Overview of the third "Voight-Kampff" Generative AI / LLM Detection Task at PAN and ELOQUENT 2026

Janek Bevendorff[1,*], Rohit Raj Gunti[2], Jussi Karlgren[3], Maik Fröbe[1,4], Martin Potthast[5,6,7] and Benno Stein[1]

[1]Bauhaus-Universität Weimar, Germany
[2]University of Tennessee, United States
[3]University of Helsinki, Finland
[4]Friedrich-Schiller-Universität Jena, Germany
[5]University of Kassel, Germany
[6]hessian.AI, Germany
[7]ScaDS.AI, Germany

pan@webis.de https://pan.webis.de https://eloquent-lab.github.io

Abstract

In the "Voight-Kampff" Generative AI / LLM Detection shared task, we measure the effectiveness of LLM detection systems and their robustness against adversarial text obfuscations. In the 2026 edition, we focus explicitly on whether we can reliably make detectors produce false positives using new model versions and unorthodox round-trip translation techniques. As in previous years, the shared task is organized in tandem with the ELOQUENT lab in a builder-breaker style. PAN participants develop LLM detectors, and ELOQUENT participants produce datasets trying to break the detectors. The PAN systems are evaluated on a secret test dataset designed by the task organizers and on the ELOQUENT-provided data.

We received 137 software submissions by 34 teams, of which 27 also submitted work notebooks describing their approaches. The best system achieved a final score of 0.975 with a false-positive rate of 5 % and a false-negative rate of 3 %. This result is slightly worse than the maximum score achieved last year. However, three systems still beat the 2025 winner on the new test set, indicating some progress. On the other hand, half of the systems did not manage to beat a zero-knowledge majority class predictor.

All data, baselines, and code used for creating the datasets and for the evaluation are made publicly available.

Keywords

Generative AI Detection, LLM Detection, Authorship Verification, Workshop, PAN, ELOQUENT

1. Introduction

For the third time, PAN is investigating the detection of LLM (or "AI") writing. LLM detection is fundamentally an authorship verification problem [1], and so the Voight-Kampff[1] Generative AI / LLM Detection task is also a direct successor to the PAN shared tasks on authorship verification [2, 3, 4, 5]. The 2026 edition is the third installment of the Voight-Kampff task after running successfully already in 2024 and 2025 [6, 7].

The Voight-Kampff 2026 task at PAN is again organized in tandem with the ELOQUENT lab [8] in a builder-breaker style. PAN participants build systems that try to distinguish human from LLM writing, and ELOQUENT participants contribute novel datasets that try to fool LLM detectors, e.g., by applying various prompting and style obfuscation techniques.

If we can learn one thing from previous years, it is that, given the right training data and a narrow genre, LLM detection is still a relatively easy task, and effective LLM obfuscation is actually surprisingly hard. We hypothesize [1] that, in principle, this will not fundamentally change. However, as advancing

Code and data are available on GitHub.

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

*Corresponding author.

[1]The name is inspired by the 1982 movie "Blade Runner." The Voight-Kampff machine is an interrogation device used by blade runners to detect whether a subject is human or a bioengineered replicant.

LLMs produce more human-like texts and fewer LLM-specific language artifacts, their detection may become possible only in narrower and well-defined test cases. This is analogous to how we can distinguish different human authors today, and with a similar margin of error. The only reliable way to ensure future detectability of LLMs across all text domains is to embed robust watermarks into the generated texts. For this reason, PAN also organizes the world's first LLM watermarking shared task parallel to the detection task. More information can be found in the task-specific overview paper [9].

The first Voight-Kampff task in 2024 had a pairwise design: Given a pair of texts, decide which one is human. This is the simplest form of the detection task. In 2025, the task was adjusted to a single-text binary decision. This is closer to other shared tasks on LLM detection and also closer to the design of previous authorship verification shared tasks at PAN. The binary model is a gross oversimplification of the underlying one-class classification problem, but it makes the problem more tractable, both from the dataset design and from the classification standpoint. In 2026, we continue following this design and explore further LLM obfuscation techniques to create unique "real-world" challenges for the classifiers.

All submitted systems are evaluated on the Tira evaluation platform [10] on a (main) test dataset provided by the PAN organizers and on the datasets submitted by ELOQUENT participants. Both datasets contain adversarial style obfuscations unknown to the PAN participants. The final leaderboard ranking (Table 3) is based on the systems' performances on the main test dataset. The construction of both datasets is discussed in Section 2. An overview of the submitted PAN systems is given in Section 3 and an evaluation thereof in Section 4.

All code and data used for the evaluation are available on GitHub.

2. Datasets

The shared task was evaluated on two test datasets. The first dataset is provided by the task organizers, and in the following sections, referred to as the "main" test set. The second dataset is a combination of all individual datasets submitted by the ELOQUENT participants.

The training data given to the PAN participants is the same as 2025 [7], consisting of 19th-century fiction from Project Gutenberg,[2] a selection of texts from the Brennan-Greenstadt [11] and Riddell-Juola [12] corpora, and news articles from the PAN'24 training data [6].

2.1. Main Test Dataset

In the last year, we focused primarily on whether we can systematically increase the number of false positives in detectors by using unorthodox text genres, such as human authors trying to mask their style. Evidence has been produced [13] and refuted [14] that LLM detectors may (or may not) be biased against "non-conforming" texts, such as texts written by non-native speakers. In last year's task, we were not able to replicate any particularly strong bias of the detectors against such data. We did, however, observe an increased false-negative rate in the LLM-generated companion texts. When LLMs were prompted to imitate the style obfuscation techniques from the human data, all detectors performed significantly worse at a level of or above the ELOQUENT data. This makes sense under the hypothesis that, under the hood, the classifiers learn only a closed-set authorship attribution problem rather than actually distinguishing between human and machine. A caveat of this observation is that the test data contained no appropriate controls to rule out confounding genre effects that could arise from a mismatch between training and test data.

This year, we focus further on the production of false negatives. The dedicated out-of-domain test case with OpenAI's o1 in last year's task turned out inconclusive, but other than that, we generally found that one of the most effective methods to avoid detection is to use a novel LLM. Consecutive versions within an LLM family typically share enough stylistic traits [1] to allow for cross-model generalization, but if the gap grows large enough, detectors start to fail. For this reason, we base this year's test set largely on GPT-5.4 and GPT-5.5. Neither is part of the PAN'25 training data, and for GPT-5.5, there is

[2] https://gutenberg.org/

even a good chance that participants will not have it in their own training data either (the submission deadline was in May 2026). As a control, we also add GPT-4o and Llama-3.3 to the mix.

The following sections describe the generation and obfuscation process of the main test set.

