Grenze journal name, Volume. 1, Issue.1, May 2015

AI VS HUMAN – ACADEMIC ESSAY AUTHENTICITY CHALLENGE

Sandhya L[1], Vennapoosa Sri Manjunath Reddy[2], Rakshitha M[3], Janga Durga Venkata Sathya Sai Mallika[4], Pothana Boyina Hari Narayana[5]

[1]Presidency University/Assistant Professor Department of CSE & IS, Bangaluru,India

[1]Email: saisandhyalax@gmail.com

[2-5] Department of Computer Science and Technology(AI & ML), Presidency University

Email: Reachmemanju15@gmail.com, shettym041@gmail.com, mallikajdvss099@gmail.com, harinarayana537@gmail.com

***Abstract*—The appearance of large language models (LLM) including GPT 3.5 and GPT-4 have called into question the integrity of academic writing, as it is becoming increasingly more difficult to distinguish between essays written by humans and those that are AIgenerated. This project seeks to create an answer to this challenge by designing a binary classification model, to identify if an essay was written by a human, or a machine. The dataset contains essays written by humans, which were gathered from the ETS Corpus of Non-Native Written English, alongside essays written by AI from seven LLMs: GPT-3.5-Turbo, GPT-4o, Gemini-1.5, Llama-3.1 (8B), Phi-3.5-mini, Claude-3.5. Utilizing natural language processing (NLP) techniques and machine learning algorithms, this system will skim for linguistic patterns in academic essays written in English and Arabic in order to separate either human essays from AI induced writing. It is a solution that will support academic authenticity and lessen academic dishonesty with the misuse of AI tools in the university classroom to help our institutions support fair academic practices.**

***Index Terms*—Authentication, Firebase.**

I. INTRODUCTION

Education has faced many issues as a result of the rapid advancement of artificial intelligence (AI), particularly with regard to large language models (LLMs). The sophistication of models such as GPT-3.5-Turbo and GPT-4o has increased to the point where they can now generate text that is human- like. The inability to tell whether a piece of writing is an essay written by a student or a response generated by a machine or artificial intelligence program has caused anxiety in academic contexts. More precisely, the capacity of an AI model to generate well-structured, cohesive writings presents moral and practical issues regarding equity in the classroom and in academic settings. One of the most pressing ethical concerns is that students may choose to act unethically and submit an essay written entirely by an LLM such that the LLM would be cited as the author. If students undermine the purpose of education - to learn - and produce written texts that are entirely the work of an AI tool - this will most definitely call into question the original in tentor quality an institution intends to measure within its curriculum. To access or evaluate students' writing, more than likely they will measure the originality or quality of the essay in its entirety. Determining if an essay in fact was sincerely authored by the student, or rather an AI-generated text is the need for this project. As such, this project, "AI vs Human: Academic Essay Authenticity Challenge," involves the design of a binary classification challenge that will assess essays based on whether they are authored by humans versus essays generated by AI. The dataset used for this project will be comprised of essays which were written by human authors, specifically including essays from the ETS Corpus of Non- Native Written English. This Corp is a robust database inclusive of a myriad of essays developed by human authors who are non-native English speakers. Since human authors from across the globe participated in writing the essays, the essays cover writing at a range of proficiency levels and an amalgamation of syntactic styles created by various authors. The essays taken form this corpus will create a more challenging assessment task for the binary challenge between the produced human essays and the AI generated responses. Alongside the human-created essays, machine-written essays will be obtained from seven different LLMs (Large language models): GPT-3.5-Turbo, GPT-4o, GPT-4omini, Gemini-1.5, Llama-3.1 (8B), Phi-3.5-mini, and Claude-3.5. The model will analyze linguistic structures and structural and stylistic features to distinguish between human and AI-essays so that academic institutions can protect the authenticity and integrity of students' work.

II. PROBLEM STATEMENT

The objective is to identify machine-generated essays to safeguard academic integrity and prevent the misuse of large language models in educational settings. The input to the system would be essays including texts authored by both native and non-native speakers, as well as essays generated by various large language models. The task is defined as follows: "Given an essay, identify whether it is generated by a machine or authored by a human." This is a binary classification task and is offered in English.

III. EASE OF USE

The simplicity of usage of essays produced by AI is one of the factors contributing to their growing popularity. AI systems can produce comprehensive, well-structured essays in a matter of seconds with very little input, usually only a subject or prompt. Students who are under academic pressure can save time with this aptitude, but it can also promote reliance and diminish critical thinking. Human-written essays, on the other hand, need extensive research, draughting, and revision; these processes take time but promote creativity, critical thinking, and a better comprehension of the subject. Human effort guarantees skill improvement and personal progress, even while AI streamlines the process.

