UTS at ELOQUENT 2026 Voight-Kampff: structural shifts in AI writing bypass state-of-the-art detectors

Dima Galat[1], Marian-Andrei Rizoiu[1]

[1]University of Technology Sydney, Australia

Abstract

We investigate which language model evasion attacks survive state-of-the-art adversarial fine-tuning, developing strategies that sweep the top 5 positions on the ELOQUENT 2026 Voight-Kampff leaderboard. While adversarial fine-tuning trivially closes the 2025 winning evasion recipes, we uncover a fundamental asymmetry in detector vulnerability: pushing generated text *out of* the detector's training distribution reliably defeats adversarial detection, whereas pulling it *into* the distribution (e.g., mimicking human training data) fails completely. Exploiting this, we introduce two novel out-of-distribution attack families—cross-decade register attacks and modernist stream-of-consciousness form. Both strategies easily bypass adversarial closure, achieving up to $\approx 50\times$ higher fool rates than previous methods while preserving naturalness. Furthermore, experiments show that the obvious deployer counter-measure (augmenting training data with period prose) fails to close the vulnerability. Our findings show that the tested detector families, including adversarially fine-tuned ones, exhibit persistent vulnerabilities under structural out-of-distribution shifts, a mechanism that directly powers our leading competition performance.

Keywords

AI-generated text detection, Voight-Kampff, PAN, ELOQUENT, adversarial fine-tuning, register-shift attacks

1. Introduction

The ELOQUENT Voight-Kampff (VK) shared task at CLEF [1, 2] asks a binary question: was a text written by a human or generated by a language model? In its 2024 edition the best-system fool rate against the PAN'25 detector panel was 0.049 [3]. The 2025 edition saw a $\sim 13\times$ jump to 0.654 [4], driven by two prompting-side recipes that did not retrain or fine-tune anything: literal Hindi back-translation plus an explicit *be-imperfect* instruction [5], and a nine-language rotation [6]. The structural question for 2026 is *which kinds of evasion attacks survive a deployer-class adversarial fine-tune that has already seen the 2025 attacks?* An attack family that survives is durable in a builder-breaker arms race; one that does not is closable inside a single fine-tune pass. Our central finding is an asymmetry along this axis.

We approach this with three coupled experiments. **First**, we replicate the 2025 attacks on four generators (Claude Opus 4.7, GPT-5.4, Gemini 2.5 Pro, Qwen3.6-35B-A3B) crossed with five strategies and 66 topics (n=1,320 generations) and train a family of deployer-class adversarial detectors (RoBERTa-large fine-tuned on our generations + PAN'25 humans) in three CV variants: *Macko-replica* (leave-one-genre-bucket-out, four folds), *Macko-LOSO* (leave-one-strategy-out, the harder deployer-pessimistic variant, five folds), and *Macko-pp* (Macko-LOSO + pre-1923 Project Gutenberg humans on the negative side, the obvious deployer counter-measure tested in §4.4). The naming follows the 2025 PAN winner's recipe of adversarial training plus OOD validation [7, 4]. **Second**, we introduce two new attack families: *cross-decade register* attacks (prompts that ask the generator to write in early-twentieth-century English literary register), and *stream-narrative* attacks (forcing a modernist stream-of-consciousness narrative form). We test whether they survive Macko-replica's and Macko-LOSO's closure. **Third**, we test the obvious deployer counter-fix (Macko-pp): augment the human-side training corpus with pre-1923 Project Gutenberg prose so period register stops being out-of-distribution.

The asymmetry is stark. While the 2025 PJS recipe inverts under adversarial detection (its effectiveness swinging from $+0.71$ to -0.71 on Claude Opus), our *cross-decade register* attack holds strong at a 0.798

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

dima.galat@student.uts.edu.au (D. Galat); Marian-Andrei.Rizoiu@uts.edu.au (M. Rizoiu)

0000-0003-3825-2142 (D. Galat); 0000-0003-0381-669X (M. Rizoiu)

Macko-LOSO fool rate on the same generator and topics. Our *stream-narrative* attack similarly achieves a robust 0.65 escape rate against an even stronger adversarial baseline. Crucially, the obvious deployer counter-measure—augmenting the human-side training data with pre-1923 prose—fails to close the cross-decade vulnerability. The underlying mechanism is clear: pushing generations *out of* the deployer's training distribution defeats adversarial fine-tuning, whereas pulling them *into* it does not.

Findings.