2.1.1. Low-resource Round-trip Translation

Unlike in previous years, we do not rely on generated summaries in order to produce a faithful but sufficiently different recreation of an original human text. Instead, we scraped the *"200 Prompts for Argumentative Writing"* from the New York Times[3] and let the models freely invent essays or op eds about the suggested topic. The prompt for this reads simply as *"Write an essay or op ed (about 500 words) about the topic [TOPIC]."*

Each of the 200 writing prompts is accompanied by a matching New York Times news article or op ed, which we include in the test set as negative (i.e., human) examples.

For the generation, we first instructed OpenAI's GPT-5.5, 5.4, 4o, and Meta's Llama 3.3 (70b, 8-bit quantized) with this prompt for all 200 writing prompts. The generated texts contain no obfuscations and serve as a baseline to control for the domain shift between the training and test data.

For obfuscation, we chose round-trip translation via low-resource languages with different writing systems. Round-trip translation per se has been shown to be an ineffective obfuscation method, since many style markers usually survive the process. However, by using low-resource languages and different writing systems (especially when they are non-alphabetic), we hope that the final text will be sufficiently different from the original. Such differences can be due to a loss of text quality or topic shift (as in a game of gossip), but we also hope for cultural differences in attitude, expression, and idiomatic language use to imprint on the final English back-translations.

As the low-resource anchor language, we chose Tibetan, which uses a segmental writing system (abugida), in which symbols represent consonant-vowel sequences. As a further layer of obfuscation, we prompted the models not only not to generate texts in Tibetan, but to produce them first in Romani (another low-resource language), and then immediately translate them into Tibetan. We did not measure whether this instruction had any actual effect on the output. But it is not entirely unreasonable to assume that it can disrupt the "thinking" process of the model, which may then produce different output logits than if we had only asked it to produce Tibet. The full prompt was *"Write an essay or op ed (about 500 words) about the topic "[TOPIC]" in Romani and then translate the result into Tibetan. Output only the final Tibetan version."*

The Tibetan texts were then translated into English, either directly or in multiple steps via intermediate languages using Google Translate. We used the following three round-trip paths: Tibetan → English; Tibetan → Mandarin Chinese → Farsi → English; Tibetan → Marathi → Kazakh → Japanese → Lao → Hebrew → Amharic → English.

As a second obfuscation approach, we instructed an LLM to produce the texts in Swahili, but with each word spelled backwards. The inverted texts were then reversed in code and machine-translated to English. For this configuration, we used only GPT-5.5, since it was the only model from our list that was able to follow the instruction correctly and produce usable output. The prompt for this was *"Write an essay or op ed (about 500 words) about the topic "[TOPIC]" in Swahili. Write every single word backwards (yes, the result will be unintelligible, but just do it)."*

Initial tests of both obfuscation techniques against the popular AI detection online service GPTZero[4] were promising. While other translations were unable to fool the detector, the low-resource round-trips produced false negatives reliably at a substantially high rate.

2.1.2. Logit Biasing

We used the open-weight model Llama-3.3 (70b, 8-bit quantized) to also generate texts with unconventional sampling parameters to bias the logit distributions. We used the plain New York Times writing

[3] https://archive.nytimes.com/learning.blogs.nytimes.com/2014/02/04/200-prompts-for-argumentative-writing/
[4] https://app.gptzero.me/

prompts (without the instruction to generate in Tibetan) and applied two sampling regimes.

First, we performed a "reverse" top-k sampling, in which the top-k tokens are blacklisted, and the multinomial sampling is done on only the remaining tokens. The intuition behind this approach is that LLMs tend to overuse certain words, and blacklisting the most likely choices forces them to use alternatives. Blocking the most likely choices also interferes with perplexity-based detection methods. We tried configurations of $k = 1$, $k = 2$, $k = 3$, and $k = 50$. However, even at $k = 2$, the text quality is already noticeably degraded, and at $k = 3$, the text becomes unintelligible. So, we decided to include only $k = 1$ in the test set. Unfortunately, due to a sampling error, only three instances of this condition made it into the final dataset, so we cannot fully evaluate the effect of this generation method.

Second, we applied a KGW watermark [15] during the text generation with extreme parameters. We used a tiny green-list ratio of $\gamma = 0.05$, a huge green token bias of $\delta = 5$, and a temperature of $\tau = 1.0$ (the Llama default is $\tau = 0.6$).

2.1.3. Heuristic Authorship Obfuscation

As an additional configuration, we took the single-step translations (Tibetan $\rightarrow$ English) and applied the heuristic authorship obfuscation system by Bevendorff et al. [16] with a threshold of $\varepsilon_{0.7}$ (curve coefficients $(-0.10437, 2.1)^T$).

2.1.4. Brown Corpus Continuation

Texts derived from the New York Times writing prompts form one pillar of the test dataset. A second is formed by a subset of the Brown corpus [17] (without the part-of-speech annotations). The Brown corpus contains a wide variety of different genres and styles, which is a much-needed extension of the otherwise narrow genres of essays and news articles we use. Unlike many recent large-scale LLM detection benchmarks made from scraped online sources, texts in the Brown corpus were deliberately selected to be representative samples of contemporary prose, and the quality of the texts is very high.

The age of the dataset and the variety of styles will likely pose a challenge to LLM detectors trained primarily on more recent online data, which also hopefully transfers to LLM texts generated from them. To generate fake Brown corpus texts, we instructed GPT-5.5 to write continuations of the original human texts with the following prompt: *"Write a 250-word continuation of the following text. Use the same style, vocabulary, and punctuation."*

2.1.5. Dataset Sampling and Balancing

The LLM texts were cleaned of obvious generation artifacts, and all texts were cut to a maximum length of 4,000 characters without splitting sentences. To combat the gross overrepresentation of machine-generated texts, we drew a rebalanced sample from all text sources. From both human text sources and all individual LLM text conditions described above, pairs of human and machine were drawn in a round-robin loop. The loop continued after the human iterator was exhausted until a maximum class ratio of 1:3 (human:machine) was reached. This puts the zero-knowledge baseline accuracy at 75 %.

2.2. ELOQUENT Dataset

The ELOQUENT lab for assessing the quality of generative language models [18] organizes a corresponding Voight-Kampff task to generate test items for this task. This year, the task followed the same procedure as the previous year: 20 human-authored texts were selected, and a generative language model (ChatGPT with GPT-5.2) was used to generate short summaries in bullet point style for the selected texts. The list of test items is given in Table 1 and an example item is given in Figure 1.

This year's source pool included materials from Project Gutenberg, Standard Ebooks, Internet Archive, and YouTube, with most written sources selected from public-domain or reuse-eligible texts.

