STEALTHRL: REINFORCEMENT LEARNING PARAPHRASE ATTACKS FOR MULTI-DETECTOR EVASION OF AI-TEXT DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce *StealthRL*, a reinforcement-learning framework that paraphrases AI-generated text to evade detection while preserving meaning and fluency, enabling systematic red-teaming of AI-text detectors. The challenge is to reduce detector confidence at strict low-FPR operating points without collapsing semantic fidelity or overfitting to a single detector family. StealthRL fine-tunes a single Qwen3-4B paraphraser with LoRA Hu et al. (2021) and a multi-objective reward combining detector evasion, semantic similarity, and fluency. On the MAGE benchmark with three detectors (RoBERTa OpenAI, Fast-DetectGPT, Binoculars), StealthRL reduces mean TPR@1%FPR from 0.27 (no attack) to 0.04 while maintaining high semantic similarity (0.953 E5 cosine), outperforming simple paraphrasing and a detector-guided baseline. We release an anonymized code package in the supplementary material with a placeholder anonymous link for reproducibility.

1 INTRODUCTION

AI-generated text detectors are now embedded in academic integrity workflows, content moderation pipelines, and misinformation defenses. Yet detector robustness remains brittle under paraphrasing attacks: minor rewording can substantially reduce detection accuracy, especially at strict low-FPR thresholds that matter in high-stakes settings. This motivates red-teaming methodologies that generate high-fidelity paraphrases while systematically probing detector failure modes.

We focus on the adversarial aspect: generating paraphrases that evade multiple detector families without semantic drift. Existing evasion methods often optimize against a single detector or rely on prompt-only rephrasing, which may not generalize across detector families. We propose **StealthRL**, a reinforcement-learning paraphraser trained against a detector ensemble with explicit semantic and fluency constraints. The goal is not to enable misuse, but to provide reproducible stress tests that reveal detector fragility and guide defensive improvements.

Contributions.

- **Multi-detector RL paraphrasing.** We train a single paraphrase policy against a detector ensemble using group-relative policy optimization with LoRA, enabling efficient adversarial fine-tuning.
- **Low-FPR, transfer-aware evaluation.** We report attack success across detector families at TPR@1%FPR with detector $\times$ method heatmaps, tradeoff curves, and bootstrap confidence intervals.
- **Reproducible evaluation pipeline.** Our evaluation stack exports standardized per-sample scores, calibrated thresholds, and paper-ready figures, and supports MAGE with optional RAID and PadBen benchmarking. Dugan et al. (2024); Zha et al. (2025)
- **Plug-and-play training harness.** The training stack is modular (YAML-configurable detectors, rewards, and base models) and supports rapid ablations without code changes.
- **Empirical results.** On MAGE, StealthRL reduces mean TPR@1%FPR to 0.04 (from 0.27 for no attack) while maintaining high semantic similarity.

Anonymized code release. An anonymized code release containing training and evaluation scripts is included in the supplementary material. A placeholder anonymous link is provided: `https://anonymous.4open.science/r/STEALTHRL`.

2 Related Work

AI-text detectors. Curvature-based detectors such as DetectGPT and Fast-DetectGPT measure probability curvature under perturbations, while paired-LM detectors such as Binoculars compare base vs instruction-tuned likelihoods. Mitchell et al. (2023); Bao et al. (2024); Hans et al. (2024) Classifier-based approaches (e.g., RoBERTa OpenAI, Ghostbuster) learn discriminative models and can be sensitive to domain shift. Verma et al. (2024)

Watermarking-based detection. Watermarking approaches embed statistical signals into model outputs to enable reliable detection, forming a complementary detector family. Kirchenbauer et al. (2023) We focus on non-watermark detectors here, but watermark-aware evaluation is an important extension.

Adversarial paraphrasing. Recent evasion methods use paraphrasing to reduce detector confidence, including Adversarial Paraphrasing and RL-based humanization such as AuthorMist. Cheng et al. (2025); David & Gervais (2025) These methods motivate our use of detector-guided rewards while highlighting the need for cross-detector transfer evaluation. Character-level attacks such as homoglyph substitutions demonstrate strong evasion but often degrade readability. Creo & Pudasaini (2025)

Prompt-based and retrieval-aware evasion. Krishna et al. show that paraphrasing substantially reduces detector accuracy and propose retrieval-based defenses. Krishna et al. (2023) Lu et al. demonstrate that large language models can be guided to evade detectors without model fine-tuning, highlighting the importance of robust evaluation under adaptive attacks. Lu et al. (2024) StealthRL complements these approaches by training a single RL policy that targets detector ensembles and evaluates transfer under low-FPR thresholds.

