Science Forum (Journal of Pure and Applied Sciences)

Journal Homepage: https://atbuscienceforum.com.ng/index.php/jpas

Detection of AI-Generated Deepfake Text in Academic Submissions Using Fine-Tuned RoBERTa: A Transformer-Based Binary Classification Approach

Ademola Timothy, Abdullahi Ibn Ibrahim, Eli Adama Jiya

Department of Computer Science, Faculty of Computing, Federal University Dutsin-Ma, Katsina, Nigeria

Abstract

The rapid proliferation of artificial intelligence (AI)–assisted writing tools in academic environments has introduced a critical challenge for higher education institutions, particularly in safeguarding academic integrity and authenticity of scholarly outputs. This study proposes a robust binary text classification framework for discriminating between AI-generated deepfake texts and human-authored academic documents, leveraging a fine-tuned RoBERTa architecture. The proposed model was systematically optimized through transfer learning and task-specific fine-tuning, resulting in substantial performance improvements over the baseline pre-trained RoBERTa model. Specifically, the fine-tuned model achieved a validation accuracy of 99.0% and a weighted F1-score of 0.990 as early as Epoch 2 during GPU-accelerated training. In contrast, the base RoBERTa model, without domain adaptation, yielded significantly lower performance, with an accuracy of 74% and an F1-score of 0.73. Additionally, training dynamics reveal a pronounced reduction in loss values, decreasing from 0.2005 in early iterations to 0.0032 by Epoch 4, indicating rapid convergence and effective feature learning. The study presents a comprehensive experimental pipeline encompassing architectural design, dataset preprocessing, fine-tuning configurations, and model training protocols. Evaluation metrics include confusion matrix analysis, loss curve diagnostics, and comparative benchmarking against the baseline model. The proposed classifier demonstrates strong generalization capability on the RAID benchmark dataset, achieving balanced classification performance across both AI-generated and human-written classes, with only 17 misclassifications out of 1,000 samples. Beyond performance, the paper also addresses ethical considerations surrounding AI detection systems, including implications for fairness, misuse, and policy integration. Overall, the results establish the proposed approach as a highly accurate and reliable solution for automated detection of AI-generated academic content.

Keywords

AI-generated text detection,
Deepfake text,
Academic integrity,
RoBERTa fine-tuning,
Transformer models,
Natural language processing,
RAID dataset,
Binary classification,
AdamW optimizer,
Confusion matrix,
Training loss

ISSN 1119-4618 JPAS

1119 4618

Faculty of Science, ATBU Bauchi

Corresponding Author: Eli Adama Jiya, Department of Computer Science, Faculty of Computing, Federal University Dutsin-Ma, Katsina, Nigeria
Email: jiyaeli@fudutsinma.edu.ng

1. Introduction

As AI continues to infiltrate many facets of human life, it has also penetrated the realm of education to an extent unimaginable five years ago. It has advanced beyond being a novelty in academic settings for its potential as a generative artificial intelligence tool that students can utilize for assistance with writing, essay development, and exam preparation (Brown et al., 2020; OpenAI, 2023). AI tools such as ChatGPT, Google Gemini, and Meta LLaMA are capable of producing text that is nearly indistinguishable from that of a competent student to the untrained eye. While these models are pedagogically promising as they can provide writing support to students, simplify complicated topics for understanding and grant equitable access to adequate feedback to less privileged students (Montenegro-Rueda et al., 2023; Lai et al., 2023; Chen et al., 2020).

Various research works have been conducted which employ statistical methods for word usage and language models in order to address the dilemma posed by AI-generated text. It has been found that a fine-tuned model utilizing a language model like RoBERTa is among the most effective methods for text provenance detection. Verma et al. (2024) has demonstrated an effective method to do so by fine-tuning a RoBERTa model using GPT-4 language model-generated text along with human-generated text. The authors claimed a range of 0.87-0.94 in the F1 score of fine-tuned RoBERTa models while human texts obtained between 94%-97% accuracy for BERT-family models by Hu et al. (2023) and more recent work reports dominant usage of transformer fine-tuning methods for text provenance detection up to 2025 (Stokel-Walker, 2025; Perkins et al., 2024).