IV. LITERATURE SURVEY

Table 1:- Data of recent research

SL. NO	Study	Methodology	Results	Key Limitations
1	Deep Fusion Model for Human vs. Machine-Generated Essay Classification	Used Bidirectional LSTM and Linguistic Features for classification.	Performance depends on the AI tool used; metrics like F1-score, Precision, and Recall were reported.	Performance varies with different AI writing tools. - Struggles with nuanced/context-heavy essays. - Potential overfitting due to limited dataset diversity. - Metrics provide limited insights into essay quality, style, and context.
2	Assessing AI Detectors in Identifying AI-Generated Code	Evaluated 5 AIGC detectors (GLTR, GPT-2 Detector, GPTZero, Sapling, DetectGPT) on AI-generated Python code.	GLTR accuracy ranged from 0.4841 to 0.7693, p-value = 0.0296. All tools struggled with false positives/negatives.	Tools showed high rates of false positives/negatives. - GPTZero struggled with syntax detection. - Overall low accuracy for distinguishing human vs. AI code.
3	Text Origin Detection: Unmasking the Source (AI vs. Human)	Combined TF-IDF features and ensemble models (Random Forest, Extra Tree Classifier) on Kaggle dataset of 419,199 submissions.	Accuracy: 80.29%, Precision: 78.02%, Recall: 81.57%, F1-score: 79.76%, MCC: 60.62%. A Tkinter GUI was developed.	Difficulty processing longer texts. - Limited cross-linguistic accuracy (e.g., English-Arabic). - More diverse datasets needed to address potential bias.
4	Detection of AI-Generated Essays in Writing Assessments	Analyzed GPT-3 AI essays (balanced dataset of 8,000 essays) using e-rater® and RoBERTa.	Achieved >95% accuracy in detecting AI-generated essays.	Limited diversity in prompts. - Grammar errors excluded, oversimplifying real-world scenarios. - Dataset size and prompt variety require expansion for better representation.
5	Hiding the Ghostwriters: Adversarial Evaluation of AI-Generated Student Essay Detection	Explored adversarial perturbation techniques (paraphrasing, word/sentence substitution) to evade AI detection on the AIG-ASAP dataset.	Perturbations significantly reduced detection rates.	Dataset limited to U.S. high school essays. - Focused on dataset creation, lacking implementation insights. - Findings may not generalize to broader contexts.

V. METHODOLOGY

A binary classification system was put into place using machine learning and natural language processing (NLP) approaches to solve the problem of differentiating essays written by humans from those produced by artificial intelligence. The approach is comprised of the following steps:

1. Data Collection

The dataset comprises two distinct sources: Human-authored essays: Essays were sourced from the ETS Corpus of Non-Native Written English. This corpus contains a wide range of writing samples from non-native English speakers, allowing the model to analyze human-authored essays with varying syntactic styles and proficiency levels. The corpus is especially valuable because of its diversity in writing patterns, making it a robust resource for training models. AI-generated essays: Texts generated by seven large language models (LLMs) were included: GPT-3.5-Turbo, GPT-4o, Gemini-1.5, Llama-3.1 (8B), Phi-3.5-mini, and Claude-3.5. Each model provided multiple samples of essays, covering various academic writing prompts. These AI-generated essays simulate responses that might be expected in a university setting.

2. Preprocessing

Both human and AI-generated essays underwent preprocessing to ensure consistency and remove noise from the text. The following steps were employed:

Tokenization: To improve linguistic analysis, each essay was tokenised into individual words or subword units.

Lowercasing: To prevent inconsistencies because to case sensitivity, all text was changed to lowercase.

Stopword Removal: To concentrate on more important topics, common English stopwords were eliminated.

Lemmatization: Words were lemmatized to their base forms, ensuring that variations of a word (e.g., "writing" vs. "write") were treated consistently.

3. Feature Extraction

A number of language characteristics that captured stylistic and structural factors were taken from articles written by humans and by machines: Inverse Document Frequency-Term Frequency (TF-IDF): This approach measured each essay's word relevance in relation to the corpus as a whole. To distinguish between common and unique word usage in essays written by humans and by AI, the TF-IDF values were employed.

Syntactic Features: The use of sentence structure, grammatical patterns, and part-ofspeech tagging were analyzed to detect differences between human and machine writing styles.8 Lexical Richness and Diversity: Measures such as word length, sentence complexity, and lexical variety were used to capture the richness of human writing versus the more repetitive nature of AI text.

