Overview of PAN 2025: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection

Janek Bevendorff,[1,13] Daryna Dementieva,[2] Maik Fröbe,[3] Bela Gipp,[4] André Greiner-Petter,[4] Jussi Karlgren[5] Maximilian Mayerl,[6] Preslav Nakov,[7] Alexander Panchenko,[8] Martin Potthast,[9,10,11] Artem Shelmanov,[7] Efstathios Stamatatos,[12] Benno Stein,[13] Yuxia Wang,[7] Matti Wiegmann,[13] and Eva Zangerle[14]

[1]Leipzig University, Leipzig, Germany,
[2]Technical University of Munich, Munich, Germany,
[3]Friedrich Schiller University Jena, Jena, Germany,
[4]Georg-August-Universität, Göttingen, Germany,
[5]University of Helsinki,
[6]University of Applied Sciences BFI Vienna, Vienna, Austria,
[7]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE,
[8]Skoltech & AIRI, Moscow, Russia,
[9]University of Kassel, Kassel, Germany,
[10]hessian.ai, Darmstadt, Germany,
[11]ScaDS.AI, Leipzig, Germany,
[12]University of the Aegean, Samos, Greece,
[13]Bauhaus-Universität Weimar, Weimar, Germany,
[14]University of Innsbruck, Innsbruck, Austria,
pan@webis.de   pan.webis.de

Abstract The goal of the PAN lab is to advance the state of the art in text forensics and stylometry through an objective evaluation of new and established methods on new benchmark datasets. In 2025, we organized four shared tasks: (1) generative AI detection, particularly in mixed and obfuscated authorship scenarios, (2) multilingual text detoxification, a continued task that aims re-formulate text in a non-toxic way for multiple languages, and (3) multi-author writing style analysis, a continued task that aims to find positions of authorship change, and (4) generative plagiarism detection, a new task that targets source retrieval and text alignment between generated text and source documents. PAN 2025 concluded successfully with 56 notebook papers.

1 Introduction

PAN is a workshop series and a networking initiative for stylometry and digital text forensics. PAN hosts computational shared tasks on authorship analysis, computational ethics, and the originality of writing. Since the workshop's inception in 2007, we organized 77 shared tasks[1] and assembled 60 evaluation datasets[2] plus nine datasets contributed by the community. In 2025, we organized four tasks that concluded in 57 notebook papers.

First, the *Voight-Kampff Generative AI Detection* task asks to distinguish between human and machine-written text, with a focus on detector sensitivity in the presence of obfuscation and mixed human-machine authorship. The subtask 1 continues the research from 2024 in collaboration with the ELOQUENT lab and frames AI detection as an authorship verification task, tested across a large number of domains and obfuscation techniques to test detector robustness. The subtask 2 asks to distinguish between 6 different forms of human-AI collaboration in a given document, ranging from fully human-written to text with deep AI intervention. The Voight-Kampff Generative AI Detection task resulted in 30 notebook submissions. The task details are described in Section 2.

Second, the continuation of the *Multilingual Text Detoxification* task asks to, given a toxic piece of text, re-write it in a non-toxic way while saving the main content as much as possible. The task was extended to include texts from 15 languages—adding to 2024 edition Italian, French, Hebrew, Hinglish, Japanese, and Tatar—and had cross-lingual and multilingual as well as supervised and unsupervised challenges. The Multilingual Text Detoxification task resulted in 12 notebook submissions. The task details are described in Section 3.

Third, the continuation of the *Multi-Author Writing Style Analysis* task asks to, given a document, determine at which positions the author changes. This task was revamped for 2023 with a new dataset and structured around topical heterogeneity as an indicator of difficulty. While the previous iterations asked to separate authors at a paragraph level, we increased the difficulty for this year and asked participants to separate at the sentence level. The Multi-Author Writing Style Analysis task resulted in 11 notebook submissions. The task details are described in Section 4.

Fourth, the new *Generated Plagiarism Detection* task asks to, given a source and an LLM-obfuscated, suspicious document, determine the positions where the suspicious document reuses text from the source. The task resulted in 3 notebook submissions. The task details are described in Section 5.

PAN is committed to reproducible research in IR and NLP, hence all participants are asked to submit their software (instead of just their predictions) through the submission software TIRA. With the recent updates to the TIRA platform [30], a majority of the submissions to PAN are publicly available as docker containers. In the following sections, we briefly outline the 2025 tasks and their results.

[1]Find PAN's past shared tasks at `pan.webis.de/shared-tasks.html`

[2]Find PAN's datasets at `pan.webis.de/data.html`

Input / Task		**Possible Assignment Patterns**
1. { [?], [?] }		1. { [A], [M] }
2. { [?], [?] }		2. { [A], [M] }, { [A], [A] }
3. { [?], [?] }	⟶	3. { [A], [M] }, { [M], [M] }
4. { [?], [?] }		4. { [A], [M] }, { [A], [A] }, { [M], [M] }
5. { [?], [?] }		5. { [A], [M] }, { [A], [A] }, { [A], [B] }
6. { [?], [?] }		6. { [A], [M] }, { [A], [A] }, { [A], [B] }, { [M], [M] }
7. [?]		7. [A], [M]

Figure 1. Hierarchy of authorship verification problems from "easiest" (1) to "hardest" (7), involving LLM-generated text. Ignoring mixed human and machine authorship, the difficulty arises from the pairing constraints imposed by the possible assignment patterns. [M] denotes LLM-generated text, while [A] and [B] denote human-authored text (same letter meaning same human author).

2 Voight-Kampff Generative AI Detection

Authorship verification is a fundamental task in author identification. PAN has continuously been organizing authorship verification tasks for years [8, 9, 10, 11] and with generative AI / LLM detection being fundamentally also an authorship verification task [15], decided to "delve" into that realm. So, in 2024 we offered, for the first time, the *"Voight-Kampff" Generative AI Authorship Verification* task [14, 3], which attracted a large number of submissions.

For the 2024 installment, we formalized different task variants and ordered them from easiest to hardest (Figure 1). To establish a baseline, we decided to start with the easiest variant, in which participants were given a pair of texts of which exactly one was of human and the other of machine origin. This year, we move on to the harder variant, in which participants are given only one text. This variant reflects a more realistic scenario of authorship verification "in the wild," aligning with the settings commonly addressed in other LLM detection shared tasks.

Moreover, we extend the task to two distinct subtasks: (1) The classic binary *"Voight-Kampff" AI Detection Sensitivity* task, and (2) a multi-class *Human-AI Collaborative Text Classification* task. The subtask 1 is organized in collaboration with the ELOQUENT Lab in a builder-breaker style similar to the previous year: PAN participants build systems to identify machine authorship, while ELOQUENT participants supply datasets to try to break the systems.

A more detailed description and analysis of the submissions and the results can be found in the joint PAN and ELOQUENT task overview paper [13].

2.1 Subtask 1: Voight-Kampff AI Detection Sensitivity

The subtask 1 is in essence the classic binary detection task known also from other LLM detection shared tasks. However, we are testing the limits of the detectors by crafting a test set with text "obfuscations" that try to evade detection. Apart from drastic text length restrictions, the obfuscations we tested or received from ELOQUENT participants in the previous year turned out to be mostly ineffective. So this year, we tested what happens when the human writers obfuscate their style and whether machines can replicate this.