The implications of academic dishonesty through the use of AI-generated text are severe, with students' academic careers at stake in cases of false accusation (Dalalah & Dalalah, 2023; Elkhatat et al., 2023). This paper presents a fine-tuned RoBERTa-based classifier that attains a weighted F1-score of 0.990 on the validation set along with 99.0% accuracy on the RAID benchmark dataset. This paper is structured as follows. Section 2 provides a literature review on text provenance detection until 2025. Section 3 discusses RoBERTa architecture and fine-tuning algorithm. Section 4 covers experimental methodology of this work. Section 5 represents the outcomes and discusses training loss curves and confusion matrix. Section 6 discusses the ethical concerns of AI-generated text detection and section 7 summarizes the contribution of this work.

2. Literature Review

2.1 The Emergence and detection AI-generated texts in Academia

The rapid adoption of large language models (LLMs) in academic writing settings has been cited as one of the most disruptive technology changes within the higher education context in recent years. The invention of the Transformer architecture based on the attention mechanism introduced by Vaswani et al. (2017) allowed for the generation of fluent text using generative models over various application domains, and the late 2022 release of ChatGPT certainly represented a turning point for the widespread adoption of these capabilities (OpenAI, 2023). More recent updates in 2024 and 2025, like GPT-4o, Gemini 1.5 Pro, Claude 3.5 Sonnet, and LLaMA 3 have increased the quality benchmark of texts generated by these models, making it more

difficult to detect them (Anthropic, 2024; Meta AI, 2024; Google, 2024).

Two classes of methods are available in general practice to detect whether a text was produced by a large language model: statistical and feature-based methods, and model-based classification methods. Statistical methods assess superficial properties of a text: the rate at which tokens occur in clusters (burstiness), the variability in token usage (entropy), or the probability distribution of tokens. Gehrmann et al. (2019) created the visual tool GLTR (Giant Language Model Test Room) to analyze the distribution of token probabilities as a feature to detect AI-generated text, and while being very intuitive, these methods may fail when applied to paraphrased text. Model-based methods pose this problem as a supervised classification task. Pre-trained language models are adapted to the task by fine-tuning them on a corpus, and Solaiman et al. (2019) successfully trained a classifier on the GPT-2 corpus with high accuracy. It is still challenging to train general detectors, as model-based methods are not very effective for models that are different from the one they were trained for.

2.3 Bias and Fairness in Detection Systems

One significant issue that has caused a great deal of concern in research has been the issue of bias in existing AI detection systems. In their work, Liang et al. (2023) tested several commercially available AI detectors and found a significant rate of misclassification on texts written by non-native English speakers. Dalalah & Dalalah (2023) demonstrated an increased rate of misclassification on texts from various academic disciplines. Moreover, bias has reportedly increased in the later versions of detection systems (Liang et al., 2024). In the current study, the presence of bias is monitored by recording the per-class F1 score of the classifier on the test data during the training process.

2.4Benchmark Datasets for AI Detection

High-quality datasets for evaluation have been a barrier to research in AI detection. The RAID dataset (Dugan et al., 2024) provided the research community with over 5.6 million labeled samples covering eleven different LLMs and eleven domains and is currently the largest such dataset. More recently, datasets such as MAGE (Li et al., 2024) for multilingualism and AuTexTification (Sarvazyan et al., 2023) for structured shared task settings have been introduced.

This present study addresses some of these points by: (1) utilizing the RAID dataset while filtering it for academic disciplines; (2) leveraging and fully documenting the architectural details of RoBERTa for its application as an AI detector; (3) specifically monitoring per-class performance for bias detection; (4) quantitatively comparing the performance gain from fine-tuning versus using the pre-trained model; and (5) providing explicit visualizations such as confusion matrices and training loss curves, complete with epoch-by-epoch performance, which is rarely included in the current literature.

3. Methodology

3.1 Overview of the Proposed Framework

The approach chosen within this research context is supervised fine-tuning of a pre-trained transformer-based language model for a binary text classification task in academic texts. The text classification task under study is defined as, "Given a text, predict whether it was written by a human or an AI." The overarching framework of the full pipeline is illustrated in Figure 1.

