Leidos at GenAI Detection Task 3: A Weight-Balanced Transformer Approach for AI Generated Text Detection Across Domains

Abishek R Edikala[†] and **Gregorios A Katsios**[†] and **Noelie V Creaghe** and **Ning Yu**
Leidos
{abishek.r.edikala, gregorios.a.katsios,
noelie.v.creaghe, ning.yu}@leidos.com
† Authors contributed equally to this work.

Abstract

Advancements in Large Language Models (LLMs) blur the distinction between human and machine-generated text (MGT), raising concerns about misinformation and academic dishonesty. Existing MGT detection methods often fail to generalize across domains and generator models. We address this by framing MGT detection as a text classification task using transformer-based models. Utilizing Distil-RoBERTa-Base, we train four classifiers (binary and multi-class, with and without class weighting) on the RAID dataset (Dugan et al., 2024). Our systems placed first to fourth in the COLING 2025 MGT Detection Challenge Task 3 (Dugan et al., 2025). Internal in-domain and zero-shot evaluations reveal that applying class weighting improves detector performance, especially with multi-class classification training. Our best model effectively generalizes to unseen domains and generators, demonstrating that transformer-based models are robust detectors of machine-generated text.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has made it increasingly difficult to distinguish between human-written and machine-generated text. This challenge poses significant risks in areas such as misinformation dissemination, academic dishonesty, and the breach of trust in online communications. Existing detection methods often rely on small datasets and struggle to generalize across different domains and generator models.

We formulate the detection task as both binary and multi-class classification, and examine the use of class weighting to investigate the impact of class imbalance on detection performance. We build and evaluate four Distil-RoBERTa-Base[1] based models trained with RAID dataset (Dugan et al., 2024), which contains over 6 million text samples from 11 generator models across 8 domains. To assess generalizability, we conduct additional cross-domain evaluations using the MGT Detection Task 1 dataset (Wang et al., 2025), which includes texts from 41 models not seen during training. Our results demonstrate that incorporating class weighting improves detection accuracy and that our models perform effectively across both familiar and unfamiliar domains and generator models. According to the COLING 2025 MGT Detection Challenge Task 3 (Dugan et al., 2025) official evaluation, our models surpass commercial MGT detection tools and achieved top rankings.

2 Related Work

Authorship attribution has a long history, and machine authorship is a recent focus. For example, IARPA's recent HIATUS program (Human Interpretable Attribution of Text using Underlying Structure)[2] takes both human and machine authorship into consideration. According to Leidos' experience in HIATUS, we found that with sufficient training data, transformer-based encoder models can effectively learn features that discriminate authorship and consistently outperformed feature-based approaches.

There are more efforts treating MGT detection as a classification problem. For example, Xiong et al. 2024 addressed multilingual MGT detection in SemEval-2024 Task 8, which includes binary classification (human vs. machine) and model attribution. Their study found that fine-tuned transformer-based models significantly outperformed traditional machine learning methods, demonstrating superior effectiveness in accurately detecting and at-

Approved for public release under eTrace approval ID# 24-leidos-1113-28544

[1] https://huggingface.co/distilroberta-base

[2] https://www.iarpa.gov/research-programs/hiatus

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pages 340–346
January 19, 2025. ©2025 International Conference on Computational Linguistics

tributing machine-generated content across various contexts. LLM-DetectAIve (Abassy et al., 2024) categorizes machine-generated texts into four types: purely human-written, entirely machine-generated, machine-generated then humanized, and human-written then machine-polished. This nuanced approach is impactful in educational and academic settings where subtle LLM edits may hide machine involvement. However, its reliance on the narrows scope of the M4GT-Bench (Wang et al., 2024) dataset may cause LLM-DetectAIve to perform accurately within familiar domains but struggle with unfamiliar ones, leading to high false positive rates and reduced accuracy in diverse real-world scenarios.

Dugan et al. 2024 introduced RAID, a comprehensive benchmark with over 6 million text samples from 11 models across 8 domains, incorporating adversarial attacks and diverse decoding strategies. Evaluating 12 detectors under a fixed 5% false positive rate revealed that open-source detectors often misclassified human-written texts and lacked robustness against minor text modifications and adversarial attacks. The study highlighted that while detectors perform well on familiar data, they struggle to generalize to unseen domains and models, underscoring the need for more resilient MGT detection methods.

3 Method

We approach MGT detection as a classification task using Transformer-based models. Our base model, Distil-RoBERTa-Base[3] (Sanh et al., 2019), is a parameter-efficient, distilled variant of RoBERTa (Liu, 2019) that enables robust detection with limited resources. We train four MGT detectors to evaluate both binary and multi-class classification, exploring the effects of class weighting to address dataset imbalance, to distinguish human-written from machine-generated text.