Evaluation benchmarks. MAGE provides a standard benchmark for machine-generated text detection with diverse domains and generators, and is widely used in detector evaluation studies. Li et al. (2024) RAID offers a large-scale robustness benchmark across domains and generators with stress tests for cross-domain generalization. Dugan et al. (2024) PadBen focuses specifically on paraphrase attacks and includes tasks that probe multi-pass paraphrasing and attack depth. Zha et al. (2025) We use MAGE as the primary evaluation benchmark in this work and report low-FPR operating point metrics consistent with prior evasion literature.

3 Threat Model

We assume **black-box access to detector scores**, i.e., the attacker can query detector confidence but does not require gradients. In practice, detectors are often open-source or deployed with a confidence API, so both black-box scoring and open-source replication are realistic. We evaluate transfer to a held-out detector family to test robustness beyond the training ensemble.

4 Method

Given AI-generated text x, we learn a paraphrase policy $\pi_\theta(y \mid x)$ that produces y with low detector confidence while preserving meaning and fluency. We optimize a composite reward:

$$R(x, y) = \alpha R_{\text{det}}(y) + \beta R_{\text{sem}}(x, y) + \gamma R_{\text{ppl}}(y), \tag{1}$$

where R_{det} is the ensemble detector score (lower AI probability), R_{sem} is E5 embedding cosine similarity, and R_{ppl} is a fluency proxy derived from a frozen LM. We train with group-relative policy optimization (GRPO) and LoRA adapters Hu et al. (2021) on Qwen/Qwen3-4B-Instruct-2507 (rank 32, $\alpha = 32$, dropout 0.05). Training uses group size 4, learning rate 2.8×10^{-4}, batch size 16, and

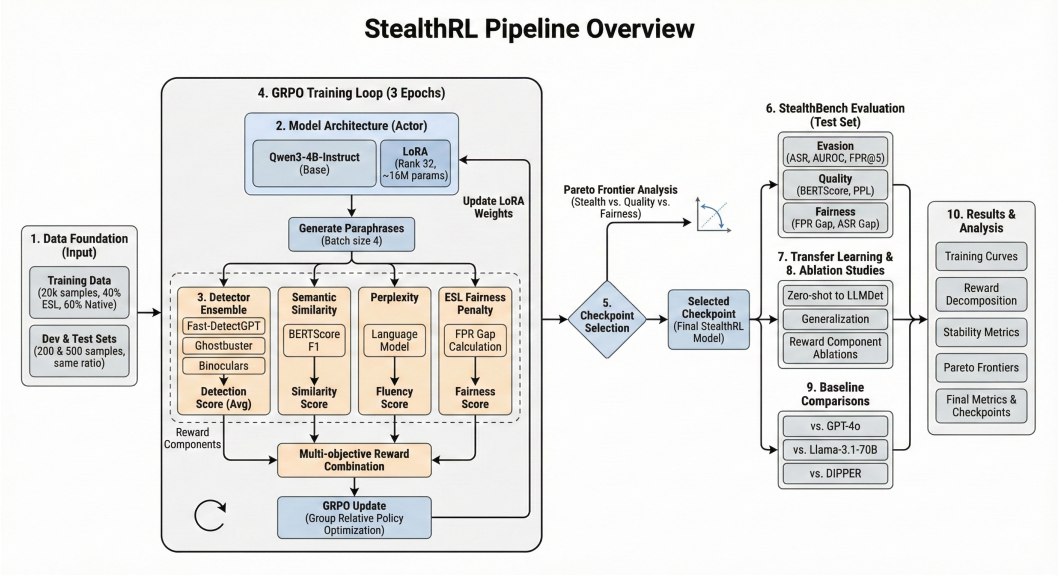

Figure 1: StealthRL training and evaluation pipeline. A single paraphrase policy is trained with detector-guided reward and quality constraints, then evaluated across multiple detector families with low-FPR metrics and tradeoff analysis.

two epochs. The detector ensemble used for training is RoBERTa OpenAI + Fast-DetectGPT with weights 0.6/0.4.

Reward normalization and constraints. Detector scores are z-score normalized online to stabilize the reward scale across training. Semantic similarity is thresholded at 0.90 and linearly mapped to $[0, 1]$; fluency rewards are similarly thresholded with a minimum quality floor to avoid rewarding degenerate outputs. We also reject empty outputs and enforce length sanity checks (minimum length and upper bound relative to the source) to prevent collapse.