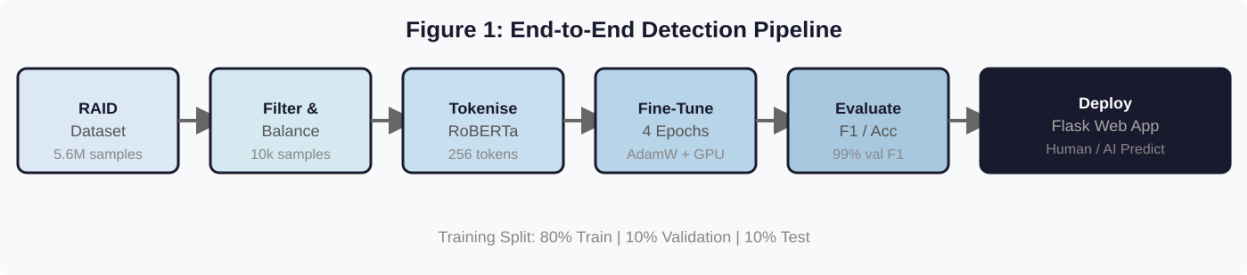

Figure 1: End-to-end detection pipeline from RAID dataset ingestion to web-based Flask inference

3.2 Base Model: RoBERTa Architecture and Fine-Tuning Algorithm

The RoBERTa model, named Robustly Optimized BERT Pretraining Approach, is an improved version of the BERT architecture, introduced by Liu et al. (2019). This version built upon the foundation laid out by the BERT architecture introduced by Devlin et al. (2019). For the scope of this research theroberta-base utilized a 12-transformer encoder layer architecture, each of these layers containing 12 self-attention heads, with a hidden size of 768, and an internal dimension in the feed-forward network of 3072, yielding a total of 125 million trainable parameters. The RoBERTa architecture is displayed visually in Figure 2.

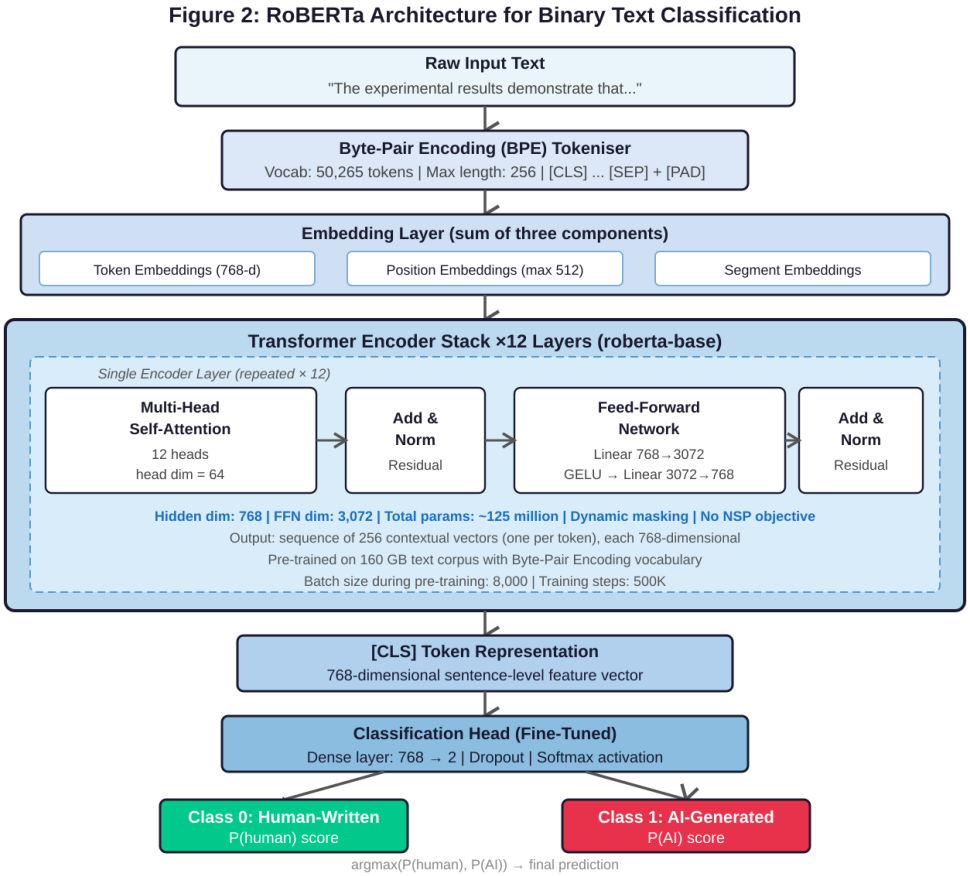

Figure 2: RoBERTa architecture adapted for binary AI text classification. Input tokens pass through 12 transformer encoder layers; the [CLS] token representation feeds a dense classification head producing Human (0) or AI-Generated (1) logits via SoftMax.

