Project Proposal: Extending AuthorMist for Cross-Detector AI Text Evasion

Sudeshna Merugu Prith Sharma Miguel Almeida
{sudeshnm, priths, malmeida}@andrew.cmu.edu
CMU Advanced NLP (11-711) – Assignment 3
April 2025

Abstract

We propose to extend the AuthorMist model (David and Gervais, 2025), a recent approach that leverages reinforcement learning to generate human-like paraphrases capable of evading AI text detectors. As AI-generated content becomes increasingly prevalent, so does the need for accurate detection and, in parallel, a better understanding of evasion strategies that exploit current detector weaknesses. AuthorMist represents a strong baseline in this space, introducing Reinforcement Learning using Group Relative Policy Optimization (GRPO) to finetune generation toward undetectability while attempting to preserve semantic meaning. For our final project, we aim to improve the generalization and robustness of AuthorMist by fine-tuning it on the CheckGPT dataset on a set of Open Source detectors, which contains diverse examples designed to test the limits of AI-generated text detectors. By training on this more challenging and detector-annotated corpus, we intend to encourage AuthorMist to develop more detector-agnostic evasion strategies and better semantic preservation across varied inputs.

1 Introduction

In the era of increasingly sophisticated AI-generated text, automatic detection tools have emerged to differentiate machine-generated from human-authored content. While these detectors aim to uphold authenticity, they simultaneously pose significant risks to author privacy and creative freedom. AI-assisted writing tools, widely adopted for their utility in improving productivity and accessibility, risk unfair scrutiny as their output may inadvertently trigger false positives, leading to undue suspicion or penalties.

Consequently, developing effective evasion techniques becomes not only a technical necessity but also a safeguard for authors using AI assistance ethically. The AuthorMist framework exemplifies recent advancements by leveraging reinforcement learning to paraphrase text, thereby effectively reducing detectability. Despite its success, AuthorMist faces challenges, particularly regarding its ability to generalize across unseen detectors and maintain stylistic coherence. Addressing these limitations is crucial for practical, real-world applicability, where content creators cannot reliably anticipate which detection systems will evaluate their texts.

2 Literature Survey

Recent advances in AI-generated text detection and evasion have rapidly evolved, creating a dynamic interplay between detectors and paraphrasing-based countermeasures. Early detectors like GLTR (Gehrmann et al., 2019) and Grover (Zellers et al., 2019) offered statistical cues for detection but quickly became vulnerable to increasingly fluent generations. Subsequent zero-shot detectors (e.g., DetectGPT (Mitchell et al., 2023)) and watermarking strategies (Kirchenbauer et al., 2023) attempted to improve robustness without retraining. Yet, many of these systems remain fragile in the face of structured paraphrasing and minimal surface-level changes (Krishna et al., 2023; Sadasivan et al., 2023).

This fragility has fueled growing interest in reinforcement learning-based paraphrasing strategies, especially models like AuthorMist (David and Gervais, 2025), which use detector feedback to evade detection while maintaining semantic integrity. Our survey deliberately focused on this class of paraphrasing models including AuthorMist, DIPPER, and LoRA-based SFT variants because they offer promising mechanisms to exploit current detector weaknesses in a structured and learnable way. We also examined detectors like RADAR and WILD because they represent modern, openly available architectures vulnerable to paraphrastic attacks, thus providing meaningful evaluation targets.

Moreover, we reviewed recent efforts in multilingual robustness (Guo et al., 2023; Liang et al., 2023), detection generalization (Guo et al., 2024; Yu et al., 2024), and adversarial resistance (Verma et al., 2024; Hans et al., 2024) to understand how real-world LLM use complicates this landscape. As usage of LLMs in academic writing increases (Liang et al., 2024; Russo Latona et al., 2024), detectors must evolve beyond simple heuristics. In light of these trends, we designed our project to extend AuthorMist through fine-tuning on CheckGPT, a diverse, detector-annotated dataset to encourage cross-detector evasion and better semantic fidelity.

Our survey shaped this direction by highlighting (1) the limited generalization of current evasion models, (2) the need for stronger evaluation against multiple detectors, and (3) the lack of benchmarks that stress-test detector resilience across paraphrasing styles. These gaps directly inform our motivation: to build a detector-agnostic, semantically consistent evasion model capable of generalizing across adversarial contexts. Through systematic evaluation of AuthorMist, DIPPER, SFT variants, and QWEN prompts against RADAR and WILD, we ground our experiments in both the theoretical and empirical needs outlined in current literature.