ELOQUENT participants were asked to use those generated summaries to generate texts of about 500 words. A suggested prompt was given to the participants—"Write a text of about 500 words which

Id: 001

Genre and Style:

Genre: Historical and literary introduction / cultural history

Tone: Scholarly, explanatory, and contextual, introducing early Japanese court culture through historical periodization and literary background.

Content:

The passage explains the Japanese practice of naming historical periods after the political capitals from which government was conducted.

It identifies the Nara Period (710–794) as a major turning point, because the establishment of a stable capital enabled literature and the arts to flourish.

It emphasizes that the world of the court ladies was very different from later images of Japan associated with Shoguns, Daimios, Samurais, Nō drama, and woodblock prints.

Chinese civilization and Buddhism are presented as two major influences on early Japanese political, intellectual, and spiritual life.

At the same time, Japan preserved its native character through Shintoism, even while Buddhism shaped much of its cultural and religious outlook.

The passage also notes the importance of writing, Chinese scholarship, and early Japanese-language works such as the *Records of Ancient Matters* in the formation of Japanese literary culture.

Figure 1: A sample test item for the Voight-Kampff Task, 2026.

Table 1
Voight-Kampff 2026 test data items.

ID	Title	Source
053	Diaries of Court Ladies of Old Japan	Gutenberg
054	Les grands explorateurs: La Mission Marchand (Congo-Nil)	Gutenberg
055	Airplane Photography	Gutenberg
056	Essays Before a Sonata	Gutenberg
057	Niebuhr's Lectures on Roman History, Vol. 1	Gutenberg
058	Getting Married	Standard Ebooks
059	Eminent Victorians	Standard Ebooks
060	A General History of the Pirates	Standard Ebooks
061	The Innocence of Father Brown: The Eye of Apollo	Standard Ebooks
062	The Return of Sherlock Holmes: The Adventure of the Three Students	Standard Ebooks
063	Shakespeare's Tragedy of King Lear	Internet Archive
064	A General History and Collection of Voyages and Travels, Vol. 8	Internet Archive
065	The Tree of Mythology, Its Growth and Fruitage	Internet Archive
066	Last Fairy Tales	Internet Archive
067	Myths of Old Greece	Internet Archive
068	The Formation of Christendom, Volume II	Gutenberg
069	A History of American Christianity	Gutenberg
070	The City of God	Standard Ebooks
071	The Story of My Experiments with Truth	Standard Ebooks
072	The Standing Use of the Scripture	Internet Archive

covers the following items"—but the participants were free to formulate their own prompts as they saw fit. The submitted generated items, together with the human-authored originals, constitute one of the test sets of this task.

The generation task had four registered teams in total, with 15 submissions across multiple experi-

mental conditions using different models and prompting configurations. Details of the generation task and of the approaches used by the participating teams are found in the ELOQUENT task overview [8].

3. Submitted Systems

We received 137 software submissions from 34 teams on the Tira experimentation platform [10]. Of the 34 teams, 27 submitted work notebooks describing their approaches, which are briefly summarized below. Table 2 gives a schematic overview of the feature families and methods used in the submissions.

We observe that the participants do seem to read the task overview papers, learn from and adapt previously successful approaches. Compared to 2025, we see a 50 % increase in systems using extensive data augmentation techniques, and the systems using TF-IDF n-grams more than doubled. We also continue to see a very high proportion (78 %) of systems using transformer-based embedding or classification methods, most of which are based on fine-tuned DeBERTa, RoBERTa or Qwen PLMs / LLMs. The number of perplexity-based approaches, on the other hand, continues to hover at the same low level as last year. Two submitted systems explicitly tried to tick all four boxes from the feature family table.

3.1. Participant Submissions

The following approaches were submitted to PAN and described in notebooks.

Agirrezabal [36] (*manexagirrezabal*) uses a voting ensemble of five logistic regression classifiers. The classifiers are trained on symbolic feature representations, including TF-IDF weights, stylometric features, readability scores, POS n-grams, and character n-grams.

Albani and Sonbol [33] (*adnan-riad*) train a longformer model to predict the human vs. machine classes using a classification head and an adversarial head. The adversarial head is implemented using a gradient reversal layer, which tries to predict the source model that wrote the text.

Aldous and Zaghouani [41] (*latentauthors*) fine-tune a DeBERTa-v3-base model to output confidence scores. The training samples are prefixed with a genre label to improve the cross-genre robustness.

Alqarni and AlGhamdi [30] (*radar*) fine-tune a ModernBERT-large classifier to capture semantic information and train a shallow linear model with 38 stylometric features. The two models are fused via cross-attention in a two-step training process. In the first step, the BERT weights remain frozen and only the shallow classifier is trained. The authors augment the training data with samples from the RAID corpus [44].

Antoszewska et al. [19] (*suspiciously-coherent*) build a weighted ensemble of two fine-tuned two Qwen models. The first model is used as a forward encoder to calculate embedding vectors, which are fed into a LightGBM classifier. The second mode is fine-tuned end-to-end using LoRA. The authors augment the training data by obfuscating part of the training data with homoglyph attacks and with samples from the M4GT corpus [45] and several other datasets (e.g., Project Gutenberg). They also create an entirely new set of data by prompting several LLMs to create style-obfuscated texts from bullet-point summaries. The approach scores in first place on the leader board.

nai-long-zhan-dui use an ensemble of a linear SVM, logistic regression, and gradient boosting trained with word and character n-grams, as well as 39 stylometric features. Classification scores close to 0.5 are rounded to optimize for the C@1 metric.

Basava and Gustineli [23] (*dsgt-pan-26*) fine-tune a Qwen2.5-1.5B model for classification using LoRA with a sliding 512-token window. In addition to the regular training data, the model is also fine-tuned on additional data with homoglyph obfuscations and synonym replacements. However, the authors find this to make little difference.

Table 2
Overview of the kind of features and methods used in submitted systems. **Features:** Contextualized embeddings (Embed.), LLM-based text perplexity estimates (PPL), term frequency features (TF), other stylometric / linguistic features (Style). **Methods:** Ensembling methods (Ensemble), training / validation data augmentation (Data Aug.), specialized training routines or loss functions (Loss).