Optimization with GRPO. GRPO computes advantages within a group of rollouts per prompt, avoiding a learned value model and reducing variance. We use group-normalized advantages and remove constant-reward groups to prevent degenerate updates. LoRA keeps the base LM frozen and updates only low-rank adapters, which is critical for efficient experimentation and stability.

5 Experimental Setup

Dataset. We evaluate on MAGE using the test split with 50 human and 50 AI samples, filtered to a 100–500 token window. Li et al. (2024) The evaluation pipeline also supports RAID and PadBen for cross-benchmark generalization analysis; these are included as optional runs in the release. Dugan et al. (2024); Zha et al. (2025)

Detectors. Evaluation uses three detector families: RoBERTa OpenAI classifier, Fast-DetectGPT (curvature-based), and Binoculars (paired-LM). Mitchell et al. (2023); Bao et al. (2024); Hans et al. (2024)

Methods. We compare six methods: M0 (no attack), M1 (simple paraphrase), M2 (StealthRL-v1 checkpoint), M3 (detector-guided selection inspired by Adversarial Paraphrasing), M4 (AuthorMist), and M5 (homoglyph substitution). Cheng et al. (2025); David & Gervais (2025); Creo & Pudasaini (2025)

Metrics. We calibrate a 1% FPR threshold per detector on human texts and report TPR@1%FPR (lower is better for the attacker), attack success rate (ASR = 1 - TPR@1%FPR), and AUROC. We

Detector	Threshold at 1% FPR
RoBERTa OpenAI	0.574
Fast-DetectGPT	0.764
Binoculars	0.504

Table 1: Detector-specific thresholds calibrated at 1% FPR on human texts (higher score indicates more likely AI).

also report semantic similarity (E5 cosine) and text quality metrics. All results use one candidate per sample (n=1).

6 Evaluation Protocol

For each detector d, we compute a threshold τ_d such that the false positive rate on human texts is 1%. We then evaluate AI samples using:

$$\text{TPR@1\%FPR}_d = \frac{1}{|\mathcal{A}|} \sum_{x \in \mathcal{A}} \mathbb{1}[s_d(x) > \tau_d], \tag{2}$$

where $\mathcal{A}$ is the AI sample set and $s_d(x)$ is the detector score. The attack success rate is $\text{ASR}_d = 1 - \text{TPR@1\%FPR}_d$. We report mean TPR/ASR across detectors to summarize transfer, and AUROC as a detector separation metric. Bootstrap confidence intervals are computed in the evaluation pipeline (500 resamples) for per-detector metrics.

Quality metrics. We compute semantic similarity using E5 embeddings, chrF for surface-form preservation, and edit rate / token change rate to quantify transformation extent. A perplexity proxy from a frozen LM serves as a fluency indicator. These metrics are reported alongside detection evasion to capture the evasion–quality tradeoff.

7 Implementation and Reproducibility

Training is implemented with a modular configuration system that defines the base model, reward weights, detectors, and optimization settings. The reward computation is asynchronous and caches detector scores to reduce repeated inference. During evaluation, we produce per-sample detector scores and per-method aggregates (including AUROC and low-FPR metrics) and export paper-ready figures and tables.

For reproducibility, the evaluation pipeline outputs (i) per-sample score tables, (ii) aggregated metrics with bootstrap confidence intervals, and (iii) standardized figures for detector transfer and quality tradeoffs. These artifacts are included in the anonymized code release and can be regenerated with a single evaluation command on the MAGE benchmark.

8 Results

Table 2 summarizes detection evasion at 1% FPR. StealthRL reduces mean TPR@1%FPR to 0.04 (vs. 0.27 for no attack) and improves mean AUROC from 0.71 to 0.43. Per-detector TPR@1%FPR drops from 0.22/0.28/0.32 (M0) to 0.08/0.02/0.02 (M2) on RoBERTa/Fast-DetectGPT/Binoculars, indicating transfer to the held-out detector family. StealthRL maintains high semantic similarity (0.953 E5 cosine), comparable to simple paraphrasing (0.959), while outperforming the detector-guided baseline (M3).

Score distributions and quality tradeoffs. Figure 6 shows that StealthRL shifts AI score distributions toward the human regime across detectors. Figures 7 and 8 visualize this effect per detector, with consistent separability reductions on the held-out families. Figure 9 highlights the expected tradeoff between evasion and fidelity. Homoglyph substitution achieves extremely low detection scores but at the cost of degraded fluency (high perplexity), while StealthRL maintains a balanced evasion–quality profile.