1. **Adversarial fine-tuning easily neutralizes superficial evasion.** Once the detector is retrained on generations that rely on simple grammar-level perturbations (like the 2025 winning recipes), their effectiveness collapses to near-zero (fool rates ≤ 0.025). Simple, superficial tweaks offer no long-term durability.
2. **Structural shifts in register and form remain highly resilient.** Prompting the generator to write in a historical register (early-twentieth-century English) or forcing a modernist stream-of-consciousness narrative style completely bypasses detector retraining. These strategies achieve up to a $50\times$ improvement in evasion over previous methods, while maintaining a natural, human-like flow.
3. **Simple data-patching defenses fail.** The obvious counter-measure for the detector builder—retraining the classifier with real historical prose added to its training pool—does not close the gap. Retraining on Project Gutenberg passages actually slightly *increases* the evasion rate of our historical-register attack to 0.846.
4. **Evasion is driven by structural out-of-distribution direction, not surface mimicry.** Detectors are not fooled because the generator successfully mimics the exact distribution of human test texts (few-shot mimicry fails completely). Instead, evasion succeeds because the text is pushed entirely *outside* of the domain the detector was calibrated to recognize, showing that the tested detector families exhibit persistent vulnerabilities under structural out-of-distribution shifts.

We release all detector checkpoints and attack prompts.

2. Related Work

The Voight-Kampff task across three editions. The PAN+ELOQUENT VK task formalises machine-generated text detection inside a builder-breaker shared evaluation [3, 8]. The first edition (2024) used the easiest verification level; only one of four ELOQUENT submissions beat the unmodified GPT-3.5 baseline. The 2025 edition [4] switched to single-text decision (level 7) and saw the $\sim 13\times$ best-system jump while the same generators were available; the improvement was driven by prompting strategy and post-processing, not generator capability. The 2026 edition, run jointly by ELOQUENT and PAN [1, 2, 9, 10], retains level 7 with a literary / historical / press text distribution.

2025-winning generation-side breakers. PJS-team [5] won 2025 with a two-stage pipeline: literal Hindi back-translation followed by a return-leg *be-imperfect* instruction. HumanAIzers [6] reached 0.642 with a related nine-language rotation. Both treat the detector panel as a black box; both manipulate the generation process alone.

Detection: supervised, zero-shot, and adversarial. The 2025 PAN detector winner [7] reached mean inverse-fool 0.989 by adversarially fine-tuning a Qwen3-14B classifier with QLoRA plus homoglyph augmentation, validated on a 2,000-instance OOD set (MIX2k). The PAN'25 organiser baseline, a TF-IDF 1–4-gram SVM, reached 0.922 and outperformed most fine-tunes on held-out validation. Earlier work on visualising machine-text statistics [11] laid the foundation; subsequent zero-shot methods include perturbation-based curvature detection [12] and the observer/performer perplexity ratio [13]. Sadasivan et al. [14] and Krishna et al. [15] argue that paraphrase attacks reduce achievable detection AUC and, in the limit, prevent reliable detection without watermarking [16]. Our cross-decade attack

falls inside that family in spirit but operates through register shift rather than paraphrase; framed as a controlled style-transfer instance [17], with the cross-decade direction chosen because it sits OOD relative to the deployer's training corpus. What we add is a rigorous test of whether the obvious deployer counter-measure (retraining on the OOD register) actually closes the gap.

3. System and Method

3.1. Generators

Four generators across two tiers: **frontier closed** — Claude Opus 4.7, GPT-5.4, Gemini 2.5 Pro; and **open-MoE** — Qwen3.6-35B-A3B in 4-bit [18]. All pinned to dated snapshots.

3.2. Strategies

The full strategy set across the 13 attacks reported in this paper falls into four families. (i) **2025-replication strategies**: `vanilla`, `imperfection`, `roundtrip`, `roundtrip_imperf` (PJS v2), `lost_in_translation` (9-language rotation, 6). Strategies 2–4 form a 2×2 factor on (`has_translation`, `has_imperfection`). (ii) **Cross-decade register** (Attack 1): `style_pre1923`, `style_translated_ru`, `style_l2_academic`. (iii) **Source-anchored, author-style, and form** (Attacks 9, 10): `synth_anchor`; named-author imitations (Borges, Sebald, Dillard); and `stream_narrative`. (iv) **Bracketing** (Attacks 11–13): the Macko-pp retraining experiment; `real_anchor`; `pan_anchor`.

We lock the round-trip translation language per generator: Hindi for the closed-frontier models (matching the PJS team's choice), and Chinese for Qwen3.6 (matching its native language capabilities). These strategies are evaluated across 66 topics from VK 2024 / 2025 / 2026, with the 20-topic 2026 test split held out exclusively for our cross-decade and bracketing experiments.

