RA at GenAI Detection Task 2: Fine-tuned Language Models For Detection of Academic Authenticity, Results and Thoughts

Rana Gharib
Arab Academy For Science and Technology
ranagharib044@gmail.com

Ahmed Elgendy
Queen's Universty
ahmed.elgendy@queensu.ca

Abstract

This paper assesses the performance of "RA" in the Academic Essay Authenticity Challenge, which saw nearly 30 teams participating in each subtask. We employed cutting-edge transformer-based models to achieve our results. Our models consistently exceeded both the mean and median scores across the tasks. Notably, we achieved an F1-score of 0.969 in classifying AI-generated essays in English and an F1-score of 0.957 for classifying AI-generated essays in Arabic. Additionally, this paper offers insights into the current state of AI-generated models and argues that the benchmarking methods currently in use do not accurately reflect real-world scenarios.

1 Introduction

Large Language Models (LLMs), as advanced artificial intelligence (AI) systems, have been trained on vast amounts of text data and can generate, summarize and comprehend human languages with impressive fluency (Naveed et al., 2024). As these models are based on deep learning architectures, primarily transformer, they can learn complex language patterns and respond end-to-end with impressions very similar to human interaction. Recently, models such as GPT-4, Mistral (Jiang et al., 2023), LLaMA (Touvron et al., 2023), etc have enabled significant advances in AI language processing for natural language understanding and generation tasks.

Their promise, however, raises serious ethical and social concerns about honesty, transparency, and misuse (Abdurahman et al., 2024). A major area where AI could create change is education. AI offers personalized learning, providing students with tailored resources that enhance effectiveness. However, the accessibility of writing tools questions academic integrity, as students may bypass essential learning processes that promote critical thinking, creativity, and problem-solving skills.

Addressing these challenges is crucial for maximizing AI's benefits while minimizing its risks. The Academic Essay Authenticity 2024 Shared Task (Chowdhury et al., 2025) focuses on creating systems to distinguish human-written text from AI-generated content and provides a validation dataset. Our model builds on recent efforts by fine-tuning multiple language models using an ensemble approach. This paper covers data analysis, pipeline, results, related work, and future directions, highlighting real-world improvements.

2 Related Work

This section examines recent advancements in detecting content generated by large language models (LLMs). With the rapid adoption of LLMs and associated challenges, researchers have increasingly focused on this area. Numerous studies have introduced systems employing both deep learning and traditional machine learning to address the authenticity and reliability concerns of AI-generated content across various fields. The widespread use of AI-generative models has fueled the development of methods to detect text generated by these models, especially to safeguard integrity in domains such as education. Broadly, classification methods fall into two categories: white-box and black-box approaches. White-box methods require direct model access and include techniques like word-level probability analysis, perturbations (Mitchell et al., 2023), and local rank perturbations. In contrast, black-box methods work without model access and include supervised learning with linguistic features (Prova, 2024), supervised learning with pretrained LMs (Wang et al., 2023), and regeneration techniques. Recent years have also seen the creation of various datasets aimed at advancing AI-generated text detection (Fraser et al., 2024), alongside shared tasks dedicated to developing novel, robust approaches (Sarvazyan et al., 2023; Fivez et al.,

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pages 312–316
January 19, 2025. ©2025 International Conference on Computational Linguistics

2024; Molla et al., 2023). However, a research gap persists in detecting AI-generated text in Arabic. The datasets for Arabic are often sourced from various resources, lacking cohesion and presenting limited challenge. Notably, a specialized dataset for Arabic was created to aid model development but contains only 1,000 examples (Alshammari and EI-Sayed, 2023).

3 Dataset & Task Description

The shared task on Academic Essay Authenticity Challenge[1] consists of two main subtasks. Each subtask will be discussed in details in the following subsections. The provided dataset primarily comprises essays created either by a human or through promopting a generative language model. The subsequent subsections will present an overview of the distribution for each dataset, emphasizing the challenges posed by imbalances and a complete description of each dataset.

3.1 Subtask A: English Academic Essay Authenticity

The first subtask is a binary classification problem where essays given are classified into two distinct classes: "Human-Generated", and "AI-Generated". Table 1 illustrates the data distribution for the different classes within the dataset. The dataset comprises essays written by both human authors and AI systems. The human-authored essays have been curated from the ETS Corpus of Non-Native Written English. For the AI-generated content, we utilized seven diverse models, both open-source and proprietary, including GPT-3.5-Turbo, GPT-4o, GPT-4o-mini, Gemini-1.5, Llama-3.1 (8B), Phi-3.5-mini, and Claude-3.5.

	Training	Validation
Human	629	1235
AI	1467	391
Overall	2096	1626

Table 1: Subtask A's Dataset Distribution.

3.2 Subtask B: Arabic Academic Essay Authenticity

The second subtask is a binary classification problem where essays given are classified into two classes: "Human-Generated", and "AI-Generated". Table 2 illustrates the data distribution for the different classes within the dataset.