Method	TPR@1%FPR ↓ (R)	F (Fast)	B (Binoc)	Mean ↓	ASR ↑	AUROC ↓
M0 No Attack	0.22	0.28	0.32	0.27	0.73	0.71
M1 Simple Para	0.14	0.06	0.02	0.07	0.93	0.54
M2 StealthRL-v1	0.08	0.02	0.02	0.04	0.96	0.43
M3 Adv. Para (guided)	0.14	0.06	0.04	0.08	0.92	0.56
M4 AuthorMist	0.02	0.00	0.00	**0.01**	**0.99**	**0.24**
M5 Homoglyph	0.02	0.00	0.00	**0.01**	**0.99**	0.41

Table 2: Main results on MAGE (TPR@1%FPR, ASR, and AUROC). Lower TPR/AUROC is better for the attacker; higher ASR is better. R/F/B denote RoBERTa/Fast-DetectGPT/Binoculars.

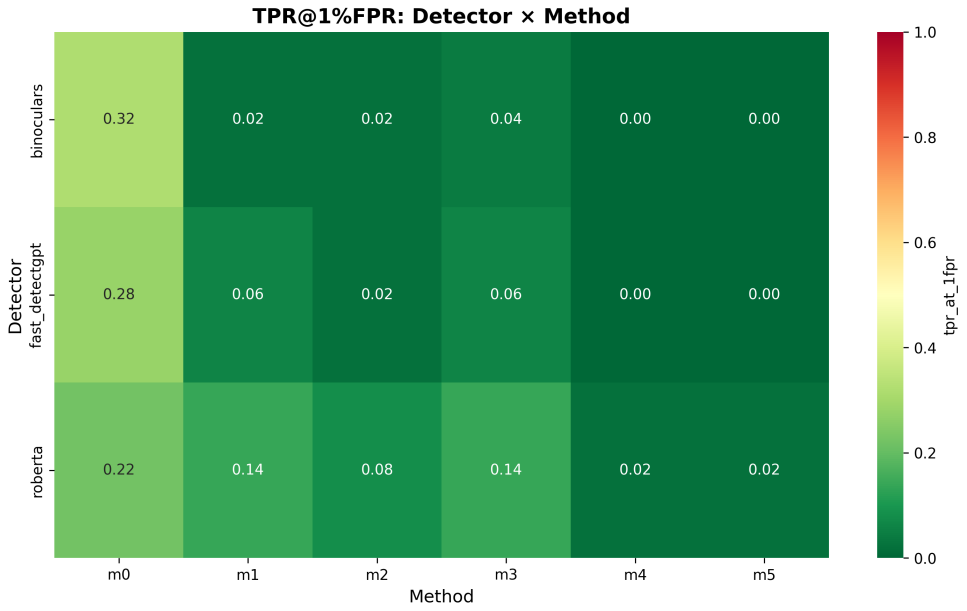

Figure 2: Detector × method heatmap of TPR@1%FPR. StealthRL shows consistently low TPR across detectors, indicating transfer beyond the training ensemble.

Ablation (implemented; results pending). We implemented a *guidance-transfer* ablation that mimics the "guidance vs. deploy" setting from Adversarial Paraphrasing: candidate selection is guided by (i) RoBERTa, (ii) Fast-DetectGPT, or (iii) the ensemble mean, and evaluated across all detectors. This isolates how guidance detector choice affects transfer. Results are pending for the final model checkpoints.

Qualitative examples.

9 Limitations and Safety

Our evaluation currently focuses on three detector families and one primary benchmark; additional detectors (e.g., watermark-based) and datasets can broaden coverage. Adversarial paraphrasing is dual-use: we present StealthRL as a red-teaming tool to expose detector fragility and inform defensive calibration, and we release code anonymized with the expectation of responsible use.

10 LLM Usage Disclosure

Large language models were used to assist with code scaffolding and with editing text for clarity. All modeling decisions, experiments, evaluations, and interpretations were designed, executed, and validated by the authors.

