Team DArgk at the 2026 ELOQUENT lab for evaluating generative language model quality: Residuals of Humanity: AI Detection Evasion via GRPO Fine-Tuning

Notebook for the Voight-Kampff Task on Eloquent Lab 2026 at CLEF 2026

Antonela Tommasel[1,2,*,†], Juan Manuel Rodriguez[3,2,*,†]

[1]Johannes Kepler University, Linz, Austria
[2]CONICET - UNICEN, Tandil, Buenos Aires, Argentina
[3]Department of Computer Science, Aalborg University, Aalborg, Denmark

Abstract

Large language models (LLMs) can generate fluent and coherent text that is increasingly difficult to distinguish from human writing, motivating the development of automatic AI-generated text detectors. However, the robustness of such detectors under adversarial generation remains uncertain. This paper presents SHADE (*Stochastic Human-like generation via Adversarial Detector Evasion*), a reinforcement learning framework that formulates detector evasion as a policy optimization problem. Instead of applying post-hoc perturbations or prompting-based rewriting, SHADE fine-tunes an instruction-tuned LLaMA model with Group Relative Policy Optimization (GRPO), using feedback from a surrogate detector based on the PAN 2025 *mdok* system. Our experiments show that full fine-tuning with a small KL regularization penalty achieves 98.5% surrogate evasion, compared to 1.5% for the base model, while LoRA-based adaptation is substantially less effective under regularization. Linguistic analysis reveals that successful evasion is associated with shorter, simpler, and less lexically diverse outputs, suggesting that high detector evasion does not necessarily correspond to more human-like writing. In the official Voight-Kampff competition setting, our submissions ranked sixth and seventh, indicating that optimization against a single surrogate detector only partially transfers to unseen evaluation classifiers. These results highlight both the potential and limitations of reinforcement learning for adversarial AI-text generation and motivate more robust, multi-detector evaluation protocols for AI-generated text detection.

Keywords

AI-generated text detection, adversarial generation, reinforcement learning, large language models, surrogate detectors

1. Introduction

The rapid advances in Large Language Models (LLMs), which are now capable of producing fluent, coherent, and human-like text across a wide range of domains and writing styles, have made it increasingly difficult to determine whether a text is written by a human or generated by an AI system [1]. This distinction is becoming important in settings where authorship, originality, accountability and trust are central, including education, journalism, scientific communication, online moderation and misinformation detection [2]. As LLM-generated content becomes easier to produce, adapt and distribute at scale, automatic AI-generated text detection has emerged as a key mechanism for identifying synthetic text and supporting decisions about content provenance. A growing body of work has therefore proposed detection methods based on perplexity, stylometric features, watermarking signals or fine-tuned transformers [3, 4, 5, 2]. However, the reliability of these detectors remains uncertain,

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
antonela.tommasel@isistan.unicen.edu.ar,antonela.tommasel@jku.at (A. Tommasel); jmro@cs.aau.dk (J. M. Rodriguez)
0000-0001-6091-8305 (A. Tommasel); 0000-0002-1130-8065 (J. M. Rodriguez)

particularly when generated texts are intentionally modified to appear less machine-generated. Their robustness under adversarial conditions is thus still an open question [6, 5].

Motivated by this challenge, the Voight-Kampff task [7] at the ELOQUENT 2026 lab [8] asks participants to generate text that evades unseen AI detectors submitted by participants of PAN 2026. In the previous edition of the Voight-Kampff task, the strongest approaches relied on text perturbation and translation-based transformations, whereas persona-based and stylistic rewriting strategies (i.e., prompting-based methods) achieved weaker results [9, 10]. Data-centric approaches have also explored fine-tuning with LoRA on human-written corpora to better mimic the distribution of human texts in the training data [11]. However, these methods largely treat evasion as either a post-processing problem or a static alignment problem, rather than as a dynamic optimization objective guided by detection feedback.

In this paper, we propose SHADE (*Stochastic Human-like generation via Adversarial Detector Evasion*), a framework that formulates detector evasion as a reinforcement learning problem. Instead of manipulating outputs post-hoc or curating training data to approximate human style, SHADE directly optimizes a language model's generation policy to reduce the confidence of an adversarial surrogate detector. Concretely, we fine-tune LLaMA 3.2 3B instruct using Group Relative Policy Optimization (GRPO), with rewards derived from a surrogate detector based on last year's Voight-Kampff task, namely *mdok* [12]. This adversarial formulation encourages the model to internalize evasion-oriented generation strategies, rather than relying exclusively on surface-level perturbations or prompting-based transformations.

The contributions of this paper are threefold: (i) we formulate AI-generated text detector evasion as a reinforcement learning problem driven by adversarial detector feedback; (ii) we introduce SHADE, a GRPO-based fine-tuning framework that optimizes an LLM against a surrogate from the previous Voight-Kampff task; and (iii) we empirically analyze the relationship between detector evasion and textual properties, showing that high evasion rates can emerge through linguistic simplification rather than through more elaborate human-style generation.

Our experiments show that full fine-tuning with a small KL regularization penalty achieves a 98.5% evasion rate against the surrogate detector, compared to only 1.5% for the base LLM. To better understand this behavior, we analyze the trade-offs between evasion and textual characteristics such as repetition, lexical richness, and readability. This analysis reveals that the model primarily evades detection by simplifying its language, suggesting that successful detector evasion does not necessarily correspond to richer or more human-like writing. We further show that parameter efficient tuning with LoRA is substantially less effective in this adversarial reinforcement learning setting, motivating full model adaptation for the proposed task.

The remainder of this paper is organized as follows. Section 2 reviews related work on AI detection evasion. Section 3 formalizes the problem and details SHADE, including its training objective, surrogate detector, and inference procedure. Section 4 analyzes the fine-tuning configurations and the linguistic properties of the generated texts. Section 5 reports results on the official Voight-Kampff task evaluation. Section 6 concludes the paper and Section 7 links the artifacts related to this work.

2. Related work

Research on AI-generated text detection and evasion lies at the intersection of two closely related questions: how such detectors can be attacked, and which evasion strategies have proven effective in shared-task settings. We therefore first discuss *AI-generated text detection*, focusing on statistical, supervised and watermarking-based approaches for distinguishing machine-generated from human-written text. We then review *attacks on AI-generated text detectors*, distinguishing input-level transformations from model-level adaptations. Finally, we discuss *evasion approaches in the Voight-Kampff task 2025*, which provides the most direct empirical context for our contribution. This structure allows us to position SHADE with respect to both the broader literature and the specific shared-task setting.