3 Reproduced Model and Task

3.1 AuthorMist Details

3.1.1 System Architecture

AuthorMist is designed to transform AI-generated text into human-like text that can evade AI text detectors. The system consists of two main components:

- A base language model (Qwen2.5-3B Instruct) that serves as the paraphrasing policy
- A reinforcement learning framework that optimizes this policy using detector feedback

The architecture follows a pipeline where AI-generated text is fed into an RL-optimized paraphrasing model that transforms it to minimize detection probability.

3.1.2 Reward Modelling

AuthorMist incorporates multiple AI text detectors (both commercial and open-source) to provide robust feedback. The system is trained against diverse detection algorithms to avoid overfitting to a single detector's weaknesses. The detector selection follows key principles:

- Diversity of detection approaches (statistical detectors, neural classifiers, hybrid systems)
- Representativeness of real-world detection systems
- API stability and reliability
- Continuous probability scores rather than binary classifications

3.1.3 Reward Function

The reward function quantitatively measures AuthorMist's success in evading AI-generated text detection. Given a set of detectors $D = \{d_1, d_2, \ldots, d_k\}$ each detector d_j outputs a probability score $P_{d_j}(Y)$ indicating the likelihood of text Y being AI-generated. The reward function is defined as:

$$R(X,Y) = 1 - \frac{1}{k}\sum_{j=1}^{k} P_{d_j}(Y)$$

This formula rewards outputs that are classified as more human-like (lower probability of being AI-generated).

3.1.4 RL Training with GRPO

AuthorMist employs Group Relative Policy Optimization (GRPO) for training. For each input text X_i, multiple paraphrased outputs $\{Y_{i1}, Y_{i2}, \ldots, Y_{iG}\}$ are sampled along with their corresponding rewards. The baseline reward b_i is calculated as:

$$b_i = \frac{1}{G}\sum_{j=1}^{G} R(X_i, Y_{ij})$$

The advantage A_{ij} for each sample is:

$$A_{ij} = R(X_i, Y_{ij}) - b_i$$

The model parameters θ are updated by maximizing the objective function:

$$J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N} \frac{1}{G}\sum_{j=1}^{G} A_{ij} \sum_{t} \log \pi_\theta(y_{ij,t}|X_i, y_{ij,<t})$$

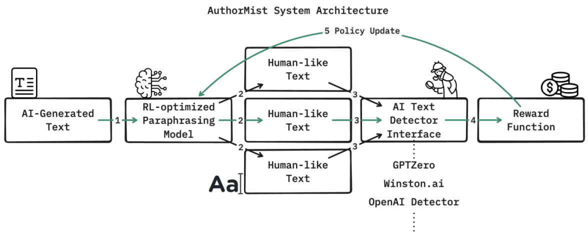

Figure 1: AuthorMist system architecture: AI-generated text is paraphrased via an RL-optimized model to reduce detector scores while preserving meaning and fluency.

3.1.5 Training Data

The training dataset consists of 10,000 human-written abstracts from the CheckGPT dataset with corresponding AI-generated versions created by paraphrasing using models like GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5. Key characteristics of the training data include:

- Multiple domains (Computer Science, Humanities, Social Sciences, Physics)
- Text lengths varying from 100 to 500 words (median 250 words) Diverse range of text lengths to account for detection algorithms' varying behaviors
- Diverse range of text lengths to account for detection algorithms' varying behaviors
- Test dataset: 300 AI-generated passages (200 from GPT-3.5 and 100 from LLaMa-2/GPT-J) + 100 human-written passages

This comprehensive approach enables AuthorMist to develop effective paraphrasing strategies that work across different contexts and text lengths.

3.2 Re-implementation Details

We selected the **AuthorMist** model as our baseline for reimplementation. AuthorMist introduces a reinforcement learning framework that uses paraphrasing as a mechanism to evade AI-generated text detectors, while maintaining semantic alignment with the original input. It achieves this via Reinforcement Learning and an optimization scheme called Group Relative Policy Optimization (GRPO), which encourages the generation of semantically similar yet undetectable outputs.

