eGrantha.ai at PAN 2026: Contrastive Representation Learning for Robust Generative AI Authorship Detection

Notebook for the PAN Lab at CLEF 2026

Anish Thapaliya[1], Shushanta Pudasaini[2] and Bal Krishna Bal[3]

[1]*Institute of Science and Technology, Tribhuvan University, Kathmandu, Nepal*
[2]*School of Computer Science, Technological University Dublin, Dublin, Ireland*
[3]*Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Nepal*

Abstract

Distinguishing human-written text from that generated by Large Language Models (LLMs), referred to as AI-generated text, has become a fundamentally harder problem with each passing year, with real consequences for academic integrity, the spread of misinformation, and the long-term reliability of web-scale training corpora. Detecting such text reliably — especially when adversarial obfuscation is involved — remains an open and practically urgent challenge. The PAN CLEF 2026 Voight-Kampff shared task directly addresses this by benchmarking detection systems against texts spanning multiple domains, including LLM outputs deliberately crafted to evade automated detectors.

Rather than framing detection as a standard binary classification problem, this paper presents **Embed-KNN**, a system that fine-tunes RoBERTa-base with a supervised contrastive objective to produce geometrically separable embeddings of human and AI-generated text. At inference time, the system classifies texts via K-Nearest Neighbors (KNN) retrieval over a FAISS index, treating detection as a *retrieval problem* over a learned embedding space rather than a direct classification task.

On the held-out validation set, Embed-KNN achieves a mean score of 0.9929 at $K{=}11$, outperforming all three official baselines across four of five metrics and yielding only 24 misclassifications out of 3,589 instances. An ablation over $K \in \{1, \ldots, 11\}$ shows that while $K{=}2$ peaks on mean score, ROC-AUC stabilizes from $K{=}8$ onward; $K{=}11$ is selected to favor robustness over validation set overfitting. These results indicate that contrastively learned embedding spaces offer a competitive and robust foundation for AI-generated text detection.

Keywords

AI-generated text detection, Contrastive Learning, KNN, PAN CLEF 2026

1. Introduction

AI-generated text refers to textual content generated by Large Language Models (LLMs) such as GPT-4 [1] and LLaMA [2]. AI-generated text has become pervasive across news, academia, and social media, raising serious concerns about misinformation, academic dishonesty, and the long-term contamination of training corpora through model collapse. Detecting whether a given text was written by a human or generated by an LLM has therefore become a critical research challenge and grows harder each year as generators improve and adversarial obfuscation strategies are deliberately employed to defeat detectors [3, 4].

A promising alternative reframes detection as a *representation learning* problem rather than a binary classification problem. DeTeCtive [5] demonstrated this convincingly at NeurIPS 2024: by training with a multi-level contrastive loss that explicitly separates writing styles in embedding space—and classifying at inference time via KNN over that space—the framework achieves state-of-the-art generalization on M4 [6], TuringBench [7], and Deepfake [8], outperforming prior methods on unseen-domain evaluation.

This study adapts the DeTeCtive contrastive framework to the PAN CLEF 2026 Voight-Kampff Generative AI Detection benchmark [4, 9], which presents a uniquely challenging setting where texts

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
anish.805522@sms.tu.edu.np (A. Thapaliya); D23129142@mytudublin.ie (S. Pudasaini); bal@ku.edu.np (B. K. Bal)
0000-0003-1612-4510 (S. Pudasaini)

are potentially obfuscated, span multiple domains, and include LLM outputs specifically crafted to mimic human writing styles. Specifically, it evaluates whether the geometric separation learned by contrastive training followed by KNN inference, referred to as **Embed-KNN**, transfers robustly to this adversarial setting.
The key contributions are:

- Demonstration that nearest-neighbor retrieval over a contrastively fine-tuned embedding space is a competitive approach for human vs. AI-generated text detection.
- A systematic ablation study across $K \in \{1, \ldots, 11\}$, revealing that while $K{=}2$ maximizes the PAN mean score on the validation set, ROC-AUC increases consistently with K and stabilizes from $K{=}8$ onward.