AI-generated Text Detection

Recent surveys characterize AI-generated text detection as a rapidly evolving field that spans several methodological families. Yang et al. [1] organize existing approaches around three main challenges: improving classifier training, exploiting intrinsic attributes of language models and embedding information such as watermarks into generated text. They further distinguish black-box detectors, which rely only on input–output behavior and observable textual characteristics, from white-box detectors, which assume access to internal model information such as architecture, parameters, logits, or generation procedures. This distinction is particularly relevant for evasion-oriented settings, where the attacker may only have access to detector outputs or surrogate feedback rather than to the target detector itself. This broader taxonomy also helps explain why detector evasion remains difficult to address with a single class of defenses. Training-based detectors depend on the representativeness of labeled data, intrinsic-attribute methods often rely on access to model probabilities or representations, and watermarking approaches require some form of information embedding during generation. As a result, detection performance can vary substantially depending on model access, computational assumptions, domains, and attack conditions.

The task of distinguishing machine-generated text from human-written text has been mainly approached from two broad directions: zero-shot statistical methods and supervised classification. Statistical detectors exploit regularities in the probability distribution of LLM-generated text, such as the tendency of generated text to exhibit lower entropy or higher likelihood under a reference model. Binoculars [3], for instance, performs classification by comparing perplexity-based scores from two language models that share a tokenizer, enabling zero-shot detection without labeled training data. AdaDetectGPT [4] extends perturbation-based detection with statistical guarantees, using adaptive thresholds to control false-positive rates across domains. Similarly, Seeliger et al. [13] proposed a statistically grounded approach for PAN 2025 [14]. Their method first computes world-level correlations from a binary-term document matrix and the associated human/AI labels, then maps each text into a sequence of correlation values interpreted as a signal. These correlation signals, combined with hand-crafted statistical features, provide a lightweight and interpretable alternative to detectors based on LLM backbones.

A second line of work frames AI-generated text detection as a supervised binary classification problem [2, 5, 12]. DetectRL [2] provides a benchmark for evaluating such detectors under more realistic conditions, including paraphrase attacks and domain shifts, and shows that many classifiers degrade substantially when exposed to adversarially modified inputs. OUTFOX [5] addresses this limitation by augmenting the training set with adversarially generated essays, i.e., texts explicitly generated to challenge the detector. In the PAN 2025 setting, mdok [12] ranked among the top-performing systems, combining an LLM backbone with robust fine-tuning to achieve strong performance in both binary and multiclass detection. The continued development of detection benchmarks at PAN 2026 [15] reflects the rapid evolution of the field. However, it also highlights an unresolved problem. As detectors become stronger, evasion methods may need to become adaptive rather than relying on fixed transformations or manually designed prompts.

Attacks on AI-generated Text Detectors

The problem of evading AI-generated text detectors can be situated withing the broader paradigm of adversarial machine learning [16], where an attacker modifies inputs of model behavior to induce misclassification. In the context of text detection, a common strategy is to apply input-level transformations to generated text. These transformations include character-level perturbations, synonym substitutions, paraphrasing and translation-based rewriting, with the goal of moving the text away from the distribution recognized by the detector [9]. Such methods can be effective because they alter superficial or statistical cues used by detectors. However, they may also introduce artifacts, reduce readability or produce unnatural phrasing. A recent survey [1] similarly identifies adversarially generated text as an emerging challenge for AI-generated text detection, noting that such texts can deceive

existing detectors and increase false-negative rates. This supports the need for evaluation settings in which detectors are tested not only on standard generated text, but also against adaptive or adversarial generation strategies.

A complementary strategy adapts the source model itself. Instead of modifying generated outputs after generation, model-level approaches fine-tune the generator so that its output distribution shifts toward texts that are less likely to be classified as machine-generated. Wang et al. [6] propose HUMPA, a proxy-attack strategy that fine-tunes a small surrogate model using Direct Preference Optimization (DPO) [17]. Their approach reduces detector performance substantially, achieving an average AUROC drop of 70.4% across multiple datasets. Importantly, this gain comes with a trade-off. more aggressive fine-tuning can reduce generation quality, making the regularization parameter β central for balancing evasion performance and text quality [6]. This finding is particular relevant to our work, as SHADE also optimizes a generator against detector feedback, but does so through reinforcement learning rather than preference optimization.

Evasion Approaches at Voight-Kampff Task 2025

The 2025 edition of the Voight-Kampff task [14] produced a range of evasion strategies that illustrate both the promise and the limitations of existing approaches. Creo et al. [9] evaluated four methods: Text Perturbation, Lost in Translation, Persona Immersion, and ADHD Writing. Text Perturbation applies character-level modifications to obscure the generated text. Lost in Translation translates the text into another language and then back into the original language, for example from English to Spanish and back to English, with the goal of shifting the distribution of the generated text. Persona Immersion and ADHD Writing, in contrast, are prompting-based methods that ask the LLM to imitate human writing either directly or through a specific cognitive or stylistic profile. The authors found that Lost in Translation achieved the best overall evasion performance, followed by Text Perturbation and ADHD Writing. These results suggest that translation-based rewriting can be more effective than direct prompt-based human imitation, likely because translation changes lexical and syntactic patterns more substantially than style instructions alone. At the same time, these methods remain post-processing strategies: they transform a generated text after it has been produced, rather than modifying the generation process itself.

Vachharajani [10] explored a related translation-based strategy using Hindi-to-Spanish re-translation. Although this approach is conceptually similar to Lost in Translation, the resulting texts exhibited awkward phrasing, repetition, grammatical errors, inconsistent idioms, and uneven sentence structure. These results suggest that translation-based evasion may succeed partly by introducing distributional shifts that detectors fail to capture, but that these shifts can come at the cost of fluency and textual quality. Finally, Gunti [11] proposed fine-tuning LLaMA 3 with LoRA on cleaned human-written corpora aiming to shift the model distribution toward human-written text. This data-centric strategy can preserve output quality because the model is trained on human-authored examples. However, it treats evasion as a static alignment problem, the model is optimized to resemble a human-text corpus rather than to adapt directly to detector feedback. Moreover, as it relies on the original PAN dataset as part of the training data, its effectiveness may depend on the match between the training distribution and the target evaluation setting.

In contrast, SHADE frames detector evasion as a dynamic optimization problem. Like model-level adaptation approaches, it fine-tunes the generator rather than perturbing outputs post hoc. However, unlike LoRA-based alignment to human corpora [11], SHADE does not require a corpus of human-written texts. Its objective is not to imitate existing human writing directly, but to optimize generation behavior against an adversarial surrogate detector. Moreover, unlike DPO-based proxy attacks [6], which optimize a generator offline using a fixed set of paired preferences, SHADE uses online reinforcement learning with detector confidence as a per-sample reward signal. Specifically, we adopt GRPO [18] and define the reward as the complement of the confidence score assigned by a static, pretrained adversarial detector. This places SHADE within the broader paradigm of Reinforcement Learning from AI Feedback, but in a setting where the feedback is used to optimize detector evasion. In this sense, our contribution is

to apply feedback-driven policy optimization to a problem that has more commonly been addressed through post-hoc transformations or static distributional alignment.