- **Dataset:** Since the original test dataset used by David and Gervais (2025) is not publicly released, we instead utilized Tufts et al. (2025). This dataset contains a mix of human-written and AI-generated texts on various tasks and languages, labeled for detectability using multiple detection systems. It provides a diverse and challenging set of examples across domains such as academic writing, fiction, and dialogue, making it well-suited for benchmarking evasion models. In our setup, this dataset serves as both the source of generation targets and the basis for evaluating detection success or failure. The responses by two advanced large language models (LLMs): **LLaMA-3** and **GPT-4o** were isolated, ensuring that each sample corresponds to a prompt answered exclusively by an LLM. These responses represent the AI-generated texts used in our evaluation. These AI-generated texts have been passed through four paraphrasing models: QWEN, Dipper, AuthorMist and SFT (Supervised Fine-Tuning on AuthorMist).

  ```
  """Please paraphrase the
  following text to make
  it more human-like while
  preserving the original
  meaning.
  {ai_text}
  Paraphrased text:"""
  ```

  We then compiled both the AI-generated paraphrased texts and the human-written texts that were already present into a unified evaluation set.

  This combined set was subsequently passed through various AI text detectors, including RADAR (Hu et al., 2023) and WILD (Li et al., 2023), to assess the effectiveness of these models at distinguishing between human and machine-generated content. By comparing the detector outputs on both AI and human texts, we were able to analyze the discriminative power of each detector and identify patterns in detection success and failure across different models.

- **Evaluation Setup:** All evaluations were conducted using NVIDIA A100 GPUs to ensure efficient inference and training at scale. We evaluated four paraphrasing systems namely

Qwen, AuthorMist, Dipper, and a supervised fine-tuned (SFT) variant of AuthorMist on their ability to transform AI-generated text from LLaMA-3 and GPT-4o into human-like outputs while evading detection. Each model was used to paraphrase the AI-generated responses, and the resulting texts were combined with human-written samples to create a binary classification dataset. The goal was to assess whether these paraphrases could reduce detectability by state-of-the-art AI text detectors. We evaluated all outputs using RADAR and WILD. For each paraphraser, we recorded both soft detection scores and hard labels.

- **Testing Metrics:** We computed key evaluation metrics including AUROC (1), Attack Success Rate (ASR) (2), F1 Score (3), Precision (4), Recall (5), and Accuracy (6) to quantify detector performance across models. QWEN and AuthorMist were evaluated using their pretrained versions, while Dipper was run via a custom generation script. The SFT-AuthorMist model was fine-tuned using LoRA in 8-bit mode with the TRL library, enabling efficient training without compromising generation quality. All evaluation outputs were visualized using score distributions and ROC curves to highlight each model's evasion effectiveness. This standardized pipeline allowed for a rigorous, comparison of paraphrasing models under consistent conditions, revealing trade-offs in detectability.

$$\text{AUROC} = \mathbb{P}(s(x_{\text{pos}}) > s(x_{\text{neg}})) \tag{1}$$

$$\text{ASR} = \frac{\text{FN}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

3.3 Reproduced Results

As shown in Table 1, AuthorMist achieves the highest Attack Success Rate (ASR) across both the RADAR and WILD detectors, demonstrating its strong evasion capabilities. On the RADAR detector, AuthorMist achieves an ASR of 0.4733, substantially higher than the other baseline models mentioned. This is also reflected in the other reduced detection metrics.

On the WILD detector too AuthorMist outperforms the baselines in ASR and on the other metrics as well. Overall, the results highlight a clear trade-off between evasion strength and general performance. AuthorMist is highly effective at bypassing detectors but these results are not as good as the results mentioned in (David and Gervais, 2025). The main reason behind this is that AuthorMist was trained to evade OpenAI, GPTZero, Hello SimpleAI, Sapling, Originality and Winston detectors, not RADAR and WILD. Also, the test dataset used by (David and Gervais, 2025) was not public, so we used a dataset that was a good representation of the test data.

4 Final Project Proposal

Working with AuthorMist, we found the Group Relative Policy Optimization a very interesting framework to finetune a model to evade detectors. Although AuthorMist does seem to have evaded RADAR and WILD, better than the QWEN baseline, our error analysis on AuthorMist evasion capabilities concluded with how the model does not generalize in evading many other open source Detector models. This is mainly due to excessive fine-tuning to a particular dataset and a particular reward function defined by a small set of detectors. This is an area for improvement where AuthorMist fails to evade the detectors that we have used. So, although we were able to reproduce the baseline results on AuthorMist, we find this to be a place for improvement. Hence, for our final project we propose to enhance the AuthorMist framework by utilizing the Group Relative Policy Optimization (GRPO) to generalize to be able to evade a wider set of detectors. We have observed GRPO's capability at optimizing policies where it evaluates performance relative to a group of policies, rather than in isolation. This approach facilitates more stable and efficient policy updates, leading to improved performance in complex tasks. In the context of AuthorMist, incorporating GRPO offers

