scores for all baselines and our system are shown in Table 3.

4.3. Training Dynamics

The model is evaluated after each of the three training epochs and the checkpoint with the highest arithmetic mean score is saved. The best checkpoint is selected at epoch 3, consistent with the model continuing to improve across all epochs without overfitting on the provided training split.

4.4. Genre-Level Analysis

Table 4 breaks down F1 and ROC-AUC by genre on the validation set, isolating which document type benefits most from genre conditioning. Performance is uniformly high across all three genres (F1 $\geq$ 0.995, ROC-AUC $\geq$ 0.999), with only marginal variation between them: essays achieve the highest scores (F1 = 0.997), while fiction—the largest and most heterogeneous genre by sample count—scores marginally lower (F1 = 0.995, ROC-AUC = 0.999). The narrow spread indicates that genre-aware conditioning generalizes consistently across document types rather than disproportionately benefiting any single genre.

Table 4
Per-genre (essays, news, fiction) F1 and ROC-AUC scores of our LatentAuthors system on the validation set.

Genre	F1	ROC-AUC	n
Essays	0.997	1.000	665
News	0.995	1.000	1087
Fiction	0.995	0.999	1837
Overall	0.996	1.000	3589

Table 5
Official PAN 2026 test-set results for LatentAuthors.

ROC-AUC	Brier	C@1	F1	$F_{0.5u}$	Mean	FPR	FNR
0.750	0.486	0.473	0.460	0.678	0.569	0.008	0.701

4.5. Official Test Results

Table 5 reports our system's official performance on the PAN 2026 test set, following the conclusion of the evaluation period. LatentAuthors ranked 31st on the official leaderboard.

Despite the near-ceiling performance observed on the validation set (Section 4.4), test-set performance is substantially lower across every metric. The most informative signal is the asymmetry between the false-positive rate (0.008) and the false-negative rate (0.701): the system rarely misclassifies human-written text as machine-generated, but fails to detect the large majority of machine-generated texts in the test set. This pattern is consistent with overfitting to the generators and obfuscation strategies present in the training and validation data, and indicates poor generalization to the previously unseen language models and obfuscation strategies introduced at test time.

5. Conclusion

We presented the LatentAuthors system for the PAN 2026 Voight-Kampff Generative AI Detection task, fine-tuning DeBERTa-v3-base with a genre-aware prefix strategy and confidence-score output. On the validation set, the system achieves a mean score of 0.996 across the five official metrics, outperforming the strongest baseline (TF-IDF SVM, 0.978) with consistently strong performance across all three genres. This performance, however, did not generalize to the official test set, where LatentAuthors achieved a mean score of 0.569 and ranked 31st. The gap is driven by a sharp asymmetry between false-positive and false-negative rates: the system rarely misclassifies human-written text but fails to detect most machine-generated texts under the unseen generators and obfuscation strategies present at test time. This suggests the near-ceiling validation scores reflect overfitting to the training-time generator distribution rather than genuine robustness, and that genre-aware conditioning alone is insufficient under distribution shift. Future work should prioritize exposure to a broader range of generators during fine-tuning and explicit robustness checks against held-out model families, rather than relying on in-distribution validation performance as a proxy for test-time robustness.

Acknowledgments

This participation was supported by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), part of the Qatar Research, Development and Innovation Council (QRDI).

Declaration on Generative AI

During the preparation of this work, the authors used Claude Sonnet 4.6 in order to: assist in drafting and structuring the paper, editing for clarity, and formatting the LaTeX source. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Bevendorff, Y. Wang, J. Karlgren, M. Wiegmann, M. Fröbe, A. Tsivgun, J. Su, Z. Xie, M. Abassy, J. Mansurov, R. Xing, M. Ta, K. Elozeiri, T. Gu, R. 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, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 3504–3534. URL: https://ceur-ws.org/Vol-4038/paper_277.pdf.

[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, 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] 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.

[4] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).

[5] B. Li, H. Qi, K. Yan, Team bohan li at pan: Deberta-v3 with r-drop regularization for human-ai collaborative text classification, Working Notes of CLEF (2025).

[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: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer-Verlag, Berlin, Heidelberg, 2023, pp. 236–241. URL: https://doi-org.qulib.idm.oclc.org/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

[7] D. Sculley, C. E. Brodley, Compression and machine learning: A new perspective on feature space vectors, in: Data Compression Conference (DCC'06), IEEE, 2006, pp. 332–341.

[8] O. Halvani, C. Winter, L. Graner, On the usefulness of compression models for authorship verification, in: Proceedings of the 12th International Conference on Availability, Reliability and Security, ARES '17, Association for Computing Machinery, New York, NY, USA, 2017. URL: https://doi-org.qulib.idm.oclc.org/10.1145/3098954.3104050. doi:10.1145/3098954.3104050.

[9] 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'24, JMLR.org, 2024.