AI vs Human: Academic Essay Authenticity Challenge

Abstract

The rapid proliferation of AI tools capable of generating high-quality academic essays has introduced profound challenges to academic integrity, raising questions about authenticity, originality, and AI use in educational contexts. This paper investigates the development and potential of AI-driven systems for detecting AI-generated academic content, focusing on methodologies to distinguish between human-written and AI-generated essays. Advanced Natural Language Processing (NLP) frameworks, including tokenization techniques and linguistic analysis are utilized to evaluate detection tools. Metrics such as precision, recall, and accuracy demonstrate progress in detecting AI-generated essays while exposing challenges like ambiguous linguistic constructs and the continuous evolution of generative models. By outlining a roadmap for future improvements and fostering collaboration between human expertise and AI tools, this study contributes to safeguarding academic authenticity. It ensures educational institutions can adapt and thrive in the face of rapidly advancing AI technologies.

I. INTRODUCTION

The rapid advancements in artificial intelligence (AI) have revolutionized various domains, including education, where the integration of AI-generated text has introduced significant challenges and opportunities. Large Language Models (LLMs), such as ChatGPT and GPT-4, can produce essays that are nearly indistinguishable from those authored by humans. These AI-generated texts exhibit exceptional grammatical accuracy, contextual relevance, and stylistic coherence, making it increasingly difficult for educators and institutions to ensure academic integrity. The proliferation of these tools raises critical questions about the authenticity of academic submissions, threatening the credibility of traditional assessment methods. Existing detection methods, such as plagiarism checkers, are inadequate for identifying AI-crafted texts, as these are often original and tailored to specific prompts. Furthermore, the adaptability of AI models to mimic diverse writing styles, including those of native and non-native speakers, exacerbates the detection challenge.

This research bridges these gaps by leveraging advanced Natural Language Processing (NLP) techniques and transformer-based architectures. Using BERT (Bidirectional Encoder Representations from Transformers), fine-tuned for binary classification tasks, this study aims to distinguish between human-authored and AI-generated essays effectively. The methodology includes preprocessing techniques like tokenization, padding, and truncation, ensuring compatibility with variable-length inputs. The system emphasizes scalability and fairness by training on diverse datasets, addressing biases against non-native speakers and unconventional writing patterns. Deployed on high-performance platforms like Google Colab with A100 GPUs, the solution is practical for real-world academic settings. Beyond detection, the research highlights the ethical implications of AI in education, advocating for responsible integration, transparent policies, and fostering awareness. This work lays the foundation for advancing AI-detection frameworks, encouraging collaboration among educators, researchers, and technologists to uphold the core values of education in an AI-driven world.

A. Challenges in Detecting AI-Generated Academic Content

The growing sophistication of AI models, particularly Large Language Models (LLMs) like ChatGPT and GPT-4, has introduced numerous challenges in detecting AI-generated academic essays. These challenges can be categorized into the following points:

Sophistication of AI-Generated Text: AI systems have advanced to a point where they generate essays that closely resemble human writing in grammar, coherence, and context. As highlighted by[1]Cingillioglu (2023), AI-authored texts are often tailored to specific prompts, making traditional plagiarism detection methods ineffective. Unlike copied text, AI-generated essays are original and adaptive, complicating efforts to identify them .