1. **Binary Classifier without Class Weighting (BC)**: Human vs. Machine, trained without applying class weights.
2. **Binary Classifier with Class Weighting (BW)**: Similar to the BC model but trained with class weights to address class imbalance.
3. **Multi-class Classifier without Class Weighting (MC)**: A multi-class classifier that predicts which generator model produced the text or if it was human-written, trained without class weights.
4. **Multi-class Classifier with Class Weighting (MW)**: The same as the MC model but trained with class weights to mitigate class imbalance.

For the BW and MW models, we compute balanced class weights using the following formula:

$$w_i = \frac{N}{C \times n_i}$$

Where N is the total number of samples in the dataset, C is the total number of classes and n_i is the number of samples in class i. This formula distributes weights evenly across classes by normalizing with the total number of classes, which helps prevent extreme weighting in cases of high imbalance.

We select this method to balance the loss contributions across classes due to its simplicity and effectiveness in enhancing the model's generalizability. Alternative strategies, such as oversampling underrepresented classes or using synthetic data augmentation techniques, can address class imbalance but add complexity and pose the risk of overfitting (Hassanat et al., 2022). Another alternative is Focal Loss (Lin, 2017) that dynamically adjusts the loss based on sample difficulty. This approach can be effective, but demands extensive hyper-parameter tuning, which is impractical given our resource constraints. Our class weighting scheme, though static and less adaptable to extreme imbalance, is resource-efficient and easy to implement.

Baseline (OD) As a baseline, we apply the RoBERTa-Large-OpenAI-Detector[4], an open-source model that is also featured as a baseline on the shared task, directly to our test sets in a zero-shot manner. This enables us to gauge the models performance on our unofficial test sets.

3.1 Data

Randomly sampling from the main training split of RAID (Dugan et al., 2024), we reserve 50K examples for validation and 400K examples for testing, with the remaining examples forming our training set.

To further assess the robustness of our MGT detectors, we also leverage Task 1 dataset (Wang et al.,

[3] https://huggingface.co/distilroberta-base

[4] https://huggingface.co/openai-community/roberta-large-openai-detector

2025) for cross-domain evaluation. Although there is minimal overlap between the generator models in RAID and those in Task 1, the datasets differ in domains and generation techniques, effectively rendering Task 1 out-of-domain (OOD) relative to a model trained solely on RAID. Additional details on the Task 1 dataset composition can be found on their official GitHub page[5]. For our evaluation, we merge the Task 1 training and development sets, excluding models with fewer than 10K examples. From the remaining models, we sample 10K examples each, ensuring equal representation across different sources. Table 2 details model and source distributions in our cross-domain evaluation.

4 Results

This section presents two sets of results: 1) performances of our four MGT detectors against RAID-derived test set, and 2) performances of the best-performing model according to 1) against out-of-domain task 1 dataset to further demonstrate model's robustness.

For each input example, a machine-likelihood score is calculated as $S = 1 - p(\text{"human"}|x)$, representing the probability that the input sample x is machine-generated. The share task adopts the same evaluation metric in RAID (Dugan et al., 2024) to measure how well each detector identifies machine-generated text while only misclassifying 5% of human-written text. Specifically, it is the true positive rate (TPR) after calibrating the decision thresholds to ensure a false positive rate (FPR) of 5%. If the FPR can be optimized to less than 5%, the evaluation script will attempt to do so.

4.1 In-Domain

Table 1 shows the overall performance of our four submissions across all domains and generator models, for both subtask A and B. Among the approaches, the Multi-class Classifier with Class Weighting (MW) detector outperforms the best when detecting MGT with adversarial attacks, more than 84% improvement over the baseline model in our self-evaluation. This suggests that training detectors with large multi-domain and multi-generator data is necessary for achieving robust performance on challenging benchmark like RAID. All four detectors achieve excellent performance when detecting MGT without adversarial attacks, exceeding 0.99 TPR at 5% FPR in the official evaluation. In all results, class-weighted detectors demonstrate a slight performance advantage over their non-weighted counterparts.

Detector	**Adversarial**		**Non-Adversarial**	
	Self	Official	Self	Official
BC(1.0.1)	0.986	0.957	0.997	0.991
BW(1.0.3)	0.989	0.972	**0.998**	**0.994**
MC(1.0.4)	0.986	0.976	0.997	0.992
MW(1.0.2)	**0.992**	**0.977**	0.997	0.993
OD	0.539	N/A	0.582	N/A

Table 1: Overall TPR at 5% FPR in our internal and the share task official evaluation, for Subtask A (Non-adversarial cross-domain MGT detection) ans Sub task B (Adversarial). Note: The detector version in '( )' corresponds to our submission 'Leidos Detector v1.0.x'.