3. SHADE: Stochastic Human-like generation via Adversarial Detector Evasion

The goal of the Voight-Kampff task is to evaluate whether text generated by a language model can be made difficult to distinguish from human-written text. In the official evaluation setting, generated texts are assessed by unseen AI-detection classifiers. Since participants do not have access to these target detectors, the task can be viewed as a black-box evasion problem in which the generator must produce prompt-consistent texts that are unlikely to be classified as AI-generated.

3.1. Problem definition

Let $\mathcal{P}$ denote the prompt space and $\mathcal{Y}$ the space of possible text outputs. We define a parameterized language model generator $\pi_\theta : \mathcal{P} \to \mathcal{Y}$, which induces a distribution over outputs $y \in \mathcal{Y}$ for a given prompt $p \in \mathcal{P}$. Let $D_\phi : \mathcal{Y} \to [0, 1]$ be an unknown AI-detection classifier, where $D_\phi(y)$ denotes the probability that text y is classified as AI-generated. The objective is to learn parameters θ such that the generated texts satisfy the semantic and stylistic constraints of the input prompt while minimizing the detector confidence. Formally, we define the evasion objective as:

$$\theta^* = \arg\max_\theta \mathbb{E}_{p\sim\mathcal{P}} \left[\mathbb{E}_{y\sim pi_\theta(p)} \left[1 - D_\phi(y)\right]\right] \tag{1}$$

Since D_ϕ is not accessible during training, we approximate this objective using a surrogate detector $\hat{D}_\phi$. The resulting optimization problem therefore seeks to adapt the generation policy so that its outputs receive low AI-generated probability scores under the surrogate, while remaining close enough to the original model to preserve generation quality.

3.2. Solution

We propose SHADE, *Stochastic Human-like generation via Adversarial Detector Evasion*, a framework that combines adversarial learning [16] and reinforcement learning to adapt a base language model for detector evasion. The central idea is to modify the weights of the base model so that evasion-oriented generation behavior is internalized by the model, rather than applied through post-hoc perturbations, translation, or manually designed prompts. To train the model, we use the datasets from the Voight-Kampff tasks of Eloquent 2024, 2025, and 2026, together with the prompt instructions described in Section 3.3.

As the official AI-detection classifiers D_ϕ are not available, we train against a surrogate detector. In particular, we use a modified version of *mdok* [12], one of the best-performing classifiers in PAN 2025. While the original *mdok* system uses `Qwen3-14BBase` [19], our surrogate replaces this backbone with `FacebookAI/roberta-base` [20] due to GPU memory constraints. We refer to this surrogate as *mdok*-`roberta`. To approximate the behavior of the original mdok detector, we trained *mdok*-`roberta` on the PAN 2025 dataset. On the PAN 2025 validation set, *mdok*-`roberta` achieved an F1 score of 99%. We therefore use it as an adversarial oracle, assuming that detectors used in PAN 2026 [15] will exploit similar distributional regularities.

Given this surrogate detector, we fine-tune `LLaMA 3.2 3B` in its instruction-tuned version[1] [21]. We use Group Relative Policy Optimization (GRPO) [18] as the optimization algorithm. In this setting, π_θ denotes the model being optimized, π_{ref} denotes the original base model before reinforcement learning, and $\hat{D}_\phi(y)$ denotes the surrogate detector score. Since $\hat{D}_\phi(y)$ estimates the probability that y is AI-generated, we define the reward as $r(y) = 1 - \hat{D}_\phi(y)$. Thus, outputs that receive lower AI-generated probability scores obtain higher rewards.

[1] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

Given a training batch, let $r_i = r(y_i)$ be the reward assigned to the i-th generated text, and let $\text{mean}(\mathbf{r})$ and $\text{std}(\mathbf{r})$ denote the mean and the standard deviation of rewards across the current batch. Following the GRPO formulation, the advantage assigned to each token t in the i-th outputs is computed by normalizing the reward within the current group:

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})} \tag{2}$$

Then, the loss function is defined as:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{\sum_{i=1}^{G}|\mathcal{O}_i|}\sum_{i=1}^{G}\sum_{t=1}^{|\mathcal{O}_i|}\left[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{[\pi_\theta(o_{i,t} \mid q, o_{i,<t})]_{\text{old}}}\hat{A}_{i,t} - \beta\mathbb{D}_{\text{KL}}\left[\pi_\theta \,\|\, \pi_{\text{ref}}\right]\right] \tag{3}$$

This formulation follows the implementation provided by the *Hugging Face TRL* library[2], which slightly differs from the original GRPO objective. For the KL-divergence, we used the same unbiased estimator as in the original implementation[3]:

$$\mathbb{D}_{\text{KL}}\left[\pi_\theta \| \pi_{\text{ref}}\right] = \frac{\pi_{\text{ref}}(o_{i,t} \mid q, o_{i,<t})}{\pi_\theta(o_{i,t} \mid q, o_{i,<t})} - \log(\frac{\pi_{\text{ref}}(o_{i,t} \mid q, o_{i,<t})}{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}) - 1 \tag{4}$$

Although the KL-divergence term is usually omitted when training models using GRPO [22], we found that using a small β improves the textual properties of the generated outputs. We analyze this effect in Section 4.

3.3. Prompt modification

For each input instance, the Voight-Kampff task provides a base prompt that specifies the text-generation request. We adapt this task-provided prompt to the chat format expected by the instruction-tuned `LLaMA 3.2 3B` model. Specifically, we wrap the original prompt in a chat template and add system-level instructions intended to exploit the model's instruction following capabilities[4]:

```
[
    {"role": "system",
     "content":  "You are a helpful assistant that generates helpful
                  answers. You will avoid pleasantries and small talk,
                  focusing on the task at hand."},

    {"role": "system",
     "content": "You will avoid short paragraphs and bullet points."},

    {"role": "user",
     "content": BASE_PROMPT + "\n\nThis is the text"},

    {"role": "assistant",
     "content": ""},
]
```

This formatting allows the prompt to match the conversational structure expected by the instruction-tuned model while preserving the original task instructions.

