DETECTING AI-GENERATED TEXT: TARGETING ACADEMIC INTEGRITY APPLICATIONS

DS6051 REPORT

Afnan Alabdulwahab
UVA School of Data Science
aa7dd@virginia.edu

Carter Day
UVA School of Data Science
yuk7du@virginia.edu

Brennan Thompson
UVA School of Data Science
xmc6rp@virginia.edu

April 26, 2025

ABSTRACT

As AI-generated text becomes increasingly indistinguishable from human writing in academic settings, ensuring fair detection systems is critical. We investigated methods for minimizing false positives when detecting AI-generated content using transformer-based models with parameter-efficient fine-tuning. While our RoBERTa-LoRA model achieved 99.6% accuracy on a Kaggle essay dataset, it showed significant bias when tested on academic abstracts—detecting AI text with 97.3% accuracy but human text with only 16.8% accuracy. By applying 10:1 class weighting and retraining on a balanced dataset, we improved human text detection to 99.3% while maintaining AI detection at 99.4%, reducing false positives from 83.2% to 0.7%.

Keywords AI text detection · Academic integrity · LLMs · RoBERTa · LoRA · RAID dataset

1 Introduction

This project addresses the burgeoning challenge of distinguishing between human-written and Large Language Model (LLM)-generated text, particularly within the context of academic integrity. Motivated by real-world incidents where students' work was wrongly flagged as AI-generated, our goal was to develop a detection system that prioritizes minimizing false positives on human-written content. Our project addresses this asymmetric error problem by developing a detection model specifically optimized to minimize false positives on human-written academic text. We focus on this priority because misclassifying legitimate student work has immediate ethical and educational implications—academic penalties for authentic work damage trust in both AI detection systems and educational institutions. Our approach involved fine-tuning and evaluating models using diverse datasets, starting with a publicly available student essay dataset and extending to the RAID benchmark, which includes adversarially modified AI-generated abstracts. By optimizing specifically for human text detection while maintaining strong AI detection capabilities, this project aims to provide educators with more reliable tools for upholding academic standards and to contribute to the broader discussion on the ethical deployment of AI detection systems in educational settings.

2 Related Works

Our research builds upon a growing body of work focused on the detection of AI-generated text. Several key publications provide context and inform our approach:

A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions [1] This paper provides a comprehensive overview of recent advancements in the field of detecting text generated by Large Language Models (LLMs). The authors categorize existing detection methods into four primary groups: watermarking-based detection, statistical-based detection, neural-based detection, and human-assisted detection. The survey further explores the various methodologies within these categories, discusses relevant datasets and existing challenges, and outlines potential future directions for enhancing detection capabilities.

DetectRL: Benchmarking LLM-Generated Text Detection in Real-World Scenarios [2] DetectRL introduces a novel benchmark specifically designed to evaluate the performance of LLM-generated text detection models in realistic, high-stakes scenarios. The benchmark tests detectors against text from critical domains such as academic writing, news articles, and creative writing. Furthermore, it incorporates various attack methods, including prompt attacks, paraphrases, perturbations, and data mixing, alongside variations in text length and considerations of human writing styles.

Detecting LLM-Generated Text in Computing Education: A Comparative Study of ChatGPT Cases [3] This study focuses on the evaluation of several AI detection tools within the specific context of computer science education. The researchers compiled a dataset of student submissions predating the widespread use of ChatGPT and generated comparable submissions using ChatGPT for evaluation. Their comparative analysis of eight different detectors identified CopyLeaks as exhibiting the highest accuracy within their specific experimental setup.

The Science of Detecting LLM-Generated Text [4] This paper provides a high-level overview of the fundamental principles underlying the detection of LLM-generated text. It categorizes existing methods into black-box detection, which relies on API-level interaction with LLMs, and white-box detection, which leverages full access to the LLM's internal mechanisms. The authors explain that while black-box methods currently exploit detectable patterns in LLM outputs, their long-term viability is questionable as LLMs become more sophisticated. White-box methods, while potentially more robust, face challenges with the increasing trend of open-sourcing LLMs.