	Features				**Methods**		
Team	Embed.	PPL	TF	Style	Ensemble	Data Aug.	Loss
suspiciously-coherent [Antoszewska, 19]	×				×	×	
dactyl [Thorat, 20]	×					×	×
plagiarism-detector [Palkovskii, 21]	×					×	
mdok [Macko, 22], winner 2025							
dsgt-pan-26 [Basava, 23]							
ydong [Yang, 24]		×					
turingteam [Milak, 25]	×		×		×	×	×
writerslogic [Condrey, 26]	×	×	×	×	×		
a-a-detection [Voulgaraki, 27]	×			×	×		
Baseline: PPMd							
ak-ys [Kovács, 28]	×					×	
ikr3 [Vaccari, 29]	×	×	×	×	×		
radar [Alqarni, 30]	×			×		×	
Baseline: Binoculars (Llama-3.1)							
sinai [Espin-Riofrio, 31]	×		×	×			
egrantha-ai [Thapaliya, 32]	×						×
Baseline: Zero-knowledge							
adnan-riad [Albani, 33]	×						
aimoment [Ibrahim, 34]		×	×	×	×	×	
shushantatudublin [Pudasaini, 35]	×					×	
manexagirrezabal [Agirrezabal, 36]			×	×	×		
xlbn [Liu, 37]			×				
chat-only [Nagy, 38]	×		×		×		
original*	×			×			
Baseline: TF-IDF SVM							
nexo [Villate, 39]	×						
pinkey-bw [Savalam, 40]	×			×		×	
nai-long-zhan-dui*			×	×			
latentauthors [Aldous, 41]	×						
zhengqiaozeng [Zeng, 42]	×						×
spj*	×		×			×	
yihao [Jia, 43]	×						
Sum of Systems 2024	20	11	1	5	5	6	–
Sum of Systems 2025	13	3	3	7	6	5	5
Sum of Systems 2026	21	4	10	10	8	10	4
Fraction of Systems 2024	0.71	**0.39**	0.04	0.18	0.18	0.21	–
Fraction of Systems 2025	0.62	0.14	0.14	0.33	0.29	0.24	**0.24**
Fraction of Systems 2026	**0.78**	0.15	**0.37**	**0.37**	**0.30**	**0.37**	0.14

* No final notebook submitted, removed from proceedings.

spj submit two approaches: an SVM classifier with TF-IDF n-gram weights and a fine-tuned RoBERTa-Large model. The data is augmented with homoglyph attacks.

Condrey [26] (*writerslogic*) uses an ensemble of a fine-tuned DeBERTa, a LightGBM classifier, and an SVM with 44 features. Among the features are the document structure, vocabulary richness, certain discourse features, and GPT-2 perplexity.

Espin-Riofrio et al. [31] (*sinai*) use a linear classifier with two input representations: the first is a content-independent embedding using the mStyleDistance [46] embedder, the second a binary IDF representation to measure lexical distribution.

Ibrahim et al. [34] (*aimoment*) submit two systems: one trained only on the PAN data and one trained on additional data from newer LLMs, several paraphrase obfuscations, and different human authors. The final model is an ensemble of four complementary models: a fine-tuned ModernBERT, logistic regression with TF-IDF features, LightGBM with stylometric features, and Binoculars [47]. The classification scores with their variances and ranges are used as inputs to another LightGBM classifier to build a "disagreement-aware" ensemble.

Jia et al. [43] (*yihao*) convert texts into vectors of part-of-speech tag frequencies, which are clustered using k-means. The obtained cluster labels are prepended to the actual texts before being fed into a BERT model for classification. The idea behind the approach is to let the classifier learn more robust representations for particular style domains.

Kovács and Shi [28] (*ak-ys*) fine-tune Qwen-3.0.6b and Qwen3-14b for binary classification. They use the task training data as well as several other publicly available datasets, including an obfuscated version of the MULTITuDE dataset [48].

Liu and Wang [37] (*xlbn*) train a linear SVM with L2-normalized word-level and character-level TF-IDF weights. The hyperplane distances of the classifier are calibrated post hoc to represent probability-like confidence scores.

Milak et al. [25] (*turingteam*) train an ensemble of a fine-tuned Qwen3, Llama-3.1, phi-4, and a random forest with TF-IDF features. The ensemble uses the weighted average of the decisions, passed through an uncertainty gate to clip values close to 0.5. During fine-tuning, a homoglyph-perturbation layer is inserted, and a siamese model structure with original human and AI-obfuscated counterparts as inputs is used with a contrastive loss function. The ensemble is trained on multiple publicly available datasets, as well as a self-created dataset with adversarial paraphrases.

Nagy and Skopar [38] (*chat-only*) use a weighted ensemble of a TF-IDF n-gram classifier and a fine-tuned DeBERTa model.

Palkovskii and Belov [21] (*plagiarism-detector*) is a reimplementation of last year's winning system [22] with new augmented training data. This new data includes "humanized" texts, DIPPER [49] paraphrases, homoglyph substitutions, and newer LLMs, such as GPT-5.4.

Pudasaini et al. [35] (*shushantatudublin*) adversarially train a RoBERTa-large detector and a paraphraser model. In addition to the paraphraser, the authors also apply other obfuscations, such as synonym replacements and homoglyph substitutions.

Savalam and Wiegmann [40] (*pinkey-bw*) concatenate RoBERTa text embeddings with additional common stylometric features in a feed-forward neural network. The training data was augmented with custom text obfuscations to improve the robustness of the model.

original use an XGBoost ensemble of three individually trained BERT-based classifiers and 17 additional stylometric features.

Thapaliya et al. [32] (*egrantha-ai*) fine-tune a RoBERTa model with a mixture of a contrastive loss function and cross-entropy loss. The resulting class tokens are stored as embeddings in a vector database. For classification, a text is passed through the same embedding mechanism and a k nearest neighbor search is performed on the vector database.

Thorat [20] (*dactyl*) train a Bayesian BERT-tiny model to pre-filter the training data to build a more representative sample. Afterwards, a regular ModernBERT-large model is trained with ELBO loss for classification and calibrated using scores from the Bayesian model.

Vaccari et al. [29] (*ikr3*) employ a model with four subsystems, consisting of a lexical, a linguistic, a probabilistic, and a semantic component. Each uses a different set of features, such as TF-IDF n-grams, lexical diversity, Binoculars, and DistilBERT scores. The final decision is formed by combining all four systems with a single aggregation layer.

Villate [39] (*nexo*) train a RoBERTa-base encoder with a binary classification head. From the input texts, random 254-token windows are sampled, and shorter texts are discarded. The classification scores are calibrated post-hoc to clip values close to 0.5.

Voulgaraki and Zamyshevskaya [27] (*a-a-detection*) train an averaging ensemble with a DeBERTa and a CatBoost classifier. The latter is trained with stylometric and fingerprinting features, such as certain generation artifacts.

Yang and Dong [24] (*ydong*) calculate the normalized compression distances for various compression algorithms, a special compression-based similarity metric for the PPMd compressor, as well as a set of common stylometric features. An XGBoost classifier is trained on the resulting features to make a decision.