[2] https://huggingface.co/docs/trl/grpo_trainer
[3] http://joschu.net/blog/kl-approx.html
[4] https://huggingface.co/docs/transformers/en/chat_templating

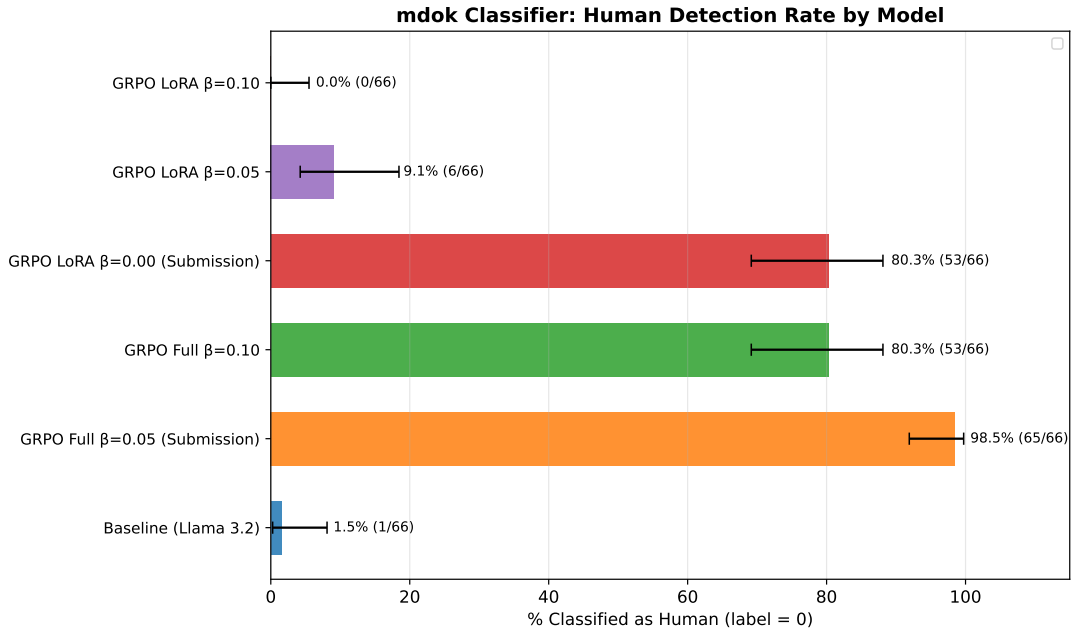

Figure 1: Generated text classified as human

3.4. Final text generation

At inference time, text generation is performed using the chat-based template described in Section 3.3. We use stochastic decoding to introduce controlled variation across candidate outputs. Sampling is controlled using temperature and top-p sampling. Temperature adjusts the randomness of the output distribution, while top-p sampling restricts generation to the smallest set of tokens whose cumulative probability exceeds the specified threshold. Generation continues until either an end-of-sequence condition is reached or the maximum number of new tokens has been produced. We set the temperature to 0.8, the top-p to 0.9, and the maximum number of new tokens to 1000.

For each prompt, we generate multiple candidate texts. Each candidate is evaluated using the surrogate detector, which assigns a score in $[0, 1]$ estimating the probability that the text was generated by an AI system. Lower scores indicate more human-like writing. The system iteratively generates candidates until either a text with an AI-generated probability below the acceptance threshold is found or the maximum number of attempts is reached. We set the acceptance threshold to 0.2 and the maximum number of attempts to 10. If no candidate satisfies the threshold, the candidate with the lowest surrogate detector score is selected as the final output.

4. Model analysis

When training SHADE, we evaluated several fine-tuning configurations. First, we compared parameter-efficient fine-tuning with LoRA against full fine-tuning, where all model weights are updated. Second, we varied the KL regularization coefficient β, which controls the penalty for deviating from the reference policy, using values $\beta \in \{0, 0.05, 0.1, 0.5, 1\}$. Configurations with $\beta = 0.5$ and $\beta = 1$ are not reported, as their KL-divergence term dominated the loss and prevented meaningful optimization of the evasion reward. We also omit the configuration with $\beta = 0$, since the absence of KL regularization led to degenerate generations characterized by repeated-token outputs.

Figure 1 reports the percentage of generated texts classified as human by the surrogate *mdok* detector for each evaluated configuration. The baseline LLaMA 3.2 model achieved a human classification rate of only 1.5% $(1/66)$, indicating that its outputs were almost always identified as AI-generated. GRPO substantially improved evasion, but its effectiveness depended strongly on the fine-tuning strategy and

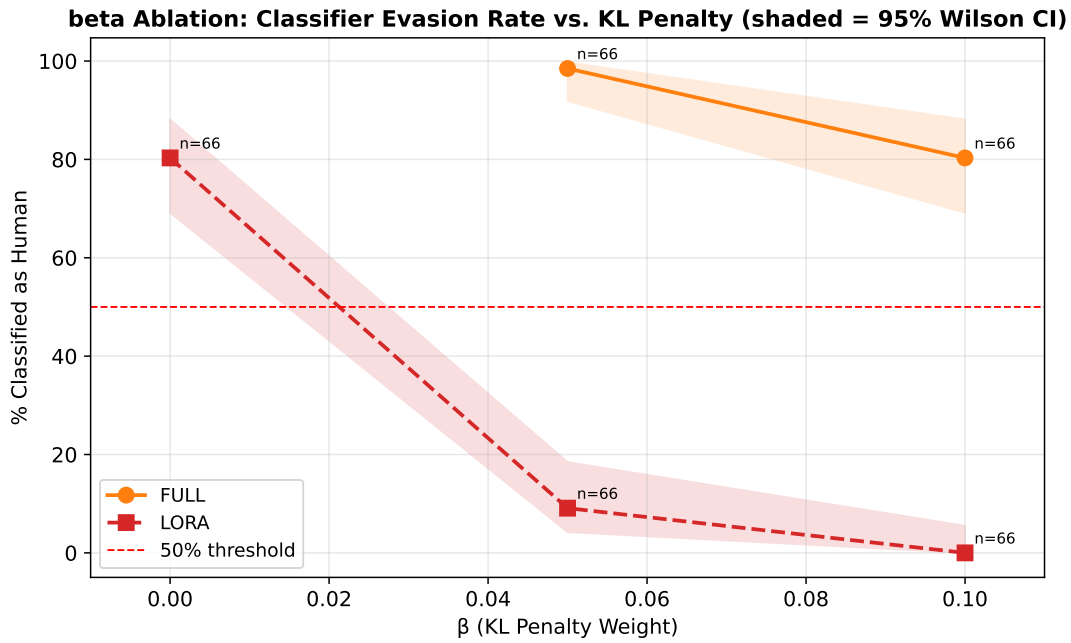

Figure 2: Beta vs. Confusion tradeoff

the KL penalty. Full fine-tuning with $\beta = 0.05$ achieved the strongest result, with 98.5% of outputs classified as human $(65/66)$. Increasing the penalty to $\beta = 0.10$ reduced the rate to 80.3% $(53/66)$, suggesting that stronger regularization limits the policy movement required for detector evasion.

The LoRA configurations followed a different pattern. Without KL regularization, LoRA achieved 80.3% human classification $(53/66)$, matching full fine-tuning with $\beta = 0.10$. However, adding regularization sharply reduced evasion, reaching only 9.1% $(6/66)$ for $\beta = 0.05$ and 0% $(0/66)$ for $\beta = 0.10$. Overall, only three configurations exceeded the 50% human-classification threshold: LoRA with $\beta = 0$, full fine-tuning with $\beta = 0.10$, and full fine-tuning with $\beta = 0.05$. Among these, full fine-tuning with $\beta = 0.05$ was the only configuration approaching near-complete surrogate evasion.