BUST: Benchmark for the Evaluation of Detectors of LLM-Generated Text [5] The authors of this paper introduce BUST, a comprehensive benchmark specifically designed to evaluate the performance of detectors for text generated by instruction-tuned LLMs. Unlike previous benchmarks, BUST emphasizes the evaluation of entire detector systems, acknowledging the influence of underlying tasks and different LLM generators. Their benchmark dataset comprises 25,000 texts from human authors and 7 different LLMs responding to instructions across 10 diverse tasks from 3 distinct sources. Using this benchmark, they evaluated five existing detectors, revealing significant performance variations across different tasks.

Among the reviewed works, two papers particularly influenced the direction of our project. The "Survey on LLM-Generated Text Detection" [1] highlighted how neural-based approaches (like our transformer models) offer strong performance but often struggle with generalization across different LLMs. This influenced our two-stage fine-tuning approach and our decision to prioritize human detection accuracy over overall accuracy. This survey also guided us toward using the RAID dataset, which contains text from multiple LLM sources (11 different models), helping our detector learn broader patterns rather than overfitting to any single AI model. Additionally, "DetectRL: Benchmarking LLM-Generated Text Detection in Real-World Scenarios" [2] demonstrated how detection models degrade when facing adversarial examples, validating our use of the RAID dataset with its various attack methods.

3 Methodology

3.1 Datasets

To conduct our research, we utilized two distinct datasets. An initial, simpler dataset was employed for preliminary fine-tuning and evaluation of three different LLM detection models. Following this preliminary evaluation, the best-performing model was subjected to further fine-tuning and robustness testing using a larger and more complex benchmark dataset. The characteristics of these datasets are detailed below.

3.1.1 Initial Evaluation Dataset: LLM - Detect AI Generated Text

The first dataset used for the initial evaluation of three LLM detection models was the "LLM - Detect AI Generated Text" dataset from Kaggle. This dataset was specifically curated for a competition focused on differentiating between essays written by students and those generated by LLMs. The essays within this dataset are responses to seven distinct prompts, the details of which are provided in the train_prompts.csv file, including the prompt text, instructions, and source materials. The primary training data, located in train_essays.csv, comprises approximately 1,378 essays, each uniquely identified and linked to a specific prompt ID. Each entry includes the full essay text and a binary label indicating its origin: 1 for LLM-generated and 0 for human-written. It is important to note a significant class imbalance within the training set, with approximately 1,375 human-written essays compared to only 3 LLM-generated examples. The initially provided test_essays.csv contained around 6,214 unlabeled essays for prediction. The sample_submission.csv outlined the expected format for competition submissions. This dataset serves as a valuable resource for the development and assessment of AI-generated text detection models, underscoring the complexities introduced by the skewed class distribution.

3.1.2 Robustness Evaluation Dataset: RAID (Robust AI Detection) Dataset

For evaluation and further fine-tuning of our top-performing model, we employed the RAID (Robust AI Detection) Dataset. This large-scale benchmark is specifically engineered to assess the robustness of AI-generated text detection models under various challenging conditions. Encompassing over 10 million documents, RAID exhibits substantial diversity across several key dimensions. It includes text generated by 11 different LLMs spanning 11 genres, such as recipes, news articles, and blog posts. Furthermore, the dataset incorporates 4 distinct decoding strategies used during text generation and features 12 types of adversarial attacks designed to intentionally mislead detection systems. Alongside this AI-generated content, RAID also includes human-written text for thorough comparative analysis. The extensive scale and multifaceted diversity of RAID, significantly surpassing previous benchmarks, offer a rigorous and realistic evaluation environment for gauging the resilience of AI detection methodologies against varied generation techniques and deliberate obfuscation, thereby representing a critical resource for advancing the field.

We targeted a subset of the RAID dataset, specifically focusing on the "abstracts" domain to align with academic integrity applications. Due to computational constraints and the limited number of human samples available, we constructed a balanced dataset utilizing all 1,766 human-written abstracts from the dataset. To maintain class balance, we sampled an equal number of regular AI-generated abstracts (distributed evenly across 11 different AI models) and adversarial AI examples designed to evade detection systems.

