Overview and Joint Report of the Voight-Kampff Task of the ELOQUENT 2026 Lab for Evaluating Generative Language Model Quality

Notebook for the ELOQUENT Lab at CLEF 2026

Rohit Raj Gunti[1], J M Shalani Dilinika[4,5], Devadasu Ghanta[6], Dima Galat[9], Narayana V A[6], Raveendrababu Pakala[6], Venkateswara Reddy Reddyvari[6], Marian-Andrei Rizoiu[8], Juan Manuel Rodriguez[7,11], Siva Skandha Sanagala[6], Antonela Tommasel[9,10], Janek Bevendorff[11] and Jussi Karlgren[2,3]

[1]*University of Tennessee, Knoxville, USA*
[2]*AMD Silo AI, Sweden*
[3]*University of Helsinki, Finland*
[4]*University of Pittsburgh, USA*
[5]*University of Kelaniya, Sri Lanka*
[6]*CMR College of Engineering and Technology, India*
[7]*Aalborg University, Denmark*
[8]*University of Technology Sydney*
[9]*Johannes Kepler University, Linz, Austria*
[10]*CONICET - UNICEN, Tandil, Buenos Aires, Argentina*
[11]*Bauhaus-Universität Weimar, Germany*

Abstract

The challenge for generative AI authorship verification is not simply distinguishing human-written text from machine-generated text, but doing so under increasingly realistic and adversarial conditions. The 2025 Voight-Kampff task emphasized that obfuscated AI text and human-AI collaborative writing remain difficult cases, with detection performance affected by domain shifts, rewriting strategies, and subtle differences between machine-polished, machine-humanized, and human-edited content. This year, we built directly on this problem by focusing on whether generated texts can become harder to detect across varied sources, genres, and generation strategies. This year's competition reflects that shift. Across five participating teams and 16 submitted responses, systems differed not only in model choice but also in how they attempted to create human-like text: through prompt engineering, controlled decoding, detector-aware generation, and fine-tuning. The results are explained in the end, which reflects a more diverse experimental setting that helps reveal where current detection systems remain robust and where they begin to fail.

1. Evaluating Generative Language Model Quality

The shared task experiment described in this report is part of the ELOQUENT lab for evaluation of generative language model quality at CLEF, the Conference and Labs of the Evaluation Forum. A full report from the lab can be found in the CLEF Conference proceedings [1]. This is the third edition of this task which has continued essentially unchanged since its inception, in a collaboration with the PAN lab. Participants in this task generate datasets of machine text, attempting to break classifier systems designed to distinguish between human-authored and machine generated text, built by participants in a corresponding task in the PAN lab [2].

In its first edition, the classifiers submitted by participants to the PAN lab handily classified the texts into human vs. machine: we found that of the submitted datasets, all were able to fool some of the classifier systems some of the time; but no generative model was consistently able to convince the better

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

classifier systems that it was human. It was clear that machine-generated texts appeared to consistently hold to certain detectable stylistic indicator features [3].

In its second year, generators started closing the gap. Several of the submissions fooled classifiers better than chance, a clear step up from the first edition and a sign of more deliberate experimentation, even as human-authored texts were misclassified only about 15% of the time. Still, the basic conclusion held: machine-generated text carried recognisable traces, and writing text that genuinely passes as human remained a gap. That edition also raised a question it could not settle, namely, why some test items were consistently easier to humanize than others [4].

Outside the lab, the detection literature points to several gaps. First, detectors are easy to fool: simple paraphrasing slips past tools like GPTZero and DetectGPT, dropping DetectGPT's accuracy from 70.3% to 4.6% without changing meaning [5], and across a range of attacks detectors lose about 35% of their accuracy on average [6]. Second, they do not generalize: even the best detector's recall falls from 90.5% to 68.4% on unfamiliar domains [7]. Third, these studies each isolate one factor at a time, so it remains unclear which part of the generation process makes text hardest to detect.

This year's competition partially addressed the gap identified in the previous edition. Teams adopted more deliberate generation strategies, including prompt-based elaboration, model variation, genre and style matching, and human-in-the-loop review. These approaches helped reduce some recognizable traces of machine-generated writing and produced stronger results than the previous edition, suggesting measurable progress in humanizing generated text. However, the gap was not fully closed. While this edition provides clearer evidence that generation choices affect detectability, further item-level analysis is still needed to explain why some test items are easier to humanize than others [4].