3.3. Detectors

Five primary detectors and three adversarial variants. **Primary** (in-distribution-tuned, not retrained on our pool): TF-IDF 1–4-gram SVM (PAN'25 organiser baseline), RoBERTa-base [19] fine-tuned on PAN'25 train (val AUC 0.9994), Binoculars [13] on a Mistral-7B pair, LogPerplexity- and LogRank-GPT-2M.

The Macko-replica family (deployer-class). Three RoBERTa-large adversarial fine-tunes on our generations + PAN'25-train humans.[1] **Macko-replica** (bucket-CV) holds out one of four genre buckets per fold; **Macko-LOSO** (strategy-CV) holds out one of the five 2025-replication strategies per fold (novel 2026 strategies fall back to the `roundtrip_imperf` fold, the deployer-pessimistic choice); **Macko-pp** adds ∼1,000 pre-1923 Gutenberg passages [20] to the human side — the obvious deployer counter-measure for the cross-decade attack of §4.4.

4. Results

4.1. The OOD inversion under the 2025-winning recipe

Table 1 reports the paired Δ fool rate of `roundtrip_imperf` (PJS v2) versus `vanilla` across the 5×5 detector-by-generator cells.

The PAN'25-trained RoBERTa is the strongest in-distribution detector we have (0.9994 val AUC). On vanilla Claude its mean $P(\text{LLM})$ is ≈ 1.00. Pass the same prompt through PJS v2 and the same detector returns $P(\text{LLM}) \approx 0.001$. The signal does not merely degrade — *it inverts*. The same input shifts Binoculars by only $+0.07$ to $+0.10$.

[1] Checkpoints released on HuggingFace: https://huggingface.co/protagonist/eloquent26-macko-replica, https://huggingface.co/protagonist/eloquent26-macko-replica-loso, and https://huggingface.co/protagonist/eloquent26-macko-pp.

Detector	Cl.	GPT	Gem.	Qw3.6
TF-IDF SVM	+.17	+.34	+.52	+.39
RoBERTa	**+.71**	**+.60**	**+.91**	**+.54**
Binoc-Mi7B	+.10	+.10	+.07	+.04
LogPerp-GPT2M	+.22	+.41	+.18	+.33
LogRank-GPT2M	+.15	+.30	+.13	+.24

Table 1
The 2025 OOD inversion. RoBERTa, our strongest in-distribution detector, is the most-fooled by PJS v2 across all four generators. Paired Δ fool rate vs. `vanilla`, n=66; bootstrap-significant at α=0.05.

	RoBERTa fool	Macko fool	Macko Δ vs. vanilla
Claude Opus	0.99	**0.025**	−0.71
GPT-5.4	0.96	**0.012**	−0.52
Gemini 2.5	0.99	**0.025**	**−0.74**
Qwen3.6	0.57	**0.014**	−0.50

Table 2
Macko-replica vs. baseline RoBERTa-PAN'25 fool rate (↑ better attack) per generator under `roundtrip_imperf`. The recipe inverts sign on every generator (sig. at α=0.05 paired bootstrap on the closed frontier and Qwen3.6); negative Δ means PJS v2 makes the text *more* detectable than vanilla.

Underlying Mechanism. The variance in strategy effectiveness does not align along the supervised vs. zero-shot boundary. Rather, it splits along the type of feature extracted: *n-gram-surface vs. contextual-distributional*. The n-gram supervised detector (TF-IDF SVM) is fooled primarily by surface-level strategies like `imperfection`. Conversely, the four contextual detectors (RoBERTa, Binoculars on Mistral-7B, LogPerp on GPT-2M, and LogRank on GPT-2M, representing three distinct model families) are fooled by `roundtrip`. The combined `roundtrip_imperf` (PJS v2) attacks both fronts. This cross-architecture alignment indicates that, within the tested families, the inversion is a property of contextual scoring rather than a quirk of any single model architecture.

4.2. Adversarial closure: the Macko-replica

We fine-tune `FacebookAI/roberta-large` on our generations + matched PAN'25 humans, leave-one-genre-bucket-out, four checkpoints (§3.3). Each text is scored only by the checkpoint that did not see its bucket during training. Table 2 reports the result.

The recipe is self-defeating once the deployer has fine-tuned on even a small subset of its outputs: applying the 2025 evasion strategy to our retrained detector makes the text *more* detectable than vanilla writing, not less. Our prediction of a post-adversarial inverse C@1 score below 0.20 is confirmed, with the empirical detection error rate dropping below 0.03 across all tested generators.