3.2 Technical Plan

In our research, we adopted a "divide and conquer" strategy to evaluate the effectiveness of different model types for detecting LLM-generated text. Each team member fine-tuned a different model on the initial Kaggle dataset: RoBERTa-base (fine-tuned fully), DistilRoBERTa-base (fine-tuned fully), and RoBERTa-base with Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning. This diverse selection allowed us to explore trade-offs across model size, tuning strategy, and computational efficiency. Based on performance on a held-out Kaggle test set, a model was then selected for more rigorous evaluation and fine-tuning using the RAID dataset. During this evaluation, we observed a significant tendency of the model to misclassify human-written content as AI-generated, raising fairness concerns — particularly in the context of academic integrity applications. To address this bias, we further fine-tuned the model on a balanced subset of the RAID dataset and introduced aggressive class weighting (human:AI = 10:1), prioritizing correct classification of human-authored texts. This approach substantially improved the model's human detection accuracy without compromising its ability to identify AI-generated content.

3.3 Evaluation Plan

The initial evaluation of our three selected models involved a standard train-test split of 70:15 on the Kaggle "Detect AI Generated Text" dataset. Each model was trained on a subset of the data and subsequently evaluated on the held-out test split from the same dataset. The primary evaluation metric for this phase was accuracy, calculated as the percentage of correct classifications out of the total number of samples. Additionally, confusion matrices were generated to provide a more detailed understanding of the types of errors made by each model.

The model that achieved the highest performance during the initial Kaggle-based evaluation was RoBERTa with Low-Rank Adaptation (LoRA). This version used parameter-efficient fine-tuning, training only a small subset (0.82%) of the full model's parameters. We added a classification head on top of the RoBERTa base model for binary classification (human vs. AI-generated text). LoRA was configured with a rank (r) of 8, a scaling factor (α) of 32, and a dropout of 0.1, applied to the query, key, and value projections of the attention mechanism. The model was trained using a learning rate of 1e-4, a batch size of 16, and AdamW optimizer, over 5 epochs. No class weighting was applied in this phase, and the evaluation relied on standard metrics such as accuracy, precision, recall, and F1-score, along with confusion matrices for deeper analysis.

After identifying RoBERTa-LoRA as the top performer, we evaluated its generalization on the RAID dataset, focusing specifically on the 'abstracts' domain to align with academic writing. This initial RAID evaluation revealed a significant bias toward over-predicting AI-generated content, resulting in a low accuracy for human-written text ($\approx$16.8%). In response, we implemented several improvements:

1. **Balanced Dataset Construction:** Created a balanced dataset with equal representation of human texts, regular AI-generated texts, and adversarial AI examples (1,766 samples of each category).
2. **Aggressive Class Weighting:** Applied a 10:1 weighting ratio (5.0 for human class, 0.5 for AI class) to heavily penalize the model for misclassifying human text as AI-generated.
3. **Human-Focused Evaluation Metrics:** Added specific metrics to track human text detection performance:

- `human_accuracy`: Measures success rate on human texts
- `human_false_positive_rate`: Tracks how often human text is wrongly labeled as AI
- `balanced_accuracy`: Provides a fairer assessment across both classes

4. **Optimized Training Configuration:** Set `human_accuracy` as the primary metric to optimize

We used a weighted cross-entropy loss function, with class weights embedded as a fixed tensor in the custom training loop. Training again used a learning rate of 1e-4, 10 epochs, and the same optimizer and batch size. The model was evaluated on a held-out RAID test set using this comprehensive set of metrics, with special emphasis on minimizing the human false positive rate.

4 Results

The results from our initial phase of evaluating the three LLM detection models on the Kaggle dataset are summarized in Table 1.

Table 1: Kaggle dataset evaluation results for three LLM detection models.

Model	Human Accuracy	AI Accuracy	Overall Accuracy	Parameter Efficiency
GPT-2	95.5%	99.8%	97.2%	100% params trained
DistilRoBERTa-base	96.9%	99.8%	98.0%	100% params trained
RoBERTa-LoRA	99.5%	99.8%	99.6%	0.82% params trained

The results from our initial phase of evaluating the three LLM detection models on the Kaggle dataset revealed strong performance across all architectures. RoBERTa-LoRA achieved the highest overall accuracy (99.6%) while requiring only 0.82% of trainable parameters compared to fully fine-tuned alternatives. All models demonstrated excellent AI text detection (99.8%) with varying human text detection accuracy (95.5–99.5%).