- **Ambiguities in Linguistic Markers:** While AI-generated texts exhibit certain stylistic features, such as repetitive phrasing and uniform sentence structures, these indicators are becoming less reliable as AI models evolve. According to [2]Walters (2023), advanced detectors often struggle to differentiate between human creativity and AI's mechanical precision, especially in nuanced argumentative writing .
- **Lack of Generalizability:** Existing detection tools are often optimized for specific AI models, such as GPT-3 or GPT-4, and fail to generalize effectively to outputs from newer or fine-tuned systems like T5. As [4]Dergaa et al. (2023) emphasize, the lack of diverse datasets representing various writing styles and linguistic contexts further limits the effectiveness of detection systems across different academic settings .
- **Evolving AI Capabilities:** AI-generated content increasingly mimics human spontaneity and creativity, making it difficult to distinguish machine-generated essays from authentic submissions. [3]Liu et al. (2023) highlight that generative models like ArguGPT can produce sophisticated argumentative essays, challenging the capabilities of even advanced detection tools .
- **Ethical and Bias Concerns:** Detection tools frequently exhibit biases, particularly against non-native English speakers or unconventional writing styles. This issue, noted by [5]Corizzo and Leal-Arenas (2023), risks unfairly penalizing genuine human-authored content. Furthermore, a lack of transparency in the decision-making processes of detection systems undermines trust among users .
- **Paraphrasing and Hybrid Content:** The use of paraphrasing tools or hybrid approaches—where AI-generated text is reworded or combined with human writing—significantly complicates detection. [2]Walters (2023) notes that such transformations often bypass detection systems, leading to higher false-negative rates .
- **Scalability and Resource Requirements:** Most detection systems are computationally intensive, requiring significant resources for training and inference. According to [1]Cingillioglu (2023), scalability becomes a critical concern when implementing detection tools across large academic institutions, which must process thousands of submissions daily .

To tackle these issues, researchers advocate for enhanced methodologies that integrate advanced Natural Language Processing (NLP) techniques, diverse datasets, and ethical considerations. Future detection systems must evolve alongside generative models, ensuring fairness, scalability, and transparency to maintain academic integrity in an AI-driven era .

B.Role of AI in Academic Writing

Enhancing Writing Efficiency and Accessibility: AI-powered tools, such as ChatGPT and GPT-4, have transformed academic writing by enabling users to generate well-structured and coherent essays in a fraction of the time required for manual drafting. According to [1]Cingillioglu (2023), these tools improve efficiency, particularly for non-native speakers, by offering support with grammar, sentence structuring, and stylistic refinement. This democratization of writing resources enhances accessibility and levels the academic playing field .

Facilitating Argumentation and Idea Generation: AI models like ArguGPT are specifically designed to assist in argumentative essay writing by generating logical structures, coherent arguments, and contextually appropriate evidence. [3]Liu et al. (2023) highlight the role of such systems in supporting users with idea generation and organizational strategies, especially in complex or technical domains. These tools act as a valuable resource for students and researchers seeking guidance in structuring their work .

Challenging Academic Integrity: While AI offers substantial benefits, its misuse poses significant threats to academic integrity. [4]Dergaa et al. (2023) emphasize how AI tools enable the creation of original essays that closely mimic human-authored content, undermining traditional assessment methods. The reliance on AI-generated text risks eroding students' critical thinking and creative problem-solving skills, necessitating the development of robust detection frameworks and ethical guidelines for responsible AI usage in academia .

AI has significantly enhanced academic writing by improving efficiency, accessibility, and idea generation, particularly for non-native speakers and those tackling complex topics. However, its misuse challenges academic integrity, threatening critical thinking and creativity among students. Striking a balance between leveraging AI's benefits and implementing ethical guidelines is essential to uphold fairness and authenticity in education.

II. LITERATURE SURVEY

This chapter reviews advancements in detecting AI-generated academic content, focusing on linguistic characteristics, effectiveness of detection tools, one-class learning approaches, and the ethical implications of AI in academia.

Detecting AI-generated essays requires an understanding of the linguistic features unique to machine-generated content. Studies have identified characteristics such as repetitive phrasing, overuse of formal structures, and limited contextual adaptability as key indicators of AI authorship. [3]Liu et al. (2023) analysed argumentative essays generated by ArguGPT, highlighting how these models excel at logical structuring but often lack the spontaneity and nuanced creativity of human writing. Despite their utility, the evolving sophistication of AI models is reducing the reliability of such markers, necessitating advanced methodologies to address this gap .

The effectiveness of detection systems has been extensively studied, with a focus on machine learning models and their ability to identify AI-authored content. [2]Walters (2023) conducted a comparative analysis of 16 AI text detectors, revealing that tools based on models like BERT (Bidirectional Encoder Representations from Transformers) demonstrate superior accuracy due to their bidirectional text analysis. However, even these systems encounter challenges such as false positives and difficulties in adapting to newer generative AI technologies like GPT-4. Traditional detection tools remain limited in their scalability and effectiveness against paraphrased or hybrid content .