4.3. Cross-decade register attacks survive Macko-replica

The obvious response to §4.2's deployer-positive finding is to ask whether *any* attack survives. We hypothesise that perturbations at the *register* level (style, vocabulary epoch, native-vs-L2 grammar) exit the training distribution further than PJS-style grammatical perturbation, and survive an adversarial fine-tune that contains modern PJS-style outputs but no pre-1923 / translated-Russian / L2-English corpus.

We add three single-pass strategies on the closed frontier: `style_pre1923` (early-twentieth-century English novelistic register), `style_translated_ru` (Constance-Garnett-era Russian-translation register), `style_l2_academic` (non-native English academic register). All three are run on the 20 held-out VK-2026 test topics on Claude Opus.

Strategy	RoB.	TF-IDF	M-bk	M-LOSO
`rt+imperf`	0.98	0.92	0.05	0.015
`style_pre1923`	0.15	0.58	0.50	0.640
`style_trans_ru`	0.07	0.45	0.42	0.483
`style_l2_acad`	0.03	0.38	0.31	0.395
`synth_anchor`	**0.57**	**0.76**	**0.70**	**0.798**

Table 3
Cross-decade register attacks on Claude Opus / 20 held-out VK-2026 test topics; fool rate ↑ better attack. M-LOSO = Macko-LOSO (strategy-CV adversarial); M-bk = Macko-replica (bucket-CV variant). Compare to `roundtrip_imperf`'s ≤ 0.025 on the same detectors.

Our first hypothesis (P1.1, predicting a fool rate above 0.20) is confirmed: the `style_pre1923` strategy hits a 0.50 fool rate against Macko-replica and 0.64 against Macko-LOSO on Claude Opus. Our second hypothesis (P1.2, predicting the style ordering) is also confirmed, with early twentieth-century register outperforming Russian translation and L2 academic styles ($0.64 > 0.48 > 0.40$). Crucially, this register-shift approach improves the evasion rate by $50\times$ compared to the 2025 baseline on the same detector, generator, and topics.

Source-anchored: the headline attack. `synth_anchor` (Attack 9) is a two-stage refinement of the cross-decade recipe. Stage 1: Claude Opus produces a self-contained ∼120-word period-prose paragraph on a subject *thematically adjacent* to the topic. Stage 2: it writes the topic in that paragraph's exact register, with the anchor visible as a few-shot stylistic template. Total cost: 2 API calls per topic.

Table 3 (last row) shows `synth_anchor` Pareto-dominates pure `style_pre1923` on every detector column: RoBERTa 0.57 vs. 0.15, TF-IDF 0.76 vs. 0.58, Macko-bk 0.70 vs. 0.50, Macko-LOSO **0.798** vs. 0.64.

The attack does not cost naturalness. Cross-judge $P(\text{human})$ from three frontier LLMs (Claude Opus, GPT-5.4, Gemini 2.5 Pro), self-family dropped (21), n=40 judgments per cell on Claude Opus / 20 VK-2026 test topics. PJS v2 sacrifices most of vanilla's naturalness ($0.54 \rightarrow 0.30$) to extract its baseline-detector evasion. The cross-decade attack *recovers* that naturalness: `style_pre1923` hits 0.500 (indifference threshold) and `synth_anchor` 0.535 (above vanilla's 0.540) while delivering the highest Macko-LOSO fool rate in the table; real-human anchor sits at 0.760. `synth_anchor` is the only strategy in the full study that crosses 0.5 fool rate against *every* detector class simultaneously *and* keeps cross-judge p_{human} above indifference.

Writing style vs. Historical register. We test whether any distinctive writing style can bypass detection, or if the evasion is specific to historical registers. Our hypothesis (P10.1, predicting a high average evasion rate across Borges, Sebald, and Dillard) is falsified, averaging only 0.26. Specifically, only W.G. Sebald's style—which relies heavily on pre-modern, period-adjacent register—succeeds in defeating the detector (0.749). Imitating modern authors (such as Borges, Dillard, Murakami, or Coelho) fails completely, showing that detectors are vulnerable to historical language distributions rather than just a "distinctive authorial voice."

4.4. The simple deployer fix fails (Attack 11)

The §4.3 mechanism has a corollary recommendation: a deployer who has identified period register as the failure direction should retrain the adversarial detector with pre-1923 prose injected on the human side, so period register stops being out-of-distribution. We formulated this prediction as **P11.1** (prior to training): the Opus / `synth_anchor` mean fool rate would fall from 0.798 (LOSO) to ≤ 0.20 on Macko-pp.