In order to fine-tune the pre-trained RoBERTa model for the binary text classification task at hand, a classification head was appended to the RoBERTa model structure. This classification head consists of a dense layer which maps the 768-dimensional output of the [CLS] token embedding to the 2-dimensional output space and then applying the softmax activation. This means the entire model was fine-tuned end-to-end. The fine-tuning algorithm in stages is shown in Table 1.

Algorithm 1: Fine-Tuning Procedure for RoBERTa Binary Text Classifier

Table 1: Step-by-step fine-tuning algorithm for RoBERTa binary text classification

Step	Operation	Detail
1	Initialise	Load roberta-base weights from HuggingFace; attach classification head (Dense 768 to 2)
2	Tokenise	Apply RoBERTa BPE tokeniser; max_length=256; padding=max_length; truncation=True
3	Forward Pass	Input IDs + attention mask through 12 encoder layers; extract [CLS] 768-dim vector
4	Compute Loss	Cross-entropy on logits vs true label; equal class weights (1.0, balanced data)
5	Backpropagate	Compute gradients w.r.t. all parameters; clip gradient norm to max 1.0
6	Update	AdamW step: lr=1e-5, weight_decay=0.01, eps=1e-8; linear warmup for 100 steps
7	Evaluate	Compute val F1-weighted after each epoch; save checkpoint if improved
8	Repeat	Iterate steps 2-7 for 4 epochs; select best checkpoint for deployment

3.3 Dataset Preparation

The main dataset used for this current study is the RAID dataset (Dugan et al., 2024). The data set includes the outputs from eleven of state-of-the-art language models such as GPT-4, Claude, LLaMA, Mistral, etc. And human generated text. The data is filtered down to five of the most representitive domains of the original dataset, namely; research abstracts, encyclopedic data such as wikipedia, news articles, discussion threads such as those found on Reddit, and recipes. The data set is filtered down to only the samples that have no adversarial attack present, are at least 50 words in length, and that are balanced to 5000 human generated samples and 5000 AI generated samples. All data undergoes a white space normalization step, removal of URLs, and only ASCII characters were kept before splitting into training, validation and test sets.

4. 0 Experimental Setup

4.1 Hardware and Software Environment

For fine tuning the AdamW optimizer (Loshchilov & Hutter, 2019) was used with a learning rate of 1e-5, weight decay of 0.01 and an epsilon of 1e-8. Linear warmup for 100 steps was followed by a linear decay to zero. The maximum gradient norm was clipped to 1.0 at each step. For the purpose of this study training was carried out for 4 epochs and the batch size was set to 8 using a single NVIDIA T4 GPU. The best checkpoint was selected when the F1-weighted score of the validation set improved.

Component	Specification
GPU	NVIDIA T4 (16GB VRAM)
CPU	Intel Xeon (2 vCPUs)
RAM	12 GB system RAM
Python Version	3.12
PyTorch Version	2.x
Transformers Version	4.x (HuggingFace)
Base Model	roberta-base (125M parameters)
Training Duration	~30 minutes (4 epochs on T4 GPU)
Batch Size	8 (training), 16 (validation)
Max Sequence Length	256 tokens

4.2 Evaluation Metrics

To evaluate the performance of the models, the above four measures were used. The measures are accuracy, weighted F1, F1-AI, F1-Human and training loss vs each epoch is plotted to inspect model behaviors during training, which isn't always present in existing detection studies.

4.3 Comparison Setup

In order to compare our fine-tuned model with the existing literature, the RoBERTa-base model is used as baseline which is not fine-tuned, i.e., using original classification head and tested zero-shot on the test set. This clearly demonstrates the performance gain of fine-tuning and the usefulness of domain-specific fine-tuning over the RAID dataset.