Dataset We created the task datasets from a selection of 19$^{\text{th}}$-century English fiction from Project Gutenberg, as well as the extended Brennan-Greenstadt [19] and Riddell-Juola [93] corpora. The latter two were constructed by collecting existing essays and then asking the authors to write another text describing their neighborhood but, in doing so, try to conceal their identity. No further instructions were given on how to achieve that. To generate LLM versions for all texts, we used the same summarize-then-expand technique as last year by prompting GPT-4o to generate bullet-point summaries of the input texts. The model was instructed to extract the main topic, a list of key points, the narrative point of view, the grammatical tense, and certain apparent style or obfuscation markers. We then used 13 LLMs to replicate both the original essays and the obfuscations from the summaries and style instructions. In addition to the neighborhood prompts, we asked the LLMs to also generate texts in the style of a 7-year-old, in subject-object-verb "Yoda" grammar, or with alliterations. Further, we added random words to the prompts which we asked the model to ignore, and we increased the temperature to the highest value that still produced sensible text.

Participants were given a training and a validation split of the dataset, which included only the original human fiction and essay texts and plain LLM versions of them. The obfuscated texts (both human and LLM) were held back for the test set. Participants were allowed to use external training and validation data, including last year's training set. The test set included both obfuscated and unobfuscated texts, as well as a small subsample of human and LLM U.S. news articles from last year's test dataset (which we never published).

Baselines We provided implementations of the following three baseline systems: As zero-shot baselines, we provided (1) Binoculars [36] (using Llama 3.1) and (2) a simple PPMd-based compression model using the compression-based cosine measure [77, 35]. The operating points for both were tuned on the validation set that was handed out to participants. As a supervised baseline, (3) we trained a linear SVM on the top-1000 TF-IDF 1–4-grams from the validation set. The TF-IDF detector and Binoculars can be considered state of the art, the compression model marks a more conservative lower baseline.

Evaluation All systems were submitted and evaluated on Tira [30]. At test time, the participants had to calculate a score between 0 and 1 for each text, indicating the likelihood that the text was LLM-generated. A score of exactly 0.5 could be given to signal a non-decision.

For each participant, we computed a confusion table and the following scores, which we used in previous authorship verification shared tasks as well:

Table 1. Arithmetic mean of all evaluation measures per submission for subtask 1.

Team	**Score**	**System**
Macko [59]	0.899	LoRA-tuned Qwen3 and data augmentation [60]
Seeliger [78]	0.880	Document-word correlations
Zaidi [99]	0.879	Fine-tuned BERT and data augmentation
Yang [98]	0.877	RoBERTa with contrastive learning
Teja [85]	0.874	Ensemble: Mixture of experts with PLMs
Marchitan [61]	0.872	Ensemble: LightGBM, XGBoost, Log. Regression, SVM with Qwen3 embeddings
Liu [57]	0.871	Ensemble: Fine-tuned PLM with contrastive loss
Valdez-V. [89]	0.869	Syntactic graphs and embeddings with GNNs
Voznyuk [92]	0.863	DeBERTa-v3 with multi-task learning (task, genre, model family classification)
TF-IDF SVM	0.856	*Baseline TF-IDF SVM*
Pudasaini [72]	0.852	Ensemble: SVM bagging of fine-tuned PLMs
Ostrower [67]	0.851	XGBoost with binoculars + stylometric features
Ochab [66]	0.844	LightGBM classifier with stylometric features
Völpel [90]	0.843	MLP with syntax n-gram features
Jimeno-G. [42]	0.838	Stacking ensemble with stylometric and word features
Sun [83]	0.835	Bi-CE [34] loss function + 25 stylometric features
Basani [6]	0.831	XGBoost classifier with token surprisal features
Titze [86]	0.827	Logistic regression on surprisal scores, entropy and JSD from two LLMs
Binoculars	0.818	*Baseline Binoculars Llama3.1 [36]*
Larson [50]	0.814	SVM with word and punctuation frequency features
Huang [38]	0.807	Fine-tuned RoBERTa + training data augmentation
Kumar [47]	0.788	Fine-tuned DistillBERT + stylometric features
PPMd CBC	0.758	*Baseline PPMd Compression-based Cosine [35, 77]*
Liang [53]	0.753	ModernBERT fine-tuning + loss-weighting based on example difficulty

- Roc-Auc: The area under the 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 [68].
- F_1: The harmonic mean of precision and recall.
- $F_{0.5u}$: A modified $F_{0.5}$ measure (precision-weighted F measure) that treats non-answers (score = 0.5) as false negatives [12].
- Mean: The arithmetic mean of all previous measures

Submitted Systems We received 20 submissions of which 7 beat the strongest baseline (TF-IDF SVM) and 9 more beat the second-strongest baseline (Binoculars). Overall, most systems had quite high mean scores above 0.9 with the best approach being almost perfect at 0.991. Table 1 shows the ranking of all participating teams ordered by their systems' Mean scores on the test set (excluding ELOQUENT submissions). If teams submitted multiple systems, only

Table 2. Subtask 2 training, development and test set distribution across six categories.

Label	Text Category	Train	Dev	Test
0	Fully human-written	75,270	12,330	34,509
1	Human-written, then machine-polished	95,398	12,289	43,154
2	Machine-written, then machine-humanized	91,232	10,137	25,234
3	Human-initiated, then machine-continued	10,740	37,170	22,802
4	Deeply-mixed text	14,910	225	12,500
5	Machine-written, then human-edited	1,368	510	2,557
Total		288,918	72,661	140,756

the highest score is shown. A more detailed break-down of how systems respond to individual obfuscations is described in the extended task overview paper [13].

In total, this subtask attracted 20 teams to submit systems in addition to the baseline systems we provided. Table 1 shows the best-performing system of each team that submitted notebook papers and a brief description of their approach.

2.2 Subtask 2: Human-AI Collaboration

The integration of AI technologies into the writing process has significantly altered traditional notions of authorship. The line between human and AI contributions has become increasingly ambiguous. AI involvement increasingly rises from *none* to *complete* [39]. From the perspective of ethical and intellectual accountability, we identify the role of humans and AIs for six types of text. Given a document collaboratively authored by humans and AIs, the subtask 2 is to classify it into one of the following six categories:

i. fully human-written;
ii. human-written, then machine-polished;
iii. machine-written, then machine-humanized (obfuscated);
iv. human-initiated, then machine-continued;
v. deeply mixed text; where some parts are written by a human and some are generated by a machine;
vi. machine-written, then human-edited.

Dataset The training and validation sets were constructed from existing datasets for fine-grained machine-generated text detection, comprising 288,918 examples for training and 72,661 for validation. For constructing the test set, we collected student essays, research papers, and peer reviews. We also incorporated several newly released datasets to comprehensively evaluate the generalization of detection systems across unseen generators and domains. The result test set consists of 140,756 instances. Detailed data distribution across six categories is shown in Table 2.

Participants were given the training and development sets. Although they were not allowed to use external training and validation data, data augmentation strategies such as back-translation, synonym replacement, random word deletion, and replacement were allowed.

Table 3. Subtask 2 evaluation results of 22 submissions, ranking by macro-recall, along with macro-F1 and accuracy, with one delayed submission.

