Generative AI Authorship Verification Based on Contrastive-Enhanced Dual-Model Decision System

Notebook for PAN at CLEF 2025

Junlang Liu, Leilei Kong*, Zhenyu Peng and Feifan Chen

Foshan University, Foshan, China

Abstract

Detecting human-written text from content produced by large language models (LLMs) remains a moving target, especially when detectors face unseen generators. We formalize the CLEF PAN 2025 Generative-AI Authorship Verification task as a text classification problem, employing a contrastive-enhanced ModernBERT-large approach, a Qwen3-based approach, and a fusion method combining both approachs. Specifically, to implement contrastive learning in the contrastive-enhanced method, we applied the large language model ChatGPT-4.1 for data augmentation, rewriting 1,000 human-written sentences.

On the official validation set, our contrastive-enhanced method achieves a 0.997 mean score, with all five PAN metrics above 0.99. On the hidden test set our submitted single-model ModernBERT-large(CE + SCL) achieves a 0.871 mean score (ROC-AUC = 0.822, F_1 = 0.855), ranking 3rd out of 24 teams. The results suggest that the contrastive-enhanced method yields competitive results, even without relying on large-scale ensemble systems.

Keywords

Generative AI Detection, Pre-trained Model, Contrastive-Enhanced, Text Classification

1. Introduction

Large language models (LLMs) have drastically lowered the cost of generating fluent text, but this progress intensifies the need to verify whether content is authored by humans or machines[1, 2]. Experience from the PAN 2024 lab shows that detectors fine-tuned on one generator family often underperform when faced with unseen models or domains[3].

Current detection methods face three core challenges:

- Models trained solely with cross-entropy loss tend to focus on surface-level features and struggle to capture subtle semantic differences between human-written and AI-generated texts.
- Previous ensemble methods, while improving accuracy, depend on multiple large models, resulting in low inference efficiency and significant deployment barriers due to hardware constraints.
- Inspired by the success of noise-based perturbation strategies in computer vision tasks—where slight input transformations help models generalize better—we explore analogous perturbation strategies for textual data to improve robustness and semantic representation learning[4, 5].

To address these issues, we propose a lightweight ensemble composed of a bidirectional encoder model (ModernBERT-large) and an autoregressive decoder model (Qwen3-4B). The encoder branch is fine-tuned using a joint cross-entropy and supervised contrastive loss to improve discrimination in borderline cases, while the decoder branch is trained with standard cross-entropy. During inference, the outputs are combined via mean-probability soft voting, incurring only minimal additional computational overhead.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

liujunlang2015@Gmail.com (J. Liu); kongleilei@fosu.edu.com (L. Kong); pengzhenyu1411@163.com (Z. Peng); chenfeifan0203@163.com (F. Chen)

0009-0009-1578-0867 (J. Liu); 0000-0002-4636-3507 (L. Kong); 0009-0003-4077-6647 (Z. Peng); 0009-0006-5292-9710 (F. Chen)

Our experiments show that the proposed method achieves state-of-the-art robustness on the PAN25 validation set, with all five official metrics—ROC-AUC, Brier, C@1, F_1, and $F_{0.5u}$—exceeding 0.99.

The remainder of this paper is organised as follows: Section 2 reviews related work; Section 3 details our model design and training; Section 4 presents results and discussion.

2. Related Work

The fast rise of LLMs has made reliably telling human- from machine-authored text a pressing NLP problem. Earlier efforts fall into (i) supervised classification, (ii) zero-shot detection, and (iii) multi-model decision aggregation.Classical lexical-feature classifiers can still rank highly—e.g. a plain SVM built on TF-IDF matched or beat neural baselines—yet their robustness drops once generators evolve. Conversely, zero-shot signals such as cross-perplexity generalise well but lag in absolute accuracy.