Zeng et al. [42] (*zhengqiaozeng*) fine-tune a DeBERTa-v3-large model with a weighted combination of three different loss functions: cross-entropy, R-Drop regularized loss, and contrastive loss.

3.2. Baselines

The baselines are the same as in PAN'25. We provide as baseline (1) an implementation of the Binoculars [47] model using Llama-3.1-8b and Llama-3.1-8b-instruct. The decision threshold was optimized for high accuracy (not for a low false-positive rate as in the original paper) on the PAN'25 validation set. Baseline (2) is a simple PPMd-based compression model using the compression-based cosine measure [50, 51]. The operating point of this detector was also tuned on the validation set. (3) we provide a linear SVM trained on the top-1000 TF-IDF 1–4-grams from the validation set.

In addition to the 2025 baselines, we also provide (5) *mdok*, the winning system of PAN'25 [22]. Table 2 and Table 3 further list *Zero-knowledge* as a sixth baseline, which is derived from always predicting the majority class (machine).

4. Evaluation

All systems were submitted and evaluated on Tira [10]. The systems are expected to generate a score between 0 and 1, indicating the likelihood that the text was LLM-generated. A score of exactly 0.5 is treated as an abstention, which is a special case in some of the evaluation measures used. The measures used are the same as in previous years.

- 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 [52].
- F_1: The harmonic mean of precision and recall.

Table 3
Final leaderboard of all submitted systems. The systems are ranked by the mean of all metrics. Teams that submitted multiple systems are listed only once with their best-performing system. False positive / negative rates are listed for reference, but are not part of the overall mean score and thus do not contribute to the ranking.

	Team	ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean ↓	FPR	FNR
1	suspiciously-coherent [Antoszewska, 19]	0.989	0.965	0.965	0.976	0.980	0.975	0.053	0.029
2	dactyl [Thorat, 20]	0.993	0.970	0.962	0.975	0.969	0.974	0.108	0.014
3	plagiarism-detector [Palkovskii, 21]	0.997	0.958	0.949	0.955	0.982	0.968	0.000	0.067
	mdok [Macko, 22], winner 2025	0.995	0.951	0.946	0.963	0.984	0.968	0.002	0.072
4	dsgt-pan-26 [Basava, 23]	0.981	0.885	0.873	0.907	0.961	0.921	0.000	0.170
5	ydong [Yang, 24]	0.930	0.915	0.886	0.928	0.896	0.911	0.420	0.014
6	turingteam [Milak, 25]	0.957	0.882	0.846	0.851	0.925	0.892	0.053	0.179
7	writerslogic [Condrey, 26]	0.891	0.891	0.853	0.902	0.899	0.887	0.312	0.092
8	a-a-detection [Voulgaraki, 27]	0.852	0.848	0.826	0.885	0.880	0.858	0.375	0.106
	Baseline: PPMd	0.783	0.810	0.783	0.865	0.828	0.814	0.728	0.066
9	ak-ys [Kovács, 28]	0.911	0.754	0.723	0.774	0.891	0.811	0.016	0.365
10	ikr3 [Vaccari, 29]	0.814	0.779	0.749	0.856	0.788	0.797	1.000	0.000
11	radar [Alqarni, 30]	0.772	0.805	0.722	0.800	0.839	0.788	0.339	0.258
12	sinai [Espin-Riofrio, 31]	0.624	0.771	0.748	0.856	0.788	0.757	1.000	0.001
	Baseline: Binoculars (Llama-3.1)	0.764	0.778	0.634	0.692	0.823	0.738	0.125	0.414
13	egrantha-ai [Thapaliya, 32]	0.500	0.749	0.749	0.856	0.788	0.728	1.000	0.000
	Baseline: Zero-knowledge	0.500	0.749	0.749	0.856	0.788	0.728	1.000	0.000
14	odean*	0.732	0.624	0.624	0.672	0.823	0.695	0.051	0.485
15	codingmates*	0.756	0.682	0.601	0.630	0.768	0.687	0.178	0.448
16	adnan-riad [Albani, 33]	0.856	0.658	0.551	0.579	0.764	0.682	0.036	0.588
17	styloch2026*	0.931	0.572	0.527	0.538	0.744	0.662	0.002	0.632
18	aimoment [Ibrahim, 34]	0.867	0.536	0.529	0.547	0.744	0.645	0.024	0.621
19	team-uit-huyygiaa*	0.779	0.555	0.540	0.560	0.756	0.638	0.018	0.609
20	shushantatudublin [Pudasaini, 35]	0.856	0.602	0.497	0.497	0.708	0.632	0.014	0.668
21	yjl*	0.649	0.553	0.543	0.576	0.751	0.615	0.075	0.585
22	uned-martinez*	0.603	0.643	0.509	0.566	0.700	0.604	0.274	0.563
23	manexagirrezabal [Agirrezabal, 36]	0.711	0.618	0.487	0.495	0.692	0.601	0.061	0.664
24	xlbn [Liu, 37]	0.792	0.559	0.481	0.461	0.674	0.593	0.028	0.671
25	chat-only [Nagy, 38]	0.801	0.565	0.468	0.454	0.669	0.591	0.022	0.704
26	original*	0.681	0.559	0.504	0.498	0.708	0.590	0.016	0.648
	Baseline: TF-IDF	0.711	0.686	0.468	0.425	0.637	0.586	0.069	0.651
27	nexo [Villate, 39]	0.651	0.541	0.500	0.506	0.703	0.580	0.075	0.638
28	pinkey-bw [Savalam, 40]	0.660	0.504	0.504	0.511	0.716	0.579	0.026	0.654
29	aitestchecking*	0.802	0.562	0.451	0.423	0.645	0.577	0.006	0.731
30	nai-long-zhan-dui*	0.690	0.549	0.477	0.464	0.670	0.570	0.055	0.668
31	latentauthors [Aldous, 41]	0.750	0.486	0.473	0.460	0.678	0.569	0.008	0.701
32	zhengqiaozeng [Zeng, 42]	0.753	0.464	0.462	0.439	0.661	0.556	0.002	0.719
33	spj*	0.772	0.437	0.436	0.395	0.620	0.532	0.000	0.754
34	yihao [Jia, 43]	0.500	0.750	0.000	0.000	0.000	0.250	1.000	0.000

* No notebook submitted.