Rank	Team	Recall	F1	Acc	System Description
1	mdok [59]	64.46	65.06	74.09	QLoRa PEFT fine-tuned Qwen3-4B-Base.
2	Bohan Li [51]	61.72	61.73	69.28	Under-sample high-frequency classes and adopt data augmentation for underrepresented classes, along with R-Drop regularization for DeBERTa-v3-base fine-tuning.
3	Advacheck [92]	60.16	60.85	69.04	Shared Transformer Encoder between several classification heads trained to distinguish the domains.
4	StarBERT [108]	57.46	56.31	66.81	Combine the deep language understanding of DeBERTa-v3-large and the high-dimensional mapping ability of Star-Block2d.
5	Atu [96]	56.87	56.45	66.30	DeBERTa enhanced by contextual and geometric attention
6	TaoLi [52]	56.74	55.39	66.27	Use DeBERTa-v3-Large
7	ReText.Ai [40]	56.11	55.25	64.79	Fine-tune Gemma-2 2B for sequence classification with multiple classification heads.
8	DetectTeam [82]	54.49	54.40	62.89	Fine-tune DeBERTa-V3-Large and combining multi-scale features.
9	WeiDongWu [95]	54.09	53.57	63.01	Combine the contextual strength of BERT with the sequence modeling capabilities of Transformer layers.
10	zhangzhiliang [107]	54.06	52.81	61.65	Fine-tune DeBERTa-V3-Large and combine it with BiLSTM and attention mechanism.
11	CNLP-NITS-PP [85]	54.05	53.49	62.23	Soft and Hard Mixture of Experts (MoE) architectures with DeBERTa-V3-Large
12	a.dusuki	52.83	51.44	60.45	–
13	Steely [78]	52.14	51.81	59.88	Cumulative sum of token-Level correlation signals
14	a.elnenaey	49.56	50.10	58.96	–
	Baseline	48.32	47.82	57.09	Fine-tune RoBerTa
15	VerbaNex AI [32]	47.15	47.15	56.24	Fine-tune Roberta with class balancing, data augmentation, and calculation of specific weights for each unbalanced class.
16	Unibuc-NLP [61]	44.33	42.76	51.42	Combine features at different layers extracted using Transformers with layer-wise projection and attentive pooling.
	Nexus Interrogators [99]	33.86	31.86	35.45	Fine-tune transformer models with data augmentation strategies on underrepresented classes.
17	johanjthomas	33.71	31.63	37.85	–
18	lza	32.90	31.98	33.20	–
19	NanMu	32.87	31.79	34.52	–
20	hkkk	32.79	31.95	34.21	–
21	YoussefAhmed21	16.48	14.98	21.22	–

Baseline To establish a baseline, we fine-tuned a pre-trained transformer-based model RoBERTa on the training set. Fine-tuning was performed using the Hugging Face `Trainer` API with the following configuration: learning rate of 2×10^{-5}, batch size of 16 for both training and evaluation, weight decay of 0.1, and a total of 3 training epochs. Checkpoints were evaluated at the end of each epoch, and the best-performing model on the development set was retained for subsequent testing. The baseline achieved a macro-recall of 68.67% on the development set, with corresponding macro-F1 and accuracy scores of 61.26% and 56.71%, respectively.

Evaluation Predictions of all systems were submitted and evaluated in CodaLab. At test time, participants assigned the predicted label among [0, 1, 2, 3, 4, 5] for each text, indicating its category. Participants in the leaderboard were ranked by macro-recall. Macro-recall is selected as the primary evaluation metric for two reasons: *(i.)* it gives equal importance to each class, preventing performance for majority classes from dominating the overall score on an unbalanced test set; and *(ii.)* macro-recall provides a more focused view on the model's ability to capture all positive instances for every class, compared with macro-F1 balancing precision and recall for each class. As additional evaluation metrics, we computed accuracy and macro-F1.

Submitted Systems 22 teams submitted their predictions to CodaLab, of which 16 submitted notebook papers [59, 78, 61, 99, 92, 85, 51, 40, 82, 96, 107, 31, 95, 52, 108, 32]. The performance of 14 teams is above the baseline, and 8 teams are below fine-tuned RoBERTa-base, as shown in Table 3. Many teams fine-tuned DeBERTa-v3-large and achieved better results than RoBERTa. Larger language models such as Qwen-3 4B and Gemma-2 2B were superior to DeBERTa and RoBERTa. The performance drop observed on the test set compared to the development set highlights the need for further improvement in fine-grained human-AI collaborative text detection.

3 Multilingual Text Detoxification

Text detoxification is a subtask of style transfer, aiming to transform toxic text into a neutral version while preserving its original meaning. With the rapid advancement of language models, concerns have intensified around their potential to generate harmful or biased content with many works developing toxicity mitigation in LLMs approaches [94]. A key challenge in this space is designing detoxification techniques that generalize effectively across languages. Building on our 2024 release of a multilingual parallel detoxification corpus covering 9 languages [27] (English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, Amharic), we now extend the task to explore both multilingual and cross-lingual generalization. This year's shared task introduces 6 additional languages—Italian, French, Hebrew, Hinglish, Japanese, and Tatar—offering new challenges for scalable and inclusive detoxification methods.

Dataset We provided several datasets for participants to train their models and enhance their approaches:

- **Multilingual ParaDetox**: Train part of parallel toxic-neutral 400 pairs per 9 languages from 2024 edition;

Table 4. Results of the final evaluation of the TextDetox test phase. Scores are sorted by the average **J**oint scores: with parallel (**P**) and without parallel (**NP**) training data. Baselines are highlighted with gray, Human References are highlighted with green.

Team	AvgP	AvgNP	System
Human References	0.854	0.847	Human paraphrases from our Multilingual ParaDetox
ducanhhbtt [23]	0.685	0.643	LoRA fine-tuning and advanced prompting with Gemma3-12B
MetaDetox [18]	0.685	0.609	CoT prompting of DeepSeek with outputs reranking
sky.Duan [97]	0.676	0.501	Combination of our mT0-detox baseline with Qwen3
Pratham [79]	0.676	0.575	Fine-tuned mT0 with lexical refining
jellyproll	0.675	0.605	mT0 baseline with improved vocab
mT0	0.675	0.572	Fine-tuned mT0 on 9 languages train ParaDetox
Jiaozipi [58]	0.656	0.607	Ensemble of LLMs with RISE framework
SVATS [44]	0.656	0.599	Combination of fine-tuned Qwen2 and Gemma2
nikita.sushko [91]	0.628	0.512	Additionally tuned mT0 with our and synthetic data
ylmmcl [48]	0.612	0.471	Combination of BART, mT0, and LLaMa3.1 for outputs ranking
Gopal [45]	0.611	0.595	Replacement of toxic spans with GPT4o-mini
d1n910 [69]	0.604	0.575	CoT with DeepSeek-R1
GPT-o3	0.562	0.484	Few-shot Prompting of GPT-o3mini
GPT-o4	0.560	0.535	Few-shot Prompting of GPT-o4
Something Awful	0.549	0.511	Llama3.1 with Reasoning with top5 selection
Delete	0.536	0.510	Elimination of toxic keywords
Backtr.	0.481	0.342	Translation of data to English+BART-detox
Duplicate	0.475	0.482	Simple duplication of toxic input

- **Multilingual Toxic Lexicon**: Collection from open corpora of toxic keywords for all 15 languages;
- **Multilingual Toxic Spans**: Toxic collocations extracted with GPT-4 from 9 languages from the train Multilingual ParaDetox dataset [26];
- **Multilingual Toxicity Classification Data**: Collection of binary toxicity classification corpora for all 15 languages.

Then, we extended our test set to 6 new languages for which no parallel training data were provided: Italian, French, Hebrew, Hinglish, Japanese, and Tatar. The language stakeholders utilized various opensourced toxicity or hate speech classification datasets then rewriting the texts into neutral version with native speakers. We provided the same annotation instructions as for 2024 edition [27]. The goal of annotation was to obtain detoxification pairs for 600 unique toxic original instances per each language to form the test set.

Phases and Tracks We structured our shared task into two phases: (i) **Development phase**: Participants were provided with the Multilingual ParaDetox parallel training data for 9 languages, alongside 600 test toxic instances for each of these languages and an additional 100 toxic instances for each of 6 new languages. (ii) **Test phase**: Participants received the full 600 toxic test instances for all 15 languages, including the newly introduced ones.