It is important to note that these results were obtained with the retry-based generation strategy described above. hen generation is performed in a one-shot setting, full fine-tuning with $\beta = 0.05$ achieves a human-classification rate of 59.1% $(39/66)$, substantially lower than with retry-based selection. A similar drop is observed for the LoRA configuration with $\beta = 0$, where only 30.3% $(20/66)$ of the generated texts are classified as human-written in the one-shot setting. These results indicate that the final evasion performance depends not only on the fine-tuned generator, but also on the inference-time selection strategy, which increases the probability of selecting outputs that better evade the surrogate detector.

Figure 2 summarizes the same results as a function of the KL coefficient, making the regularization trade-off more explicit. Full fine-tuning remains effective under moderate KL regularization, whereas LoRA rapidly loses evasion capability as β increases. This suggests that low-rank adapters are less able to simultaneously preserve proximity to the reference model and optimize the adversarial reward. The Wilson confidence intervals reflect the limited evaluation size ($n = 66$), but they do not alter the qualitative trend. Taken together, these results indicate that a small, non-zero KL penalty is beneficial in the full fine-tuning setting because it prevents degenerate policy drift while still allowing sufficient capacity for adversarial adaptation.

Figure 3 and Figure 4 illustrate the training dynamics for the full and LoRA fine-tuning strategies, respectively. In the full fine-tuning setting, both configurations rapidly increase the reward during the first epochs, indicating that the model learns to reduce the surrogate detector's AI-generated probability. The run with $\beta = 0.05$ achieves a slightly higher and more stable reward than $\beta = 0.10$, which is consistent with its stronger evasion performance in Figure 1. The KL divergence shows a large initial

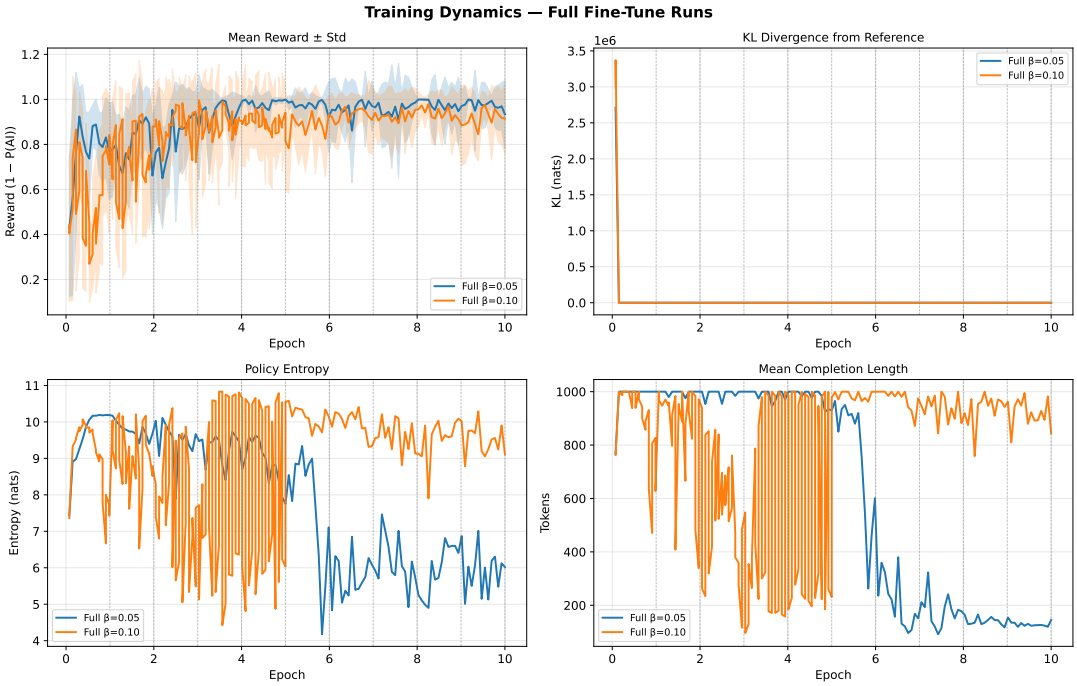

Figure 3: Training metrics full model

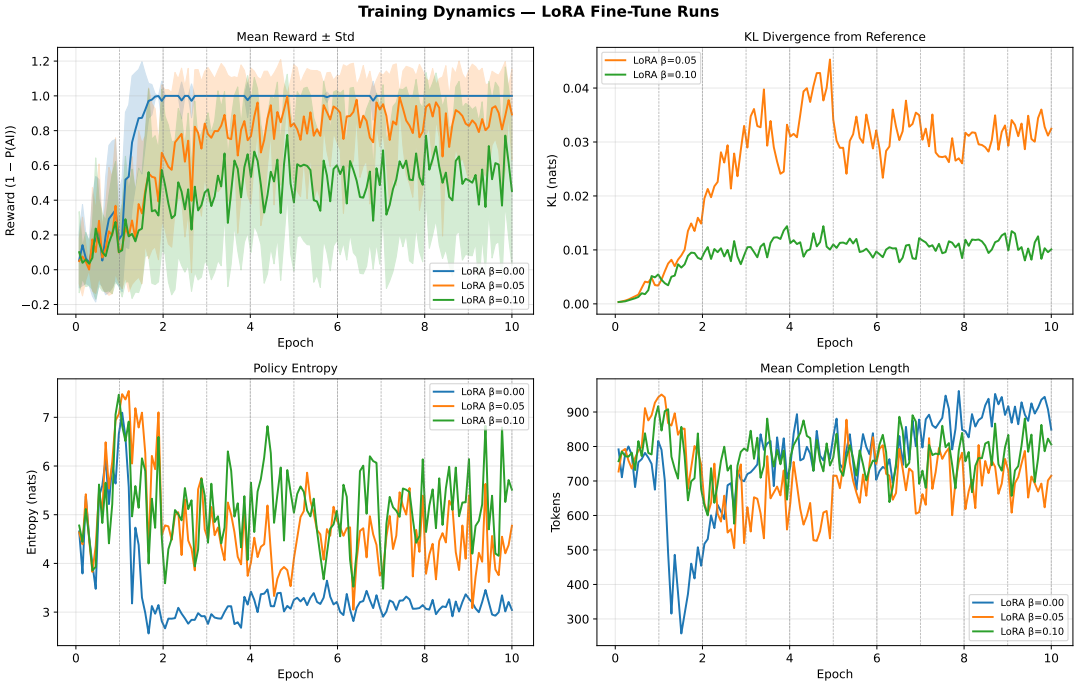

Figure 4: Training metrics LoRA model

spike and then quickly drops by several orders of magnitude, suggesting that most of the policy shift occurs early in training. After this initial adaptation, the model continues optimizing within a more stable region of the parameter space.