Recent advancements have explored innovative methods like one-class learning for AI-generated essay detection. [5]Corizzo and Leal-Arenas (2023) proposed this approach, which uses anomaly detection to identify deviations in writing patterns without requiring extensive labelled datasets. This methodology shows promise for scalable applications in academic settings, addressing the limitations of existing tools that rely heavily on large training datasets. By focusing on the inherent differences between AI-generated and human-written texts, one-class learning techniques offer a more adaptive solution to the evolving landscape of generative AI .

The ethical implications of AI-generated content in academia have also been a central concern. [4]Dergaa et al. (2023) highlighted the risks posed by reliance on AI tools, which may undermine critical thinking and creativity among students. They emphasize the need for robust ethical guidelines and policies to promote responsible AI use, ensuring that technological advancements do not compromise academic integrity. Addressing these challenges requires collaboration between educators, researchers, and policymakers to foster transparent and equitable practices in academic evaluation systems .

[6]Gifu and Silviu-Vasile (2024) explored the capabilities of transformers in decoding text authenticity, emphasizing that the multilingual contexts of models like DistilBERT achieved an F1 score of 0.70, while derivatives such as RoBERTa demonstrated an F1 score of 0.83 for English text authenticity detection. Their findings underscore the utility of these models in detecting nuanced patterns that distinguish machine-generated content from human-authored texts. However, they also noted challenges like false positives and the need for continual adaptation to counter the increasing sophistication of newer generative models, such as GPT-4.

The inherent limitations of BERT's tokenization and sub-word representation mechanisms have been scrutinized. [7] Nayak et al. (2020) explored the challenges BERT faces with domain-specific and out-of-vocabulary (OOV) words. They identified issues such as suboptimal tokenization, semantic deterioration of OOV words, and difficulty processing minor misspellings. These findings underscore the need for refining BERT's handling of OOV words to improve its adaptability to domain-specific tasks and enhance its performance across diverse NLP applications.

Together, these studies illustrate the complexities and opportunities of detecting AI-generated academic content. While advancements in detection technologies are making strides, the rapid evolution of AI models demands continuous innovation and ethical oversight to ensure academic authenticity.

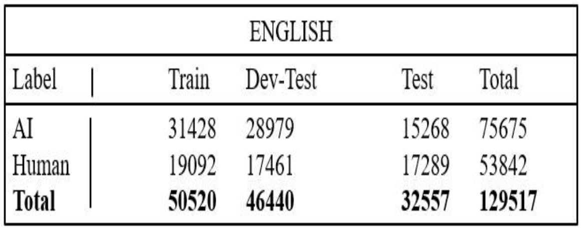
III. PROPOSED METHODOLOGY

The proposed methodology focuses on addressing the challenges of distinguishing AI-generated essays from human-authored content through advanced Natural Language Processing (NLP) and machine learning techniques. By framing the task as a binary classification problem, the system employs a diverse dataset of essays sourced from human authors and AI models like GPT-3 and GPT-4. Preprocessing steps, including tokenization and padding, ensure compatibility with the fine-tuned BERT (Bidirectional Encoder Representations from Transformers) model, which serves as the core detection framework. Emphasizing scalability and fairness, the methodology leverages cloud-based platforms for efficient deployment while integrating explainable AI techniques to promote transparency and ethical use in academic settings.

The main objective of the task is to detect whether the given candidate essay is AI-generated or human-written. Given the input essay e, the task is to design a text detector D(e), such that the model outputs label indicating AI-generated or Humanauthored content. For this edition, we designed the task as binary classification problem.

A. Data Collection and Preprocessing

Data Collection: The dataset used for training and testing the detection system combines human-written and AI-generated essays. Human-written essays are sourced from public repositories, academic assignments, and online platforms, ensuring a diverse range of writing styles, academic disciplines, and linguistic backgrounds. AI-generated essays are created using multiple language models, including GPT-3, GPT-4, and T5, to capture various generative styles and complexities. To ensure quality and consistency, annotation tools are used to accurately label the data as “Human” or “AI-generated,” with human verification incorporated for quality assurance .