In a heatmap, Figure 1 illustrates the performance of our best model (MW) for Subtask B. The model generally maintains high performance cross domains and generators, except for the Cohere generator model with Reviews domain text. As evidenced by the leaderboard, most submissions see performance drop on this subset, highlights a challenging aspect of Cohere generated review text and necessitates further analysis.

Figures 2 and 3 depict the impact of generator decoding strategies and specific adversarial attacks on the performance of the MW detector. While random sampling, a common technique to increase the diversity of generated text, marginally hinders detection compared to greedy decoding; a repetition penalty, another diversity-enhancing method, shows no significant impact on our detector. Overall, our model remains robust against different forms of text manipulation: most types of adversarial attack did not impact MGT detection, except for paraphrase (2.5% drop) and zero-width space (0.5% drop) attacks. In the RAID data, LLM paraphrased human-written text remains label as "human", which we argue is a gray area and may potentially contributes to our detector's decreased performance for this attack.

4.2 Cross-Domain

Figure 4 illustrates the aggregate performance of our best-performing model (MW) on the cross-domain dataset derived from Task 1 (Wang et al., 2025). We observe that our detector maintains high performance across both new domains and never-before-seen generator models. The results suggest that our detector effectively generalize to different application domains and generator models

[5] https://github.com/mbzuai-nlp/COLING-2025-Workshop-on-MGT-Detection-Task1/

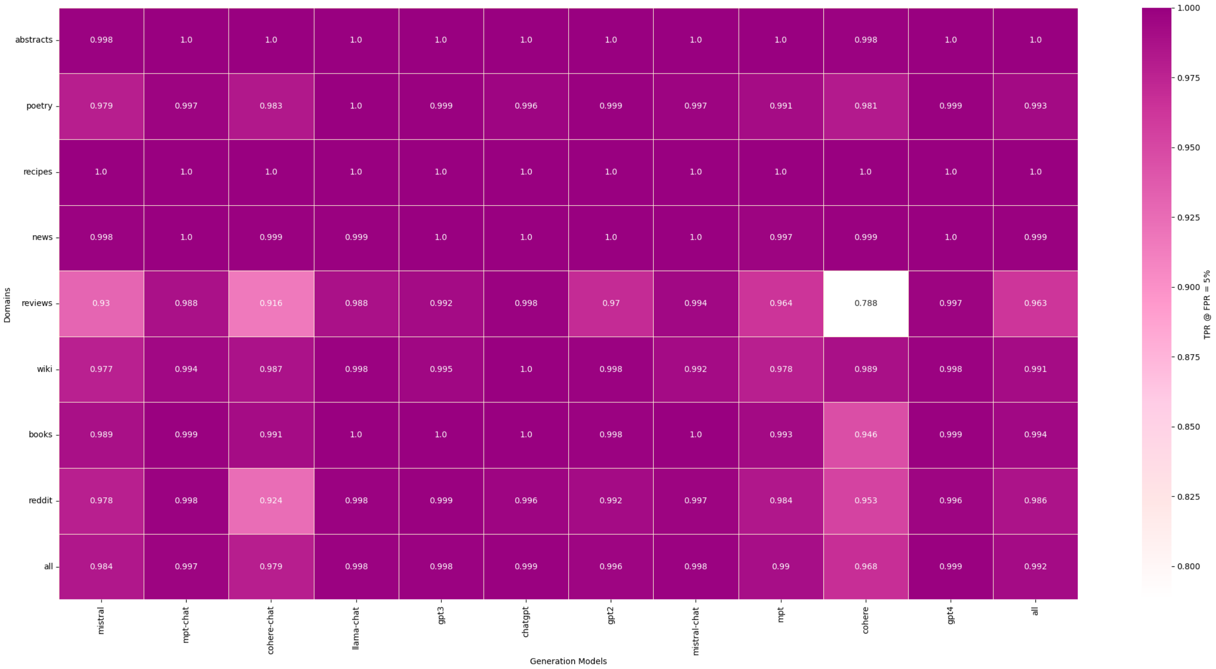

Figure 1: Heatmap measuring TPR at 5% FPR of our Multiclass Weighted (MW) detector across generators and domains, using our RAID-derived test set.