2. Participation

By the deadline, the task had four registered teams in total with 15 submissions across multiple experimental conditions using different models and prompting configurations, as shown in Table 1. Compared to last year, this year's shows a more actively engaged participant pool with a higher number of submitted runs relative to the number of registered teams.

The experimental setting followed the same general builder-breaker logic: participating teams ("breakers") produced AI-generated or modified texts intended to challenge human-vs-machine detection systems developed by teams ("builders") in the corresponding PAN lab task [8]. The submissions varied by response strategy, approach, and model choice, allowing the evaluation to compare not only team-level performance but also the effectiveness of different generation, rewriting, obfuscation, or humanization conditions. This setting is important because the goal is not simply to count participation, but to understand which experimental designs make generated text harder to distinguish from human-authored text. Thus, the increased number of submitted conditions suggests broader experimentation within a smaller participant group, rather than only a simple increase in team participation.

We asked participating teams (a) to report the language of their submission, (b) whether they used system prompts or modified the suggested prompt, (c) which model, training data were associated with the model, whether the data were publicly available, and (d) whether any instruction tuning, alignment training, fine-tuning, decoding modification, or other adaptation was applied. These questions were designed to make the submissions comparable beyond their final outputs. Since the task evaluates whether generated texts can evade human–machine detection, it is important to understand not only which model produced the text, but also how the model was prompted. Especially, whether it was trained or adapted for the task, what data informed the approach, and whether the strategy relied on model-level training, inference-time control, or simple prompt engineering.

Across the four teams, the 15 submissions show substantial diversity in experimental design.

Team UTS used seven English submissions based on Claude Opus 4.7 for the final runs, without generator fine-tuning, quantization, or chain-of-thought scaffolding. Their strategy centered on task-specific prompt layering and adversarial analysis: they created a 1,650-text generated corpus from five generators, five strategies, and 66 topics, paired it with PAN/CLEF 2025 human texts, and trained

RoBERTa-large adversarial detectors to study robustness and guide attack design. Their work also included variants using additional pre-1923 Project Gutenberg passages, highlighting a data-centric and detector-aware approach.

Team CMR followed a more direct prompt-based strategy. Their submissions used Gemini 3 Fast, College Copilot-365, and Personal Copilot-365, with instructions such as elaborating each item into approximately 500 words. Their approach was mainly based on using available generative systems with lightweight prompt modifications, including one Telugu and four English submissions.

Team Darkg used Meta's Llama-3.2-3B-Instruct as the base model and applied LoRA adapters trained with GRPO. Their training prompts were drawn from Voight-Kampff task data from 2024, 2025, and 2026, and the reward function was based on reducing AI-detection probability as estimated by their own Mdok2 classifier. At inference time, they generated up to ten candidates per prompt and selected the candidate with the lowest detector score, turning generation into a classifier-guided search process. This represents a more explicitly optimization-driven strategy, where model adaptation and candidate selection were both aligned toward evading detection.

Team Anlam-Turning used Microsoft Phi-3.5-mini-instruct in 4-bit NF4 quantization without fine-tuning. Instead of training the model, they introduced a complex inference-time controlled decoding pipeline, including anti-cliché logit masking, repetition and n-gram penalties, sentence-rhythm control, per-paragraph temperature scheduling, a quantum top-3 sampler, Binoculars-style reranking, and post-hoc syntactic rewriting. This team, therefore, represents a decoding-control approach: the foundation model was kept frozen, while the generation process was heavily engineered to reduce recognizable LLM patterns.

Taken together, the four teams and 15 submissions demonstrate a broad methodological spectrum: prompt-only generation, proprietary model prompting, open-source model fine-tuning, classifier-guided reinforcement-style adaptation, adversarial detector training, and inference-time controlled decoding. This diversity is important because it shows that participation was not limited to comparing model families: teams explored different assumptions about where "human-likeness" can be introduced—through the prompt, the training data, the generator, the decoder, post-processing, or a detector-guided selection loop.