Embeddings: Word embeddings like GloVe or BERT were employed to capture contextual and semantic nuances in the essays.

4. Model Selection and Training

To determine the best classifier, a mix of deep learning models and conventional machine learning methods were tested.

Dataset:

The dataset for this project consists of tokenized essays sourced from both human authors and AI-generated content,providing a solid foundation for distinguishing between the two through natural language processing (NLP) techniques.

5. Human-Authored Essays:

The human-authored essays were tokenized from the ETS Corpus of Non-Native Written English. This corpus includes a variety of essays written by non-native English speakers, representing a broad spectrum of proficiency levels and syntactic styles. The tokenization process involved splitting the text into individual tokens—essentially breaking down each essay into smaller units, such as words or subword segments. This method allows for more granular analysis of linguistic patterns, aiding in the identification of subtle differences in sentence structure, word choice, and grammar commonly found in human writing. By tokenizing these essays, we ensured that the linguistic features could be effectively compared to machine-generated text, even when the writing deviated from standard forms of native English.

***AI-Generated Essays**:*

The text from seven cutting-edge large language models (LLMs) was tokenised to create the AI-generated essays. These models included:

GPT-3.5 Turbo
The GPT-4o
The Gemini-1.5

Llama 3.1 (8B)
Mini Claude-3.5
Phi-3.5

Tokenizing the AI-generated essays ensures consistency in how both human and machine-written texts are processed. By breaking the AI- generated content into tokens, we were able to capture specific patterns and repetitions in sentence construction, which are often characteristic of machine-generated text. The tokenization process is crucial for comparing the linguistic output of AI models with human writing and allows for better analysis of syntactic structures, stylistic variations, and lexical diversity.

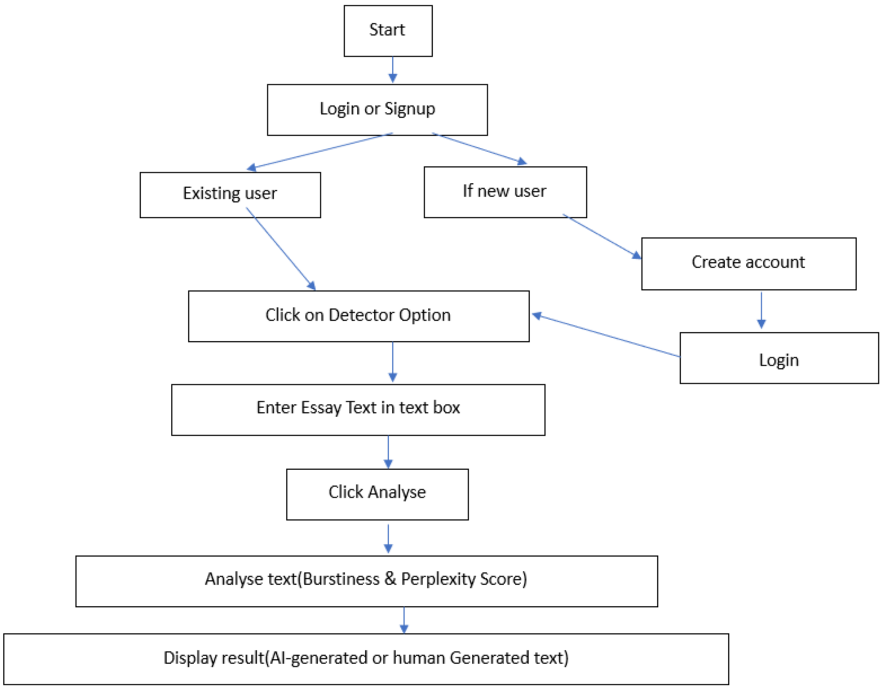
Tokenization Process:

The essays from both human and AI sources were tokenized using a standard NLP approach that breaks the text into smaller units based on word boundaries, punctuation, and subword patterns. This process involved: Word-level tokenization: Splitting essays into individual words, removing punctuation, and standardizing case(lowercasing).

Subword-level tokenization: Using techniques like Byte-Pair Encoding (BPE), which helps in handling rare or unseen words by breaking them down into smaller, more frequent subword units. The resulting tokenized data is stored in a structured format, with each essay represented as a sequence of tokens. This consistent tokenization is key to ensuring that the machine learning model can effectively analyze and compare human and AI-generated essays. By focusing on tokenized data from human and machine sources, the project leverages the power of linguistic granularity to capture differences in writing patterns. This enables the binary classification system to more accurately differentiate between human authored and AI-generated essays, even when the distinctions are subtle.