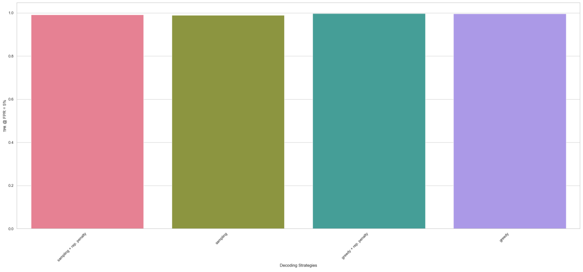

Figure 2: Impact of decoding strategy (x-axis) on MW detector performance (y-axis).

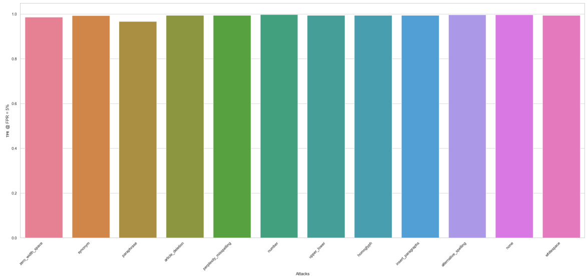

Figure 3: Impact of adversarial attack strategy on MW detector performance.

not encountered during training.

Our detector's strong performance may result not only from its generalization capabilities but also from training data overlap among generator models. LLMs are trained on massive datasets, that often overlap, causing them to generate similar outputs by drawing from the same underlying data (a phenomenon known as memorization or data leakage). This similarity might lead our detector to recognize common patterns across different generators.

5 Conclusion

In this paper, we address the challenges of detecting machine-generated text (MGT) from multiple domains and different generators. Four variants of the DistilRoBERTa-Base model are trained on large and diverse RAID dataset, all achieving highly promising results. Specifically, the multi-class classifier with class weighting (MW) performs the best in both multi-domain and cross-domain evaluations despite of adversarial attacks. This suggests that our approach generalizes well across multiple different domains, unseen generator models, and text manipulations.

Our study demonstrates that transformer-based models with class weighting are effective for MGT detection, representing a significant step toward robust and generalizable detection techniques. However, the strong performance may be influenced by factors limiting true generalizability. Specifically, shared training data among LLM generators might lead detectors to recognize common patterns rather than genuinely generalize across different generators. Additionally, the current evaluation method sets decision thresholds based on a 5% FPR during testing, whereas in practice, thresholds are learned and fixed during training. These aspects require further investigation to ensure detection reliability. In future work, we aim to address these limitations and extend experiments to quantify the impact of training data size.

6 Ethical Considerations and Limitations

Our models may inherit biases present in the training data, which could result in unfair performance across various demographics or content types. This bias poses an ethical concern, as it may lead to higher rates of misclassification of human-written text, especially for underrepresented groups. To mitigate these issues, it is important that datasets are high-quality and diverse. Evaluating model performance across various subgroups and implementing techniques to detect and reduce bias in both models and datasets are essential for model development.

MGT detectors could be misused to infringe on privacy or suppress free speech. Broad adoption may also discourage creative or assistive uses of language models if content is misclassified as machine-generated. While misclassifications may be rare with highly accurate detectors, bias can persist if the majority of the training data isn't written by professional authors. To prevent misuse and protect individual rights, establishing ethical guidelines and usage policies is crucial. Clear policies are needed to differentiate between unethical practices and acceptable uses, governing the ethical deployment of MGT detectors.

The complexity of transformer-based models poses challenges for transparency and explainability. Incorporating explainable AI techniques can help users understand and trust the detector's decisions. These methods can make model decisions more interpretable and are important as they enable accountability and encourage human-machine collaboration.

Although our models performed well on the evaluation datasets, they may not generalize to all future models or domains due to the quick evolution of language models. Data overlap among language models may contribute to the detectors recognizing patterns rather than truly generalizing, potentially inflating performance metrics. Continuous updates and retraining are necessary to maintain performance. Minimizing data overlap is important to better assess true generalization capabilities.

Our goal is to responsibly contribute to the development of MGT detection technologies that are fair, transparent, and beneficial to society.

References

Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, et al. 2024. Llm-detectaive: a tool for fine-grained machine-generated text detection. *arXiv preprint arXiv:2408.04284*.

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

Liam Dugan, Andrew Zhu, Firoj Alam, Preslav Nakov, Marianna Apidianaki, and Callison-Burch Chris. 2025. Genai content detection task 3: Cross-domain machine generated text detection challenge. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Ahmad B Hassanat, Ahmad S Tarawneh, Ghada A Altarawneh, and Abdullah Almuhaimeed. 2022. Stop oversampling for class imbalance learning: A critical review. *arXiv preprint arXiv:2202.03579*.