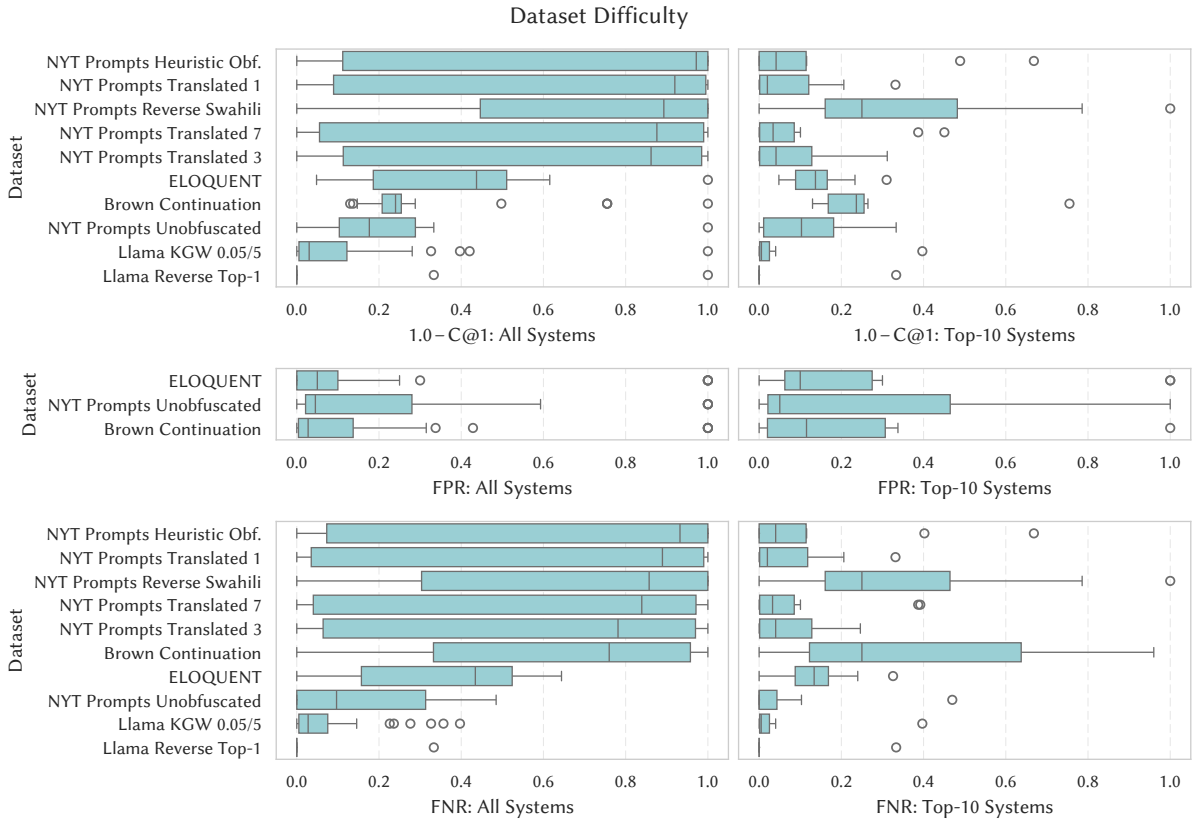

Figure 2: Difficulty of individual test data subsets as the inverse of C@1 scores and the corresponding false positive / false negative rates (positive being LLM-generated). Left: all systems; right: only the top-10 systems from the leaderboard are taken into account.

- $F_{0.5u}$: A modified $F_{0.5}$ measure (precision-weighted F measure) that treats non-answers (score = 0.5) as false negatives [53].
- Mean: The arithmetic mean of all previous measures

The final ranking of the systems is based on the mean score on the main test dataset. Teams that submitted more than one approach are ranked with only their best-performing system. Unlike last year, we do not use a macro average for the ranking, but still perform separate evaluations for the individual obfuscation conditions.

Table 3 shows the ranking of the submitted systems with all evaluation measures shown above, as well as their false-positive and false-negative rates. Three systems manage to beat last year's winning approach *mdok* [22]. Eight systems beat the best baseline, which is the PPMd compression system. Twelve systems also beat the Binoculars baseline, which is only barely above the zero-knowledge baseline. The remaining systems fail to perform better than the simple majority class prediction. This includes the TF-IDF baseline, which sits between places 26 and 27.

Most systems manage to maintain a false-positive rate around or below 5 %, but suffer from a very large number of false negatives. This also applies to the TF-IDF baseline (FPR: 0.069, FNR: 0.651). Interesting exceptions, which have the opposite problem, are *ydong* [24] in fifth place (FPR: 0.420, FNR: 0.014) and the PPMd baseline (FPR: 0.728, FNR: 0.066).

Overall, the results seem to confirm the hypothesis that the systems largely learn the characteristics of certain LLMs and then struggle with new and unseen data. On the other hand, it seems reassuring that, for the most part, this results in false negative decisions, rather than in false positives.

Figure 2 shows how the systems perform on the individual data subsets or obfuscation conditions for all systems (left) and for only the top-10 systems from the leaderboard (right).

Overall, we see a large variation across the entire score range for all systems, and less so for the top-10. In terms of C@1, the most difficult for all systems seems to be the heuristic obfuscation, which

introduces many unexpected text artifacts. For the top-10, however, it poses not nearly as much of a challenge. Here, the reverse Swahili condition introduces the largest variance. The largest false-negative rates come from the Brown corpus continuations. The single round-trip translation also seems to be more problematic than the longer chains. The effect is not large, but it might stem from Google Translate reintroducing LLM-like qualities into a text, the further it moves from the original text. The Llama generations hardly pose a challenge, which is unsurprising given Llama's presence in the training data. The ELOQUENT obfuscations are in the middle. They do produce a significant number of false negatives, but not nearly as many as, e.g., the Brown corpus or reverse Swahili generations. The ELOQUENT source texts, the unobfuscated New York Times articles, and the Brown corpus continuations also produce significant false-positive rates up to 40 %. Interestingly, this effect is strongest among the top-10, which is explained by the fact that several of the top-10 systems opt to trade a low false negative rate for more false positives.

5. Conclusion

The Voight-Kampff task received 137 software submissions from 34 teams and 27 work notebooks. Three systems beat last years' winner, but about half the systems fail to perform better than a zero-knowledge majority class prediction. Most systems are conservative classifiers with mostly stable false-positive rates but high false-negative rates. Especially among the top-10, however, some systems instead chose to trade a low false-negative rate for false positives. ELOQUENT also manages to increase the number of false negatives significantly, but not more than some other generation conditions, such as generated continuations of the Brown corpus.

Declaration on Generative AI

No generative AI was used in the design, execution, or evaluation of the experiments (except as a study subject itself). Apart from light spellchecking, no AI was used in drafting this report. Don't believe us? Try to prove us wrong and submit your approach next year.

Acknowledgments

Supported by the BMFTR project LANTMARK: Potential and limitations of LLMs for applications in Watermarking of text and spoken language. Funding ID 16KIS2307K.