2.1. Supervised Classification Models

Traditional machine-learning methods remain remarkably competitive. Lorenz et al. employ a linear SVM trained on TF-IDF features and achieve performance close to the top[6]. Meanwhile, several teams fine-tune Transformer-based classifiers. Cao et al. enhance their model by augmenting the training set with additional human-written samples[7]. The Tri-Sentence Analysis method splits each long document into three shorter segments and averages their individual predictions to stabilise the final decision[8]. Lin et al. incorporate R-Drop regularisation to reduce the variance caused by dropout during inference[9]. Overall, supervised models achieved some of the highest mean scores in the PAN-24 competition. However, despite strong results on validation sets during training, these models often show reduced robustness when applied to out-of-domain test data, leading to noticeable performance drops in generalisation scenarios.

2.2. Zero-Shot Detection Models

Unsupervised techniques avoid costly annotation by exploiting statistical irregularities in machine text. Compression-based detectors such as PPMd-CDM treat lower entropy as an AI signature and require only a generic compressor[10]. The Binoculars framework measures the ratio between an observer model's perplexity and that of a performer model to expose hidden over-repetition in generated text[11]. However, their average performance in PAN-24 competition was notably lower than that of supervised systems, underscoring an inherent trade-off between broad generality and fine-grained accuracy.

2.3. Multi-model Decision Aggregation Models

To enhance robustness, some teams opted to combine multiple detection strategies. BinocularsLLM integrates two QLoRA-fine-tuned language models with Binoculars-style perplexity scoring, applying soft voting across all components to reach a final decision[12]. This ensemble achieved the top rank in the competition. LAVA takes a different approach by training separate adapters for different families of generative models and employs a conservative "unanimous agreement" rule—only predicting human authorship when all modules concur—effectively reducing false positives[13]. These ensemble-based systems demonstrated high mean scores in the evaluation, but their improved performance comes at the cost of increased inference time and memory usage, highlighting the trade-off between speed and accuracy.

3. System Overview

To build a robust generative-AI authorship verifier, three strategies are developed:

1. ModernBERT-large is fully fine-tuned as a classifier using both cross-entropy and supervised contrastive loss.

2. Qwen3-4B is fully fine-tuned with cross-entropy loss.
3. ModernBERT-large and Qwen3-4B are fused via weighted soft voting.

Our design aims to achieve the following main goals:

- Compare the performance of two different model architectures on the generative AI detection task after supervised fine-tuning.
- To enhance the overall robustness and generalization of the system by incorporating two structurally different models.