	Training	Validation
Human	1145	182
AI	925	299
Overall	2070	481

Table 2: Subtask B's Dataset Distribution.

3.3 Data Preprocessing

For both subtasks, no additional data preprocessing steps were applied beyond those inherent to the models themselves. This decision was based on the rationale that AI models, unlike humans, exhibit distinctive patterns in their writing, such as the frequency of punctuation marks and spelling accuracy, among other aspects, which can serve as discriminative features for our models.

4 Methodology

4.1 Language Models

Several language models were experimented with through the process of fine-tuning, driven by their remarkable performance in the context of our specific topic, We finetuned RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020), mBERT and DeBERTa (He et al., 2021) for subtask A. All of the models showed similar performance on the validation set but for mBERT which has a slightly less performance. As for subtask B, we fine-tuned AraBERT (Antoun et al., 2020), ArBERT and MarBERT (Abdul-Mageed et al., 2021). AraBERT showed superior performance in terms of F1-score on all of the subtask as will be shown in the results section.

4.2 Loss Function

In the experimentation with various loss functions to optimize model performance, several options were tested, including Cross-Entropy Loss, Focal Loss, Tversky Loss, and Dice Loss. Each of these loss functions was evaluated based on their ability to handle class imbalance and improve the model's predictive accuracy. It was found that Focal Tversky Loss (Abraham and Khan, 2018) and Dice Loss (Li et al., 2020) produced the best results in terms of balancing sensitivity and specificity. However, Dice Loss was ultimately chosen for its superior performance, as it consistently outperformed the

[1] https://codalab.lisn.upsaclay.fr/competitions/20118

others in handling overlapping classes and achieving higher overall performance during the validation phase. Therefore, Dice Loss was selected as the final loss function for the model.

$$\text{Dice Loss} = 1 - \frac{2 \cdot \sum y_{\text{true}} \cdot y_{\text{pred}} + \epsilon}{\sum y_{\text{pred}}^2 + \sum y_{\text{true}}^2 + \epsilon} \quad (1)$$

4.3 Majority Voting

Majority voting is an ensemble technique where multiple classifiers make predictions, and the final prediction is based on the most frequent class label. This method helps mitigate issues like overfitting and bias by combining the strengths of different models, leading to improved accuracy and robustness. It reduces the impact of errors from individual classifiers, providing a more reliable and generalized prediction. Equation 2 illustrates majority voting.

$$\hat{y} = \arg \max_{c \in \{c_1, c_2, \ldots, c_k\}} \sum_{i=1}^{n} \delta(y_i = c) \quad (2)$$

4.4 Experiment Settings

The training procedure was conducted using Kaggle's [2] free-to-use platform, which provides 29 GB of RAM, a 16 GB NVIDIA P100 GPU, and Python. The autofit functionality from ktrain (Maiya, 2022) was utilized, incorporating a triangular learning rate policy (Smith, 2017).

Hyperparameter	Task 3	Task 6
Epochs	5	5
Learning Rate	2e-5	1e-5, 2e-5
Batch Size	8	8, 4
Max length	350	350
Optimizer	Adam	Adam
Early Stopping Patience	3	3
Reduce On Plateau	2	2
Loss Function	Dice Loss	Dice Loss

Table 3: Training Hyperparameters. Parameters shown for RoBERTa, DeBERTa and XLM-RoBERTa for tasks A and AraBERT for task B, respectively.

5 Results

5.1 Subtask A

Table 4 illustrates our ensemble-based model's performance on the test set . The Ensemble-based model used a majority voting scheme for DeBERTa, Roberta and XLM-RoBERTa. Our approach ranked 12th in the overall rankings leaderboard.

Model	Precision	Recall	F1-Score
Top-3 Ensemble	0.975	0.964	0.969

Table 4: Results For Subtask A.

This straightforward, quick-to-train, and easy-to-implement online learning upon approach secured 12th place in Subtask A. We opted for a relatively simple model to demonstrate that current basic methods can effectively handle datasets, though they may encounter challenges in real-world applications.

5.2 Subtask B

Table 5 illustrates our ensemble-based model's performance on the test set . The Ensemble-based model used a majority voting scheme for different fine-tuned version of AraBERT. Our approach ranked 6th in the overall rankings leaderboard. This straightforward, quick-to-train, and easy-to-implement online learning upon approach secured 6th place in Subtask B. We opted for a relatively simple model to demonstrate that current basic methods can effectively handle datasets, though they may encounter challenges in real-world applications.

Model	Precision	Recall	F1-Score
Top-3 Ensemble	0.956	0.959	0.957

Table 5: Results For Subtask A.