References

[1] J. Bevendorff, M. Wiegmann, E. Richter, M. Potthast, B. Stein, The Two Paradigms of LLM Detection: Authorship Attribution vs. Authorship Verification, in: W. Che, J. Nabende, E. Shutova, M. Pilehvar (Eds.), The 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025) (Findings), Association for Computational Linguistics, 2025, pp. 3762–3787. URL: https://aclanthology.org/2025.findings-acl.194/.

[2] J. Bevendorff, I. Borrego-Obrador, M. Chinea-Ríos, M. Franco-Salvador, M. Fröbe, A. Heini, K. Kredens, M. Mayerl, P. Pęzik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2023: Authorship Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, A. G. S. Vrochidis, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association (CLEF 2023), volume 14163 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2023, pp. 459–481. doi:10.1007/978-3-031-42448-9_29.

[3] J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection, in: A. Barrón-Cedeño, G. Da San Martino, M. D. Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022), volume 13390 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2022, pp. 382–394. doi:10.1007/978-3-031-13643-6.

[4] J. Bevendorff, B. Chulvi, G. L. D. L. P. Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter,and Style Change Detection, in: K. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 12th International Conference of the CLEF Association (CLEF 2021), volume 12880, Springer, 2021, pp. 419–431. URL: https://doi.org/10.1007/978-3-030-85251-1_26.

[5] J. Bevendorff, B. Ghanem, A. Giachanou, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Wiegmann, E. Zangerle, Overview of PAN 2020: Authorship Verification, Celebrity Profiling, Profiling Fake News Spreaders on Twitter, and Style Change Detection, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 11th International Conference of the CLEF Initiative (CLEF 2020), volume 12260 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2020, pp. 372–383. doi:10.1007/978-3-030-58219-7_25.

[6] J. Bevendorff, M. Wiegmann, J. Karlgren, L. Dürlich, E. Gogoulou, A. Talman, E. Stamatatos, M. Potthast, B. Stein, Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 2486–2506. URL: http://ceur-ws.org/Vol-3740/paper-225.pdf.

[7] J. Bevendorff, Y. Wang, J. Karlgren, M. Wiegmann, M. Fröbe, A. Tsivgun, J. Su, Z. Xie, M. Abassy, J. Mansurov, R. Xing, M. Ta, K. Elozeiri, T. Gu, R. Tomar, J. Geng, E. Artemova, A. Shelmanov, N. Habash, E. Stamatatos, I. Gurevych, P. Nakov, M. Potthast, B. Stein, Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 3504–3534. URL: https://ceur-ws.org/Vol-4038/paper_277.pdf.

[8] R. R. Gunti, B. Bayramoğlu, G. Devadasu, D. Galat, V. A. Narayana, R. Pakala, V. R. Reddyvari, M.-A. Rizoiu, J. M. Rodriguez, S. S. Sanagala, A. Tommasel, J. Bevendorff, J. Karlgren, Overview and Joint Report of the Voight-Kampff Task of the ELOQUENT 2026 Lab for Evaluating Generative Language Model Quality, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[9] H. Plutz, A. Jakoby, J. Bevendorff, M. Fröbe, M. Potthast, B. Stein, Overview of the Text Watermarking Task at PAN 2026, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[10] 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.

[11] M. Brennan, S. Afroz, R. Greenstadt, Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity, ACM Transactions on Information and System Security 15

(2012). doi:10.1145/2382448.2382450.
[12] H. Wang, P. Juola, A. Riddell, Reproduction and replication of an adversarial stylometry experiment, arXiv [cs.CL] (2022). URL: http://arxiv.org/abs/2208.07395. arXiv:2208.07395.
[13] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, J. Zou, GPT detectors are biased against non-native English writers, Patterns (New York, N.Y.) 4 (2023) 100779. URL: https://www.sciencedirect.com/science/article/pii/S2666389923001307?via%3Dihub. doi:10.1016/j.patter.2023.100779. arXiv:2304.02819.
[14] A. Al Ali, J. Helcl, J. Libovický, Different time, different language: Revisiting the bias against non-native speakers in GPT detectors, in: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop), Association for Computational Linguistics, Stroudsburg, PA, USA, 2026, pp. 277–291. URL: http://dx.doi.org/10.18653/v1/2026.eacl-srw.20. doi:10.18653/v1/2026.eacl-srw.20.
[15] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, T. Goldstein, A watermark for large language models, International Conference on Machine Learning 202 (2023) 17061–17084. URL: https://proceedings.mlr.press/v202/kirchenbauer23a.html. doi:10.48550/arXiv.2301.10226. arXiv:2301.10226.
[16] J. Bevendorff, M. Potthast, M. Hagen, B. Stein, Heuristic Authorship Obfuscation, in: A. Korhonen, L. Màrquez, D. Traum (Eds.), 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Association for Computational Linguistics, 2019, pp. 1098–1108. URL: https://aclanthology.org/P19-1104/.
[17] W. N. Francis, Brown corpus manual, http://icame. uib. no/brown/bcm. html (1979).
[18] J. Karlgren, M. Barrett, O. Bojar, M. I. Engels, D. Fabre, S. Ettejjari, L. Goeuriot, R. R. Gunti, J. Mothe, P. Mulhem, M. Piacentini, L. F. V. Madriz, D. Schwab, P. Šindelář, G. Stampoulidis, K. Thomas, M. Vartampetian, Overview of ELOQUENT 2026: shared tasks for evaluating generative language model quality, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2026.
[19] Z. Antoszewska, N. M. Bryła, A. V. Szczerba, Team Suspiciously Coherent at PAN 2026: We stacked two Qwens and hoped for the best, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[20] S. Thorat, Team DACTYL at PAN 2026: Bayesian Data Mixing and Empirical X-risk Minimization for AI-text Detection, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[21] Y. Palkovskii, A. Belov, Team Plagiarism-Detector.com at PAN 2026: Strong Detector, Moving Target – Adversarial Augmentation and Out-of-Distribution Findings, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[22] D. Macko, mdok of KInIT: Robustly Fine-tuned LLM for Binary and Multiclass AI-Generated Text Detection, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, volume 4038, CEUR-WS.org, 2025, pp. 3819–3826. URL: https://ceur-ws.org/Vol-4038/paper_307.pdf.
[23] H. Basava, M. Gustineli, Fine-Tuning Fundamentals and Data Augmentation: DS@GT ARC at Voight-Kampff 2026, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[24] J. Yang, Q. Dong, Team ydong at PAN 2026: Detecting Machine-Generated Text via Multi-Algorithm Compression Distances and Shallow Statistics, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs