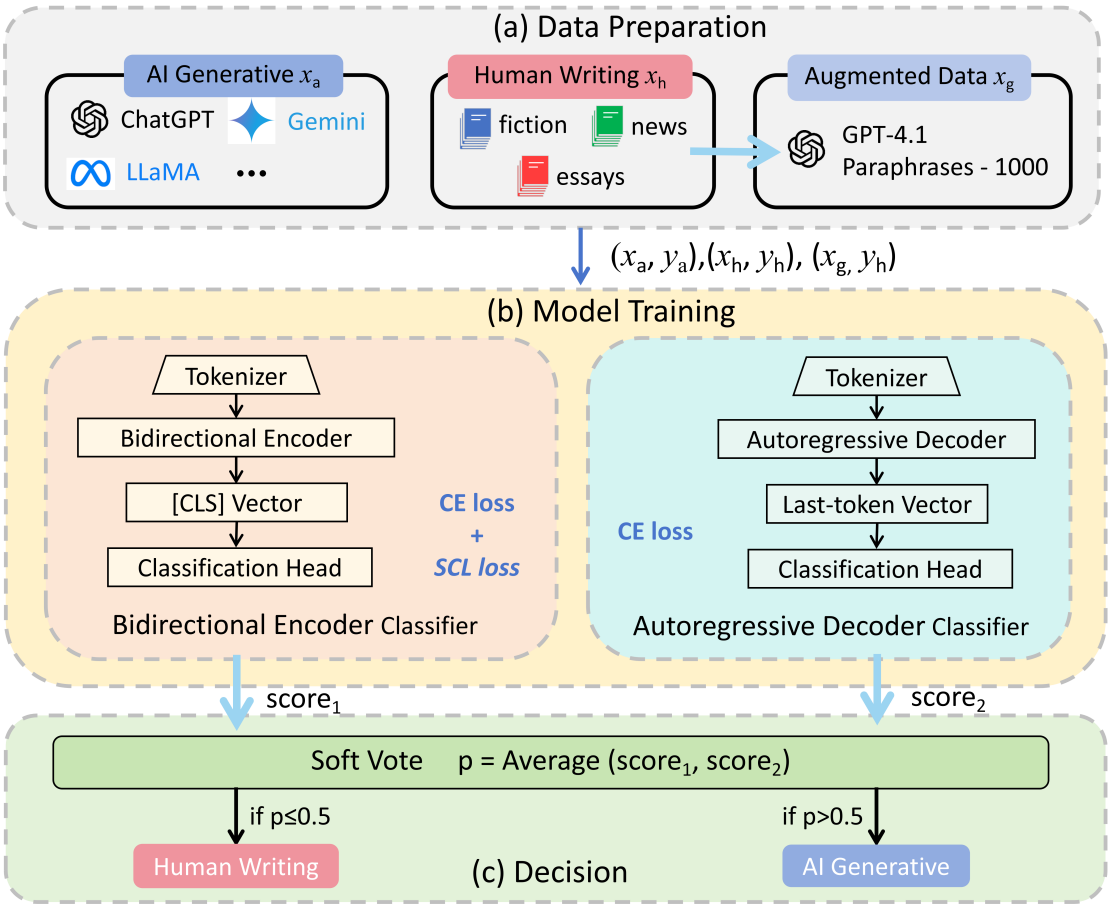

Figure 1: The overall architecture of the system.

Let $\mathcal{H} = \{h_i\}_{i=1}^{N}$ be the set of human-written texts ($h_i \in \Sigma^*$). Let $\mathcal{A} = \{a_j\}_{j=1}^{M}$ be the set of AI-generated texts. For 1000 texts $h_i \in \mathcal{H}$ we obtain an augmented paraphrase g_i using the GPT-4.1 model. The set of all augmented texts is $\mathcal{G} = \{g_i\}_{i=1}^{N}$. We assign the paraphrase set $\mathcal{G}$ the machine-generated label (1), while the corresponding original texts in $\mathcal{H}$ retain the human-written label (0). Unless stated otherwise we denote the complete corpus by $\mathcal{D} = \mathcal{H} \cup \mathcal{A} \cup \mathcal{G}$ and a generic sample by $x \in \mathcal{D}$.

3.1. Contrastive-Enhanced ModernBERT-large

3.1.1. Data Augmentation

To expose the detector to challenging near-human counterfeits, we first sampled 1000 sentences from the human class and then asked ChatGPT-4.1 (04-01-2025) to rewrite each sentence in its own words

while preserving the original meaning. We call these rewrites *paraphrases* and assign them the label 1 (machine-generated); their source sentences retain label 0 (human). Because the two versions of every sentence convey the same idea yet belong to opposite classes, they form hard positive-negative pairs that sharpen the contrastive objective.

Balanced mini-batches. Purely shuffling the data can yield mini-batches containing only positives or only negatives, which dilutes the contrastive signal. Therefore, we deterministically interleave samples in the order `human` → `paraphrase` → `human` → `machine`, aiming to keep the class ratio within each batch as close to 1:1 as possible.

Prompt

System Prompt:
This is a piece of text generated by a human. I want to express the same meaning as this sentence, but without changing its writing style. Please help me rephrase it. Just output the rephrased sentence directly.

User Prompt:
I approach a corner in the hallway as the door to a classroom in front of me opens and a girl steps out. She is wearing a form fitting black shirt with ...

Answer:
I round a corner in the hallway just as the door of a classroom ahead swings open and a girl steps out. She's dressed in a fitted black shirt, snug yet ...

For the augmented dataset, we first separated human-written texts and AI-generated texts. We then alternately inserted them one by one into the training dataset.Additionally, the remaining AI-generated texts were randomly inserted, and the 1000 augmented samples were ensured to be included in the same training batches as the original human-written texts. The final statistics of the training data are presented in Table 1.