Model	AUROC	F1 Score	Precision	Recall	Accuracy	ASR
Evaluation on RADAR Detector						
QWEN	0.9732	0.9699	0.973154	0.966667	0.955	0.033333
Dipper	0.960667	0.930556	0.971014	0.893333	0.900	0.106667
SFT	0.9700	0.976744	0.973510	0.980000	0.965	0.020000
AuthorMist*	**0.8887**	**0.6781**	**0.951807**	**0.526667**	**0.625**	**0.473333**
Evaluation on WILD Detector						
QWEN	0.9360	0.9404	0.887574	1.000000	0.905	0.000000
Dipper	0.974333	0.9201	0.883436	0.960000	0.875	0.040000
SFT	0.971467	0.9404	0.887574	1.000000	0.905	0.000000
AuthorMist*	**0.8512**	**0.9132**	**0.881988**	**0.946667**	**0.865**	**0.053333**

Table 1: Comprehensive evaluation of all models on the RADAR and WILD detectors. AuthorMist exhibits strong evasion capabilities (high ASR) but with reduced AUROC and accuracy, suggesting trade-offs in robustness and semantic fidelity. (* implies baseline model)

several advantages:

- **Enhanced Stability in Training:** By assessing policies in relation to a group, GRPO mitigates the high variance often encountered in policy gradient methods, resulting in more stable and reliable training outcomes.

- **Improved Generalization Across Detectors:** GRPO's group-based evaluation promotes the development of paraphrasing strategies effective against a diverse set of AI detectors, thereby enhancing the model's ability to generalize across different detection frameworks.

- **Efficient Exploration of Paraphrasing Strategies:** The relative performance assessment inherent in GRPO encourages the exploration of a broader range of paraphrasing techniques, potentially uncovering novel methods to evade detection while preserving semantic integrity.

To rigorously evaluate the effectiveness of the GRPO-enhanced AuthorMist model, we plan to test it against several state-of-the-art AI text detectors:

- **Binoculars:** A zero-shot detection method that contrasts outputs from closely related language models to distinguish between human and machine-generated text (Hans et al., 2024).

- **Fast-DetectGPT:** An efficient zero-shot detector that utilizes conditional probability curvature to identify AI-generated content, offering significant improvements in detection speed and accuracy (Bao et al., 2023).

- **PHD (Persistent Homology Dimension):** A detection approach that estimates the intrinsic dimensionality of text embeddings to robustly differentiate between human and AI-generated texts (Tulchinskii et al., 2023).

By assessing the GRPO-enhanced AuthorMist model against these detectors, we aim to validate its effectiveness in producing paraphrases that maintain semantic fidelity while successfully evading detection across multiple frameworks. This comprehensive evaluation will provide insights into the model's robustness and its potential applicability in real-world scenarios where diverse detection methods are employed.

References

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*.

Isaac David and Arthur Gervais. 2025. Authormist: Evading ai text detectors with reinforcement learning. *arXiv preprint arXiv:2503.08716*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Qian Guo, Chen Henry Wei, Linyi Yang, Bowen Zhou, Zhilin Yang, and Yue Zhang. 2024. Biscope: Accurate and generalizable detection of ai-generated text via bidirectional prediction dynamics. *OpenReview preprint*. Https://openreview.net/pdf?id=Hew2JSDycr.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in neural information processing systems*, 36:15077–15095.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 17061–17084.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John F. Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 51008–51025.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023. Mage: Machine-generated text detection in the wild. *arXiv preprint arXiv:2305.13242*.

Weixin Liang, Yining Mao, and James Zou. 2024. A corpus-level detection framework to estimate the prevalence of llm usage in peer reviews. *arXiv preprint arXiv:2410.03019*.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7):100779.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Alan Russo Latona, Tade Raji, Jacob Jordan, and Yacine Jernite. 2024. How many peer reviews are written with ai assistance? *arXiv preprint arXiv:2405.02150*.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Brian Tufts, Xuandong Zhao, and Lei Li. 2025. A practical examination of ai-generated text detectors for large language models. *arXiv preprint:2412.05139*.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Serguei Barannikov, Irina Piontkovskaya, Sergey Nikolenko, and Evgeny Burnaev. 2023. Intrinsic dimension estimation for robust detection of ai-generated texts. *arXiv preprint arXiv:2306.04723*.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of NAACL*.