6 Discussion and Future Work

The relative success of our model highlights the potential for language models to serve as effective tools for detecting AI-generated text. However, we believe that the current benchmarking and fine-tuning approaches have certain limitations, particularly because they overlook the complexities present in real-life scenarios. Unlike controlled experimental settings, practical applications of AI detection face a range of unpredictable variables that make straightforward classification difficult.

In recent years, various tactics have emerged among internet users to bypass AI detectors. Some

[2] https://www.kaggle.com/

of these strategies involve adding "human" features to the text, such as intentional spelling mistakes, varied linguistic complexity, shifts between active and passive voice, or even missed punctuation marks. Other methods aim to modify the generated text from the model's perspective, employing techniques like repetitive paraphrasing, contextual word substitutions, random alterations (including word swaps or deletions), and sentence-level rearrangements. More advanced strategies include combining outputs from multiple models or utilizing auto-completion to produce hybrid texts, adding further layers of complexity.

This phenomenon has been explored extensively in the literature concerning English essays (Perkins et al., 2024). Yet, to our knowledge, it remains largely unexplored in the context of Arabic language detection. Given the linguistic richness and structural complexity of Arabic, this language poses unique challenges for models fine-tuned on existing datasets, potentially requiring new and specialized approaches for effective detection.

Developing a comprehensive dataset that encompasses all these approaches and beyond is an exciting direction for us. Additionally, we believe that multilingual language models, despite their impressive capabilities, exhibit a distinct linguistic signature. This opens up opportunities for research into the multilingual aspect, where data from various sources—such as generative models and multiple languages—can be utilized to train our detectors, allowing us to observe the effects of multilingual data.

7 Conclusion

This paper introduces an approach for detecting academic authenticity using an ensemble of language models. Despite its simplicity, the method achieves high performance after only a few epochs. While this is advantageous, it also has drawbacks. The straightforward nature of the approach, when trained on benchmark datasets, may not accurately represent its performance in real-world scenarios. We discuss several factors that could challenge the model's effectiveness and call on researchers to address these challenges in future work.

Acknowledgments

This research is supported by the Vector Scholarship in Artificial Intelligence, provided through the Vector Institute as the second author is a receipt of the Vector Scholarship for the year 2024-2025.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. Arbert marbert: Deep bidirectional transformers for arabic. *Preprint*, arXiv:2101.01785.

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, 3(7).

Nabila Abraham and Naimul Mefraz Khan. 2018. A novel focal tversky loss function with improved attention u-net for lesion segmentation. *Preprint*, arXiv:1810.07842.

Hamed Alshammari and Ahmed EI-Sayed. 2023. Airabic: Arabic dataset for performance evaluation of ai detectors. *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 864–870.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Shammur Absar Chowdhury, Hind Al-Merekhi, Mucahid Kutlu, Kaan Efe Keleş, Fatema Ahmad, Tasnim Mohiuddin, Georgios Mikros, and Firoj Alam. 2025. GenAI content detection task 2: AI vs. human – academic essay authenticity challenge. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Pieter Fivez, Walter Daelemans, Tim Van de Cruys, Yury Kashnitsky, Savvas Chamezopoulos, Hadi Mohammadi, Anastasia Giachanou, Ayoub Bagheri, Wessel Poelman, Juraj Vladika, Esther Ploeger, Johannes Bjerva, Florian Matthes, and Hans van Halteren. 2024. The clin33 shared task on the detection of text generated by large language models. *Computational Linguistics in the Netherlands Journal*, 13:233–259.

Kathleen C. Fraser, Hillary Dawkins, and Svetlana Kiritchenko. 2024. Detecting ai-generated text: Factors influencing detectability with current methods. *Preprint*, arXiv:2406.15583.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Arun S. Maiya. 2022. ktrain: A low-code library for augmented machine learning. *Preprint*, arXiv:2004.10703.

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

Diego Molla, Haolan Zhan, Xuanli He, and Qiongkai Xu. 2023. Overview of the 2023 ALTA shared task: Discriminate between human-written and machine-generated text. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association*, pages 148–152, Melbourne, Australia. Association for Computational Linguistics.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models. *Preprint*, arXiv:2307.06435.

Mike Perkins, Jasper Roe, Binh H. Vu, Darius Postma, Don Hickerson, James McGaughran, and Huy Q. Khuat. 2024. Simple techniques to bypass GenAI text detectors: implications for inclusive education. *International Journal of Educational Technology in Higher Education*, 21(1).

Nuzhat Prova. 2024. Detecting ai generated text based on nlp and machine learning approaches. *Preprint*, arXiv:2404.10032.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *Preprint*, arXiv:2309.11285.

Leslie N. Smith. 2017. Cyclical learning rates for training neural networks. *Preprint*, arXiv:1506.01186.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Zecong Wang, Jiaxi Cheng, Chen Cui, and Chenhao Yu. 2023. Implementing bert and fine-tuned roberta to detect ai generated news by chatgpt. *Preprint*, arXiv:2306.07401.