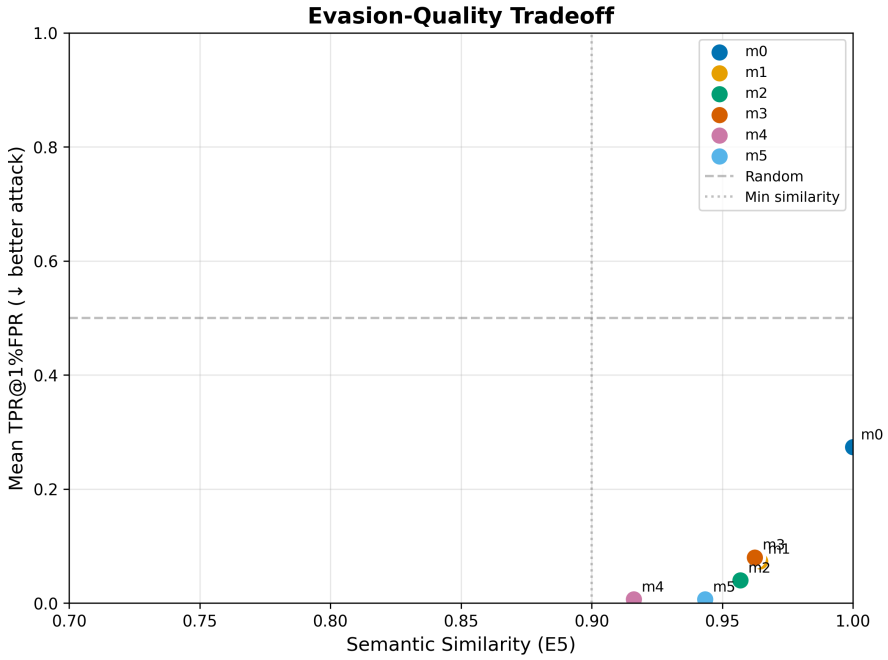

Figure 3: Evasion–quality tradeoff (mean TPR@1%FPR vs. semantic similarity). StealthRL achieves strong evasion while preserving meaning; homoglyph attacks improve evasion but degrade fluency and readability.

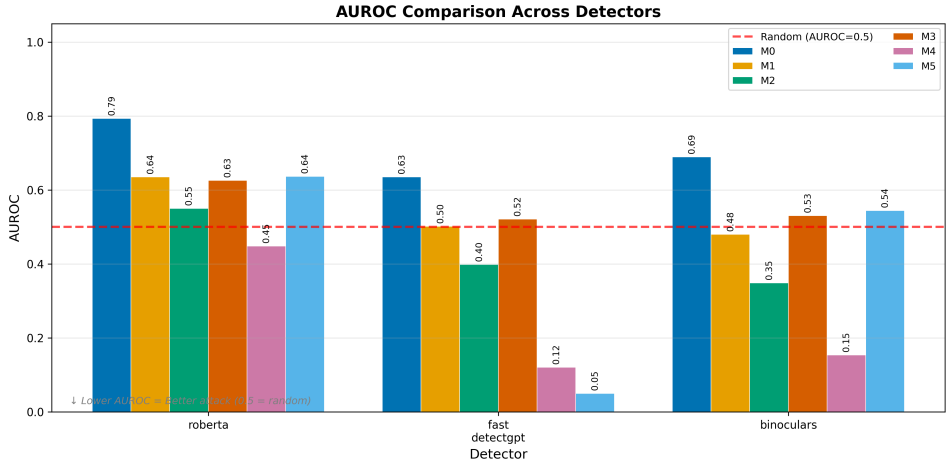

Figure 4: Per-detector AUROC comparison across methods. Lower AUROC indicates weaker detector separation for the attacker.

11 Conclusion

StealthRL provides a practical adversarial RL framework for generating high-fidelity paraphrases that evade multiple detector families under strict low-FPR constraints. Our results show strong transfer to a held-out detector while preserving semantic fidelity, and our released evaluation pipeline supports reproducible red-teaming of AI-text detectors.