Table 1
Text counts per model in our dataset.

Model	Count	Model	Count
human	9101	gpt-3.5-turbo	1374
falcon3-10b-instruct	879	gpt-4-turbo	272
qwen1.5-72b-chat-8bit	271	gpt-4-turbo-paraphrase	276
gemini-1.5-pro	1072	gpt-4o-mini	1358
gemini-2.0-flash	1079	gpt-4o	1336
gemini-pro	276	o3-mini	1075
gemini-pro-paraphrase	265	gpt-4.5-preview	278
deepseek-r1-distill-qwen-32b	901	llama-2-7b-chat	262
text-bison-002	265	llama-2-70b-chat	269
mistral-7b-instruct-v0.2	266	llama-3.1-8b-instruct	1063
ministral-8b-instruct-2410	1100	llama-3.3-70b-instruct	405
mixtral-8x7b-instruct-v0.1	264	ChatGPT-4.1	1000
Total (human)	**9101**	**Total (AI)**	**15606**

3.1.2. Supervised Fine-Tuning with Joint Loss

To better capture nuanced semantic differences, we adopt a supervised fine-tuning strategy combined with contrastive learning, training the LLM directly on labeled data. Specifically, we attach a fully

connected classification head to the hidden representation of the [CLS] token, allowing the model to output a binary label given an input text—where 0 denotes a human-written text and 1 denotes a machine-generated one.

To enhance the model's ability to discriminate between subtle semantic patterns, we incorporate supervised contrastive learning following the formulation proposed by Beliz Gunel et al[14]. The overall training objective is a weighted combination of cross-entropy loss and supervised contrastive loss. The final loss function is defined as follows:

$$\mathcal{L} = (1 - \lambda) \cdot \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{SCL}} \tag{1}$$

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \log p_\theta(y_i \mid \mathbf{x}_i) \tag{2}$$

$$\mathcal{L}_{\text{SCL}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp\left(\mathbf{z}_i^\top \mathbf{z}_j / \tau\right)}{\sum_{\substack{k=1 \\ k \neq i}}^{N} \exp\left(\mathbf{z}_i^\top \mathbf{z}_k / \tau\right)} \tag{3}$$

Specifically, $P(i)$ denotes the set of positive samples that share the same class label as the anchor sample i, z represents the hidden representation (feature vector) extracted by the model, and $\tau \in \mathbb{R}^+$ is a temperature hyperparameter that controls the concentration level of the similarity distribution. This formulation encourages the model to bring semantically similar samples closer in the representation space while pushing apart dissimilar ones, thereby improving class-level discrimination.

Equation (1) represents the overall loss, Equation (2) corresponds to the standard cross-entropy loss, and Equation (3) denotes the contrastive learning loss.

3.2. Supervised Fine-Tuning with LLMs

For the decoder-based model, we adopt Qwen3-4B as our backbone. The fine-tuning strategy is similar to that used in the encoder-based model. Specifically, we add a fully connected classification head to the output vector of the last token after decoding, and perform binary classification—predicting whether a given input text is human-written or machine-generated.

Unlike the encoder-based model, this decoder-only model is trained using only the standard cross-entropy loss, as the limited GPU memory prevented us from incorporating the contrastive-learning loss.

3.3. Contrastive-enhanced Dual-Model Decision(CeDD)

To combine the strengths of both the encoder-based and decoder-based models, we aggregate their prediction outputs using a soft voting strategy. Specifically, the final prediction probability is computed as the mean of the individual classification probabilities:

$$p_{\text{final}} = \frac{1}{2} \left(p_{\text{encoder}} + p_{\text{decoder}} \right) \tag{4}$$

A classification decision is then made based on a fixed threshold:

$$\text{Prediction} = \begin{cases} 1 & \text{if } p_{\text{final}} \geq 0.5 \quad \text{(machine-generated)} \\ 0 & \text{if } p_{\text{final}} < 0.5 \quad \text{(human-written)} \end{cases}$$

This simple yet effective fusion mechanism leverages the complementary inductive biases of the two model architectures. It improves prediction robustness without introducing significant computational overhead and helps mitigate model-specific errors on borderline or ambiguous samples.