When evaluating our best-performing RoBERTa-LoRA model on the more challenging RAID dataset, focusing exclusively on the "abstracts" domain to align with academic writing tasks. The first evaluation phase used the RAID subset without any further fine-tuning, and revealed a significant bias: while the model achieved 97.3% AI accuracy, it only achieved 16.8% human accuracy, with an overall accuracy of 57% and a precision of 53.8%. This was confirmed by the confusion matrix, which showed the model often misclassified human-written texts as AI-generated.

To address this bias, we re-fine-tuned RoBERTa-LoRA on a balanced subset of the RAID dataset (1,766 samples each from human, AI, and adversarial AI texts), incorporating aggressive class weighting (human:AI = 10:1). This helped penalize misclassifications of human-written content, which is particularly important in academic integrity applications.

After reweighting and re-fine-tuning, performance improved dramatically. On a held-out RAID test set, results are shown in Table 2.

Table 2: RoBERTa-LoRA performance on RAID test set before and after reweighting.

Model	Human Accuracy	AI Accuracy	Overall Accuracy	Precision	F1-Score
RoBERTa-LoRA (before)	16.8%	97.3%	57%	53.8%	69.3%
RoBERTa-LoRA (after)	99.3%	99.4%	99.4%	99.6%	99.5%

The confusion matrices in Figure 1 visually highlight the impact of reweighting on the model's performance. Prior to reweighting, the model misclassified a large number of human-written texts as AI-generated, contributing to the high false positive rate. After applying class weighting and retraining, the updated model correctly identified nearly all human texts, confirming the dramatic reduction in misclassification and improved fairness.

5 Discussion

The initial results from our evaluation phase on the Kaggle dataset showed strong performance across all three models, with RoBERTa-LoRA achieving the highest overall accuracy and parameter efficiency. However, when this top-performing model was evaluated on a more diverse external dataset—specifically a subset of RAID focused on academic abstracts—it exhibited a severe drop in human classification performance. This revealed a substantial bias toward

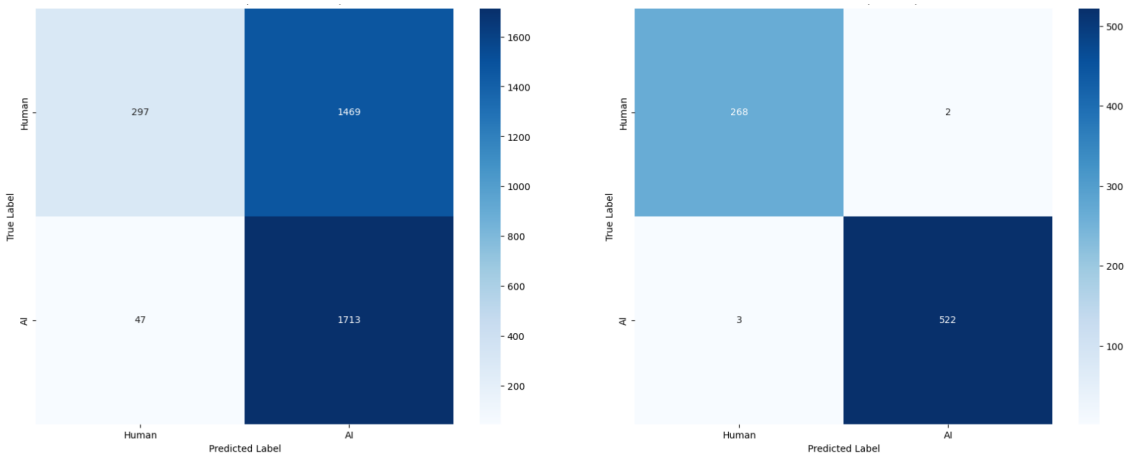

Figure 1: Confusion matrices from RAID evaluation: Left — before reweighting; Right — after reweighting.

predicting text as AI-generated, suggesting that the model had likely overfit to the structure and patterns present in the original training data.