VI. ARCHITECTURE

Fig 1


Start
Login or Signup
Existing user
If new user
Create account
Click on Detector Option
Login
Enter Essay Text in text box
Click Analyse
Analyse text(Burstiness & Perplexity Score)
Display result(AI-generated or human Generated text)

VII. RESULTS AND DISCUSSION

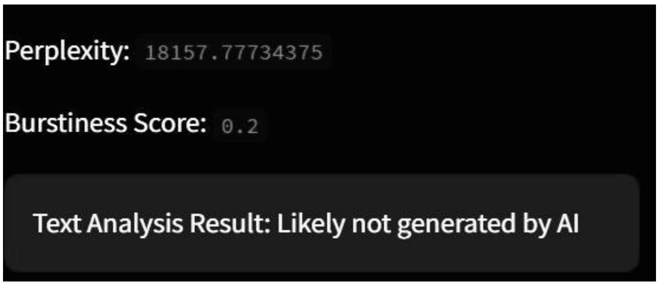
This study examined the differences between literature created by AI and text written by humans, paying particular attention to ideas of authorship and voice. The investigation clarifies the difficulties and restrictions that come with using ChatGPT and other AI techniques to produce complex and cohesive academic writing. The challenges AI-generated text faces in maintaining a distinct and identifiable authorial presence are brought to light by both quantitative and qualitative authorship studies. The technology lacks true self- identification, even though it can use personal pronouns like "I." Instead of an instinctive understanding of the tone or meaning of the text, the usage of "I" only indicates a prompted posture. On the other hand, the type-token ratio suggests that the student's writing is more lexically diverse and has a more intricate and distinctive style. The student's article is written in a distinctive, nuanced, and customised style, and it exudes a strong feeling of authorship. Effective utillisation of sources results in successful grading since they fit the academic format. Using hedges, boosters, and a mix of active and passive voice constructions carefully results in a thorough and well-rounded discourse.

Strategic repetition preserves the main concept without compromising language diversity or complexity. The first person pronoun is used a lot in the opening, but the AI- generated language lacks originality. Repetition of phrases, sparing use of hedges and boosters, and excessive use of active voice constructions all contribute to a straightforward and less nuanced voice. In addition to lacking diversity, the instrument does not produce a sufficiently thorough examination of the subject. The study highlights how difficult it may be to create scholarly literature using AI tools, particularly when it comes to maintaining a sophisticated authorial presence. AI is still unable to replicate complicated speech, subtle voice, and genuine self-identification, despite its ability to copy certain style elements.

AI-generated work should be handled carefully by students and academics, who should be aware of its limits in capturing the complexities of academic writing. This study highlights the distinct advantages of human-authored material in the balance between AI-generated and human-authored language. These advantages include a nuanced voice, a varied vocabulary, and the skillful application of rhetorical devices. While useful for some jobs, AI-generated prose is unable to capture the richness and uniqueness of academic writing produced by humans. As technology advances, it is essential to comprehend these constraints in order to preserve the authenticity and integrity of scholarly discourse. The findings of the present investigation, which underscore the constraints of artificial intelligence-generated text in scholarly composition, align with previous studies.

In terms of essay quality, writing speed, or authenticity, all of these findings suggest that, at the time of the study, content produced by artificial intelligence (AI) might not be any superior to content created by humans. It's critical to acknowledge the dynamic nature of AI technologies, particularly in light of the quick development of ChatGPT and related Large Language Models since the start of the study.

Their talents could be impacted by further advancements in these models, which could have an effect on how well they perform on academic writing assignments. Keeping up with these advancements requires regular reappraisal through periodic studies like the one being conducted now .The study highlights the difficulties in upholding academic integrity when using material produced by artificial intelligence in assignments.