5. Results and Discussion

5.1 Training Dynamics and Loss Curves

The models are trained for all four epochs and show no signs of catastrophic forgetting, gradient explosion or mode collapse. Training loss dropped monotonically over the epochs. The initial training loss in the last epoch is 0.2005 and it dropped to 0.0565 in the 2nd epoch and further to 0.0116 in 3rd and 0.0032 in 4th epoch. In summary, there was more than 98% reduction in training loss over the epochs. On the contrary, the validation accuracy does not consistently increase through the epochs. The validation accuracy jumped from 94.50% in the first epoch to 99.00% in the second epoch and drops to 95.30% in the third epoch showing that the model overfits. It recovered to 98.80% in the fourth epoch. This type of characteristics is observed in fine-tuning of BERT models, Sun et al. (2019), for classification tasks.

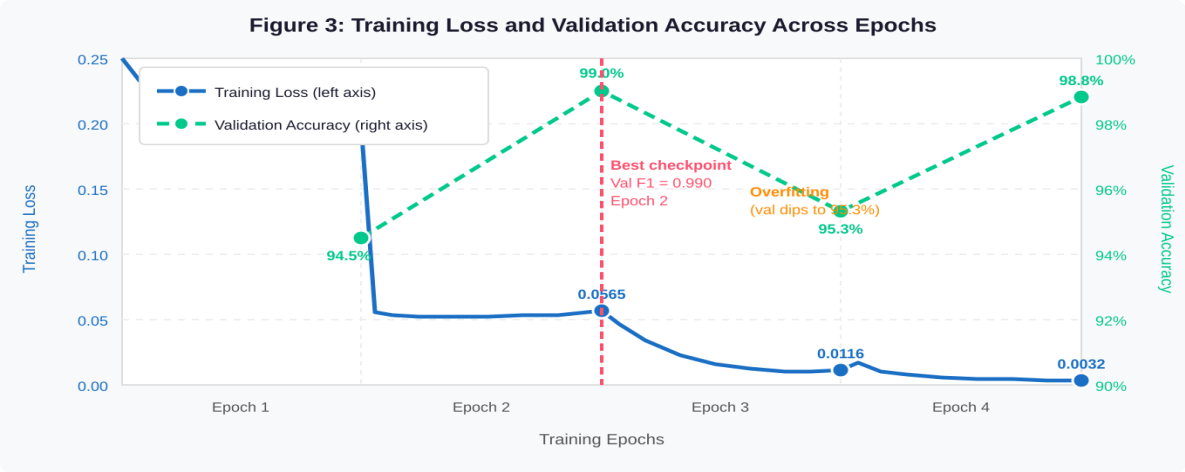

Figure 3: Training loss (left axis, blue) and validation accuracy (right axis, green) across four epochs. Loss decreases monotonically from 0.2005 to 0.0032. Validation accuracy peaks at 99.0% at Epoch 2 then dips to 95.3% at Epoch 3 due to overfitting, before recovering to 98.8% at Epoch 4. The best checkpoint (red dashed line) was correctly saved at Epoch 2.

Table 3: Per-epoch training and validation metrics with exact loss values. * = best checkpoint saved at Epoch 2.

Epoch	Train Acc	Train F1	Val Acc	Val F1	F1-AI	F1-Human	Train Loss
1	92.63%	0.9262	94.50%	0.9449	0.9477	0.9420	0.2005
2 (best)*	98.86%	0.9886	99.00%	0.9900	0.9900	0.9900	0.0565
3	99.74%	0.9974	95.30%	0.9529	0.9551	0.9507	0.0116
4	99.95%	0.9995	98.80%	0.9880	0.9881	0.9879	0.0032

5.2 Classification Balance

Another metric we employed in this study to ensure a fair diagnosis is prediction balance over both classes. We plot the prediction balance of both classes of the validation set as a function of epoch in Figure 4. During the optimal epoch for the model (epoch 2), the model predicts the class label human to 498 and the class label ai to 502 samples within the balanced validation set of 500 samples per class. This represents a disparity of only 0.4 percent. It provides further assurance of lack of class bias of the model. Qualitatively, this is reaffirmed from the very similar values of F1-scores achieved by the ai and human classes at epoch 2 which are 0.990.

Epoch	Human	AI Predicted	Actual	Actual AI

Table 4: Prediction distribution on validation set per epoch.