The entropy and completion-length curves further reveal how full fine-tuning adapts the generation policy. For $\beta = 0.05$, policy entropy decreases sharply after approximately five epochs, while the average completion length drops from near the maximum generation length to much shorter outputs. This suggests that the best-performing configuration converges toward a more specialized generation

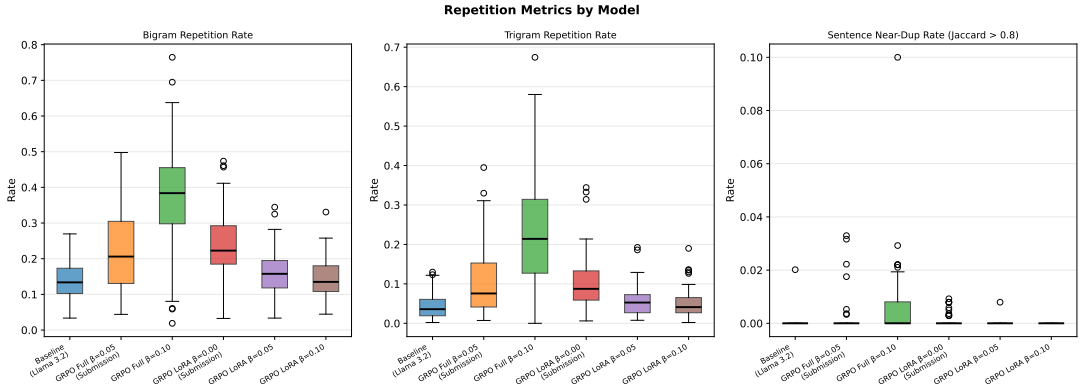

Figure 5: Repetitions within the generated text

strategy that the surrogate detector tends to classify as human-written. In contrast, the $\beta = 0.10$ configuration maintains higher entropy and longer completions throughout training, indicating that stronger KL regularization preserves more of the original generation behavior but also limits evasion effectiveness.

The LoRA runs exhibit a different pattern. The unregularized LoRA configuration ($\beta = 0$) reaches a high reward early in training, but this is accompanied by rapid entropy collapse and large fluctuations in completion length, suggesting a less stable policy. When KL regularization is introduced the LoRA configurations maintain higher entropy and more stable completion lengths, but their rewards remain substantially lower, especially for $\beta = 0.10$. Moreover, the KL divergence for the regularized LoRA runs remains small throughout training, indicating that the adapters induce only limited deviations from the reference policy. This helps explain the reduced evasion performance observed in Figure 1 and Figure 2. Under KL regularization, LoRA appears unable to modify the generation policy enough to consistently evade the surrogate detector.

Taken together, the training curves suggest that successful evasion requires a pronounced shift in generation behavior. Full fine-tuning provides enough capacity for such a shift, particularly when a small KL penalty prevents degenerate policy drift while still allowing adaptation. In contrast, LoRA offers insufficient adaptation capacity once proximity to the reference model is enforced. This supports the interpretation that detector evasion is not simply a matter of preserving fluent generation while lowering detector confidence. Rather, the model appears to discover specific generation regimes that exploit the surrogate detector.

Figure 5 analyzes repetition patterns in the generated texts using bigram repetition, trigram repetition, and sentence near-duplication metrics. The baseline model exhibits the lowest repetition levels overall, suggesting that the original instruction-tuned model produces more locally diverse outputs. In contrast, the full fine-tuned GRPO models introduce substantially higher repetition rates, particularly for $\beta = 0.10$, which shows the highest median bigram and trigram repetition among all configurations. This indicates that full-policy adaptation can alter local lexical structure in ways that increase repeated n-gram patterns. Interestingly, this effect is stronger for $\beta = 0.10$ than for the best-performing $\beta = 0.05$ configuration, suggesting that higher repetition is not itself sufficient to explain surrogate evasion.

The LoRA models generally preserved lower repetition levels than full fine-tuning, especially when KL regularization is applied. This is consistent with the earlier observation that regularized LoRA remains closer to the reference policy, but also achieves substantially lower evasion rates. Sentence-level near-deduplication remains close to zero across most configurations, with only isolated outliers. Thus, the observed degeneration mainly occurs at the local lexical level through repeated bigrams and trigrams, rather than through full sentence reuse. The contrast between the two full fine-tuning runs is particularly informative. Although $\beta = 0.10$ produces more repetitive text, it achieves lower evasion than $\beta = 0.05$. This suggests that SHADE's strongest evasion performance is not simply caused by repetition, but by a broader shift in generation behavior.

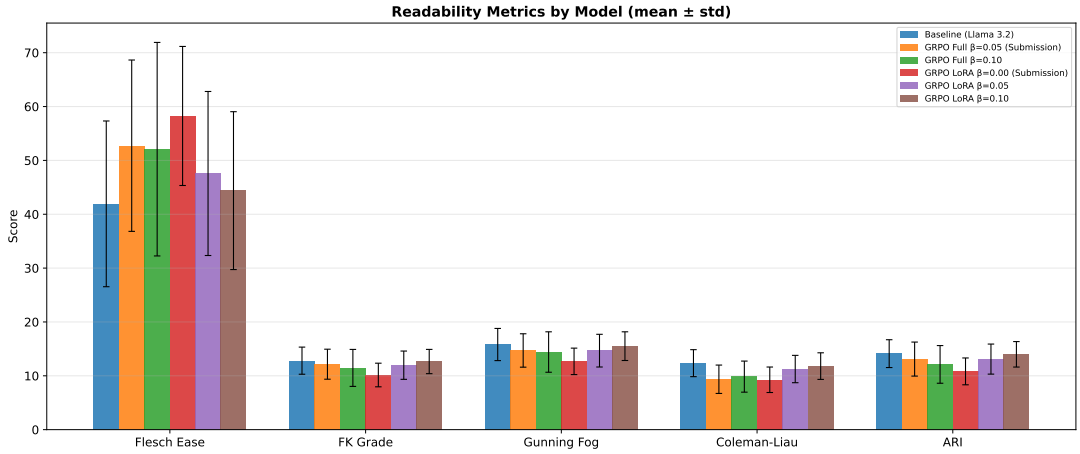

Figure 6: Generated text readability metrics

Table 1
Comparison of human-likeness and linguistic statistics across baseline and GRPO-trained models.