3. Data

The test set consists of 20 items, summarized and exemplified in Table 2. Following the design of the 2025 Voight-Kampff ELOQUENT task, we selected human-authored source materials and converted them into topic-and-style summaries for participants to use as generation prompts. In the 2025 edition, the organizers selected 20 human-written texts of 350–700 words, summarized them using OpenAI's ChatGPT service, and shared the summaries with participants using the suggested prompt "Write a text of about 500 words which covers the following items." In the current edition, the same summary-based generation setup was retained, but the source pool was expanded to include materials from Project Gutenberg, Standard Ebooks, Internet Archive, and YouTube, with most written sources selected from public-domain or reuse-eligible texts. The source passages were summarized using ChatGPT with GPT-5.2, with prompts asking for bullet-point summaries and descriptions of genre and style. We also varied the formatting of the summaries across items, including paragraph, bullet, numbered, alphabetical, and roman-numbered formats. These summaries were then shared with participants, together with the suggested instruction to generate a text of about 500 words covering the listed topics, while allowing teams to modify the prompt or apply their own generation strategies.

Table 1
Participation

Team	Submissions	Approach	Model
UTS	Submission 1	Task-specific prompt layering; source-anchored generation	Claude Opus 4.7
	Submission 2	Task-specific prompt layering; cross-register generation	Claude Opus 4.7
	Submission 3	Task-specific prompt layering; n-pass refinement	Claude Opus 4.7
	Submission 4	Named-author imitation strategy	Claude Opus 4.7
	Submission 5	Adversarial generation guided by detector analysis	Claude Opus 4.7
	Submission 6	Detector-aware generation using PAN'25 human negatives	Claude Opus 4.7
	Submission 7	Detector-aware generation with additional Gutenberg-style human data	Claude Opus 4.7
CMR	Submission 1	Prompt-based elaboration to approximately 500 words; Telugu output	Gemini 3 Fast
	Submission 2	Prompt-based elaboration to approximately 500 words	College Copilot-365
	Submission 3	System-style instruction for accurate, clear, and relevant 500-word elaboration	Personal Copilot-365
	Submission 4	Direct prompt instruction to elaborate each item to 500 words	Personal Copilot
	Submission 5	Prompt-based elaboration of each text to 500 words	College Copilot-365
DArgk	Submission 1	Fine-tuned weights trained with GRPO; detector-guided best-of-10 candidate selection	Llama-3.2-3B-Instruct + GRPO
	Submission 2	LoRA/GRPO variant with classifier reward and best-of-10 inference selection	Llama-3.2-3B-Instruct + LoRA/GRPO
Anlam-Turing	Submission 1	Frozen model with controlled decoding, anti-cliché masking, rhythm control, reranking, and syntactic rewriting	Phi-3.5-mini-instruct
PTUK	Submission 1	Followed a direct prompt-based strategy with human review without fine-tuning, quantization, adapter training, or chain-of-thought scaffolding	Claude Opus 4.8 (Anthropic)

Test Item (ID: 054)

Genre and Style:
Genre: Detective Fiction / Mystery Narrative Opening.
Tone: Suspenseful, controlled, and observant. It uses concise descriptive narration and rising tension to introduce a sensitive intellectual crime, blending Holmesian deduction with an atmosphere of secrecy, urgency, and institutional anxiety.

Content:
* Back in 1895, a series of circumstances led Sherlock Holmes and Dr. Watson to spend several weeks in a major university town.
* The exact college and individuals involved are deliberately left unnamed.
* The scandal surrounding the case was painful enough that discretion seemed wiser than publicity.
* Holmes was staying in modest furnished lodgings near a library.
* He was deeply absorbed in research on early English charters.
* The work was demanding but promising—important enough that Watson hints it may deserve its own story someday.
* One evening, they received an unexpected visitor: Mr. Hilton Soames.
* He was a tutor and lecturer at a college referred to as St. Luke's…….

Figure 1: Illustrative Voight-Kampff 2026 test item: a detective fiction scenario involving a scholarship examination fraud investigated by Sherlock Holmes.

Table 2
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

4. Submissions

4.1. UTS

We entered ELOQUENT 2026 with a sharper version of last year's problem: can an attack remain effective after detectors have already seen and adapted to earlier attacks? That question shaped every design choice in our submissions.