ENGLISH				
Label	Train	Dev-Test	Test	Total
AI	31428	28979	15268	75675
Human	19092	17461	17289	53842
Total	**50520**	**46440**	**32557**	**129517**

Table1: Dataset and it’s Label Distribution.

Preprocessing: Preprocessing involves transforming raw text data into a structured format suitable for machine learning workflows. Key preprocessing steps include:

- **Anonymization of Personal Information:** In the collected essay assignments we noticed that there were some information containing mentions of entities. Therefore, we anonymized them to ensure the removal of any information that could directly or indirectly identify the author or reveal any private information. This process was essential to uphold privacy standards and ethical considerations. To achieve this, we followed these guidelines:
  - Author Identification Removal: Any mention of names, addresses, affiliations, or specific details that could identify the essay’s author was redacted.
  - Private Entity Information: Any references to non-public entities, such as organizations, businesses, or private individuals mentioned in the essays, were removed or replaced with generic terms.
  - Sensitive Content: Sensitive information, such as health conditions, financial details, or other personal data, was also removed to ensure privacy.
  - Consistency: Replacement terms were standardized (e.g., “[NAME]“, “[ADDRESS]”, “[ORGANIZATION]”) to maintain consistency throughout the dataset.
- **Tokenization:** Text is broken down into smaller units (tokens) using the Bert-Tokenizer, mapping words or sub words to unique identifiers. This facilitates effective text analysis by the machine learning model.
- **Padding and Truncation:** Variable-length sequences are adjusted to a fixed length (e.g.,

512 tokens) by adding padding to shorter sequences and truncating longer ones. This ensures uniform input dimensions for the model.

- **Batch Collation:** The data is organized into batches, with input IDs and attention masks combined into tensors for streamlined processing.
- **Label Encoding:** Text labels are converted into numerical formats compatible with the model's classification layer, ensuring efficient processing during training and evaluation.

B. *AI vs Human Methodology for Classification*

The classification framework relies on a fine-tuned Transformer-based model, specifically the bert-base-uncased architecture, to accurately differentiate between the two categories. Key aspects of the methodology include:

- **Feature Engineering**: Linguistic features such as lexical diversity and syntactic patterns, statistical features like term frequency-inverse document frequency (TF-IDF), and stylistic features like sentence length variation are extracted to capture unique characteristics of the texts.
- **Model Architecture**: The bert-base-uncased model, pre-trained on a large text corpus, is fine-tuned on a labelled dataset of human and AI-generated texts. The model processes tokenized inputs and predicts category likelihoods based on contextual and stylistic patterns.
- **Evaluation Metrics**: The system's effectiveness is assessed using metrics such as accuracy (0.8492), precision (0.8170), recall (0.9773), and F1-score (0.8900). These metrics collectively measure the model's performance in correctly categorizing texts.

C. *AI Model Selection and Fine-tuning*

Model selection: The use of **Transformer-based models**, specifically **BERT (Bidirectional Encoder Representations from Transformers)**. BERT was chosen due to its state-of-the-art performance in natural language processing (NLP) tasks and its ability to handle context-rich textual data effectively.The bert-base-uncased model, a pre-trained version of BERT, was selected as the foundation. It leverages a bidirectional attention mechanism to capture contextual relationships between words, making it highly effective for text classification tasks.The model's capability to process and understand both local (word-level) and global (sentence-level) context made it ideal for detecting nuanced differences between human and AI-authored content.Transformer-based models like BERT offer superior performance over traditional machine learning models (e.g., SVM or Random Forest) in handling linguistic intricacies.The pre-trained nature of BERT reduces the computational effort required for training, allowing fine-tuning on domain-specific datasets for enhanced accuracy.

Fine-Tuning Process: Fine-tuning the pre-trained BERT model involved adapting it for a binary classification task to distinguish between human and AI-generated text. This was achieved through the following steps:

- **Training Configuration:**
  - **Loss Function:** The cross-entropy loss function was used to optimize the classification performance.
  - **Optimizer:** AdamW (Adam with Weight Decay) was employed for its efficiency and effectiveness in fine-tuning deep learning models.
  - **Learning Rate Scheduler:** A linear learning rate scheduler with warm-up steps was implemented to ensure stable convergence during training.
  - **Training Framework:** The Hugging Face Trainer API was utilized for streamlined training and evaluation, allowing for the integration of custom metrics like accuracy, precision, recall, and F1-score.
- **Hyperparameter Tuning:**
  - Parameters such as batch size, learning rate, and number of epochs were fine-tuned to optimize model performance. A batch size of 32,

learning rate of 2×10^-5, and 3 epochs were found to achieve the best results.