This is particularly concerning within an academic context, as it raises significant questions about the fairness and reliability of such detection tools in evaluating student work. The higher likelihood of a false positive (incorrectly classifying human work as AI) compared to a false negative (incorrectly classifying AI work as human) undermines the credibility of the model for high-stakes academic integrity assessments.

When we targeted improving human text detection through a balanced subset of RAID with human, AI, and adversarial AI texts, and applied aggressive class weighting during re-fine-tuning, the retrained model achieved remarkable improvements, reducing the false positive rate from 83.2% to just 0.7% while maintaining high AI detection accuracy (99.4%). However, it's important to note that we did not test our final model on any additional unseen dataset. As such, it is possible that the new model is now overfitting to the RAID distribution, raising questions about how well it would generalize to other real-world writing domains.

Despite promising results, our work reflects broader limitations in current AI-generated text detection. Our models rely on binary classification without incorporating deeper linguistic or stylistic features that might provide more robust detection signals. As LLMs continue to improve in fluency and diversity, distinguishing features are likely to become more subtle. Additionally, while our model performed well on the RAID dataset, its generalizability to other domains or LLMs remains untested, underscoring the need for future research on robustness and cross-domain evaluation.

6 Conclusion and Future Work

While the reweighted model achieved a high overall detection accuracy, it is crucial to acknowledge the uneven distribution of false positives and false negatives. In situations where accusations of academic dishonesty carry severe consequences, the disproportionately higher likelihood of misclassifying human-written work as AI-generated remains a significant concern, necessitating a degree of skepticism when interpreting the results. While these models can achieve classification accuracy approaching 99%, it's crucial for educators to recognize their limitations—particularly the risk of false positives, where student work may be incorrectly flagged as AI-generated. Acknowledging this bias is essential to avoid unfair academic consequences and ensure that such tools are used responsibly, even as they offer promising support in promoting academic integrity. Nevertheless, with accuracy metrics so high, these tools are nevertheless still valuable within the realm of academic integrity.

Looking ahead, further research is warranted to evaluate the model's ability to handle more nuanced scenarios, such as AI-generated content integrated within human-written text or AI-generated text that has undergone human editing. Our investigation suggests that addressing these complexities represents the next critical challenge for AI detection systems. To this end, future work could explore the integration of more sophisticated analytical techniques beyond standard classification metrics:

- **Perplexity and Burstiness Analysis:** Implementing techniques similar to GPTZero, which measure text predictability and sentence-to-sentence variation, could offer more granular detection capabilities.

- **Stylometric Feature Integration:** Incorporating lexical diversity measures, Shannon entropy calculations, and sentence complexity metrics could contribute to a more robust hybrid detection model.
- **Vocabulary Richness Metrics:** Employing advanced measures such as Type-Token Ratio (TTR) and Hapax Legomenon Rate could help identify subtle differences in vocabulary usage between human and AI writers.

References

[1] Yuxia Wang, Haonan Li, Xudong Han, Yao Zhang, Timothy Baldwin, and Preslav Nakov. A survey on llm-generated text detection: Necessity, methods, and future directions. *arXiv preprint arXiv:2310.14724*, 2023.

[2] Haibin Huang, Yixuan Zhao, Yuxia Wang, Xudong Han, Timothy Baldwin, and Preslav Nakov. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[3] Yao Cheng, Jie Yang, Dong Wang, Xiaoqing Wang, and Qi Zhao. Detecting llm-generated text in computing education: A comparative study of chatgpt cases. *arXiv preprint arXiv:2307.07411*, 2023.

[4] Percy Liang, Kevin Eykholt, Haotian Zhang, et al. The science of detecting llm-generated text. *Communications of the ACM*, 2023.

[5] Han Zhao, Jindong Wang, Yicheng Liu, Dongyu Wang, Runxin Yang, Haoyi Li, Zhou Yu, and Wayne Xin Zhao. Bust: Benchmark for the evaluation of detectors of llm-generated text. In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.