The answer came down to an asymmetry. Pushing generated text out of the detector's training distribution defeats adversarial fine-tuning. Pulling it into that distribution, for instance, by mimicking the human training data, does not.

We checked the negative case first. Replicating the 2025 winning recipes (literal Hindi back-translation with a "be-imperfect" instruction, and the nine-language rotation) across four generators, 5 strategies, and 66 topics, we trained our own deployer-class RoBERTa-large detectors on those generations. Once the detector has seen even a small sample of the 2025 outputs, the recipe collapses: its fool rate inverts sign and drops to 0.025 or below. The surface-level perturbation does not last.

All seven of our top submissions run on Claude Opus 4.7 and differ only in how the prompt steers the generation. Submission 1 (source-anchored generation) is a two-call recipe: the generator first writes a short period-prose paragraph on a subject close to the topic, then writes the topic itself in that exact early twentieth-century register, using the paragraph as a stylistic template. Submission 2 (cross-register generation) drops the anchor and prompts the historical register directly. Submission 3 (n-pass refinement) iterates a draft several times. Submission 4 (named-author imitation) asks for the voice of a specific writer; only the period-adjacent voices, such as Sebald, survive, while modern ones like Borges fail. That split is the tell: the detector is blind to historical language, not to a distinctive authorial voice.

Submission 5 (adversarial generation guided by detector analysis) is the top entry. It fixes a narrative form we chose by analyzing which out-of-distribution axes beat both detector classes simultaneously, then selects, per topic, the candidate that minimizes the worst-case p(LLM) across an eight-detector committee covering supervised classifiers, zero-shot perplexity, and LLM judges. That worst-case rule keeps the entry safe, regardless of which detector the organizers actually run.

The last two submissions test the inverse direction, and both confirm the asymmetry by failing. Submission 6 (detector-aware generation using PAN'25 human negatives) few-shots Claude with real PAN'25 human passages and asks it to match their voice; against the adversarial detector it lands at essentially zero. Submission 7 (detector-aware generation with additional Gutenberg-style human data) anchors on real pre-1923 prose instead of synthetic, and underperforms the synthetic anchor of Submission 1. Pulling text toward the human training data does not help.

Against our strengthened adversarial detector, Submission 1 escapes at 0.846 and Submission 5 at 0.65, roughly 50 times the 2025 baseline, with no loss of naturalness. The obvious detector countermeasure, adding real pre-1923 prose to the human side of training, does not close the gap, and if anything, it nudges the fool rate up. Out-of-distribution structure, not surface mimicry, is what current adversarial detectors cannot handle. Our full method, results, and ablations are in our working notes.

4.2. CMR - Prompt-Based Text Elaboration Using Gemini and Copilot

We submitted five runs using a prompt-based generation strategy across Telugu and English. The first submission was in Telugu and used Google Gemini 3 Fast, with prompts asking the model to elaborate the given content into approximately 500 words. The remaining submissions were primarily in English and used Microsoft Copilot variants, including College Copilot-365, Personal Copilot-365, and Personal Copilot. Across the submissions, the we relied on the provided task-vk-test-2026 / Voight-Kampff 2026 data as the input source rather than reporting any separate fine-tuning or additional training dataset.

Our approach can be characterized as a lightweight prompt-elaboration method. Instead of modifying model weights through fine-tuning, instruction tuning, adapter training, or alignment procedures, the submissions used direct instructions such as "elaborate each ID for 500 words" or "elaborate the text in 500 words." Some submissions also included a general assistant-style system prompt asking the model

to produce accurate, clear, helpful, safe, and relevant responses. Overall, our five submissions explored different publicly available or productivity-oriented generative AI systems while keeping the main generation strategy consistent: expanding each test item into a longer 500-word response based on the provided competition inputs.

4.3. DArgk – SHADE: Stochastic Human-like generation via Adversarial Detector Evasion

Our submission, SHADE (*Stochastic Human-like generation via Adversarial Detector Evasion*) [9], frames the generation task as an adversarial reinforcement learning problem. Rather than applying post-hoc perturbations, translation or prompt-only rewriting, we adapt the generation policy of an instruction-tuned Llama 3.2 3B model so that its outputs receive low AI-generated probabilities from a surrogate detector.