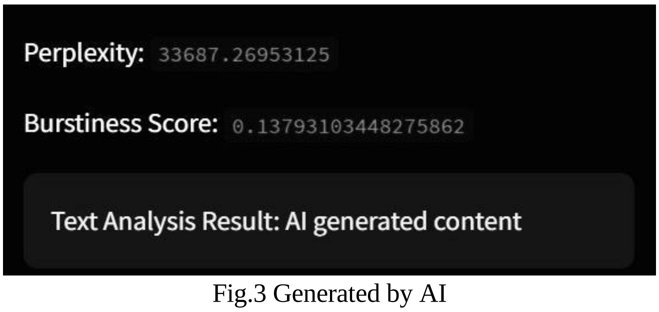


Perplexity: 18157.77734375

Burstiness Score: 0.2

Text Analysis Result: Likely not generated by AI

Fig.2 Not generated by AI



Perplexity: 33687.26953125

Burstiness Score: 0.13793103448275862

Text Analysis Result: AI generated content

Fig.3 Generated by AI

VIII. CONCLUSION

The link between AI-generated writing and human-written language in academic settings is still a dynamic area of research as technology develops. Understanding the role AI can play in academic writing and the ongoing pursuit of academic integrity will be made possible by a more thorough ongoing investigation, flexibility in responding to changing AI capabilities, and investigation of novel assessment techniques. The study highlights the complexity of using ChatGPT and the difficulties encountered while attempting to produce the intended results. It makes the argument that the art of writing itself is similar to the skill of successfully provoking ChatGPT. ChatGPT's shortcomings become clear in an English literature lesson, especially when it comes to its incapacity to offer reliable quotes and sources and its propensity to insert factual inaccuracies. Currently, ChatGPT is a difficult path to follow because it requires careful cross-referencing and proofreading in order to produce meaningful material.

The study explores how authorship, voice, and technology interact in scholarly writing. The study acknowledges that AI-generated literature can provide fairly coherent outputs, but it also emphasises how difficult it is for AI to replicate the subtle authorship traits that are present in human writing, such as contextual appropriateness and precise referencing. Even while the writing seems human, a deeper look exposes problems with register, overused vocabulary, and a lack of subtlety.

The study explores how authorship, voice, and technology interact in scholarly writing. The study acknowledges that AI-generated literature can provide fairly coherent outputs, but it also emphasises how difficult it is for AI to replicate the subtle authorship traits that are present in human writing, such as contextual appropriateness and precise referencing. Even while the writing seems human, a deeper look exposes problems with register, overused vocabulary, and a lack of subtlety.

The study highlights how important it is to have conversations about authorship, plagiarism, and originality as AI-generated writing becomes more common. The findings caution against assuming that AI-generated literature always satisfies the standards for intricate and authentic human writing, even though it seems coherent. This has implications for the ongoing discussion on the ethical application of AI in academic contexts. Future studies should consider conducting similar experiments when ChatGPT and other AI models advance. Comparisons using recognised evaluation criteria could be used to examine the efficacy of AI-generated text vs human-authored content on a variety of topics. This approach would offer a more sophisticated understanding of the evolving environment and the potential applicability of findings to other academic contexts.

According to the study, AI-enabled software that can recognise and categorise material as either generated or human-written needs to be critically examined. It is crucial to look into the techniques used to make these distinctions. Furthermore, it would be useful to determine whether a person could identify the originality and place of origin of a text with equivalent accuracy. If it is possible, this feature could greatly help teachers assess the reliability of written materials.

IX. REFERENCES

[1] W.H. Pan,M.J.Chok, J. L. S. Wong, Y. X. Shin, Z. Yang, Y. S. Poon, C. Y. Chong, D. Lo, and M. K. Lim, "Assessing AI Detectors in Identifying AI- Generated Code: Implications for Education," IEEE, 2024.

[2] V. Pandita, A. M. Mujawar, T. Norbu, V. Verma, and P. Patil, "Text Origin Detection: Unmasking the Source – AI vs Human," IEEE , 2024.

[3] X. Peng, Y. Zhou, B. He, L. Sun, and Y. Sun, "Hiding the Ghostwriters: An Adversarial Evaluation of AI- Generated Student Essay Detection," arXiv:2402.00412v1 [cs.CL], Feb. 2024.

[4] Zhang,Y., & Li, H. ,"AI-Generated Text: A New Frontier in Content Creation" , IRJET,2024.

[5] L Mindner, T Schlippe, K Schaaff, “Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT” Springer link.

[6] Jiao, W., Wang, W., Huang, J.-T., Wang, X., Tu, Z.: Is ChatGPT a good translator? A preliminary study. arXiv preprint arXiv:2301.08745 (2023).

[7] Jeblick, K., et al.: ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. arXiv e-prints (2022).

[8] Wankhade, M., Rao, A., Kulkarni, C.: A survey on sentiment analysis methods, applications, and challenges. Artif. Intell. Rev. 55, 5731–5780 (2022).

[9] Rakhmanov, O., Schlippe, T.: Sentiment analysis for Hausa: classifying students’ comments. In: SIGUL 2022, Marseille, France (2022).

[10] Shijaku, R., Canhasi, E.: ChatGPT generated text detection (2023).