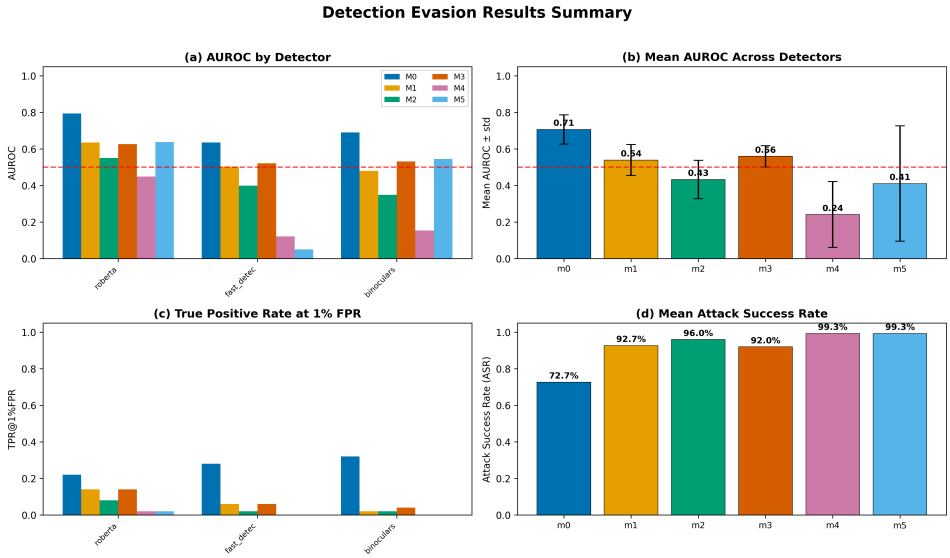

Figure 5: Aggregate method comparison across detectors. StealthRL consistently outperforms simple paraphrasing and detector-guided selection under low-FPR thresholds.

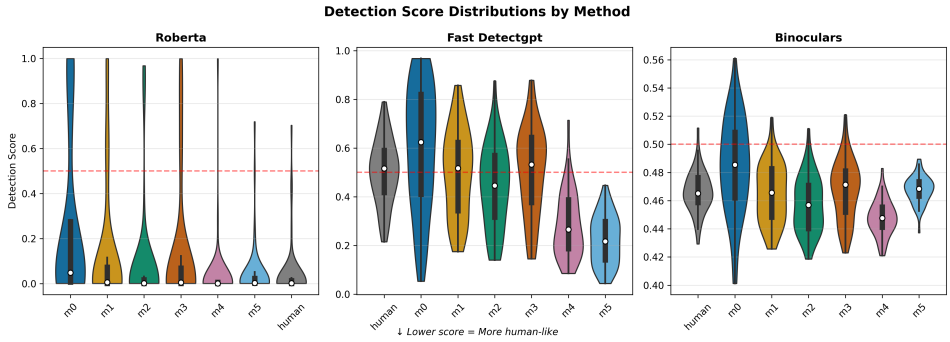

Figure 6: Detector score distributions for human vs AI across methods. StealthRL shifts AI distributions toward human scores across detector families.

REFERENCES

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature, 2024. URL https://arxiv.org/abs/2310.05130.

Yize Cheng, Vinu Sankar Sadasivan, Mehrdad Saberi, Shoumik Saha, and Soheil Feizi. Adversarial paraphrasing: A universal attack for humanizing ai-generated text, 2025. URL https://arxiv.org/abs/2506.07001.

Aldan Creo and Shushanta Pudasaini. Silverspeak: Evading ai-generated text detectors using homoglyphs, 2025. URL https://arxiv.org/abs/2406.11239.

Isaac David and Arthur Gervais. Authormist: Evading ai text detectors with reinforcement learning, 2025. URL https://arxiv.org/abs/2503.08716.

Liam Dugan, Alyssa Hwang, Filip Trhlik, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. Raid: A shared benchmark for robust evaluation of machine-generated text detectors, 2024. URL https://arxiv.org/abs/2405.07940.

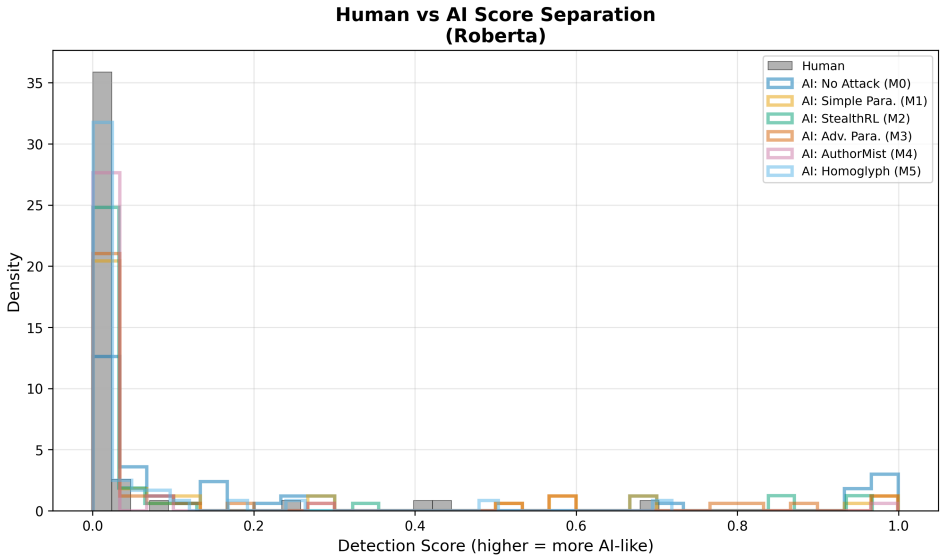

Figure 7: Human vs AI score separation for the RoBERTa detector across methods. StealthRL reduces separability more than simple paraphrasing while maintaining fidelity.