As the official detectors are unseen, we trained a surrogate detector based on the PAN 2025 *mdok* system. Due to hardware constraints, we replaced the original large backbone with `FacebookAI/roberta-base` and trained this detector on the PAN 2025 data. This surrogate was then used as the reward model during optimization.

We fine-tuned the generator with Group Relative Policy Optimization (GRPO). For each prompt, the model generated candidate outputs, which were scored by the surrogate detector. The reward was defined as the complement of the detector confidence, so texts classified as less AI-generated received a higher reward. A KL penalty against the original reference model was used to limit policy drift and reduce degenerate generations. We explored both full fine-tuning and LoRA-based adaptation, as well as different KL coefficients.

At inference time, we used a chat-formatted version of the task prompt and stochastic decoding with temperature and top-p sampling. For each input prompt, multiple candidates were generated and evaluated by the surrogate detector. If any candidate obtained an AI-generated probability below a fixed threshold, it was selected. Otherwise, the candidate with the lowest detector score was returned.

4.4. Anlam-Turing

Team Anlam-Turing used a controlled inference-time generation approach rather than fine-tuning a model. They used Microsoft Phi-3.5-mini-instruct in 4-bit NF4 quantization, with the model weights kept unchanged. Their main contribution was a custom decoding pipeline designed to reduce common LLM writing patterns, including repetition penalties, anti-cliché masking, sentence-rhythm control, temperature scheduling, a quantum top-3 sampler, discriminator-style reranking, and post-processing with syntactic rewriting. The organizers' prompt was used with an added system preamble that encouraged more varied sentence structure, concrete phrasing, and less template-like writing. Overall, the approach focused on making generated text appear more human through controlled decoding and stylistic post-processing, without any additional training or model adaptation.

4.5. PTUK - Human-in-the-Loop Prompting for Genre- and Style-Controlled Generation

PTUK followed a direct prompt-based strategy with human review. Their submission used Claude Opus 4.8 (Anthropic), without fine-tuning, quantization, adapter training, or chain-of-thought scaffolding. They combined the organizers' suggested prompt with each item's Content and Genre and Style fields, adding a system-level instruction to match the specified genre and register and to avoid dash characters as a known stylistic marker of machine-generated text. Texts were generated one item at a time. Each output was reviewed for (a) coverage of Content points, (b) adherence to the specified genre and tone, (c) word count around 500 words, and (d) absence of dash characters, with under-length or off-register outputs regenerated or revised. This represents a lightweight prompt-and-review approach, relying on prompt design and human-in-the-loop revision.

Table 3
Evaluation Results for Eloquent/Voight-Kampff Test Dataset

Model	Team	Brier	C@1	F1	F0.5u	Mean	FPR	FNR
Claude opus 4.7	UTS	0.342	0.228	0.361	0.585	0.379	0.000	0.743
Claude opus 4.7	UTS	0.356	0.283	0.430	0.554	0.431	0.000	0.594
Claude opus 4.7	UTS	0.448	0.344	0.499	0.714	0.501	0.000	0.534
claude opus 4-7 roundtrip imperf	UTS	0.415	0.351	0.509	0.722	0.499	0.000	0.530
Qwen	UTS	0.555	0.470	0.525	0.806	0.514	0.000	0.511
meta-llama/Llama-3.2-3B-instruct	Dargk	0.567	0.480	0.527	0.808	0.520	0.000	0.493
claude-opus-4-7 rejection sample opus	UTS	0.590	0.485	0.535	0.813	0.531	0.000	0.493
claude-opus-4-7 source anchored synth	UTS	0.503	0.509	0.558	0.828	0.549	0.000	0.471
meta-llama/Llama-3.2-3B-instruct	Dargk	0.524	0.561	0.597	0.852	0.583	0.000	0.417
microsoft/Phi-3.5-mini-instruct (3.8 B parameters)	Anlam Turing	0.773	0.741	0.834	0.826	0.818	0.000	0.249
Claude (Anthropic), Opus 4.8 model, accessed via Anthropic's chat interface	PTUK	0.804	**0.882**	0.000	0.000	0.447	0.147	0.000
Gemini 3 Fast	CMR	0.855	0.870	**0.871**	**0.888**	**0.871**	0.000	0.029
College-Co-Pilot-365	CMR	0.855	0.870	**0.871**	**0.888**	0.871	0.000	0.029
College-Co-Pilot-366	CMR	0.854	0.870	**0.871**	**0.888**	0.871	0.000	0.029
Personal-Co-Pilot-365	CMR	0.852	0.870	**0.871**	**0.888**	0.870	0.000	0.029
Personal Copilot-365	CMR	0.850	0.870	**0.871**	**0.888**	0.870	0.000	0.029