To emphasize both multilingual and cross-lingual generalization, we reported results across two evaluation tracks in each phase:

– **AvgP**: The average performance across the 9 languages with available *Parallel* training data according (hence the name). This track focuses on building *multilingual* detoxification models that generalize well across multiple high-resource settings.
– **AvgNP**: The average performance on the 6 new languages for which *No Parallel* training data was released—only test sets were provided. This track presents a *cross-lingual* challenge, encouraging participants to develop approaches that transfer knowledge from the training languages or leverage other external resources to perform well in low-resource settings.

Evaluation For both phases, we provided the leaderboard based on an automatic evaluation setup. We evaluate the outputs based on three parameters—style of text, content preservation, and conformity to human references—combining them into the final **J**oint score:

– **Style Transfer Accuracy (STA)** ensures that the generated text is indeed more non-toxic. It was estimated with XLM-R [22] `large` instance fine-tuned for the binary toxicity classification task for our target languages. We compared the non-toxicity scores of models outputs with human references.
– **Content Similarity (SIM)** is the cosine similarity between LaBSE embeddings [29] of both the toxic source and human references and the generated texts.
– **Fluency** is used to estimate the proximity of the detoxified texts to human references and their fluency estimated with xCOMET [49].

Final Joint Score (J) was the aggregation of the three above metrics:

$$\mathbf{J} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{STA}(x^{ref}, y_i) \cdot (0.4 * \mathbf{SIM}(x_i, y_i) + 0.6 * \mathbf{SIM}(x_i^{ref}, y_i)) \cdot \mathbf{FL}(x_i, x^{ref}, y_i)$$

We calculated all the metrics separately per each language. In the end, we calculated the **Average** score of **J**oint scores per all languages in the track.

Baselines We provided several both unsupervised and more modern baselines. For the easy start, we provided:

i. Duplicate: a simple duplication of the toxic input.
ii. Delete: elimination of a toxic keywords based on a predefined dictionary for each language.

iii. Backtranslation: tranlsation of any input to English and detoxification with BART-detox model and translation back.
iv. LLMs prompting: GPT-4o and GPT-o3-mini zero-shot prompting.

For supervised approaches, we provided mBART [26] and mT0 [75] models fine-tuned on 9 languages training ParaDetox.

Submitted Systems Per both *development* and *test* phases, we got 31 systems submitted that resulted in 12 notebooks submissions [97, 18, 79, 58, 87, 28, 23, 44, 48, 91, 45, 69]. While there is indeed a very big tendency of LLMs prompting solutions, still, many submissions were based on various improvements over *seq2seq* generative models or LLMs. Thus, many participants tried chain-of-thoughts or other advanced prompting techniques over recent powerful LLMs like DeepSeek [25], LLaMa3 [1], Qwen [4], and Gemma [24], as well as special fine-tuning and cross-lingual inference with mT0 [65].

Results The results of the most interesting submissions are presented in Table 4. First, only five submissions outperformed our strongest baseline, mT0, and even these remained well below human reference performance. Additionally, many systems showed imbalanced results between languages with and without training data. Nevertheless, several creative approaches demonstrated that effective cross-lingual text detoxification is feasible with modern language models.

4 Multi-Author Writing Style Analysis

Writing style analysis serves as the cornerstone for authorship identification. The multi-author writing style analysis task within PAN@CLEF has continuously advanced this essential research domain by developing challenges. The task has undergone substantial transformation across multiple iterations: beginning with the identification and clustering of individual authors [74], progressing to distinguishing between single-author and multi-author documents [88, 43, 106], advancing to determining the precise number of contributing authors [105], and paragraph-level detection of style changes within documents [100, 101, 102, 103].

In the 2025 edition of the PAN multi-author writing style analysis task, we asked participants to identify positions of writing style changes within a set of documents. Building on previous editions that focused on the detection of paragraph-level style changes, this year's task advances to detecting style changes at the sentence level, making the setting more realistic.

The dataset provided to participants consists of three datasets varying in the difficulty of detection style changes: *Easy:* Each document covers a variety of topics, allowing participants to leverage topic information as a cue for detecting changes in writing style. Furthermore, the stylistic similarity between sentences in the document is rather low. *Medium*: The topics within a document are more homogeneous, requiring approaches to rely more heavily on stylistic features rather than topic differences to identify style changes. The stylistic similarity between sentences is moderate. *Hard*: All sentences within a document are of a single topic and stylistically similar.

We control for topical diversity across the datasets to ensure that the focus is on stylistic changes. In particular, the hard dataset eliminates topical differences as a proxy signal for authorship, requiring the use of writing style analysis to detect changes.

Data Set and Evaluation

We leverage data from the Reddit platform[3] for the multi-author writing analysis task. In particular, we select user posts from topic-specific subreddits, including *r/worldnews*, *r/politics*, *r/askhistorians*, and *r/legaladvice*. This diverse selection of sources allows for curating documents with varying levels of topical coherence. To construct individual documents, we extract posts from these subreddits, apply preprocessing steps (such as removing quotes, whitespace, emojis, and hyperlinks), and then split the posts into individual sentences.

Based on this data, we construct documents by extracting sentences from a single Reddit post, authored by two to four users. For each sentence, we compute semantic and stylistic feature vectors, enabling the computation of topical (semantic) and stylistic similarity between individual sentences. Based on these similarities, we apply a mixing approach for all sentences of the individual authors of the given Reddit post. We then concatenate sentences based on their topical and stylistic similarity, allowing us to control for the difficulty of the style detection task. For the three datasets, we configure the similarity threshold for consecutive sentences to be (1) relatively high for the *easy* dataset, (2) moderate for the *medium* dataset, and (3) small for the *hard* dataset. Each of the easy, medium, and hard datasets contains 6,000 documents. We provided participants with training, validation, and test splits for all three datasets. The training sets contain 70% of the documents in each dataset, while the validation and test sets contain 15% each. The test sets were withheld for the evaluation phase of the competition.

The submitted approaches are evaluated on each dataset using the macro-averaged F1-score calculated across all documents.

Results

The task received twelve valid software submissions and working notes papers. The F1-scores for each task achieved by the participants are shown in Table 5. The best average F1-score across the three datasets was achieved by team wqd, reaching a score of 0.870. For the easy dataset, Team stylospies achieved a marginally better result, while scoring the fifth and third best results for the medium and hard datasets, respectively. For the medium dataset, xxsu-team achieved a marginally higher score. Generally, we observe that the individual approaches perform quite differently on the three datasets. For instance, teams

[3] https://www.reddit.com/

Table 5. Overall results for the multi-author writing style analysis task, ranked by average F_1 performance across all three datasets. Best results are marked in bold.

Team	Easy F_1	Medium F_1	Hard F_1
wqd [55]	0.958	0.823	**0.830**
xxsu-team [54]	0.955	**0.825**	0.829
stylospies [17]	**0.959**	0.786	0.791
team-tmu [37]	0.950	0.792	0.792
better-call-claude [76]	0.929	0.815	0.731
openfact [46]	0.919	0.771	0.752
cornell-1 [16]	0.909	0.793	0.698
batatavada-pict [73]	0.823	0.766	0.667
hhu [62]	0.761	0.666	0.642
ksu [2]	0.507	0.747	0.467
hellojie [20]	0.461	0.583	0.484
team-of-bf [56]	0.486	0.443	0.473
Baseline Predict 1	0.178	0.177	0.147
Baseline Predict 0	0.439	0.440	0.453

cornell-1 and better-call-claude perform better on the medium dataset than on the easy and the hard datasets. Most submissions were able to outperform the two simple baselines: one baseline that predicted a style change for each pair of sentences, and one that predicted no style change for each pair of sentences. Further details on the approaches taken can be found in the overview paper [104].