Zheng Yu, Bing Li, Bowen Zhou, Zhilin Yang, and Yue Zhang. 2024. Text fluoroscopy: Layer-wise disentanglement of human and llm representations. *EMNLP 2024*. Https://aclanthology.org/2024.emnlp-main.885.pdf.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 9051–9065.

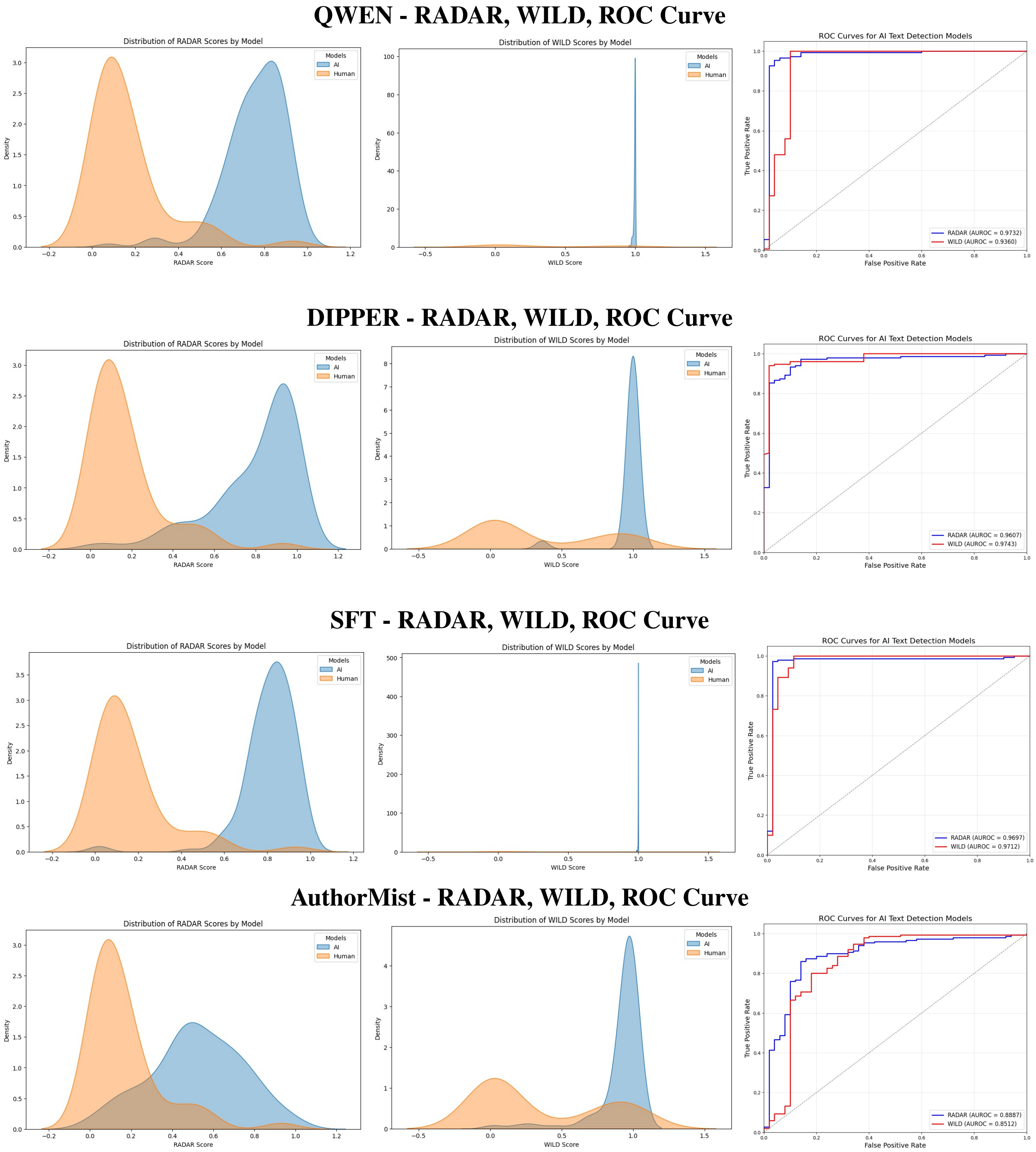

Figure 2: Detector score distributions and ROC curves for QWEN, DIPPER, SFT and AuthorMist paraphrased outputs on RADAR and WILD.