2. Related Work

AI-generated text detection methods can be broadly categorized into two paradigms: **statistical zero-shot approaches** and **supervised learning approaches**. Statistical approaches exploit the underlying textual and distributional properties of language model outputs to identify AI-generated patterns without requiring labeled training data. Gehrmann et al. [10] introduced GLTR, a forensic tool that applies statistical tests over token rank, entropy, and top-k probability distributions to assist humans in identifying machine-generated text. Building on this foundation, Mitchell et al. [11] proposed DetectGPT, which identifies AI-generated text by measuring the curvature of the log-probability function under perturbations, exploiting the tendency of LLMs to generate text near local maxima of the model's probability distribution. Hans et al. [12] further advanced zero-shot detection through Binoculars, which contrasts perplexity scores across two closely related language models to produce a more discriminative detection signal, achieving strong performance without any task-specific training.

Supervised learning approaches, by contrast, fine-tune a pre-trained transformer model on labeled corpora of human and AI-generated text. Chen et al. [13] introduced GPT-Sentinel, which fine-tunes RoBERTa and T5 on a curated dataset to distinguish human-written from GPT-generated content, demonstrating that transformer-based classifiers can achieve robust detection performance. Hu et al. [14] proposed RADAR, which couples a paraphrasing model with an adversarially trained detector, improving robustness against paraphrase-based obfuscation attacks that commonly fool standard classifiers.

More recent supervised approaches reframe detection as a *representation learning* problem, moving away from direct classification in favor of learning a discriminative embedding space. Guo et al. [5] introduced DeTeCtive, which employs multi-level contrastive learning to construct a discriminative embedding space, framing detection as a nearest-neighbor retrieval problem, and achieving strong generalization across unseen models and domains. He et al. [15] proposed Detree, which models the relationships among different human-AI text generation processes as a Hierarchical Affinity Tree structure, enabling fine-grained detection of collaborative texts beyond simple binary classification.

This study applies the retrieval-based detection paradigm, evaluating its effectiveness against the official baselines on the Voight-Kampff Generative AI Detection shared task at PAN CLEF 2026 [4, 9].

3. Task and Datasets

3.1. Task

Similar to the previous year's PAN CLEF 2025 Generative AI Authorship Verification task [16], the PAN CLEF 2026 Voight-Kampff Generative AI Detection [4, 9] requires participants to assign a continuous score $s \in [0, 1]$ to each test document, where $s = 1$ indicates the text is AI-generated and $s = 0$ indicates that it is human-written. The performance of the system is evaluated based on the arithmetic mean of the five metrics mentioned below:

- **ROC-AUC Score**
- **Brier Score Complement**: Evaluates the accuracy of probabilistic predictions using the complement of the mean squared error.
- **C@1**: A modified accuracy metric that assigns a score of 0.5 to undecided predictions while averaging accuracy over the remaining cases.
- **F1-score**
- **F0.5u**: A variation of the $F_{0.5}$ metric that treats undecided predictions ($score = 0.5$) as false negatives.

3.2. Datasets

The competition data were provided via Zenodo [1] which reused the same training and validation splits from the previous year, i.e., PAN CLEF 2025 [16], and were evaluated on different test data consisting of new models and obfuscation methods [9]. Table 1 summarizes the dataset statistics.

Table 1
Class distribution of the training and validation splits across human-written and AI-generated documents.

Split	Total	Human	AI	Domains
Train	23,707	9,101	14,606	3
Validation	3,589	1,277	2,312	3

The training corpus comprises **23,707** documents spanning three domains (essays, fiction, news). The majority class is AI-generated text (14,606 samples, 61.6%), with human-written samples comprising the remaining 9,101 (38.4%), resulting in a moderately imbalanced distribution. The validation set contains 3,589 documents with a 35.58%/64.42% human-to-AI split. The held-out test split was not publicly released and is therefore excluded from Table 1. Test-set performance was evaluated directly via the TIRA platform [17].

4. System Overview

This section describes the system named **Embed-KNN** in detail. The overall architecture follows three sequential stages: (1) contrastive fine-tuning of a pre-trained encoder, (2) embedding extraction and FAISS indexing, and (3) KNN-based inference. The full training and inference pipeline is illustrated in Figure 1.

4.1. Contrastive Fine-Tuning of Text Encoder

An encoder-only transformer, specifically `RoBERTa-base` [18], was adopted as the backbone. RoBERTa-base was selected for its strong sentence-level representation capabilities and its established effectiveness in prior AI-generated text detection work [5]. The `[CLS]` token representation from the final hidden layer was used as the document embedding $\mathbf{z} \in \mathbb{R}^{768}$, which was subsequently ℓ_2-normalized before being passed to the contrastive loss. All documents exceeding the maximum context length of 512 tokens were truncated to the first 512 tokens.