5 Generative Plagiarism Detection

Plagiarism detection has a long-standing tradition in PAN, with main tasks running from 2009 [71] to 2015 [80]. Over time, the focus gradually shifted toward more specialized intrinsic tasks, such as the still active authorship analysis challenges. However, the recent breakthrough of generative AI has dramatically transformed the landscape of plagiarism detection. For the first time in history, LLMs can serve as so-called automatic plagiarists [5]. This shift inspired us to revive a classic plagiarism detection task for 2025, this time centered on automatically generated plagiarism using LLMs.

For the 2025 edition, we adhered to the well-established foundations of the 2015 plagiarism detection task, particularly in evaluation methodology and dataset formatting [5]. Participants received an annotated synthetic dataset of pairs of documents (S, P), where S is a source document and P is the plagiarism document in which the paragraphs p were replaced with paraphrased versions of paragraphs s in S using LLMs without citation. This setup closely mirrors the 2015 PAN text alignment task[4], allowing us to evaluate how well past approaches have aged.

[4]http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/plagiarism-detection.html

5.1 Dataset

The synthetic dataset was constructed by first identifying the most semantically similar document pairs on arXiv, using embeddings from the SPECTER model [21] applied to the 2025 release of ar5iv[5]. We then sampled a subset of 100,000 documents with an even distribution across all arXiv categories (also known as archives), to ensure a wide variety of topics. For each remaining document pair (S, P), we aligned the most semantically similar paragraphs s and p from S and P, respectively, based on three criteria. The alignment score was computed as a weighted aggregate: 50% semantic similarity via SciBERT sentence embeddings [7], 40% lexical similarity using TF-IDF vector similarity, and 10% section title similarity using SciBERT embeddings. The inclusion of similarity in the title of the section helped discourage the alignment of paragraphs from unrelated sections of the documents.

For each pair (S, P), we selected one of three LLMs: LLaMA-3 [1] (3.3 70B Instruct), DeepSeek-R1 [25] (Distill-Qwen-32B) or Mistral [63] (7B Instruct v0.3), and replaced all p in each aligned paragraph (s, p) with LLM-paraphrased versions s' derived from paragraphs s in S. To support a more detailed analysis of system performance, we established several categories of document pairs. First, 5% of the 100,000 pairs remained unchanged, i.e., both S and P are original arXiv documents. An additional 20% of pairs do not contain any plagiarism, but some paragraphs in P have been paraphrased by an LLM independently of S. These examples are useful for evaluating systems that aim to detect LLM-generated content rather than plagiarism specifically. The remaining 75% of document pairs were constructed as described above.

We further classified the severity of plagiarism in P into three levels: low, medium, and high. These refer to the proportion of paragraphs in P that were replaced with paraphrased versions from S. In 30% of the document pairs, the severity was *low*, with 20% to 40% of paragraphs replaced. In 40% of the pairs, severity was *medium*, with 40% to 60% replaced. The remaining 30% had *high* severity, where 70% to 100% of paragraphs in P were substituted.

Paraphrasing Prompts Each LLM used three types of prompts to generate paraphrased plagiarism. These were distributed across document pairs as follows. 60% of the pairs used a *simple prompt*:

```
Paraphrase the given paragraph for a professional audience.
```

30% used a *medium prompt*:

```
Reformulate the given paragraph in a sophisticated manner
while preserving its meaning. Modify sentence structure,
reword phrases, and incorporate elements of general
knowledge to ensure coherence. The less token overlap, the
better.
```

[5] https://ar5iv.labs.arxiv.org/

Table 6. Plagiarism alignment dataset and LLM splits.

Splits / LLMs	Llama-3		DeepSeek-R1		Mistral		Altered	Original	Total
Train	18,423	79.80%	18,452	79.46%	6,265	79.65%	15,101	3,918	62,159
Validation	2,353	10.19%	2,383	10.26%	802	10.20%	1,919	518	7,975
Test	2,310	10.01%	2,386	10.28%	799	10.16%	1,919	490	7,904
Total	23,086	42.62%	23,221	42.86%	7,866	14.52%	18,939	4,926	78,038

The final 10% used a *hard prompt* that incorporated immediate context to help the generated paragraph blend into its surrounding text. The prompt took the following form:

```
Completely rephrase the given paragraph in your own words.
Feel free to incorporate elements from general knowledge to
ensure coherence, flow, and better understanding.

{context_before}
```

All prompts included additional instructions to output only the paraphrased content, avoiding any explanatory text. Special tokens were used to suppress verbose output, tailored to each LLM. For DeepSeek-R1, a custom <thinking>...</thinking> block was used to suppress the model's internal reasoning steps, which would otherwise significantly slow down the generation. It is worth noting that Mistral performed poorly in following prompt instructions. It often produced explanatory content, hallucinated facts, or entered repetitive output loops, an issue reminiscent of neural network architectures before the attention mechanism era. In total, the final dataset consisted of 78,038 document pairs, divided into training, validation, and test subsets. The training and validation sets were provided to participants, while the test set was kept private for the evaluation phase. The data splits and sizes is given in Table 6.

5.2 Evaluation

All systems were submitted and evaluated on the TIRA platform. The participants were tasked with identifying all the paragraphs s' in P and aligning each with the corresponding paragraph s in S. The training and validation sets contained all alignments (s, s') for each pair of documents (S, P), together with the full text of both documents. The evaluation was carried out using the original scripts from the 2015 PAN plagiarism detection task. The metrics included micro and macro F1 scores as well as the established `plagdet` metric [70].

Four teams participated in the task by submitting software. Table 7 shows the aggregated evaluation results for all submissions that we also compared to the PAN baseline from 2012. We report the arithmetic mean of all evaluation measures (micro precision, macro precision, micro recall, macro recall, micro plagdet, and macro pladget) as main evaluation score. All submissions substantially improve upon the PAN-12 baseline that used lexical near-duplicate detection. All submissions used some form of semantic similarity embeddings.

Table 7. Arithmetic mean of all evaluation measures per submission for the plagiarism detection alignment task.

Team	Score	System
chi-zi-zhi-xin-dui [81]	0.440	Sentence-BERT, MPNet, TF-IDF
jrluo [41]	0.263	E5 and MiniLM-L6
foshan-university [84]	0.400	TF-IDF and BERT classifier
yukino [64]	**0.471**	Glove embeddings
Baseline PAN-12	0.233	Lexical near-duplicate detection
Baseline Llama-3.3 [1]	0.269	Llama-3.3 70B embeddings
Baseline Qwen2 [4]	0.375	Qwen2 7b Instruct embeddings

Therefore, we added two additional baselines relying upon two typical embedding models: Llama-3.3 70B and Qwen2 7B instruct. Team Yukino achieving the highest score relying on Glove embeddings closely followed by Team Su, which used an ensemble of multiple semantic embeddings combined with lexical TF-IDF similarity. An extended evaluation will be available in the task overview [33].

Acknowledgments

The work of Janek Bevendorff, Matti Wiegmann, Maik Fröbe, Martin Potthast, and Benno Stein has been funded as part of the OpenWebSearch project by the European Commission (OpenWebSearch.eu, GA 101070014). The work of André Greiner-Petter has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 554559555.