5. Results

The PAN builder task had 66 registrations, with 34 teams submitting 129 classifier experiments. For each team, only the best classifier is reported here, to simplify tables.

Given 34 number of classifiers, 16 submissions, and 20 test items:

a) For each generator submission, how many times did it fool the classifiers? Score goes from 0 (zero) to $K \times 25$.

b) For each test item, how difficult is it? How many times did the N generated submissions succeed in fooling the classifiers? Score from 0 (zero) to $K \times N$.

table-voightkampff-evaluation-results

The final evaluation contained 16 scored submissions (see Table 3), including the human-original reference condition. Since the goal of the ELOQUENT side of the task is to generate texts that are difficult for PAN detectors to classify as machine-generated, we interpret lower detector scores as stronger generation-side performance. In particular, we report the inverse of C@1, $1 - \mathrm{C@1}$, as a fooling-oriented measure: higher values indicate that the submitted texts were more difficult for the detectors to classify correctly.

Compared with the 2025 edition [4], the strongest 2026 submissions show a clear advance in adversarial effectiveness. In 2025, the most difficult ELOQUENT-generated dataset reached an inverse C@1 score

of approximately 0.63, with the strongest results coming from translation, back-translation, and related rewriting strategies. In the present edition, the best-performing generated condition, `Claude Opus 4.7`, reached an inverse C@1 of 0.228 against all PAN detection systems, with a false negative rate of 0.743. This means that nearly three quarters of the generated texts in this condition were classified as human-authored by the full detector pool. Even when evaluated only against the top-10 PAN systems, the same condition achieved an inverse C@1 of 0.367 and a false negative rate of 0.618, roughly matching or slightly exceeding the most difficult condition reported in the previous year. This suggests that the best 2026 generation strategies did exploit weaker detectors, and they remained challenging even for the strongest systems.

The improvement, however, was not uniform across all submissions. The most successful conditions were concentrated in the detector-aware and adversarially designed submissions. UTS's Claude-based runs produced the strongest overall results, especially `claude-opus` run in the round-trip imperfection condition. These approaches moved beyond simple prompting by using task-specific prompt layering, source anchoring, cross-register variation, rejection sampling, and detector-informed analysis. UTS's best inverse C@1 scores ranged from 0.228 to 0.470, showing that carefully engineered generation strategies can substantially reduce detector reliability.

Dargk submissions represent a second important direction of progress: model adaptation guided by detector feedback. The LoRA/GRPO variant achieved an inverse C@1 of 0.480 and a false negative rate of 0.493, while the second Dargk condition achieved an inverse C@1 of 0.561. Against the top-10 detector systems, these results dropped to 0.685 and 0.808, respectively, showing that the strongest detectors were more resilient to this optimization strategy. Still, these results are significant because they demonstrate a more systematic generator training setup, where the detector is incorporated into the generation loop as a reward signal or selection criterion.

Anlam Turing's and PTUK (prompt-only) submissions were less effective. Anlam Turing's condition achieved an inverse C@1 of 0.741 and a top-10 inverse C@1 of 0.833, indicating that detectors were usually able to identify the generated texts. The CMR submissions (prompt-only), based mainly on direct elaboration prompts using Gemini or Copilot-style systems, were detected almost completely: their inverse C@1 ranged over 0.9, with a false negative rate of only 0.029. These results provide a useful contrast with the stronger adversarial submissions. These findings suggest that simply asking a model to elaborate a topic into a 500-word text is no longer sufficient to fool current detectors; effective evasion now appears to require more deliberate control over style, rhythm, source grounding, candidate selection, or detector-aware optimization.