Method	E5 Sim ↑	PPL ↓	EditRate ↑	chrF ↑
M0 No Attack	1.00	27.35	0.00	1.00
M1 Simple Para	0.96	31.09	0.83	0.56
M2 StealthRL-v1	0.95	35.91	0.84	0.53
M3 Adv. Para (guided)	0.96	30.27	0.83	0.56
M4 AuthorMist	0.90	25.07	0.91	0.53
M5 Homoglyph	0.94	146.32	0.65	0.65

Table 3: Quality metrics on MAGE (E5 similarity, perplexity proxy, edit rate, chrF). Higher similarity and chrF are better; lower perplexity indicates more fluent outputs.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: Zero-shot detection of machine-generated text, 2024. URL https://arxiv.org/abs/2401.12070.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models, 2023. URL https://arxiv.org/abs/2301.10226.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense, 2023. URL https://arxiv.org/abs/2303.13408.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. Mage: Machine-generated text detection in the wild, 2024. URL https://arxiv.org/abs/2305.13242.

Ning Lu, Shengcai Liu, Rui He, Qi Wang, Yew-Soon Ong, and Ke Tang. Large language models can be guided to evade ai-generated text detection, 2024. URL https://arxiv.org/abs/2305.10847.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature, 2023. URL https://arxiv.org/abs/2301.11305.