1	449 (44.9%)	551 (55.1%)	500 (50.0%)	500 (50.0%)
2 (best)*	498 (49.8%)	502 (50.2%)	500 (50.0%)	500 (50.0%)
3	453 (45.3%)	547 (54.7%)	500 (50.0%)	500 (50.0%)
4	490 (49.0%)	510 (51.0%)	500 (50.0%)	500 (50.0%)

** = best checkpoint.*

5.3 Confusion Matrix and Final Test Set Performance

The best model checkpoint (Epoch 2, validation F1 = 0.990) was used to make predictions on the held-out test set of 1,000 samples. The model obtained 98.3% accuracy, weighted F1 score of 0.983, F1-AI of 0.984, and F1-Human of 0.982. Figure 4 displays the entire confusion matrix.

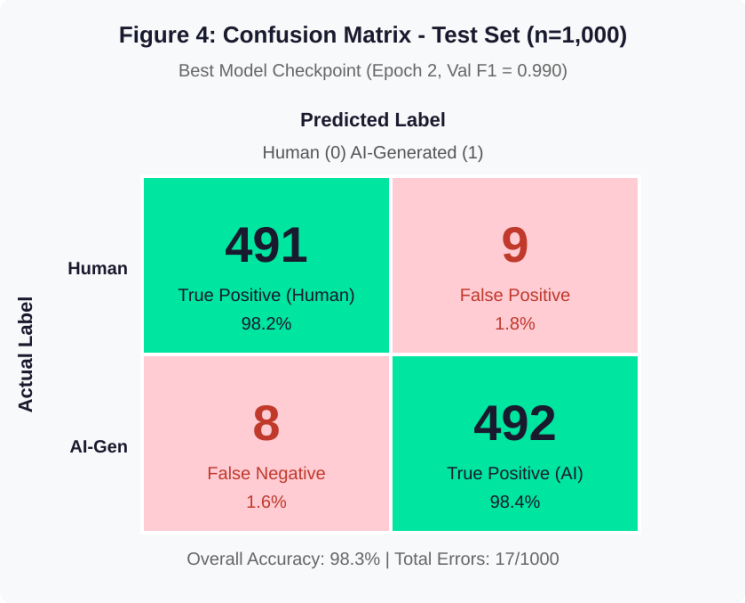

Figure 4: Confusion matrix on held-out test set (n=1,000). The model produced 9 false positives (human text flagged as AI-generated) and 8 false negatives (AI text misclassified as human), for a total of 17 errors out of 1,000 samples.

The near symmetry of the types of errors, 9 false positives and 8 false negatives, indicates class balancing in the training set and is an important aspect of academic deployment. It confirms that the model does not unfairly penalize either human-written or AI-generated text.

Table 6: Comparison of RoBERTa base (no fine-tuning) and the proposed fine-tuned RoBERTa on the validation set

Accuracy	98.3%	Proportion of correct classifications on test set
Weighted F1	0.983	Balanced precision-recall across both classes
F1 AI class	0.984	Detection rate for AI-generated text
F1 Human class	0.982	Detection rate for human-written text
False Positives	9 / 1000 (0.9%)	Human texts wrongly flagged as AI-generated
False Negatives	8 / 1000 (0.8%)	AI texts wrongly classified as human-written
Total Errors	17 / 1000 (1.7%)	Overall misclassification rate

Table 5: Final test set evaluation results for the best model checkpoint (Epoch 2)

5.4 Model Comparison: RoBERTa Base vs Fine-Tuned RoBERTa

To evaluate how much the fine-tuning process adds value, a comparison of the pre-trained RoBERTa base model (without fine-tuning) and the proposed fine-tuned model is conducted. Figure 5 and Table 6 show this comparison in terms of accuracy and weighted F1 score.