6. Lessons Learnt, Observations, Thoughts

The human-original reference condition provides an important continuity check with the previous edition. In 2025, human-authored texts were misclassified as machine-generated at an average rate of about 15%, largely due to a small number of difficult source texts. PTUK's Human-in-the-Loop Prompting submission yielded a false positive rate of 0.147 against all submitted PAN detection systems and 0.155 against the top-10 detection systems. These results indicate that even genuine human-authored texts were occasionally misclassified as AI-generated, providing an important baseline for interpreting the fooling rates achieved by the generated submissions.

Overall, the 2026 results show that the task has advanced from testing generic obfuscation methods toward evaluating more sophisticated adversarial generation pipelines. Last year's findings showed that all generated datasets could fool some classifiers some of the time, but that only a few approaches approached strong fooling performance. This year, the best submissions exceeded that earlier ceiling against the full PAN detector pool and remained competitive against the top-10 systems. At the same time, the wide gap between adversarially designed submissions and prompt-only submissions shows that progress depends less on model size alone and more on how generation is controlled, selected, and evaluated. The central finding is therefore twofold: detectors have become stronger and more numerous, but generation-side strategies have also become more targeted, turning the task into a clearer

builder–breaker contest between robust detection and deliberate human-likeness optimization.

7. Plans for Next Year

For the next edition, we plan to continue the Voight-Kampff task in broadly the same builder-breaker format. The present results suggest that the task remains useful as a shared evaluation setting for studying both AI-text detection and adversarially human-like generation. At the same time, future editions should increase the difficulty in a principled way. One promising direction is to include more modern source material, since contemporary writing styles, topics, and online genres may create a more realistic and challenging testbed than relying primarily on older public-domain texts. We also plan to (a) explore whether the task can be extended through more diverse genres, (b) develop stronger human-authored baselines, (c) explore newer generative models, and (d) explore more controlled forms of obfuscation or human-AI collaboration. The goal for the next edition is therefore not only to repeat the task, but to make it a stronger benchmark for evaluating whether detection systems can generalize to current, varied, and deliberately difficult text-generation settings.

Acknowledgments

The ELOQUENT lab is partially supported by the European Commission through the OpenEuroLLM (101195233-DIGITAL-2024-AI-06) and the DeployAI (101146490-DIGITAL-2022-CLOUD-AI-B-03) projects as part of their activities on building, evaluating, and disseminating generative language models. Views and opinions expressed are those of the authors and do not necessarily reflect those of the European Union or the Digital Europe programme. Neither the European Union nor the programme can be held responsible for them.

References

[1] 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, S. MacAvaney, J. M. Struß, E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer, 2026.

[2] J. Bevendorff, R. Gunti, J. Karlgren, M. Fröbe, M. Potthast, B. Stein, Overview of the third "Voight-Kampff" Generative AI / LLM Detection Task at PAN and ELOQUENT 2026, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR-WS, 2026.

[3] 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, M. Vlachos, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740, CEUR-WS.org, 2024.

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

[5] K. Krishna, Y. Song, M. Karpinska, J. Wieting, M. Iyyer, Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense, Advances in neural information processing systems 36 (2023) 27469–27500.
[6] Y. Wang, S. Feng, A. Hou, X. Pu, C. Shen, X. Liu, Y. Tsvetkov, T. He, Stumbling blocks: Stress testing the robustness of machine-generated text detectors under attacks, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 2894–2925.
[7] Y. Li, Q. Li, L. Cui, W. Bi, Z. Wang, L. Wang, L. Yang, S. Shi, Y. Zhang, Mage: Machine-generated text detection in the wild, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 36–53.
[8] J. Bevendorff, M. Fröbe, A. Greiner-Petter, A. Jakoby, M. Mayerl, P. Nakov, H. Plutz, M. Potthast, B. Stein, M. N. Ta, Y. Wang, E. Zangerle, Overview of pan 2026: Voight-kampff generative ai detection, text watermarking, multi-author writing style analysis, generative plagiarism detection, and reasoning trajectory detection, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer, 2026.
[9] A. Tommasel, J. M. Rodriguez, Team DArgk at the 2026 ELOQUENT lab for evaluating generative language model quality: Residuals of humanity: AI detection evasion via GRPO Fine-Tuning, in: TBD (Ed.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR-WS, 2026.