Bibliography

[1] AI@Meta: Llama 3 Model Card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md (2024), accessed: 2024-12-14

[2] Alsheddi, A., El Bachir Menai, M.: Style Change Detection in Multi-authored English Texts Based on Graph Convolutional Networks. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[3] Ayele, A.A., Babakov, N., Bevendorff, J., Casals, X.B., Chulvi, B., Dementieva, D., Elnagar, A., Freitag, D., Fröbe, M., Korenčić, D., Mayerl, M., Moskovskiy, D., Mukherjee, A., Panchenko, A., Potthast, M., Rangel, F., Rizwan, N., Rosso, P., Schneider, F., Smirnova, A., Stamatatos, E., Stakovskii, E., Stein, B., Taulé, M., Ustalov, D., Wang, X., Wiegmann, M., Yimam, S.M., Zangerle, E.: 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. 15th International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, vol. 14959, pp. 231–259, Springer,

Berlin Heidelberg New York (Sep 2024),
https://doi.org/10.1007/978-3-031-71908-0_11

[4] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)

[5] Barrón-Cedeño, A., Potthast, M., Rosso, P., Stein, B.: Corpus and evaluation measures for automatic plagiarism detection. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta, European Language Resources Association (2010), URL
http://www.lrec-conf.org/proceedings/lrec2010/summaries/35.html

[6] Basani, A.R., Chen, P.: DivEye at PAN 2025: Diversity Boosts AI-Generated Text Detection. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[7] Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, pp. 3613–3618, ACL (2019), https://doi.org/10.18653/V1/D19-1371

[8] Bevendorff, J., Borrego-Obrador, I., Chinea-Ríos, M., Franco-Salvador, M., Fröbe, M., Heini, A., Kredens, K., Mayerl, M., Pęzik, P., Potthast, M., Rangel, F., Rosso, P., Stamatatos, E., Stein, B., Wiegmann, M., Wolska, M., Zangerle, E.: Overview of PAN 2023: Authorship Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association (CLEF 2023), Lecture Notes in Computer Science, vol. 14163, pp. 459–481, Springer, Berlin Heidelberg New York (Sep 2023), https://doi.org/10.1007/978-3-031-42448-9_29

[9] Bevendorff, J., Chulvi, B., Fersini, E., Heini, A., Kestemont, M., Kredens, K., Mayerl, M., Ortega-Bueno, R., Pezik, P., Potthast, M., Rangel, F., Rosso, P., Stamatatos, E., Stein, B., Wiegmann, M., Wolska, M., Zangerle, E.: Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022), Lecture Notes in Computer Science, vol. 13186, Springer, Berlin Heidelberg New York (Sep 2022),
https://doi.org/10.1007/978-3-031-13643-6

[10] Bevendorff, J., Chulvi, B., Sarracén, G.L.D.L.P., Kestemont, M., Manjavacas, E., Markov, I., Mayerl, M., Potthast, M., Rangel, F., Rosso, P., Stamatatos, E., Stein, B., Wiegmann, M., Wolska, M., Zangerle, E.: Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter,and Style Change Detection. In: 12th International Conference of the CLEF Association (CLEF 2021), Springer (Sep 2021), URL
https://doi.org/10.1007/978-3-030-85251-1_26

[11] Bevendorff, J., Ghanem, B., Giachanou, A., Kestemont, M., Manjavacas, E., Markov, I., Mayerl, M., Potthast, M., Rangel, F., Rosso, P., Specht, G.,

Stamatatos, E., Stein, B., Wiegmann, M., Zangerle, E.: Overview of PAN 2020: Authorship Verification, Celebrity Profiling, Profiling Fake News Spreaders on Twitter, and Style Change Detection. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 11th International Conference of the CLEF Initiative (CLEF 2020), Lecture Notes in Computer Science, vol. 12260, pp. 372–383, Springer, Berlin Heidelberg New York (Sep 2020), https://doi.org/10.1007/978-3-030-58219-7_25

[12] Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: Generalizing Unmasking for Short Texts. In: 14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019), pp. 654–659, Association for Computational Linguistics (Jun 2019), URL https://aclanthology.org/N19-1068/

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

[14] Bevendorff, J., Wiegmann, M., Karlgren, J., Dürlich, L., Gogoulou, E., Talman, A., Stamatatos, E., Potthast, M., Stein, B.: Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2024. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, pp. 2486–2506, CEUR Workshop Proceedings, CEUR-WS.org (Sep 2024), URL http://ceur-ws.org/Vol-3740/paper-225.pdf

[15] Bevendorff, J., Wiegmann, M., Richter, E., Potthast, M., Stein, B.: The Two Paradigms of LLM Detection: Authorship Attribution vs. Authorship Verification. In: The 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025) (Findings), Association for Computational Linguistics (Jul 2025)

[16] Boloni-Turgut, D., Verma, D., Cardie, C.: Team cornell-1 at PAN: Ensembling Fine-Tuned Transformer Models for Writing Style Analysis. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[17] Boriceanu, I., Băltoiu, A.: Style Change Detection Using Graph and Structural-Linguistic Features for Multi-Author Writing Analysis. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[18] Bourbour, S., Kelishami, A.S., Gheysari, M., Rahimzadeh, F.: Cross-Lingual Detoxification with Few-Chain Prompting: A Competitive System for TextDetox 2025. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[19] Brennan, M., Afroz, S., Greenstadt, R.: Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. ACM Transactions on Information and System Security **15**(3) (Nov 2012), https://doi.org/10.1145/2382448.2382450

[20] Chen, D., Li, J., Qi, H.: Llama-3 with 4-bit Quantization and IA^3 Tuning for Multi-Author Writing Style Analysis. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[21] Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D.S.: SPECTER: document-level representation learning using citation-informed transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 2270–2282, Association for Computational Linguistics (2020), https://doi.org/10.18653/V1/2020.ACL-MAIN.207

[22] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 8440–8451, Association for Computational Linguistics (2020), https://doi.org/10.18653/V1/2020.ACL-MAIN.747

[23] Dang, T.D.A., D'Elia, F.P.: GemDetox: Enhancing a massively multilingual model for text detoxification on low-resource languages. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[24] DeepMind: Gemma Model Card. https://github.com/google-deepmind/gemma (2024), accessed: 2025-06-09

[25] DeepSeek-AI: Deepseek-v3 technical report (2024), URL https://arxiv.org/abs/2412.19437

[26] Dementieva, D., Babakov, N., Ronen, A., Ayele, A.A., Rizwan, N., Schneider, F., Wang, X., Yimam, S.M., Moskovskiy, D., Stakovskii, E., Kaufman, E., Elnagar, A., Mukherjee, A., Panchenko, A.: Multilingual and explainable text detoxification with parallel corpora. In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 7998–8025, Association for Computational Linguistics, Abu Dhabi, UAE (Jan 2025), URL https://aclanthology.org/2025.coling-main.535/

[27] Dementieva, D., Moskovskiy, D., Babakov, N., Ayele, A.A., Rizwan, N., Yimam, S.M., Ustalov, D., Stakovskii, E., et al.: Overview of the multilingual text detoxification task at pan 2024 (2024)

[28] Farid, H., Ahmad, Z., Mahmood, A., Ameer, I.: HF_Detox at PAN 2025 TextDetox: Prompt-Driven Multilingual Detoxification. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[29] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic BERT sentence embedding. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, pp. 878–891, Association for Computational Linguistics (2022), https://doi.org/10.18653/V1/2022.ACL-LONG.62

[30] Fröbe, M., Wiegmann, M., Kolyada, N., Grahm, B., Elstner, T., Loebe, F., Hagen, M., Stein, B., Potthast, M.: Continuous Integration for Reproducible Shared Tasks with TIRA.io. In: Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), pp. 236–241, Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York (Apr 2023)