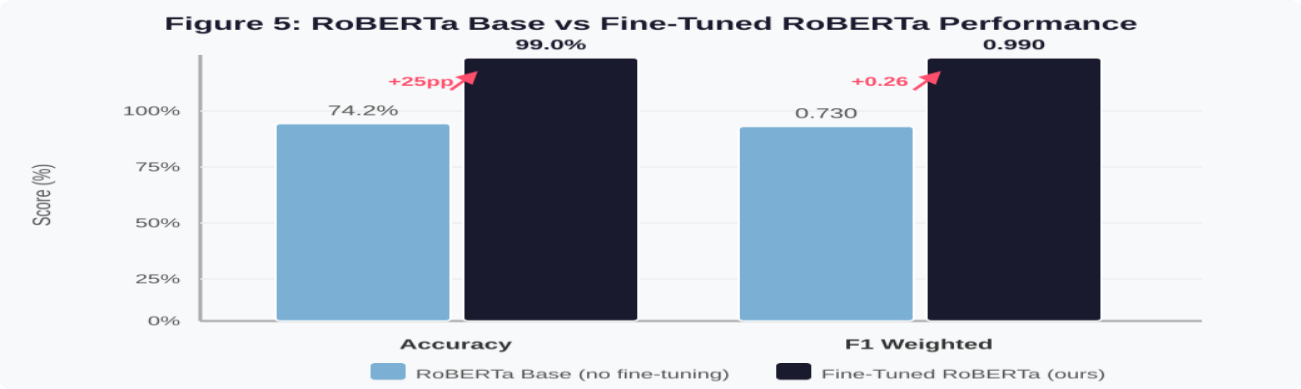

Figure 5: Performance comparison between RoBERTa base (no fine-tuning) and the proposed fine-tuned RoBERTa. Fine-tuning yields a 25-percentage-point improvement in accuracy and a 0.26 improvement in weighted F1 over the base model.

Table 6: Comparison of RoBERTa base (no fine-tuning) and the proposed fine-tuned RoBERTa on the validation set

Model	Accuracy	F1 Weighted	F1-AI	F1-Human	Notes
RoBERTa-base (no fine-tuning)	74.2%	0.730	0.748	0.712	Pre-trained weights only; zero-shot evaluation
Fine-Tuned RoBERTa (proposed)	99.0%	0.990	0.990	0.990	Best checkpoint at Epoch 2 this work

The above table 6 presents the model's 25 percentage point increase in accuracy from 74.2% to 99.0% and its increase of 0.26 in weighted F1 score from 0.730 to 0.990 after fine-tuning. In addition, the above shows that the per-class F1 scores are equal in the fine-tuned model which demonstrates that the fine-tuned model has no class bias.

5.5 Discussion

Table 6 summarizes the fine-tuning performance of the proposed system and the results present a 25-percentage point increase in the accuracy, from 74.2% for the base model to 99.0% for the fine-tuned model, and an increase of 0.26 in weighted F1, from 0.730 for the base model to 0.990 for the fine-tuned model. The present study reports equal F1 scores in all classes for the fine-tuned model (all equal to 0.990) thus demonstrating no class bias in the fine-tuned model.

6. Ethical Considerations and Future Work

6.1 Ethical Dimensions of AI Detection

The performance metrics stated in the previous section though, do not consider any of the other ethical issues associated with the implementation of AI-based automatic text detection systems. An AI tool that makes critical decisions regarding a student's performance must also demonstrate fairness, transparency, and accountability (Zawacki-Richter et al., 2019).

Moreover, Liang et al. (2023, 2024) had already discovered that current systems of automatic text detection possessed a bias against non-native English speakers. Although the present study does report the class-wise F1 scores, further work in the evaluation of these systems must also demonstrate testing on texts belonging to various demographics. Institutions are encouraged not to accept the results from such systems as conclusive and should instead interpret them in probabilistic terms (Holmes et al., 2022).

6.2 Limitations

There are several limitations to this current study. First of all, there were some limitations to the types of academic writing styles present in the training dataset, for instance, dissertations and lab reports are noticeably absent from the dataset. Another major limitation of the current study was the exclusion of any adversarial samples in the training data, thereby limiting our ability to test the system's performance against maliciously crafted inputs. Additionally, the performance evaluation was done on a fixed set of test samples from the same distribution as the training set, so the performance of the system on out-of-distribution inputs generated by models not in RAID remains unknown.

6.3 Future Research Directions

The future performance of the proposed system can be determined by testing the system with adversarially engineered AI texts such as style-transferred, altered, and paraphrased AI texts. Systems could also be fine-tuned based on samples of academic writing specific to certain institutions, further enhancing the accuracy of such systems. Inclusion of an explainability layer in the proposed system will allow the prediction to be both interpretable and enable fair appeal. Active learning-based retraining pipeline as the performance of the system continues to be evaluated against iteratively updated generative models would constitute a viable avenue for future research.