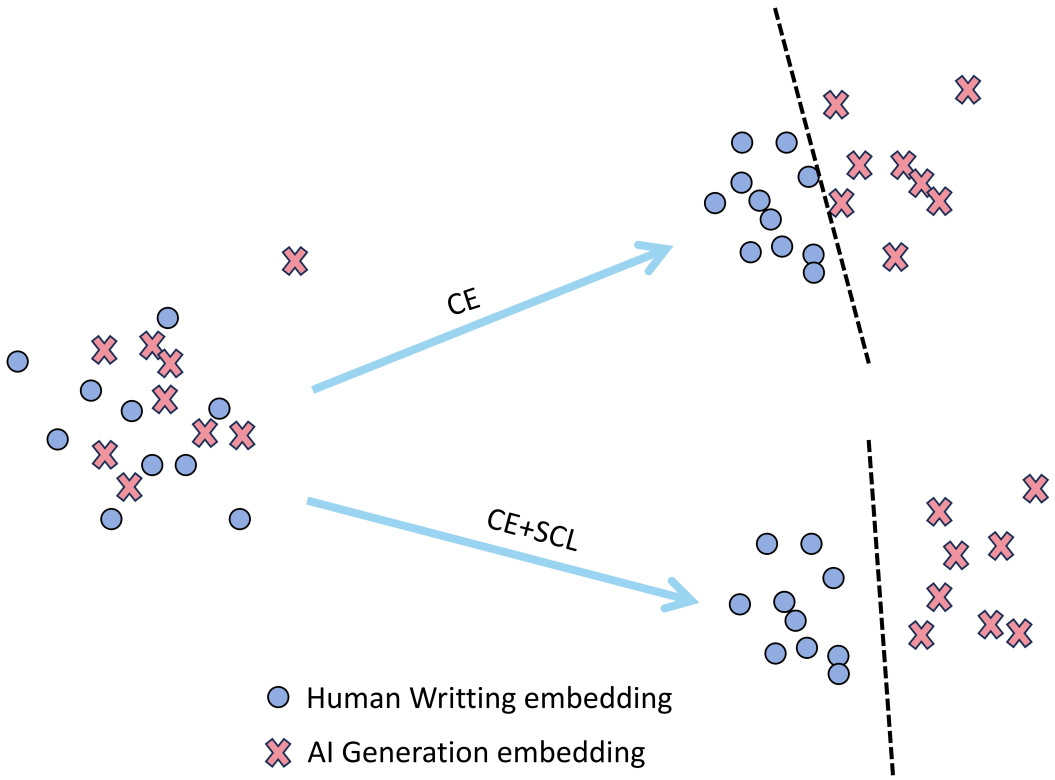

Figure 2: Visualization of embedding distributions under different training objectives. The embeddings represent texts with different labels. The model trained with CE+SCL encourages tighter intra-class clustering and larger inter-class separation. This illustrates that contrastive learning helps pull apart samples with different labels in the embedding space.

Algorithm 1 Contrastive-Enhanced Dual-Model Decision

Input: Text sample T
Output: Authorship label (0 or 1)

Initialize ENCODER_MODEL $\leftarrow$ ModernBERT-large
Initialize DECODER_MODEL $\leftarrow$ Qwen3-4B
$score_1 \leftarrow$ `Predict`(ENCODER_MODEL, T)
$score_2 \leftarrow$ `Predict`(DECODER_MODEL, T)
$final\_score \leftarrow \frac{score_1 + score_2}{2}$
if $final\_score \geq 0.5$ **then**
    $label \leftarrow$ "1"
else
    $label \leftarrow$ "0"
end if
return $label$

4. Results and Discussion

In this section, we present the implementation details, evaluation metrics, and provide a comprehensive analysis of the results. We utilize the TIRA platform to evaluate our three methods using test datasets[15].