[31] Fuchuan, Y., Cao, H., Zhongyuan, H.: Sentence-Level AI-Generated Text Detection with Fine-Tuned BERT. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[32] Gómez Sánchez, D., Jimenez, J., Ramírez, M., Martinez, J.: RoBERT-IA: Human-AI Collaborative Text Classification. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[33] Greiner-Petter, A., Fröbe, M., Wahle, J.P., Ruas, T., Gipp, B., Aizawa, A., Potthast, M.: Overview of the Generative Plagiarism Detection Task at PAN 2025. In: CLEF 2025 Working Notes, CEUR-WS.org (2025)
[34] Guo, H., Cheng, S., Jin, X., Zhang, Z., Zhang, K., Tao, G., Shen, G., Zhang, X.: Biscope: Ai-generated text detection by checking memorization of preceding tokens. Advances in Neural Information Processing Systems **37**, 104065–104090 (2024)
[35] Halvani, O., Winter, C., Graner, L.: On the usefulness of compression models for authorship verification. In: Proceedings of the 12th International Conference on Availability, Reliability and Security, vol. Part F1305, ACM, New York, NY, USA (29 Aug 2017), https://doi.org/10.1145/3098954.3104050
[36] Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J., Goldstein, T.: Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. International Conference on Machine Learning **abs/2401.12070**, 17519–17537 (22 Jan 2024), https://doi.org/10.48550/arXiv.2401.12070
[37] Hosseinbeigi, S.B., Mehrani, A.: Team TMU at PAN 2025: An Ensemble of Fine-Tuned LaBSE and Siamese Neural Network for Multi-Author Writing Style Analysis. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[38] Huang, J., Cao, H., Lin, X., Han, Z.: Application and Analysis of Roberta-base Model Fine Tuning Based on Data Enhancement in AI Text Detection. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[39] Hutson, J.: Human-ai collaboration in writing: A multidimensional framework for creative and intellectual authorship. International Journal of Changes in Education (2025)
[40] Ignatenko, D., Zaitsev, K., Shkriaba, O.: ReText.Ai Team at PAN 2025: Applying a Multiple Classification Heads to a Transformer Model for Human-AI Collaborative Text Classification. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[41] Jieren, L., Mancheng, H., Biao, L., Zhongyuan, H.: Two-Stage Generative Plagiarism Detection: From TF-IDF/Jaccard Filtering to Transformer Classification. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[42] Jimeno-Gonzalez, M., Martínez-Cámara, E., Noelia Fernandez, P.G., na López, L.A.U.: Team SINAI-INTA at PAN 2025: Uncovering machine generated text with linguistic features. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[43] Kestemont, M., Tschuggnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the Author Identification Task at PAN 2018: Cross-domain Authorship Attribution and Style Change Detection. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2018)
[44] Kozlovskiy, V., Ploskin, A., Tantry, S., Matveeva, T., Savelyeva, S.: Can Small Models Outperform Large Ones in Text Detoxification? In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[45] Krishna, N., Sai Teja, L., Mishra, A.: Team Detox at PAN: Multilingual Text Detoxification using LLM. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[46] Księżniak, E., Węcel, K., Sawiński, M.: OpenFact at PAN 2025: Punctuation-Guided Pretraining for Sentence-Level Style Change Detection. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[47] Kumar, R., Trivedi, A., Varshney, O.: Voight-Kampff AI Detection Sensitivity : IIITS@CLEF'25. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[48] Lai-Lopez, N., Yuan, S., Wang, L., Zhang, L.: Lexicon-Guided Detoxification and Classifier-Gated Rewriting: A PAN 2025 Submission. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[49] Larionov, D., Seleznyov, M., Viskov, V., Panchenko, A., Eger, S.: xCOMET-lite: Bridging the gap between efficiency and quality in learned MT evaluation metrics. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 21934–21949, Association for Computational Linguistics, Miami, Florida, USA (Nov 2024), URL https://aclanthology.org/2024.emnlp-main.1223
[50] Larson, J.: Generative AI detection using simple Feature Selection and SVM. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[51] Li, B., Qi, H., Yan, K.: Team Bohan Li at PAN: DeBERTa-v3 with R-Drop regularization for Human-AI Collaborative Text Classification. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[52] Li, T.: Fine-Grained Human-AI Collaborative Text Classification Using DeBERTa. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[53] Liang, Z., Sun, K., Cao, H., Luo, J., Han, Z.: Research on Text Author Classification Based on ModernBERT and Gradient Loss Function. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[54] Lin, K., Liu, C., Ye, F., Han, Z.: SCL-DeBERTa: Multi-Author Writing Style Change Detection Enhanced by Supervised Contrastive Learning. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[55] Lin, X., Han, Z., Liu, C., Duan, X.: Style Change Detection in Multi-Author Writing: A Deep Learning Approach Based on DeBERTa. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[56] Liu, B., Yang, L., Qi, H.: Integrating Adversarial-Contrastive Learning and Large Language Model for Multi-Author Writing Style Analysis. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[57] Liu, J., Kong, L., Peng, Z., Chen, F.: Generative AI Authorship Verification Based on Contrastive-Enhanced Dual-Model Decision System. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[58] Liu, X., Yi, Y., Chen, Z., Xu, S., Ke, Z., Guo, X., Huang, Y., Zhang, W., Chen, J., Han, Y.: Jiaozipi at CLEF 2025:A Multilingual Text Detoxification Method Based on Large Language Model-Based Ensemble Learning. In: Working Notes

of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[59] Macko, D.: mdok of KInIT: Robustly Fine-tuned LLM for Binary and Multiclass AI-Generated Text Detection. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[60] Macko, D., Moro, R., Srba, I.: Increasing the robustness of the fine-tuned multilingual machine-generated text detectors. arXiv preprint arXiv:2503.15128 (2025)
[61] Marchitan, T., Creanga, C., Dinu, L.: Unibuc - NLP at "Voight-Kampff" Generative AI Detection PAN 2025. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[62] Meier, P., Boland, K., Kallmeyer, L., Dietze, S.: Team HHU - An Ensemble-Based Approach to Multi-Author Writing Style Analysis Combining Experts for Different Difficulty Levels. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[63] MistralAI: Mistral 7b instruct v0.3 Model Card. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3 (2024), accessed: 2025-02-14
[64] Mo, D., Zhang, H., Zhang, X., Kong, L.: Using GloVe for Fragment Feature Matching and Overlap Ratio Optimized Generated Plagiarism Detection Method. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[65] Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T.L., Bari, M.S., Shen, S., Yong, Z.X., Schoelkopf, H., Tang, X., Radev, D., Aji, A.F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., Raffel, C.: Crosslingual generalization through multitask finetuning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pp. 15991–16111, Association for Computational Linguistics (2023), https://doi.org/10.18653/V1/2023.ACL-LONG.891
[66] Ochab, J., Matias, M., Boba, T., Walkowiak, T.: StylOch at PAN: Gradient-boosted trees with frequency-based stylometric features. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[67] Ostrower, B., Doongare, P., Unnikrishnan, M.: Binoculars, BART, and Adversaries: Multi-Faceted AI Text Detection for PAN 2025. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[68] Peñas, A., Rodrigo, Á.: A Simple Measure to Assess Non-response. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 1415–1424 (2011), URL https://aclanthology.org/P11-1142.pdf
[69] Peng, J., Kaiyin, S., Kaichuan, L., Zhankeng, L., Zhongyuan, H.: A Multilingual Text Detoxification Method Based on Chain-of-Thoughts Prompting Approach. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[70] Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An evaluation framework for plagiarism detection. In: COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China, pp. 997–1005, Chinese Information Processing Society of China (2010), URL https://aclanthology.org/C10-2115/