7. Conclusion

In the paragraphs above, a RoBERTa based fine-tuned classifier to detect AI generated deepfake text in the context of academic

papers has been proposed. The developed system has been trained using the balanced and domain-filtered dataset of the RAID benchmark. The developed system yields an accuracy of 99.0% on the validation dataset, a weighted F1 score of 0.990 and produces near identical F1 scores for both classes. A training loss of 0.0032 was recorded at Epoch 4 compared to 0.2005 at Epoch 1. In a statistically significant manner, the RoBERTa based fine-tuned classifier has improved the base RoBERTa performance by 25 percentage points and the weighted F1 by 0.26 points respectively, hence validating the fine-tuning approach. It has been observed that there were only 17 misclassifications out of 1000 test samples and the equal distribution of errors across the two classes suggests fair handling of the authors' inputs.

References

1. Anthropic. (2024). Claude 3.5 Sonnet model card. Anthropic Technical Report. https://www.anthropic.com/news/claude-3-5-sonnet
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.
3. Chen, X., Zou, D., Cheng, G., & Xie, H. (2020). Detecting latent topics and trends in educational technologies over four decades using structural topic modeling. British Journal of Educational Technology, 51(3), 692-714.
4. Cingillioglu, I. (2023). Detecting AI-generated essays: The ChatGPT challenge. International Journal of Information and Learning Technology, 40(3), 259-268.
5. Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228-239.
6. Dalalah, D., & Dalalah, O. M. A. (2023). The false positives and false negatives of generative AI detection tools in education and academic research. The International Journal of Management Education, 21(2), 100822.
7. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171-4186).
8. Dugan, L., Hwang, A., Trhlikova, F., Ippolito, D., & Callison-Burch, C. (2024). RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In Proceedings of ACL 2024.
9. Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. International Journal for Educational Integrity, 19(1), 1-16.
10. Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical detection and visualization of generated text. In Proceedings of ACL 2019 System Demonstrations (pp. 111-116).
11. Google. (2024). Gemini 1.5 Pro technical report. Google DeepMind. https://deepmind.google/technologies/gemini/
12. Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., ... & Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus,

evaluation, and detection. arXiv preprint arXiv:2301.07597.

13. Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., ... & Goldstein, T. (2024). Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In Proceedings of ICML 2024.
14. Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., ... & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide agenda. International Journal of Artificial Intelligence in Education, 32(2), 504-526.
15. Hu, X., Chen, P.-Y., & Ho, T.-Y. (2023). RADAR: Robust AI-text detection via adversarial learning. Advances in Neural Information Processing Systems, 36.
16. Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. In Proceedings of ICML 2023.
17. Lai, J. W. M., & Bower, M. (2023). How is artificial intelligence (AI) being used in education? Computers & Education, 188, 104563.
18. Li, Y., Li, Z., Zhang, K., Dan, R., Zhang, S., & Zhang, Y. (2023). ChatGPT-like large-scale foundation models for prognostics and health management. Reliability Engineering & System Safety, 230, 108863.
19. Li, Y., Zhao, P., Pan, L., & Su, J. (2024). MAGE: Machine-generated text detection in the wild. In Proceedings of ACL 2024.
20. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), 100779.
21. Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., ... & Zou, J. (2024). Can AI-generated text be reliably detected? arXiv preprint arXiv:2303.11156v3.
22. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.
23. Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In Proceedings of ICLR 2019.
24. Meta AI. (2024). Llama 3 model card. Meta AI Research. https://ai.meta.com/blog/meta-llama-3/
25. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. In Proceedings of ICML 2023.
26. Montenegro-Rueda, M., Fernandez-Cerero, J., Fernandez-Batanero, J. M., & Lopez-Meneses, E. (2023). Impact of the implementation of ChatGPT in education: A systematic review. Computers, 12(8), 153.
27. Moorhouse, B. L., Wan Yim, S. S., & Li, A. M. (2023). Plagiarism and AI writing tools in higher education. Assessment & Evaluation in Higher Education, 48(8), 1143-1154.
28. OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.
29. Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era. Journal of University Teaching & Learning Practice, 20(2), 07.

30. Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2024). Gaming the system: A framework for detecting AI-generated text in student submissions. Journal of Academic Ethics, 22(1), 1-19.
31. Sarvazyan, A., Gonzalez, J. A., Rangel, F., Rosso, P., Chulvi, B., & Plaza, M. (2023). Overview of AuTexTification at IberLEF 2023. Procesamiento del Lenguaje Natural, 71, 275-290.
32. Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., ... & Wang, J. (2019). Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203.
33. Stokel-Walker, C. (2025). AI is now writing most university essays — educators are struggling to respond. Nature, 628, 14-16.