4.1. Implementation Details

In this research, the training CeDD was implemented in PyTorch and executed on a single Nvidia A800 GPU. The model was trained using full bf16 precision to ensure numerical stability and training efficiency. The fine-tuning process lasted for 3 epochs, using the AdamW optimizer with a learning rate of 2e-5. The batch size was set to 32 without employing gradient accumulation. For the ModernBERT-large model, the training objective combined standard cross-entropy loss with supervised contrastive loss, with a lambda weight of 0.9 and a temperature of 0.3. The warm-up ratio was set to 0.1, and training logs were recorded every 50 steps. An independent validation set was used during training for evaluation. All experiments were conducted under a fixed random seed and employed cosine learning rate scheduling to ensure reproducibility.

4.2. Evaluation Metrics

To evaluate the performance of our proposed model, we used the evaluation metrics provided by PAN25, which include the following metrics:

- ***ROC–AUC***: The area under the ROC (Receiver Operating Characteristic) curve.
- ***Brier***: The complement of the Brier score (mean squared loss).
- ***C@1***: A modified accuracy score that assigns non-answers (score = 0.5) the average accuracy of the remaining cases.
- $\boldsymbol{F_1}$: The harmonic mean of precision and recall.
- $\boldsymbol{F_{0.5u}}$: A modified $F_{0.5}$ measure (precision-weighted F measure) that treats non-answers (score = 0.5) as false negatives.

4.3. Validation-set Results

As mentioned earlier, we compare three approaches for detecting AI-generated text: classification using an encoder-based model, classification using a decoder-based model, and a contrastive-enhanced dual-model decision strategy that combines both. The performance of various LLMs under these approaches is summarized in Table 2, based on evaluations on the validation dataset.

Table 2
Evaluation results of our final models on the *pan25-generative-ai-detection-val-20250428-training* dataset.

Model	ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
ModernBERT-large(CE)	0.998	0.996	0.996	0.997	0.998	0.997
ModernBERT-large(CE+SCL)	1	0.998	0.997	0.998	0.999	0.998
Qwen3-4B(CE)	0.999	0.991	0.990	0.992	0.994	0.993
CeDD	1	0.996	0.995	0.996	0.997	0.997
Baseline-SVM	0.996	0.951	0.984	0.980	0.981	0.978
Baseline-PPMD	0.786	0.799	0.757	0.812	0.778	0.786
Baseline-Binoculars	0.918	0.867	0.844	0.872	0.882	0.877

Upon analyzing the results shown in Table 2, it is evident that ModernBERT-large delivers the most stable and consistent performance across all evaluation metrics. Notably, it achieves an F_1 score of 0.998 and an $F_{0.5u}$ score of 0.999, highlighting its efficiency and accuracy in text classification tasks.

Qwen3-4B also performs competitively, especially in the Brier and mean scores, reflecting its strength in handling order-sensitive or generative-context inputs. This supports the effectiveness of the decoder-only architecture.

Our final system CeDD integrates both models and demonstrates near-optimal results across all six metrics. This confirms the effectiveness of our CeDD in enhancing the robustness, stability, and accuracy of generative authorship verification.

4.4. Test-set Results

Table 3 reports the score released by the PAN 2025 organisers for our submitted run.

Table 3
Performance of ModernBERT on the test set.

Model	ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean	FPR	FNR
ModernBERT-large (CE+SCL)	0.962	0.891	0.889	0.923	0.963	**0.928**	0.005	0.120
Baseline TF-IDF SVM	0.838	0.871	0.836	0.827	0.862	0.856	0.201	0.153
Baseline Binoculars Llama3.1	0.760	0.835	0.793	0.802	0.831	0.818	0.314	0.206
Baseline PPMd CBC	0.636	0.795	0.735	0.763	0.771	0.758	0.784	0.129

The single-model run ranked seventh overall, which we attribute to ModernBERT's simpler decision boundary potentially generalising better to unseen domains.