4.1.1. Contrastive Training Objective

Following the supervised contrastive learning framework of Khosla et al. [19] and DeTeCtive [5], the encoder was fine-tuned with a supervised contrastive loss that explicitly pulls embeddings of the same

[1] https://zenodo.org/records/14962653

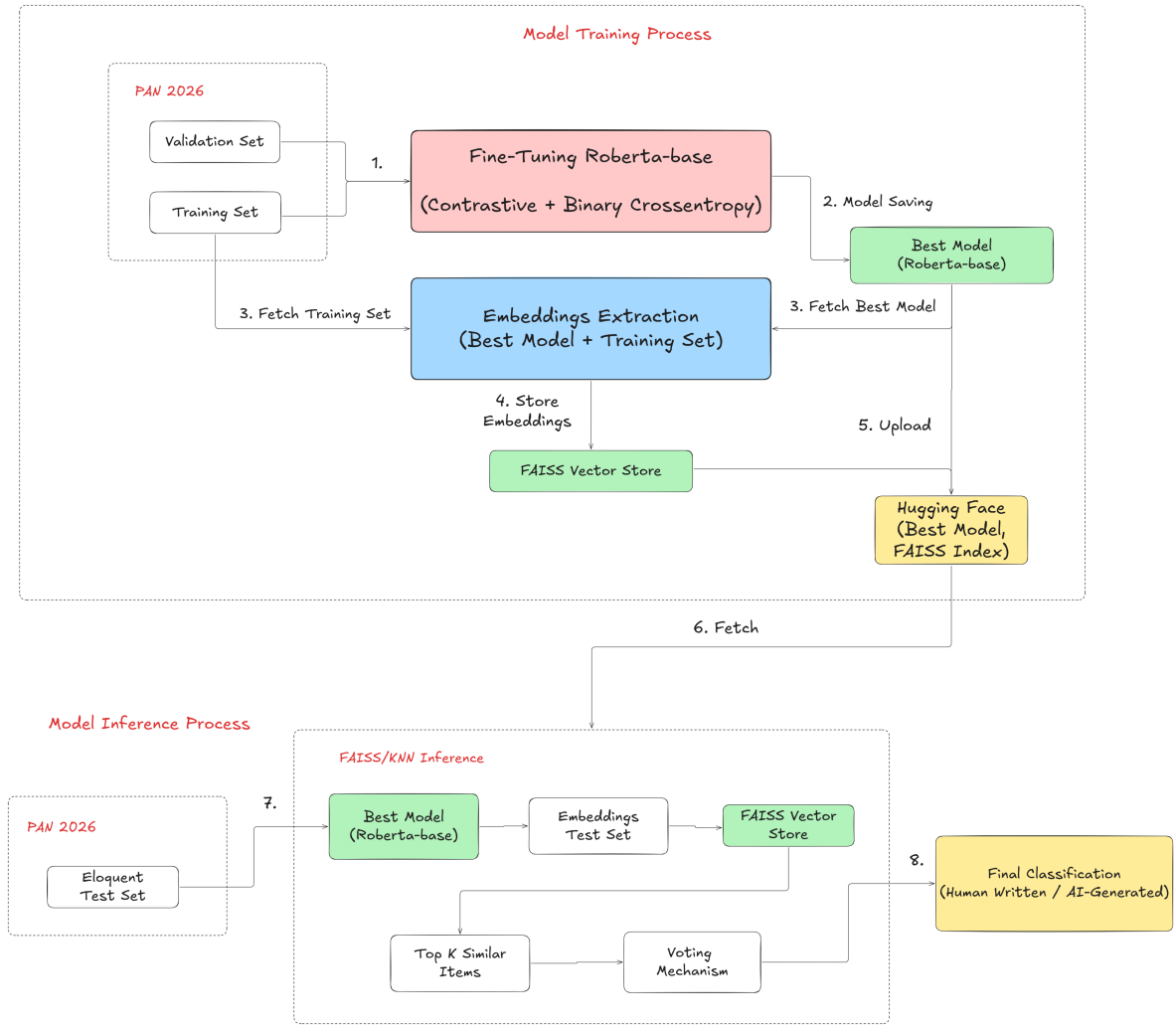

Figure 1: Overview of the training and inference pipeline. During training, RoBERTa-base was fine-tuned with a supervised contrastive loss. At inference time, the trained encoder produced document embeddings that were queried against a FAISS index using KNN classification.

class (human or AI-generated) together in the representation space while pushing embeddings of different classes apart.