Strategy	Macko-LOSO	Macko-pp
synth_anchor	0.798	**0.846**
style_pre1923	0.640	0.657
style_sebald	0.749	0.798
style_borges	0.030	0.095
caveman	0.000	0.000
rewrite_k5	0.028	0.011
roundtrip_imperf	0.015	0.000

Table 4
Macko-LOSO vs. Macko-pp fool rates (↑ better attack) on the held-out 20 VK-2026 test topics, generator = Claude Opus. Adding pre-1923 prose to the human side of training does *not* close the cross-decade attack.

We trained Macko-pp: same architecture and recipe as Macko-LOSO, five strategy folds, but with ∼1,000 ∼500-word passages from `deepmind/pg19` (Project Gutenberg books published before 1919) mixed into the human side at ∼40% pg19 / ∼60% PAN'25. Table 4 reports the result.

Our main prediction (P11.1, which posited that the Gutenberg augmentation would reduce the fool rate to ≤ 0.20) is cleanly falsified: the fool rate actually rises to 0.846, slightly higher than the baseline. Similarly, the average evasion rate across all historical styles remains high (0.767). Meanwhile, our control strategy (`rewrite_k5`) stays caught at 0.011, confirming that the detector has not simply decalibrated, but remains specifically blind to the period register.

A control points to register-specificity. `style_borges` is in the same OOD-distinctive-author family as `style_sebald` but is not period register. Its Macko-pp fool rate is 0.095, so the detector handles it normally. The augmentation closure fails specifically along the period-register axis it ostensibly trained on, while leaving other OOD axes untouched. We discuss candidate mechanisms in §5.

4.5. Bracketing synthetic-anchor from both directions

`synth_anchor` is the headline attack. Two follow-up experiments, both designed before generation, attempt to beat it from opposite directions of the human-text distribution. Both fall short, in informative directions.

Attack 12: real anchor (does provenance help?). `synth_anchor` uses a *synthetic* anchor: Claude generates the ∼120-word period-register paragraph from its prior. We hypothesized (P12.1, prior to generation) that a *real* pre-1923 paragraph as anchor would strictly beat the synthetic version: real `pg19` prose is genuinely human and should pull the conditioning distribution further from Claude's prior. This hypothesis is cleanly falsified: using real period prose actually underperforms the synthetic anchor (0.778 vs. 0.846 on Macko-pp, and 0.730 vs. 0.798 on Macko-LOSO). The likely reason is thematic relevance: the synthetic anchor generator is prompted to write about a theme related to the target topic, whereas the real anchors are pulled from random literary passages. This indicates that style-matching is most effective when the style template is thematically coherent with the target text, making the synthetic anchor a purposeful construct rather than just a proxy for real historical writing.

Attack 13: PAN-mimicry (the inverse direction). Every other attack in this paper pulls Claude's output *out of* the deployer's training-time human distribution. Attack 13 (P13.1) inverts: few-shot Claude with five PAN'25-train passages per topic and instruct it to match their voice, lexical patterns, and natural imperfections. The detector's "human" decision region is shaped by exactly that pool; mimicry should land where the detector has the lowest confidence in flagging anything machine. We predicted Macko-pp fool rate ≥ 0.85. This strategy fails completely: the evasion rate drops to 0.000 on Macko-pp and 0.006 on Macko-LOSO. Although few-shot coaching allows the model to mimic human

Strategy	RoB-L	RoB+homo	DeB-v3	mean
caveman	0.90	0.90	0.98	**0.93**
style_dillard	0.84	0.86	0.94	**0.88**
stream	0.63	0.65	0.78	**0.69**
style_murakami	0.53	0.59	0.75	0.62
style_borges	0.31	0.29	0.51	0.37
synth_anchor	0.18	0.11	0.28	0.19
fragments	0.16	0.17	0.23	0.19
dialect_appalachian	0.09	0.09	0.18	0.12
style_sebald	0.06	0.05	0.15	0.09
style_pre1923	0.03	0.04	0.08	0.05
style_translated_ru	0.01	0.01	0.07	0.03

Table 5
Fool rate against the neutral PAN'25-trained committee, Claude Opus pool, n=6,500 across 11 strategies on 20 VK-2026 test topics. The ordering inverts Macko-pp's: strategies that beat Macko-pp (period-register; bottom rows) are caught here; strategies that beat this committee (register reduction, literary mimicry, modernist stream; top rows) are caught by Macko-pp.

surface features (like vocabulary and grammar quirks), it does not alter the model's underlying token probability signature, to which our adversarially retrained classifier is highly sensitive.