[71] Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: PAN plagiarism corpus 2009 (PAN-PC-09) (version 1) (Sep 2009), https://doi.org/10.5281/zenodo.3250083
[72] Pudasaini, S., Miralles-Pechuán, L., Lillis, D., Salvador, M.L.: Enhancing AI Text Detection with Frozen Pretrained Encoders and Ensemble Learning. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[73] Rohra, H., Shah, N., Sonawane, S.: Team BatataVada at PAN: Sentence-Level Style Change Detection with RoBERTa for Multi-Author Writing Style Analysis. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[74] Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., Stein, B.: Overview of PAN'16—New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative (CLEF 16) (2016)
[75] Rykov, E., Zaytsev, K., Anisimov, I., Voronin, A.: Smurfcat at PAN 2024 textdetox: Alignment of multilingual transformers for text detoxification. In: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, vol. 3740, pp. 2866–2871, CEUR-WS.org (2024), URL https://ceur-ws.org/Vol-3740/paper-276.pdf
[76] Schmidt, G., Römisch, J., Halchynska, M., Gorovaia, S., Yamshchikov, I.: better_call_claude: Sequential Style Shift Model for Fine-Grained Multi-Author Style Change Detection. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[77] Sculley, D., Brodley, C.E.: Compression and machine learning: A new perspective on feature space vectors. In: Data Compression Conference (DCC'06), pp. 332–341, IEEE (2006), ISBN 9780769525457, ISSN 1068-0314,2375-0359, https://doi.org/10.1109/dcc.2006.13
[78] Seeliger, M., Styll, P., Staudinger, M., Hanbury, A.: Human or Not? Light-Weight and Interpretable Detection of AI-Generated Text. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[79] Shah, P., Shah, V., Kale, S.: Multilingual Text Detoxification via Prompted MT0-XL and Lexical Filtering. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[80] Stamatatos, E., Potthast, M., Pardo, F.M.R., Rosso, P., Stein, B.: Overview of the PAN/CLEF 2015 evaluation lab. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings, Lecture Notes in Computer Science, vol. 9283, pp. 518–538, Springer (2015), https://doi.org/10.1007/978-3-319-24027-5_49
[81] Su, Z., Han, Y., Jia, Y., Kong, L.: Hierarchical Generative Plagiarism Detection Method. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)
[82] Sun, Q., Ma, L., Yang, W., Xian, T., Xie, M., Wu, W., Zhang, Z., Zheng, M.: DeBERTa-FPN: Fusion Feature Pyramid Network for Human-AI Collaborative Text Classification. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[83] Sun, Y., Afanaseva, S., Stowe, K., Patil, K.: Bi-directional Cross-entropy loss and Stylometric Feature combined Classifier. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[84] Tang, J., Hu, Q., Han, Z.: Efficient Plagiarism Detection via Sentence Embeddings and FAISS-based Retrieval. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[85] Teja, L.S., Yadagiri, A., Pakray, P.: Team CNLP-NITS-PP at PAN: Advancing Generative AI Detection: Mixture of Experts with Transformer Models. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[86] Titze, S., Halvani, O.: LOG-AID: Logit-Based Statistical Features for AI Text Detection. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[87] Totok, A., Ermolaev, A., Izyumova, A., Finogeev, E.: The Evolution of Methods for Text Detoxification: The Role of Language in Method Selection. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[88] Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the author identification task at PAN 2017: style breach detection and author clustering. In: CLEF 2017 Labs and Workshops, Notebook Papers (2017)

[89] Valdez-Valenzuela, A., Gómez-Adorno, H.: AI-Generated Text Detection using ISGraphs and Graph Neural Networks. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[90] Völpel, F., Halvani, O.: Adept: AI-Generated Text Detection Based on Phrasal Category N-Grams. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[91] Voronin, A., Moskovsky, D., Sushko, N.: PAN 2025 Textdetox: Exploring a Sage-T5-like approach for text detoxification. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[92] Voznyuk, A., Gritsai, G., Grabovoy, A.: Team Advacheck at PAN: Multitasking Does All the Magic. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[93] Wang, H., Juola, P., Riddell, A.: Reproduction and replication of an adversarial stylometry experiment. arXiv [cs.CL] (15 Aug 2022), URL http://arxiv.org/abs/2208.07395

[94] Wang, M., Zhang, N., Xu, Z., Xi, Z., Deng, S., Yao, Y., Zhang, Q., Yang, L., Wang, J., Chen, H.: Detoxifying large language models via knowledge editing. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3093–3118, Association for Computational Linguistics, Bangkok, Thailand (Aug 2024), https://doi.org/10.18653/v1/2024.acl-long.171

[95] Wu, W., Yang, W., Zhang, Z., Xie, M., Zheng, M., Xian, T., Sun, Q.: Bert_T for Human-AI Collaborative Text Classification. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[96] Xian, T., Zhong, Y., Liu, F., Xie, M., Sun, Q., Zheng, M., Wu, W., Zhang, Z.: DBG: Human-AI Collaborative Text Classification with DeBERTa-enhanced Contextual and Geometric Attention. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[97] Xianbing, D., Zhongyuan, H., Jiangao, P., Kaiyin, S.: Multilingual Text Detoxification System Based on Parallel Architecture: An Intelligent Approach Integrating Local Models and Large Language Models. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[98] Yang, J., Yan, K.: Genre-Aware Contrastive Learning for AI Text Detection: A RoBERTa-Based Approach. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[99] Zaidi, S., Ahmed, H., Akbar, S., Shakeel, Z., Alvi, F., Samad, A.: Team Nexus Interrogators at PAN: Voight-Kampff Generative AI Detection. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[100] Zangerle, E., Mayerl, M., , Potthast, M., Stein, B.: Overview of the Style Change Detection Task at PAN 2021. In: CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org (2021)

[101] Zangerle, E., Mayerl, M., , Potthast, M., Stein, B.: Overview of the Style Change Detection Task at PAN 2022. In: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org (2022)

[102] Zangerle, E., Mayerl, M., , Potthast, M., Stein, B.: Overview of the Style Change Detection Task at PAN 2023. In: CLEF 2023 Labs and Workshops, Notebook Papers, CEUR-WS.org (2023)

[103] Zangerle, E., Mayerl, M., Potthast, M., Stein, B.: Overview of the Multi-Author Writing Style Analysis Task at PAN 2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2024)

[104] Zangerle, E., Mayerl, M., Potthast, M., Stein, B.: Overview of the Multi-Author Writing Style Analysis Task at PAN 2025. In: CLEF 2025 Working Notes, CEUR-WS.org (2025)

[105] Zangerle, E., Mayerl, M., Specht, G., Potthast, M., Stein, B.: Overview of the Style Change Detection Task at PAN 2020. In: CLEF 2020 Labs and Workshops, Notebook Papers (2020)

[106] Zangerle, E., Tschuggnall, M., Specht, G., Stein, B., Potthast, M.: Overview of the Style Change Detection Task at PAN 2019. In: CLEF 2019 Labs and Workshops, Notebook Papers (2019)

[107] Zhang, Z., Yang, W., Wu, W., Xie, M., Zheng, M., Sun, Q., Xian, T.: DBA: A Hybrid Neural Network Model for Generative Human-AI Collaborative Text Classification. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)

[108] Zheng, M., Zhong, Y., Liu, F., Xian, T., Xie, M., Wu, W., Zhang, Z., Sun, Q.: StarBERT: A Hybrid Neural Network Model for Human-AI Collaborative Text Classification. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (Sep 2025)