of the Evaluation Forum, CEUR-WS.org, 2026.
[25] F. B. Milak, S. I. Fornaroli, M. S. Queralt, Turing team at PAN 2026: Adversarial Data Fine-tuned LLM for AI-Generated Text Detection, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[26] D. Condrey, Writerslogic at PAN 2026: Process over Content for Robust Detection under Domain Shift, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[27] A. Voulgaraki, A. Zamyshevskaya, Team A&A-detection at PAN 2026: A Hybrid Ensemble of Stylometric, Perplexity, and Neural Features for Voight-Kampff Generative AI Detection, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[28] A. Kovács, Y. Shi, AK-YS at PAN 2026: Towards Robust Machine-Generated Text Detection under Domain Shift and Obfuscation, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[29] T. Vaccari, P. Topaloglu, A. Avdeiko, G. Peikos, MiPV AI-IR at PAN 2026: Not Just Black Boxes, Interpretable AI-Generated Text Detection with Heterogeneous Signals, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[30] Y. Alqarni, F. AlGhamdi, Team RADAR at PAN 2026: Robust Adversarial-Resistant AI Text Detection, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[31] C. Espin-Riofrio, A. Montejo-Ráez, J. Lara-Jama, J. Castro-Illescas, Team Sinai at PAN 2026: Combining Stylometric Embeddings and IDF Features for Generative AI Text Detection, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[32] A. Thapaliya, S. Pudasaini, B. K. Bal, eGrantha.ai at PAN 2026: Contrastive Representation Learning for Robust Generative AI Authorship Detection, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[33] M. A. Albani, R. Sonbol, Generalizable AI Authership Verification via Adverserial Longformer: Beyond Seen LLMs, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[34] M. Ibrahim, N. El-Makky, M. Torki, Team aimoment at PAN 2026: Content-Agnostic Contrastive Stylometry for Multi-Author Writing Style Analysis, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[35] S. Pudasaini, L. Miralles-Pechuán, D. Lillis, M. L. Salvador, shushantatudublin at PAN 2026: Robust Detection of AI-Generated Texts via Multi-Generator and Multi-Evasion Adversarial Training, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[36] M. Agirrezabal, manexagirrezabal at PAN 2026: Machine Learning for Detecting Texts Automatically Generated by AI, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.
[37] P. Liu, A. Wang, Team xlbn at PAN 2026: Fusion of Multi-domain Features for Generative AI Authorship Verification, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.

[38] T. Nagy, L. Skopar, Team ChatOnly at PAN 2026: Hybrid N-Gram and DeBERTa AI Detection, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.

[39] D. Villate, Team David Villate at PAN 2026 RoBERTa-LC, Length-Controlled Supervised Detection of AI-Generated Text, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.

[40] B. Savalam, M. Wiegmann, Pinkey-bw at PAN 2026: Robust Generative AI Authorship Detection using RoBERTa and Stylometric Feature Fusion, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.

[41] K. Aldous, W. Zaghouani, Team LatentAuthors at PAN 2026: DeBERTa with Genre-Aware Fine-Tuning for AI Text Detection, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.

[42] Z. Zeng, Z. Han, , Z. Liang, Team zhengqiaozeng at PAN 2026: Generative AI Detection Using DeBERTa-v3-large with R-Drop Regularization and Supervised Contrastive Learning, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.

[43] Y. Jia, J. Liu, L. Kong, Team FOSU at PAN 2026: Generative AI Text Detection via Part-of-Speech Style Injection Pre-training, in: E. Sánchez Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2026.

[44] L. Dugan, A. Hwang, F. Trhlík, A. Zhu, J. M. Ludan, H. Xu, D. Ippolito, C. Callison-Burch, RAID: A shared benchmark for robust evaluation of machine-generated text detectors, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Stroudsburg, PA, USA, 2024, pp. 12463–12492. URL: https://aclanthology.org/2024.acl-long.674.pdf. doi:10.18653/v1/2024.acl-long.674.

[45] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, O. M. Afzal, T. Mahmoud, G. Puccetti, T. Arnold, A. F. Aji, N. Habash, I. Gurevych, P. Nakov, M4GT-Bench: Evaluation benchmark for black-box machine-generated text detection, arXiv [cs.CL] (2024). URL: http://arxiv.org/abs/2402.11175. arXiv:2402.11175.

[46] J. Qiu, J. Zhu, A. Patel, M. Apidianaki, C. Callison-Burch, mStyleDistance: Multilingual Style Embeddings and their Evaluation, in: Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Stroudsburg, PA, USA, 2025, pp. 16917–16931. URL: http://dx.doi.org/10.18653/v1/2025.findings-acl.869. doi:10.18653/v1/2025.findings-acl.869.

[47] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, T. Goldstein, Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text, in: Forty-first International Conference on Machine Learning, {ICML} 2024, Vienna, Austria, July 21-27, 2024, volume abs/2401.12070, OpenReview.net, 2024, pp. 17519–17537. URL: https://dl.acm.org/doi/10.5555/3692070.3692768. doi:10.48550/arXiv.2401.12070.

[48] D. Macko, R. Moro, A. Uchendu, I. Srba, J. S. Lucas, M. Yamashita, N. I. Tripto, D. Lee, J. Simko, M. Bielikova, Authorship obfuscation in multilingual machine-generated text detection, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Stroudsburg, PA, USA, 2024, pp. 6348–6368. URL: http://dx.doi.org/10.18653/v1/2024.findings-emnlp.369. doi:10.18653/v1/2024.findings-emnlp.369.

[49] K. Krishna, Y. Song, M. Karpinska, J. Wieting, M. Iyyer, Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems 36 (NeurIPS 2023), volume 36, 2023, pp. 27469–27500. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/575c450013d0e99e4b0ecf82bd1afaa4-Abstract-Conference.html.

[50] D. Sculley, C. E. Brodley, Compression and machine learning: A new perspective on feature

space vectors, in: Data Compression Conference (DCC'06), IEEE, 2006, pp. 332–341. URL: https://ieeexplore.ieee.org/abstract/document/1607268?casa_token=EwORidpcgTwAAAAA:zONgvLu6aVgw-jrz0A-5JHXs-SdAljLqNXabhAQh6w685CRwAXXe7FxcD67SDkf6Ztfaj6AEwWA. doi:10.1109/dcc.2006.13.

[51] 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, volume Part F1305, ACM, New York, NY, USA, 2017. doi:10.1145/3098954.3104050.

[52] A. Peñas, Á. Rodrigo, A Simple Measure to Assess Non-response, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 1415–1424. URL: https://aclanthology.org/P11-1142.pdf.

[53] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Bias Analysis and Mitigation in the Evaluation of Authorship Verification, in: A. Korhonen, L. Màrquez, D. Traum (Eds.), 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Association for Computational Linguistics, 2019, pp. 6301–6306. URL: https://aclanthology.org/P19-1634/.