The asymmetry, and what it implies. Cross-decade attacks defeat adversarial fine-tuning by pulling output into a distribution region the detector's training corpus did not cover; PAN-mimicry attacks fail because the detector's training corpus did cover that region densely, and its calibration there is sharp. The detectors we test are not fooled by *what the text looks like at the surface*; they are fooled by *what its underlying token distribution looks like*. Surface-style coaching can be matched via few-shot but does not rewrite the distribution. Out-of-distribution shifts do.

The bracketing strengthens the synthetic-anchor finding from two sides: the attack is not substitutable with simpler real-prose conditioning (real-anchor underperforms), and it does not work merely because Claude can imitate human prose (PAN-mimicry fails).

4.6. A neutral PAN'25-trained committee inverts the picture

A more realistic 2026 condition is a deployer who trained on PAN'25 alone — no exposure to our generations. We model this with three independent supervised detectors, each fine-tuned on PAN'25 train: RoBERTa-large [19]; RoBERTa-large with Macko's homoglyph augmentation [7]; DeBERTa-v3-large [22]. We re-generate the 20-topic 2026 test split at N=30 on cross-decade and author-style families, N=20 elsewhere, plus a round-two extension probing register-shift variants (`fragments`, `dialect_appalachian`, `stream`). Total n=6,500 Claude Opus candidates.

The ordering symmetrically inverts Macko-pp's: strategies that survive Macko-pp (`synth_anchor` 0.846, `style_pre1923` 0.657, `style_sebald` 0.798) are caught by the neutral committee ($\leq$ 0.19); strategies that beat the neutral committee (`caveman` 0.93, `style_dillard` 0.88, `stream` 0.69) are caught by Macko-pp ($\leq$ 0.09). No strategy in this pool beats both detector classes simultaneously — a hard Pareto frontier. The round-two failures (`fragments`, `dialect_appalachian`) probe alternative OOD axes and fail in the same way, ruling out a generic "any OOD axis works" interpretation.

4.7. Stream-narrative defeats both classes

The Pareto frontier from the previous section presents one notable exception: `stream` cells on 7 narrative-genre topics escape Macko-pp at a $\geq 30\%$ rate, while expository topics are caught uniformly. This suggests that *form-content match* is the operative variable, pointing to a second co-equal OOD mechanism alongside period register: radical shifts in narrative form.

class	detector	$\mu\, p_{\text{LLM}}$	fool@0.5
supervised	RoB-L/PAN'25	0.03	0.97
supervised	RoB-L + homo	0.02	0.98
supervised	DeB-v3/PAN'25	0.002	1.00
zero-shot	LogPerp-GPT2M	0.41	0.97
zero-shot	LogRank-GPT2M	0.37	1.00
judge	GPT-5.4	0.46	0.51
judge	Gemini-2.5-Flash	0.02	1.00
adversarial	**Macko-pp**	**0.34**	**0.65**

Table 6
`stream_narrative`: per-detector mean $P(\text{LLM})$ and fool rate at threshold 0.5, n=300 Claude Opus cells on 20 VK-2026 test topics. The decisive row is bottom: Macko-pp's escape rate jumps from 9% on regular `stream` (Table 4) to **65**% once narrative form is forced.

We test this by forcing a narrative voice onto every topic. `stream_narrative` introduces a named witness (e.g., a court lady, a young secretary, an apprentice) whose perspective shapes the entire output. It sustains a modernist stream-of-consciousness form ($\leq$ 3 full stops in $\sim$500 words) and demands concrete particulars once per $\sim$40 words. We generated 300 cells on Claude Opus using this strategy.

`stream_narrative` escapes Macko-pp at 65%, a 7$\times$ improvement over regular `stream` (9%), while keeping all three supervised fool rates $\geq$ 0.97. The binding constraint is judge_gpt54 right at threshold (0.46 mean, 0.51 fool); Gemini Flash accepts $\geq$ 99% as human. Per-topic minimax across the 8-detector committee picks stream_narrative in 19 of 20 topics (mean worst-case 0.42; 17 of 20 below 0.5 on *every* detector). The single exception is topic 067 (*Aesthetic Philosophy / Music Criticism Essay*): all 14 stream_narrative cells score $P(\text{LLM})$=1.0 on Macko-pp despite near-zero supervised. Argumentative content does not survive the narrative-witness frame; the argument structure leaks through.

5. Discussion

The OOD-inversion paradox is not a paradox. A standard practice is "train the strongest supervised detector you can on the latest benchmark, ship it." Our PAN'25 RoBERTa is exactly that: 0.9994 in-distribution AUC. Under PJS v2 it returns $P(\text{LLM}) \approx 0.001$ on the same generations whose vanilla counterparts it scored ≈ 1.00. The mechanism in §4.1 explains why: contextual / distributional detectors learn the distribution of fluent native English. Literal back-translation flattens that distribution in a direction the supervised model has never seen, so its representation of the text drifts into the human half of its decision boundary, and the more confidently so the better the detector was tuned. The high in-distribution AUC and the high out-of-distribution fool rate are the same number twice, in opposite directions.