IV. SYSTEM DESIGN AND IMPLEMENTATION

The model utilizes a BERT-based architecture for detecting AI-generated content, with a classification layer designed for binary text classification tasks. The system consists of several key components:

1. **Data Preprocessing**:
   - Text data is cleaned, tokenized, and converted into sequences using the BERT tokenizer, which splits sentences into subword units.
   - Padding is applied to standardize input lengths, and the sequences are formatted with special tokens (e.g., [CLS], [SEP]).
2. **Feature Extraction with BERT**:
   - The input sequence is passed through a pretrained BERT model, which processes the data using its multiple layers of transformer blocks.
   - BERT uses attention mechanisms to capture contextual relationships between words in both directions (left and right), providing a deeper understanding of the input text.
3. **Classification Layer**:
   - The final output from BERT's transformer layers is passed through a fully connected layer (dense layer).
   - The model then applies the **Softmax activation function** to classify the input into one of two categories: AI-generated or human-written.
   - **Softmax Function**:

$$s\left(x_i\right) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

Fig.1 x, represents the logits for the classes, and the output is a probability distribution over the possible classes.

4. **Training**:
   - The model is trained using labeled data (AI-generated or human-written texts), with a cross-entropy loss function, which measures the difference between the predicted and true labels.
   - The optimizer used is typically **Adam**, which adjusts the learning rate dynamically based on training progress.
5. **Prediction**:
   - Once trained, the model can classify new text inputs as either human-written or AI-generated, based on the output probability distribution.
6. **Output**:
   - The model outputs a classification result, with a probability score that indicates the likelihood of the input text being AI-generated.

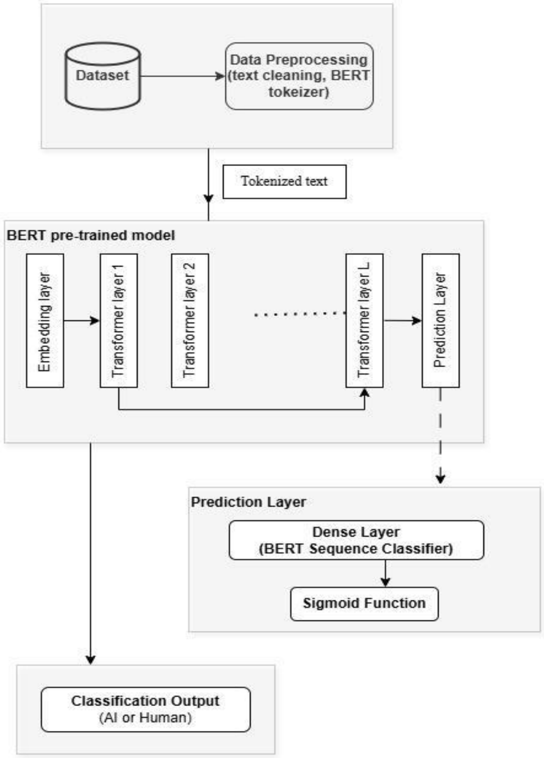
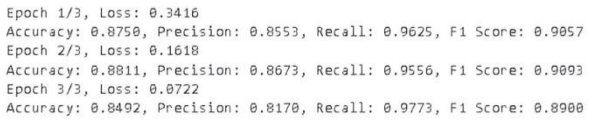