Model	Cls Err	Words	Sentences	TTR	Hapax	Flesch Ease	FK Grade	Bigram Rep	Trigram Rep
Baseline (LLaMA 3.2)	1.5%	551 ± 26	22.4	0.478	0.339	41.9	12.8	0.1376	0.0426
GRPO Full $\beta = 0.05$	98.5%	374 ± 136	13.8	0.414	0.258	52.7	12.2	0.2231	0.1060
GRPO Full $\beta = 0.10$	80.3%	434 ± 159	18.6	0.329	0.189	52.1	11.5	0.3605	0.2140
GRPO LoRA $\beta = 0.00$	80.3%	559 ± 22	25.0	0.377	0.227	58.3	10.1	0.2387	0.1057
GRPO LoRA $\beta = 0.05$	9.1%	554 ± 34	23.5	0.454	0.311	47.6	12.0	0.1579	0.0550
GRPO LoRA $\beta = 0.10$	0.0%	557 ± 23	22.2	0.473	0.336	44.4	12.7	0.1438	0.0503

Figure 6 presents readability statistics across the evaluated models. Overall, the GRPO-based configurations tend to produce texts with higher Flesch Reading Ease scores and lower FK Grade values than the baseline, indicating a shift toward simpler and more accessible language. This pattern is especially visible for the high-evasion configurations. GRPO full fine-tuning with $\beta = 0.05$ increases Flesch Ease from 41.9 to 52.7 while slightly reducing FK Grade from 12.8 to 12.2, and GRPO LoRA with $\beta = 0$ achieves the highest Flesch Ease score and (58.3) and the lowest FK Grade (10.1). However, readability alone does not fully explain evasion performance. For instance, the LoRA $\beta = 0$ configuration is the most readable but reaches only 80.3% evasion, while full fine tuning with $\beta = 0.05$ achieves 98.5% evasion with less extreme readability changes.

Table 1 provides a broader view of these shifts. The best-performing configuration, GRPO full fine-tuning with $\beta = 0.05$, produces shorter texts than the baselines (374 ± 136 vs. 551 ± 26 words), with fewer sentences, lower type-token ratio, and lower hapax proportion. These changes suggest that detector evasion is associated with a reduction in lexical richness and structural complexity. At the same time, the comparison between full fine-tuning with $\beta = 0.05$ and $\beta = 0.10$ shows that excessive repetition is not sufficient for stronger evasion. The $\beta = 0.10$ model has substantially higher bigram and trigram repetition rates, but lower evasion. Taken together, the readability and lexical metrics suggest that SHADE succeeds by moving the generation policy toward simpler, shorter, and less lexically diverse outputs, while avoiding the strongest forms of local degeneration.

5. Voight-Kampff Task Result

In addition to the surrogate-detector evaluation, we also analyze the results provided by the Voight-Kampff competition organizers. Unlike the previous experiments, which evaluate generated texts against our mdok-based surrogate detector, the competition setting measures how closely each submission resembles source-like behavior under the official evaluation procedure. Therefore, these results provide a direct estimate of how well the proposed evasion strategy transfers from the surrogate detector to the unseen evaluation setting. Our two submitted runs, *DArgk - Full fine-tuned* $\beta = 0.05$ and *DArgk - LoRA*

$\beta = 0$, ranked in the middle of the ranking in positions sixth and seventh, respectively. The *DArgk - Full fine-tuned* $\beta = 0.05$ submission obtained a mean score of 0.320, with 337 lower, 432 equal, and 1811 higher comparisons. The *DArgk - LoRA* $\beta = 0$ submission obtained a slightly worse mean score of 0.338, with 269 lower, 402 equal, and 1909 higher comparisons.

The gap between our surrogate results and the competition ranking highlights an important limitation of detector-specific reinforcement learning. While SHADE substantially reduces the confidence of the surrogate detector, this advantage does not fully transfer to the official evaluation setting. This suggests that the model may have learned generation patterns that exploit the surrogate detector in particular, rather than detector-invariant properties of human-written text. The *DArgk - Full fine-tuned* $\beta = 0.05$ run slightly outperformed the submitted *DArgk - LoRA* $\beta = 0$ run, which is consistent with our internal finding that full fine-tuning provides more effective adaptation than parameter-efficient tuning. However, the relatively small difference between the two submitted runs in the official ranking suggests that the advantages observed against the surrogate detector are attenuated when evaluated against unseen classifiers. Overall, the competition results indicate that SHADE is effective as a surrogate-targeted evasion method, but that improving cross-detector generalization remains a central challenge.

6. Conclusion

This paper presented SHADE, a reinforcement learning framework for generating texts that evade AI-generated text detectors. Instead of relying on post-hoc perturbations, translation, or manually designed prompting strategies, SHADE formulates detector evasion as a policy optimization problem. Using a surrogate detector based on the PAN 2025 *mdok* system, we fine-tuned an instruction-tuned LLaMA model with GRPO and optimized the generator to reduce the surrogate detector's confidence that the generated text was AI-authored. Overall, SHADE demonstrates both the potential and the risks of reinforcement learning for detector evasion. The approach shows that LLM generation policies can be optimized to exploit detector feedback, but also that high surrogate evasion does not guarantee robust transfer to unseen evaluation settings or human-like text quality. These findings highlight the need for more robust, adaptive, and transparent evaluation protocols for AI-generated text detection.

Several limitations of this study suggest directions for future work. First, our reinforcement learning objective relies on a single surrogate detector, which may bias the generator toward detector-specific artifacts. Future work should reduce this risk by training against ensembles of heterogeneous surrogate detectors, including statistical, neural, and watermark-aware models, or by periodically updating the surrogate detector in an adaptive adversarial training loop. Second, the evaluation set used for the internal analysis is relatively small, which limits the statistical strength of some comparisons. This limitation could be addressed by evaluating the method on larger benchmark collections and by testing whether the observed trends hold across multiple domains and datasets. Third, the reward function focuses primarily on detector evasion and does not directly optimize for semantic fidelity, fluency, coherence, or human judgments of naturalness. Although the KL penalty partially constrains the model, future work should incorporate multi-objective rewards that jointly balance evasion, semantic preservation, readability, lexical diversity, and fluency, complemented by human evaluation to determine whether the generated outputs are genuinely perceived as human-written. Finally, our analysis is limited to a specific base model, detector family, and shared-task setting. Future work should therefore investigate the transferability of learned evasion strategies across different LLMs, languages, domains, and unseen detector families.

Ethical considerations. This work addresses AI-generated text detector evasion, a topic with clear dual-use implications. While the proposed method can help evaluate the robustness of current detectors, similar techniques could be misused to conceal AI authorship in contexts where originality, accountability or trust are important. We therefore frame SHADE as a robustness evaluation approach and evaluate it within the controlled setting of the Voight-Kampff shared task. Our analysis reports not only evasion performance but also limitations and side effects. These results emphasize that high detector

evasion should not be interpreted as evidence of genuinely human-like writing. More broadly, the findings caution against relying on single-detector decisions in high-stakes settings such as education, scientific publishing, journalism or content moderation. Future work should emphasize multi-detector robustness evaluation, transparent uncertainty reporting, human oversight and safeguards against deceptive deployment.