Formally, $\mathbf{z}_i$ denotes the ℓ_2-normalized embedding of sample i. The set of positive samples for anchor i within the batch is defined as $\mathcal{P}(i) = \{p \mid y_p = y_i,\ p \neq i\}$, and the set of all other samples is $\mathcal{A}(i) = \{a \mid a \neq i\}$. The supervised contrastive loss [19] is:

$$\mathcal{L}_{\text{sCL}} = -\sum_{i=1}^{N} \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p \,/\, \tau)}{\sum_{a \in \mathcal{A}(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a \,/\, \tau)}, \tag{1}$$

where τ is a temperature. The dot products above are equivalent to cosine similarity given the ℓ_2 normalization applied to all embeddings.

The total training objective combines the contrastive loss with an auxiliary binary cross-entropy loss $\mathcal{L}_{\text{CE}}$ appended via a two-layer classification head:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sCL}} + \lambda\, \mathcal{L}_{\text{CE}}, \tag{2}$$

where $\lambda = 1.0$ weights the cross-entropy term equally with the contrastive term. The cross-entropy head was used only during training to provide an auxiliary classification signal; it was discarded at inference time in favor of KNN classification over the learned embedding space.

Table 2
Training hyperparameters and hardware configuration for Embed-KNN.

Hyperparameter	Value
Max sequence length	512 tokens
Optimizer	AdamW
Learning rate	2×10^{-5}
LR schedule	Linear warmup (6%) + decay
Weight decay	0.01
Batch size	32
Training epochs	3
Temperature τ	0.07
Loss weight λ	1.0
KNN neighbours k	1 to 11
Hardware	NVIDIA RTX 3060
Random seed	0

4.2. Embedding Extraction and FAISS Indexing

Upon completion of training, the encoder was used in inference mode to generate ℓ_2-normalized embeddings for all documents in the training corpus. These embeddings were stored in a FAISS flat inner-product index, which supports exact nearest-neighbor search via cosine similarity (given ℓ_2-normalized vectors, inner product is equivalent to cosine similarity). The index, including the best model, was persisted to disk and uploaded to Hugging Face Spaces.

4.3. KNN-based Inference

At inference time, the saved FAISS index and best model, i.e., the fine-tuned RoBERTa-base, were downloaded from Hugging Face Spaces, and a query document $\mathbf{x}_q$ from the validation/test set was encoded by RoBERTa-base to produce the normalized embedding $\mathbf{z}_q$. The k nearest neighbours of $\mathbf{z}_q$ were retrieved from the FAISS index by cosine similarity. The final prediction score $s \in [0, 1]$ was computed via a **majority voting** mechanism using the k retrieved neighbours.

4.4. Training Configuration

Table 2 summarizes the full training configuration. The AdamW optimizer was used with a peak learning rate of 2×10^{-5}, linear warmup over 6% of total training steps, and weight decay of 0.01 to regularize the encoder weights. Training proceeded for three epochs with a per-device batch size of 32 on a single NVIDIA RTX 3060. A fixed random seed of 42 was used throughout for reproducibility.

5. Experiments and Results

5.1. Effect of K on Detection Performance

Table 3 reports performance across $K \in \{1, \ldots, 11\}$. While K=2 achieves the highest PAN mean score on the validation set, the ROC-AUC increases consistently with K, peaking at 0.9955 for $K \geq 8$. Furthermore, the PAN CLEF 2026 test set was explicitly designed to include novel models and obfuscation strategies unseen during training [4, 9], making robustness to distribution shift a primary concern. Small values of K are inherently more sensitive to local embedding perturbations and neighborhood noise, whereas a larger K produces smoother, more stable confidence estimates and decision boundaries. Therefore, K=11 was selected for the final system, where all five PAN metrics stabilized (Table 3), and ROC-AUC was maximized, prioritizing generalization over validation-set overfitting.