T Lin. 2017. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohanned Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, et al. 2024. M4gt-bench: Evaluation benchmark for black-box machine-generated text detection. *arXiv preprint arXiv:2402.11175*.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Elozeiri, Saad El Dine Ahmed, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Alexander Aziz, Nurkhan Laiyk, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2025. Genai content detection task

1: English and multilingual machine-generated text detection: Ai vs. human. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, Abu Dhabi, UAE. Association for Computational Linguistics.

Feng Xiong, Thanet Markchom, Ziwei Zheng, Subin Jung, Varun Ojha, and Huizhi Liang. 2024. Fine-tuning large language models for multigenerator, multidomain, and multilingual machine-generated text detection. *arXiv preprint arXiv:2401.12326*.

A Appendix

A.1 Cross-Domain Evaluation Composition

Table 2 provides the exact number of examples per model and source used in our corss-domain evaluation.

Model	Source	# Examples
Bloomz	M4GT	10000
Cohere	M4GT	10000
Davinci	M4GT	10000
Dolly	M4GT	10000
Gemma-2-9B-it	M4GT	10000
Gemma-7B-it	M4GT	10000
GPT-3.5	HC3	10000
GPT-3.5-Turbo	M4GT	5000
	Mage	5000
GPT-4	M4GT	10000
GPT-4o	M4GT	10000
Human	HC3	3333
	M4GT	3333
	Mage	3333
LLaMa-3-70B	M4GT	10000
LLaMa-3-8B	M4GT	10000
Mixtral-8x7B	M4GT	10000
Text-Davinci-002	Mage	10000

Table 2: Number of examples per model and source in our Task-1-derived test set.

A.2 Cross-Domain Evaluation Results

Figure 4 illustrates the aggregate performance of our best-performing model (MW) on the cross-domain dataset derived from Task 1.

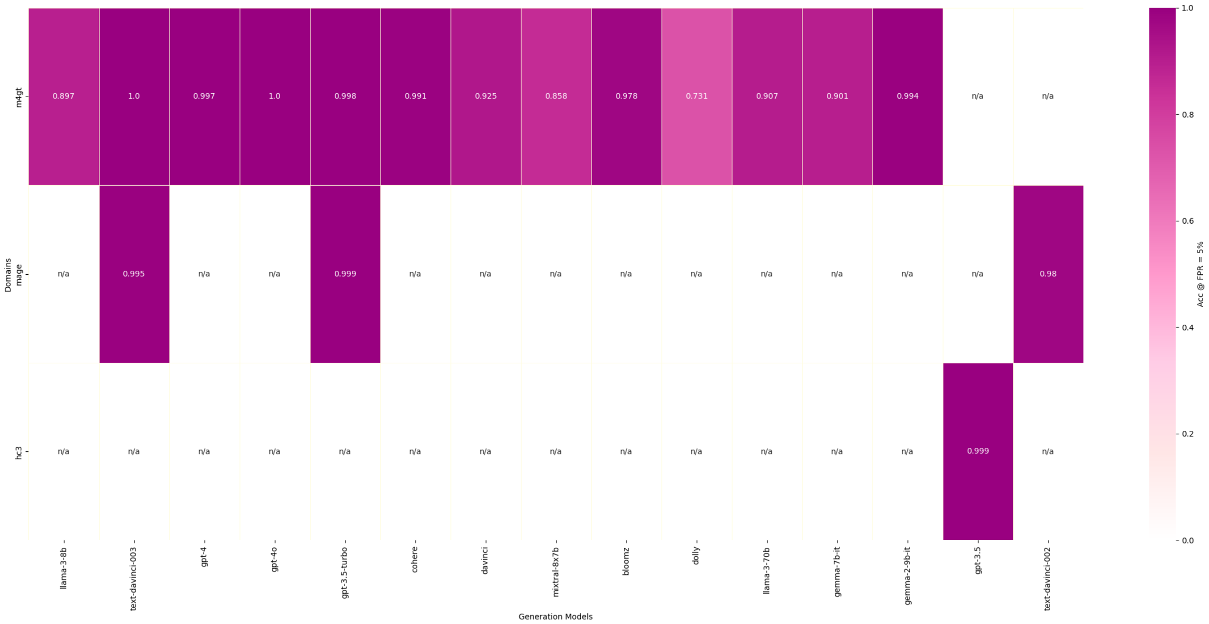

Figure 4: Heatmap measuring Accuracy at FPR 5% of our Multiclass Weighted (MW) detector across generators and domains using our Task-1-derived test set. Empty cells of the heatmap ("n/a") correspond to model and source combinations that are not present in the COLING 2025 MGT Detection Challenge Task 1 dataset.