7. Artifacts

The following are the required artifacts to replicate the submission:

- Source code for training and generating the submission, including pretrained weights for the mdok-roberta model: https://github.com/knife982000/voight-kampff-2026.
- Pretrained generative model Full fine-tuned weights $\beta = 0.05$: https://huggingface.co/jmrodri/Llama-3.2_voight-kampff_beta_005.
- Pretrained generative model LoRA weights $\beta = 0$: https://huggingface.co/jmrodri/Llama-3.2_voight-kampff_lora_beta_000.
- PAN 2025 Voight-Kampff AI Detection dataset: https://zenodo.org/records/14962653
- Eloquent 2024, 2025, and 2026 Voight-Kampff dataset: https://huggingface.co/datasets/Eloquent/Voight-Kampff

8. Declaration on Generative AI

Generative AI tools were used solely to assist with grammar correction and language polishing. All conceptual content, analysis, and final decisions remain the authors' responsibility.

Acknowledgments We are grateful for the support of CLAAUDIA through AI-Cloud for providing the computational infrastructure. This research was funded in whole or in part by the Austrian Science Fund (FWF): 1255776/COE12.

References

[1] Z. Yang, Z. Feng, R. Huo, H. Lin, H. Zheng, R. Nie, H. Chen, The imitation game revisited: A comprehensive survey on recent advances in ai-generated text detection, Expert Systems with Applications 272 (2025) 126694. URL: https://www.sciencedirect.com/science/article/pii/S0957417425003161. doi:https://doi.org/10.1016/j.eswa.2025.126694.

[2] J. Wu, R. Zhan, D. Wong, S. Yang, X. Yang, Y. Yuan, L. S. Chao, Detectrl: Benchmarking llm-generated text detection in real-world scenarios, in: A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (Eds.), Advances in Neural Information Processing Systems, volume 37, Curran Associates, Inc., 2024, pp. 100369–100401. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/b61bdf7e9f64c04ec75a26e781e2ad51-Paper-Datasets_and_Benchmarks_Track.pdf. doi:10.52202/079017-3186.

[3] 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: R. Salakhutdinov, Z. Kolter, K. A. Heller, A. Weller, N. Oliver, J. Scarlett, F. Berkenkamp (Eds.), Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024, Proceedings of Machine Learning Research, PMLR / OpenReview.net, 2024, pp. 17519–17537. URL: https://proceedings.mlr.press/v235/hans24a.html.

[4] H. Zhou, J. Zhu, P. Su, K. Ye, Y. Yang, S. Gavioli-Akilagun, C. Shi, Adadetectgpt: Adaptive detection of llm-generated text with statistical guarantees, in: D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, N. Chen (Eds.), Advances in Neural Information Processing Systems,

volume 38, Curran Associates, Inc., 2025, pp. 89077–89118. URL: https://proceedings.neurips.cc/paper_files/paper/2025/file/808bc5e3d076c1125a87f81b42d5e52d-Paper-Conference.pdf.

[5] R. Koike, M. Kaneko, N. Okazaki, OUTFOX: llm-generated essay detection through in-context learning with adversarially generated examples, in: M. J. Wooldridge, J. G. Dy, S. Natarajan (Eds.), Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, AAAI Press, 2024, pp. 21258–21266. URL: https://doi.org/10.1609/aaai.v38i19.30120. doi:10.1609/AAAI.V38I19.30120.

[6] T. Wang, Y. Chen, Z. Liu, Z. Chen, H. Chen, X. Zhang, W. Cheng, Humanizing the machine: Proxy attacks to mislead llm detectors, in: Y. Yue, A. Garg, N. Peng, F. Sha, R. Yu (Eds.), International Conference on Learning Representations, volume 2025, 2025, pp. 68339–68367. URL: https://proceedings.iclr.cc/paper_files/paper/2025/file/ab1ee157f7804a13f980414b644a9460-Paper-Conference.pdf.

[7] R. Gunti, B. Bayramoğlu, J. M. S. Dilinika, 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 at the ELOQUENT 2026 lab for evaluating generative language model quality, in: E. S. Salido, A. B.-C. no, 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.

[8] 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.

[9] A. Creo, M. Hormazábal-Lagos, H. Cerezo-Costas, P. A. Doval, Fake it 'til you make it human, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 1381–1390. URL: https://ceur-ws.org/Vol-4038/paper_109.pdf.

[10] P. Vachharajani, Literal re-translation as a method for AI text disguise and detection evasion, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 1443–1448. URL: https://ceur-ws.org/Vol-4038/paper_116.pdf.

[11] R. R. Gunti, The distinctive data-centric approach for the voight kampff task, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 1391–1400. URL: https://ceur-ws.org/Vol-4038/paper_110.pdf.

[12] 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 the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 3819–3826. URL: https://ceur-ws.org/Vol-4038/paper_307.pdf.

[13] M. Seeliger, P. Styll, M. Staudinger, A. Hanbury, Human or not? light-weight and interpretable detection of ai-generated text, Working Notes of CLEF (2025).

[14] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), volume 16089

of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2025, pp. 388–411. doi:10.1007/978-3-032-04354-2_21.

[15] J. Bevendorff, M. Fröbe, A. Greiner-Petter, A. Jakoby, M. Mayerl, P. Nakov, H. Plutz, M. 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 – Extended Abstract, in: R. Campos, A. Jatowt, Y. Lan, M. Aliannejadi, C. Bauer, S. MacAvaney, Z. Ren, S. Verberne, N. Bai, M. Mansoury (Eds.), Advances in Information Retrieval. 48th European Conference on IR Research (ECIR 2026), volume 16485 of *Lecture Notes in Computer Science*, Springer Nature, Cham, Switzerland, 2026, pp. 225–232. doi:10.1007/978-3-032-21321-1_32.

[16] D. Lowd, C. Meek, Adversarial learning, in: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD '05, Association for Computing Machinery, New York, NY, USA, 2005, p. 641–647. URL: https://doi.org/10.1145/1081870.1081950. doi:10.1145/1081870.1081950.

[17] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, C. Finn, Direct preference optimization: your language model is secretly a reward model, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[18] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, D. Guo, Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL: https://arxiv.org/abs/2402.03300. arXiv:2402.03300.

[19] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[21] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[22] J. Hu, Y. Zhang, Q. Han, D. Jiang, X. Zhang, H.-Y. Shum, Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model, in: D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, N. Chen (Eds.), Advances in Neural Information Processing Systems, volume 38, Curran Associates, Inc., 2025, pp. 162239–162262. URL: https://proceedings.neurips.cc/paper_files/paper/2025/file/ed873d79e7c268c020c4b4db13a2812a-Paper-Conference.pdf.