Table 3
Embed-KNN validation performance across $K \in \{1, \ldots, 11\}$ on the PAN CLEF 2025 validation set [16]. **Bold row** indicates the highest PAN mean score (K=2); ROC-AUC peaks and stabilizes from K=8 onward.

K	AUC	Brier	C@1	F1	F0.5u	Mean	FPR	FNR	Acc
1	0.9935	0.9941	0.9941	0.9918	0.9920	0.9931	0.0043	0.0086	0.9941
2	**0.9943**	**0.9941**	**0.9947**	**0.9925**	**0.9910**	**0.9933**	**0.0048**	**0.0070**	**0.9944**
3	0.9943	0.9941	0.9939	0.9914	0.9914	0.9930	0.0048	0.0086	0.9939
4	0.9947	0.9941	0.9939	0.9914	0.9911	0.9930	0.0048	0.0086	0.9939
5	0.9951	0.9941	0.9933	0.9906	0.9911	0.9928	0.0048	0.0102	0.9933
6	0.9951	0.9940	0.9933	0.9906	0.9911	0.9928	0.0048	0.0102	0.9933
7	0.9953	0.9941	0.9933	0.9906	0.9911	0.9929	0.0048	0.0102	0.9933
8	0.9955	0.9941	0.9933	0.9906	0.9911	0.9929	0.0048	0.0102	0.9933
9	0.9955	0.9941	0.9933	0.9906	0.9911	0.9929	0.0048	0.0102	0.9933
10	0.9955	0.9941	0.9933	0.9906	0.9911	0.9929	0.0048	0.0102	0.9933
11	0.9955	0.9941	0.9933	0.9906	0.9911	0.9929	0.0048	0.0102	0.9933

5.2. Comparison with Baselines

Table 4 presents the performance of Embed-KNN against the three official validation-set baselines: TF-IDF SVM, Binoculars, and PPMd Compression-based Cosine (PPMd CBC) [16]. Against the strongest baseline, TF-IDF SVM (mean $= 0.9780$), Embed-KNN achieves the highest mean score of 0.9929, outperforming all three baselines in four of the five metrics, and marginally trailing TF-IDF SVM only in ROC-AUC. Against Binoculars (mean $= 0.8770$), Embed-KNN achieves an absolute improvement of 0.1159 in mean score, and against PPMd CBC (mean $= 0.7860$), the margin widens to 0.2069. These results suggest that the contrastively learned embedding space captures discriminative features that generalize well, outperforming both perplexity-based and compression-based approaches by a considerable margin.

Table 4
Comparison of Embed-KNN against official baselines on the validation set [16]. Best value per column is **bold**

System	Type	AUC	Brier	C@1	F1	F0.5u	Mean
TF-IDF SVM	Supervised	**0.9960**	0.9510	0.9840	0.9800	0.9810	0.9780
Binoculars	Zero-shot	0.9180	0.8670	0.8440	0.8720	0.8820	0.8770
PPMd CBC	Zero-shot	0.7860	0.7990	0.7570	0.8120	0.7780	0.7860
Embed-KNN (K=11)	Supervised	0.9955	**0.9941**	**0.9933**	**0.9906**	**0.9911**	**0.9929**

5.3. Error Analysis using Confusion Matrix

Table 5 presents the confusion matrix for Embed-KNN at K=11. Of the $3{,}589$ validation instances, only 24 were misclassified in total — 11 false positives (FP) and 13 false negatives (FN) — yielding an overall accuracy of 0.9933. The asymmetry between FP and FN is noteworthy: the false negative rate exceeds the false positive rate, indicating that the system is marginally more prone to misclassifying AI-generated texts as human-written than to wrongly flagging human-written texts as AI-generated. The low FPR further suggests that the contrastively learned embedding space effectively separates genuine human-authored texts from AI-generated ones, with very few human texts falling within the neighbourhood of AI embeddings.

Table 5
Confusion matrix for Embed-kNN (K=11) on the PAN CLEF 2025 validation set ($n = 3{,}589$). Rows represent true labels; columns represent predicted labels.