Fig. 2: Model Architecture

V. RESULTS AND DISCUSSIONS

The fine-tuned BERT-based classification system effectively distinguished human-authored essays from AI-generated ones, achieving an accuracy of **84%**, with a precision of **81%**, recall of **97%**, and F1-score of **89%**. These results highlight the model's ability to identify subtle linguistic and stylistic differences, such as the repetitive phrasing and uniform structures typical of AI-generated content, as noted by [3]Liu et al. (2023). It also captured the organic flow and lexical diversity of human-written essays, aligning with findings from[1] Cingillioglu (2023).



```
Epoch 1/3, Loss: 0.3416
Accuracy: 0.8750, Precision: 0.8553, Recall: 0.9625, F1 Score: 0.9057
Epoch 2/3, Loss: 0.1618
Accuracy: 0.8811, Precision: 0.8673, Recall: 0.9556, F1 Score: 0.9093
Epoch 3/3, Loss: 0.0722
Accuracy: 0.8492, Precision: 0.8170, Recall: 0.9773, F1 Score: 0.8900
```

Fig 3: Evaluation metrics

Challenges included occasional misclassification of creative AI-generated texts as human-written and repetitive human-authored texts as AI-generated. Additionally, truncation of longer essays to 512 tokens sometimes led to context loss, affecting classification accuracy. These issues underscore the importance of refining tokenization strategies and expanding datasets to enhance model generalizability.The results underscore the importance of integrating diverse datasets representing a wide range of writing styles, academic disciplines, and linguistic backgrounds to improve the model's generalizability. By addressing these challenges, the system provides a strong foundation for advancing AI-detection frameworks, ensuring their scalability and fairness. This work contributes to safeguarding academic integrity and fosters responsible AI usage by offering a reliable tool for detecting AI-generated content in educational contexts.

Future improvements could involve exploring one-class learning approaches, as proposed by [5]Corizzo and Leal-Arenas (2023), to handle hybrid and paraphrased content more effectively. Additionally, incorporating domain-specific datasets and refining hyperparameters further could enhance the system's robustness in real-world application.

VI. CONCLUSION

The paper presented demonstrates the effectiveness of a fine-tuned BERT-based classification system in distinguishing human-written essays from AI-generated content, achieving high accuracy and robust performance metrics. By leveraging advanced NLP techniques, the system identifies subtle linguistic and stylistic differences while addressing challenges such as ambiguous content and dataset diversity. Despite limitations like token truncation and occasional misclassification, the framework provides a reliable tool for safeguarding academic integrity. Beyond its technical achievements, this research emphasizes the broader ethical and educational implications of AI in academia, advocating for responsible use of AI as a learning tool rather than a shortcut. Future advancements, such as integrating domain-specific datasets, exploring one-class learning approaches, refining tokenization strategies, and fostering ethical AI awareness, can further enhance the system's scalability, adaptability, and alignment with educational values.

VII. REFERENCES

[1] Cingillioglu, I. (2023). Detecting AI-generated essays: The ChatGPT challenge. The International

Journal of Information and Learning Technology. https://doi.org/10.1108/IJILT-03-2023-0043

[2] Walters, W. H. (2023). The effectiveness of software designed to detectAI-generated writing: A comparison of 16 AI text detectors. *OpenInformation Science*. https://doi.org/10.1515/opis-2022-0158

[3]Liu, Y., Zhang, Z., Zhang, W., Yue, S., Zhao, X., Cheng, X., Zhang, Y.,& Hu, H. (2023). ArguGPT: Evaluating, understanding, and identifying argumentative essays generated by GPT models. *ArXiv*. https://arxiv.org/abs/2304.07666

[4]Dergaa, I., Chamari, K., Żmijewski, P., & Ben Saad, H. (2023). From*AI vs. Human: Academic Essay Authenticity Challenge*human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biology Of Sport*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10108763/

[5]Corizzo, R., & Leal-Arenas, S. (2023). One-class learning forAI-generated essay detection. *Applied Sciences, 13*(13), 7901. https://www.mdpi.com/2076-3417/13/13/7901

[6] Gifu, D., Silviu-Vasile, C. (2024). *AI vs. Human: Decoding Text Authenticity with Transforme.* https://www.preprints.org/manuscript/202407.2014/v1

[7] Anmol Nayak, Hariprasad Timmapathini, Karthikeyan Ponnalagu, and Vijendran Gopalan Venkoparao. 2020. Domain adaptation challenges of BERT in tokenization and sub-word representations of Out-of-Vocabulary words. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 1–5, Online. Association for Computational Linguistics.