The 2025 recipe is closable; the cross-decade recipe is not (by the obvious fix). A small adversarial fold collapses 2025-attack fool rate by 30–50$\times$ at a cost any deployer can absorb (§4.2). The same playbook does not close the cross-decade attack (§4.4: P11.1 falsifies at 0.846). The bracketing experiments favour the interpretation that pg19-augmentation widens the "human" decision region without giving the classifier signal to discriminate human period prose from machine-emulated period prose: surface-style coaching alone (PAN-mimicry) works only weakly against the baseline (0.702) and not at all against the adversarial fine-tune (0.000).

What survives is OOD direction, not perturbation depth. Surface mimicry, perturbation depth (n-pass refinement), entropy reduction (caveman), and format shift (genre-dialogue) all fail against Macko-LOSO at $\leq$ 0.14 fool rate. Only register-OOD attacks, and only those whose register is structurally pre-modern, defeat the adversarial detector at meaningful fool rates. The mechanism finding is therefore

Rank	Participant / Submission	Mean Score
1	**uts-claude-v11**	0.144
2	**uts-claude-v4**	0.164
3	**roundtrip_imper**	0.197
4	**uts-claude-v7**	0.211
5	**uts-qwen-v7**	0.308
6	Dargk - "Dargk"	0.320
7	Dargk - "LoRA beta0"	0.338
8	**rejection_sample_opus**	0.462
9	**source_anchored_synth**	0.499
10	Anlam-Turing - "direct"	0.926

Table 7
ELOQUENT 2026 Voight-Kampff leaderboard. Our submissions are highlighted in bold. Submissions from UTS sweep the entire top 5 positions. Our leading entry `uts-claude-v11`, which relies heavily on the `stream_narrative` strategy, secured the highest rank, confirming that OOD narrative form translates to leading source-like behavior against the live competition detector panel.

narrower than "attackers can always escape": it specifies a particular axis (period register $\times$ theme coherence) along which a present-day adversarial detector is structurally vulnerable.

The OOD-direction mechanism has at least two survivable axes. The same strategy can flip sign across deployer postures. `caveman` fails Macko-LOSO at 0.000 but wins the neutral committee at 0.93 (Table 5); the gap is training-data exposure. `stream_narrative` (§4.7) goes further: 65% Macko-pp escape plus ≥ 0.97 fool against three independent neutral supervised detectors. `synth_anchor` and `stream_narrative` differ in surface (period register vs. modernist stream form) but share structure: each fits a distinct human-prose distribution that PAN-pool training never densely covers. The mechanism has at least two survivable axes, and likely more — the present finding bounds what Macko-pp can close, not what attackers can generate.

Future work: term distributions under translation. Translation studies has long observed that translations age faster than their source texts; to our knowledge this ageing has not been characterised as term-distribution drift. Our `style_translated_ru` attack — a Garnett-era translation register that holds a 0.483 Macko-LOSO fool rate (Table 3) — suggests translation-inherited term distributions are a productive feature axis, both for attack design and for auditing which human registers a detector's training pool actually covers.

Limitations. Single-language (English); closed-API generators will be deprecated (we pin dated snapshots for reproducibility, but exact reproduction expires with the provider); $n{=}20$ topics on the 2026-test split gives wide CIs on per-topic effect estimates. We do not test non-period-register specialised genre OOD (legal opinions, medical case reports) and cannot rule out further survivable axes.

Broader impact. The cross-decade attack evades adversarial detection by prompting alone, without retraining or model access. We release the attack and detector checkpoints so a next-year deployer can train against it directly, and report the pg19 counter-fix's failure so the community does not pursue it as a known-effective recipe.

6. Conclusion

Our comprehensive evaluation—spanning a 4-generator $\times$ 5-strategy $\times$ 66-topic factorial, plus 12 novel 2026 attacks on Claude Opus—reveals a clear asymmetry in evasion attacks. While adversarial fine-tuning definitively closes the 2025 evasion recipes (driving Macko-replica fool rates down to ≤ 0.025),

two novel OOD strategies (`synth_anchor` and `stream_narrative`) easily survive that closure, achieving 0.846 and 0.65 escape rates against the strengthened Macko-pp baseline.