	Predicted: Human	**Predicted: AI**
True: Human	2301 (TN)	11 (FP)
True: AI	13 (FN)	1264 (TP)

6. Conclusion

This research presented Embed-KNN, a contrastive learning-based nearest-neighbour retrieval system for AI-generated text detection applied to the PAN CLEF 2026 Voight-Kampff shared task, framing detection as a retrieval problem over a contrastively learned embedding space.

Evaluated on the validation set, Embed-KNN achieved a PAN mean score of 0.9929 at the submitted configuration of K=11, outperforming all three official baselines across four of the five individual PAN metrics, with particularly strong confidence calibration reflected in a Brier score of 0.9941 compared to 0.9510 for the strongest baseline. Error analysis revealed only 24 misclassifications out of $3{,}589$ instances, with a marginally higher false negative rate than false positive rate. A post-hoc ablation across $K \in \{1, \ldots, 11\}$ further revealed that while K=2 achieved the highest PAN mean score, ROC-AUC peaked and stabilized from K=8 onward, with K=11 selected to prioritize robustness and confidence calibration over validation-set overfitting.

Reframing detection as a retrieval problem over a contrastively learned embedding space proved to be a practically effective and computationally accessible approach, achievable on a single consumer GPU. The observation that $K = 2$ maximized the PAN mean score while ROC-AUC stabilized only from $K = 8$ onward highlights that different metrics can favour different configurations, and that no single metric tells the full story when selecting hyperparameters. Looking ahead, evaluating robustness against unseen LLM families and more sophisticated obfuscation strategies remains the most promising directions for future work.

Acknowledgments

The author thanks the PAN CLEF 2025/2026 organizers for their efforts in constructing the dataset and organizing the Voight-Kampff Generative AI Detection 2026 shared task.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) for paraphrasing and to assist with grammar and spelling checks of the manuscript. The author reviewed and edited all content as needed, and takes full responsibility for the publication's content.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models. arxiv 2023, arXiv preprint arXiv:2302.13971 10 (2023).

[3] J. Wu, S. Yang, R. Zhan, Y. Yuan, L. S. Chao, D. F. Wong, A survey on LLM-generated text detection: Necessity, methods, and future directions, Computational Linguistics 51 (2025) 275–338. URL: https://aclanthology.org/2025.cl-1.8/. doi:10.1162/coli_a_00549.

[4] 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.

[5] X. Guo, S. Zhang, Y. He, T. Zhang, W. Feng, H. Huang, C. Ma, Detective: detecting ai-generated text via multi-level contrastive learning, in: Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24, Curran Associates Inc., Red Hook, NY, USA, 2024.

[6] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, C. Whitehouse, O. M. Afzal, T. Mahmoud, T. Sasaki, et al., M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection, in: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 1369–1407.

[7] A. Uchendu, Z. Ma, T. Le, R. Zhang, D. Lee, TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2001–2016. URL: https://aclanthology.org/2021.findings-emnlp.172/. doi:10.18653/v1/2021.findings-emnlp.172.

[8] Y. Li, Q. Li, L. Cui, W. Bi, Z. Wang, L. Wang, L. Yang, S. Shi, Y. Zhang, MAGE: Machine-generated text detection in the wild, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 36–53. URL: https://aclanthology.org/2024.acl-long.3/. doi:10.18653/v1/2024.acl-long.3.

[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] S. Gehrmann, H. Strobelt, A. Rush, GLTR: Statistical detection and visualization of generated text, in: M. R. Costa-jussà, E. Alfonseca (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Florence, Italy, 2019, pp. 111–116. URL: https://aclanthology.org/P19-3019/. doi:10.18653/v1/P19-3019.

[11] 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'23, JMLR.org, 2023.

[12] 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.

[13] Y. Chen, H. Kang, V. Zhai, L. Li, R. Singh, B. Raj, Gpt-sentinel: Distinguishing human and chatgpt generated content, 2023. URL: https://arxiv.org/abs/2305.07969. arXiv:2305.07969.

[14] X. Hu, P.-Y. Chen, T.-Y. Ho, Radar: robust ai-text detection via adversarial learning, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[15] Y. He, S. Zhang, Y. Cao, L. Ma, P. Luo, DETree: DEtecting Human-AI Collaborative Texts via Tree-Structured Hierarchical Representation Learning, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2025. URL: https://arxiv.org/abs/2510.17489.

[16] 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.

[17] 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.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[19] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, Advances in neural information processing systems 33 (2020) 18661–18673.