5. Conclusion

This work presents a supervised contrastive learning approach built upon the ModernBERT-large model for the CLEF PAN 2025 Generative-AI Authorship Verification task. By jointly optimizing cross-entropy loss and supervised contrastive loss, our method improves the model's ability to distinguish between human-written and AI-generated texts.

- On the official validation set, ModernBERT-large (CE+SCL) achieved a near-perfect mean score of 0.998 across all PAN metrics.
- On the hidden test set, this single-model approach obtained a mean score of 0.871, ranking 3rd out of 24 teams, confirming the effectiveness of our design.

In addition to the above results, we summarize the following key insights: (i) Supervised contrastive learning substantially enhances class separability and semantic discrimination; (ii) A single well-regularized encoder model can outperform complex ensembles while remaining efficient and scalable; (iii) Paraphrased data generated by GPT-4.1 serves as highly effective contrastive pairs during training, especially in narrowing the gap between human-like machine outputs and real human writing.

Overall, our findings show that a contrastively fine-tuned ModernBERT encoder can achieve strong performance on generative authorship verification, even without relying on large-scale ensemble systems or decoder-based large language models.

Acknowledgments

This work is supported by the National Social Science Foundation of China (No. 22BTQ101).

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-o3 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] 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: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

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

[3] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[4] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: H. D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 1597–1607. URL: https://proceedings.mlr.press/v119/chen20j.html.

[5] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Commun. ACM 64 (2021) 107–115. URL: https://doi.org/10.1145/3446776. doi:10.1145/3446776.

[6] L. Lorenz, F. Z. Aygüler, F. Schlatt, N. Mirzakhmedova, BaselineAvengers at PAN 2024: Often-Forgotten Baselines for LLM-Generated Text Detection, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, CEUR-WS.org, 2024, pp. 2761–2768. URL: http://ceur-ws.org/Vol-3740/paper-262.pdf.

[7] H. Cao, Z. Han, J. Ye, B. Liu, Y. Han, Enhancing Human-Machine Authorship Discrimination in Generative AI Verification Task with BERT and Augmented Data, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, CEUR-WS.org, 2024, pp. 2536–2540. URL: http://ceur-ws.org/Vol-3740/paper-230.pdf.

[8] J. Huang, Y. Chen, M. Luo, Y. Li, Generative AI Authorship Verification Of Tri-Sentence Analysis Base On The Bert Model, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, CEUR-WS.org, 2024, pp. 2632–2637. URL: http://ceur-ws.org/Vol-3740/paper-243.pdf.

[9] Z. Lin, Z. Han, L. Kong, M. Chen, S. Zhang, J. Peng, K. Sun, A Verifying Generative Text Authorship Model With Regularized Dropout, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, CEUR-WS.org, 2024, pp. 2728–2734. URL: http://ceur-ws.org/Vol-3740/paper-257.pdf.

[10] 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/10.1145/3098954.3104050. doi:10.1145/3098954.3104050.

[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, 2024. URL: https://arxiv.org/abs/2401.12070. arXiv:2401.12070.

[12] E. Tavan, M. Najafi, MarSan at PAN: BinocularsLLM , fusing Binoculars' Insight with the Proficiency of Large Language Models for Machine-Generated Text Detection, in: G. Faggioli, N. Ferro,

P. Galuščáková, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, CEUR-WS.org, 2024, pp. 2901–2912. URL: http://ceur-ws.org/Vol-3740/paper-281.pdf.

[13] Z. Chen, Y. Han, Y. Yi, Team chen at PAN: Integrating R-Drop and Pre-trained Language Model for Multi-author Writing Style Analysis, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, CEUR-WS.org, 2024, pp. 2547–2553. URL: http://ceur-ws.org/Vol-3740/paper-232.pdf.

[14] B. Gunel, J. Du, A. Conneau, V. Stoyanov, Supervised contrastive learning for pre-trained language model fine-tuning, 2021. URL: https://arxiv.org/abs/2011.01403. arXiv:2011.01403.

[15] 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.