Furthermore, we show that the obvious counter-measure of augmenting the training set with period prose fails to close the vulnerability (with P11.1 falsified at 0.846 vs. predicted ≤ 0.20). Ultimately, the operative mechanism is OOD direction rather than surface mimicry: pulling generations *out of* the deployer's training distribution (via historical register or modernist stream form) consistently defeats adversarial fine-tuning, whereas pulling them *into* it (e.g., PAN-mimicry) fails entirely. These OOD strategies form the core of our submissions, which sweep the top 5 positions on the ELOQUENT 2026 leaderboard. Our detector checkpoints and attack prompts are publicly available.

Artifacts and ELOQUENT 2026 submission

Released: detector checkpoints and attack prompts.[2] Macko-replica's false-positive rate on PAN'25 held-out humans is ≤ 0.04, so the calibration overhead of adversarial fine-tuning is small. We submitted five entries to the ELOQUENT 2026 leaderboard on the 20-topic 2026 test split, spanning different selection criteria: a Sebald-style cross-decade entry; two selection-rule variants over a portfolio of register-shift attacks (supervised-argmin and supervised+judge-minimax); the 8-detector-minimax entry (UTS-Claude-v11) dominated by `stream_narrative`; and an open-weight Qwen3.6-27B entry.

Competition Results. The final results validate the strength of these OOD evasion strategies in the true deployment condition: submissions from UTS sweep the entire top 5 positions on the leaderboard. Table 7 reports the top-10 submissions.

References

[1] 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. 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.

[2] J. Bevendorff, R. Gunti, J. Karlgren, 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: 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: https://ceur-ws.org/Vol-3740/paper-225.pdf.

[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: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-4038/paper_277.pdf.

[5] P. Vachharajani, Literal re-translation as a method for AI text disguise and detection evasion, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 1443–1448. URL: https://ceur-ws.org/Vol-4038/paper_116.pdf.

[2] Adversarial checkpoints as in §3.3; the PAN'25-baseline detector is at https://huggingface.co/protagonist/roberta-eloquent.

[6] A. Creo, M. Hormazábal-Lagos, H. Cerezo-Costas, P. Alonso-Doval, Fake it 'til you make it human, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-4038/paper_109.pdf.

[7] D. Macko, R. Moro, I. Srba, mdok of KInIT: Robustly fine-tuned LLM for binary and multiclass AI-generated text detection, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-4038/paper_307.pdf.

[8] J. Karlgren, E. Artemova, O. Bojar, M. I. Engels, V. Mikhailov, P. Šindelář, E. Velldal, L. Øvrelid, Overview of ELOQUENT 2025: Shared tasks for evaluating generative language model quality, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, 2025.

[9] 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.

[10] 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.

[11] S. Gehrmann, H. Strobelt, A. M. Rush, GLTR: Statistical detection and visualization of generated text, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019. URL: https://aclanthology.org/P19-3019/.

[12] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, DetectGPT: Zero-shot machine-generated text detection using probability curvature, in: International Conference on Machine Learning (ICML), 2023, pp. 24950–24962.

[13] 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: International Conference on Machine Learning (ICML), 2024. URL: https://arxiv.org/abs/2401.12070.

[14] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, S. Feizi, Can AI-generated text be reliably detected?, 2023. URL: https://arxiv.org/abs/2303.11156. `arXiv:2303.11156`.

[15] 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.

[16] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, T. Goldstein, A watermark for large language models, in: Proceedings of the 40th International Conference on Machine Learning (ICML), 2023. URL: https://arxiv.org/abs/2301.10226.

[17] E. Reif, D. Ippolito, A. Yuan, A. Coenen, C. Callison-Burch, J. Wei, A recipe for arbitrary text style transfer with large language models, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2022. URL: https://aclanthology.org/2022.acl-short.94/.

[18] Qwen Team, Qwen3.6-35B-A3B, 2026. URL: https://huggingface.co/Qwen/Qwen3.6-35B-A3B, moE 35B total / 3B active. 4-bit DWQ MLX build via `mlx-community/Qwen3.6-35B-A3B-4bit-DWQ`.

[19] 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, arXiv preprint arXiv:1907.11692

(2019). URL: https://arxiv.org/abs/1907.11692.

[20] J. W. Rae, A. Potapenko, S. M. Jayakumar, C. Hillier, T. P. Lillicrap, Compressive transformers for long-range sequence modelling, in: International Conference on Learning Representations (ICLR), 2020. URL: https://arxiv.org/abs/1911.05507.

[21] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-bench and chatbot arena, in: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2023. URL: https://arxiv.org/abs/2306.05685.

[22] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, 2021. URL: https://arxiv.org/abs/2111.09543. arXiv:2111.09543.