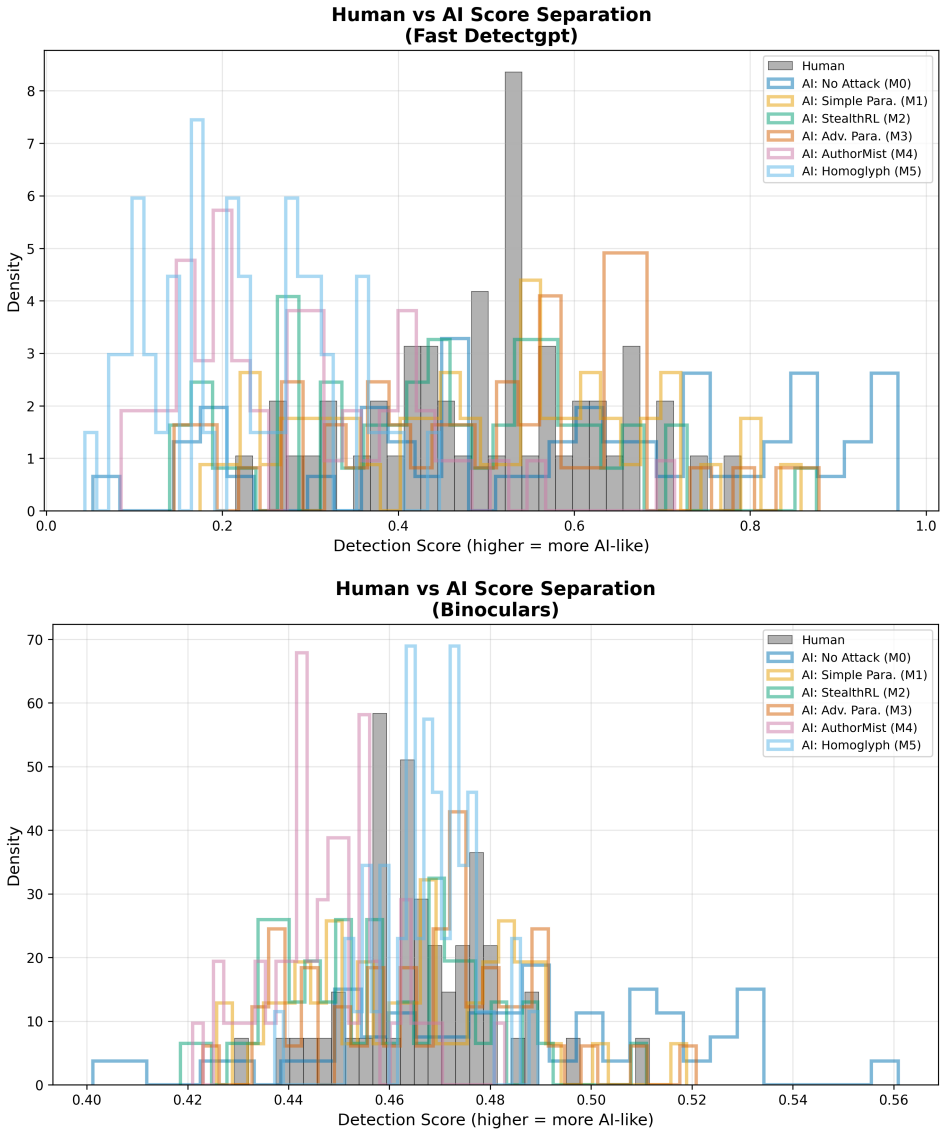

Figure 8: Human vs AI score separation for Fast-DetectGPT (top) and Binoculars (bottom) across methods. StealthRL shifts AI scores toward the human regime on both detectors, indicating transfer beyond the training ensemble.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. Ghostbuster: Detecting text ghostwritten by large language models, 2024. URL https://arxiv.org/abs/2305.15047.

Yiwei Zha, Rui Min, and Shanu Sushmita. Padben: A comprehensive benchmark for evaluating ai text detectors against paraphrase attacks, 2025. URL https://arxiv.org/abs/2511.00416.

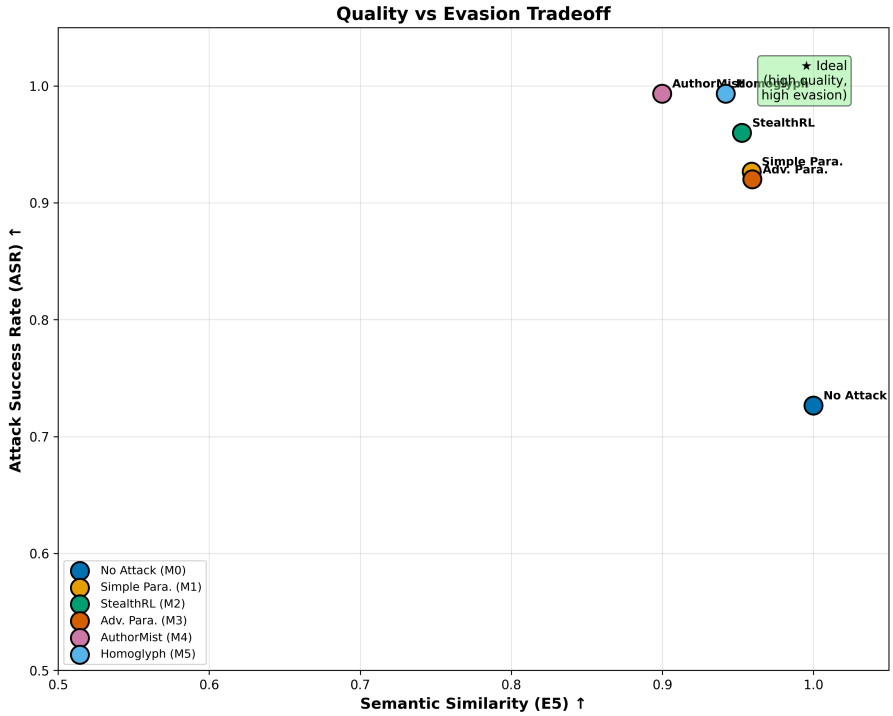

Figure 9: Quality vs evasion scatter. StealthRL balances semantic similarity and detector evasion across samples.

Original (M0)	**StealthRL (M2)**
During cardio the heart increases its workload and all the body's other systems adjust to help support that endeavor. The blood vessels dilate, the muscles do their best to help pump blood back to the heart, and the lungs work harder to take in oxygen and remove waste gases like carbon dioxide.	In cardio, the heart ramps up its workload, prompting the body's other systems to adapt and assist. Blood vessels widen, muscles strive to push blood back to the heart, and lungs intensify their efforts to absorb oxygen and expel waste gases such as carbon dioxide.
The engine has to endure the torque of powering two axels and a drive shaft generally the transfer casing connects to the drive shaft with a ujoint and same for power distribution. The electricity goes into the battery first then is sent to the alternator where it generates voltage that powers all other electrical components like the lights, radio.	The engine must handle the torque from two axles and a drive shaft, typically linked via a universal joint in the transfer casing for power distribution. Electricity first charges the battery, then flows to the alternator, which converts it into voltage to power essential electrical systems such as lights and the radio.

Table 4: Representative paraphrases from the evaluation run (MAGE test split).