AI vs AI: Evaluating the Reliability of Detectors and Transformer Models for Human–AI Text Classification

Haleemah Amisu
Towson University
Towson, Maryland, USA
hamisu1@students.towson.edu

Ayandayo Adeleke
Towson University
Towson, Maryland, USA
aadelek7@students.towson.edu

Gabriella Akenn Musa
Towson University
Towson, Maryland, USA
gakennm1@students.towson.edu

***Abstract*—Large language models have become increasingly important across domains, with applications ranging from knowledge base answering and code generation to copywriting, text generation, and classification. Their effectiveness has fueled widespread reliance, but also raised concerns among employers, educators, publishers and other users about the authenticity and reliability of the generated contents. To address this, various AI detection tools have been developed to distinguish between human and machine generated texts. We propose evaluating the performance and reliability of existing AI content detectors using Kaggle and HC3 data sets. Focusing on various domains, the study will compare multiple detectors against both human and AI-generated content. This will assess accuracy, identify cross-domain performance, and provide insights into the practical reliability of AI detection tools.**

***Index Terms*—Large Language Models, AI-generated content, AI content detector, reliability, accuracy**

I. Introduction

Arising from the continual development in Artificial Intelligence (AI), the application of its content, including dependency across many important sectors, is on the increase. However, with this increase in usage and dependency comes the critical need to evaluate the reliability and trustworthiness of AI content, particularly as it spans various domains. The questions of authenticity, accountability, and credibility arise even though the growth of AI enhances efficiency and accessibility. The challenge raised by AI-generated content has given rise to AI detectors designed to assess the validity and authenticity of such content. This study aims to examine the reliability of these AI-generated content detectors to ensure their efficacy and dependability in real-world applications. The significance of this research arises from the weighty impact AI content detectors have on decision-making processes in fields such as education, medicine, journalism, cybersecurity, law, and policymaking.

Overall, this research study addresses an important gap in AI technology, providing a fundamental understanding necessary for advancing AI applications reliably across domains. Also, the study is driven by the recognition that current discussions about AI detection are divided and regularly focused on only technical standards without effectively tackling real-world situations where false positives and false negatives carry uneven risks. The study's main objective is to evaluate its AI-generated content and detector's reliability and performance, investigate their limitations, explore causes of failure rates, and rectify existing gaps by offering solid theoretical insights and practical guidance, including attendant implications on future human-AI relationships.

II. Description of the Dataset

For this study, we will use two publicly available datasets that provide labeled examples of both human and AI-generated text. The first dataset is the "LLM Detect AI Generated Text" [1] dataset, which was released on Kaggle as part of a competition. The dataset contains approximately 10,000 essays across 7 different essay prompts, with some essays written by students and others generated with a large language model. Since the dataset does not state which model created the AI-generated essays, we will treat them as representative examples of text produced by modern large language models (LLMs). The essay format makes this dataset particularly relevant to the education domain, where concerns about plagiarism and AI-assisted writing have become most prominent.

The second dataset is the "Human–ChatGPT Comparison Corpus (HC3)" [2], hosted on Hugging Face. HC3 contains paired responses to questions, with one response written by a human and the other by ChatGPT. The Questions cover domains such as finance, medicine, law, and open-domain knowledge. There are 37,000 question and answer pairs in this dataset. The Question and Answer structure of this dataset is different from the essay dataset both in length, and format. This will allow us to evaluate detection performance on both essay format data, and on shorter information based text.

We have decided to select these two different datasets because they provide complementary perspectives. Using both allows the study to assess whether detection accuracy remains

consistent across different kinds of text, strengthening the reliability and applicability of the findings

III. PRELIMINARY LITERATURE REVIEW

The growing reliance on large language models (LLMs) such as GPT-3 and GPT-4 has raised concerns about distinguishing between human- and AI-generated content. While artificial intelligence(AI) offers efficiency and productivity across domains, its widespread use has made it increasingly difficult to verify its authenticity and reliability. This challenge has fueled the development of AI detection tools, and several studies have sought to evaluate their reliability and effectiveness. Elkhatat et al. investigated the capabilities of various AI content detection tools, including OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag. Using 15 paragraphs generated by ChatGPT models 3.5 and 4, alongside five human written responses, the study found that common tools like GPTZero and Copyleaks were more accurate in detecting GPT-3.5 outputs than GPT-4, but often misclassified human-written text. [3] Similarly, Halaweh et al reported that detectors such as Turnitin and ZeroGPT failed to consistently identify ChatGPT-generated text, particularly after paraphrasing. [4] Weber et al evaluated 12 publicly available AI-detector tools and found that the widely used detectors are neither accurate nor reliable, with a tendency to classify outputs as human-written rather than accurately detecting AI text. [5] Chaka also evaluated the accuracy of five AI-detection tools and found that they were not ready to accurately and consistently detect AI-generated content across different contexts. [6] Additional concerns have been raised regarding AI detector tools in the context non-native English writers. A study by Liang et al found that detectors exhibited high false-positive rates, misclassifying non-native English writing as AI-generated, while native English samples were more reliably identified [7].

While many studies emphasize the inaccuracy of AI-detection tools, a few have reported promising results. Kar et al evaluated the accuracy of five free AI-detector tools and found four of them to precisely detect content to be 100% AI generated while the fifth tool proved inaccurate [8].

Collectively, these studies suggest that while AI detectors show promise, they face serious limitations, including false positives, inaccuracy, vulnerability to paraphrasing, and language bias. This highlights the need for further research to systematically evaluate AI-detector performance, particularly across domains.

PROPOSED METHODOLOGY

The study will employ a two-stage methodology to evaluate the detection of AI-generated text, that combines the use of existing open-source detectors with the training of transformer-based classification models. The purpose of this design is to first establish baseline performance using detectors already released in the research community and then compare these results with fine-tuned models developed specifically for the two datasets under study.

The first stage focuses on benchmarking open detectors against the Kaggle essay [1] dataset and the HC3 corpus [2]. Tools such as DetectGPT, Fast-DetectGPT, and the RoBERTa-based classifier provided with the HC3 release will be applied directly to the raw text in each dataset [2]. These detectors operate on different underlying principles: DetectGPT and Fast-DetectGPT rely on language model perplexity and probability curvature, while the HC3 RoBERTa classifier has been trained to distinguish human responses from ChatGPT outputs [9].

By evaluating these detectors across both datasets, the study will provide a practical baseline of current detection capability. The predictions generated by each detector will be compared with the ground-truth labels in the datasets, and performance will be assessed using accuracy, precision, recall, and F1-score. Confusion matrices will be generated to highlight patterns of false positives and false negatives, with particular emphasis on the risk of human text being misclassified as AI.

The second stage will involve the development of custom transformer-based detectors. Pre-trained architectures such as DistilBERT and RoBERTa will be fine-tuned on the Kaggle and HC3 datasets for binary classification of human versus AI text. Data preparation will be deliberately minimal, limited to removing corrupted or empty entries while preserving all stylistic characteristics of the text. This ensures that the model evaluates the texts exactly as they are presented by either the human or the LLM. Each dataset will be divided into training, validation, and test sets using a 70–15–15 split. Since there are 7 distinct prompts for the kaggle essay dataset, prompt identifiers will be used to control the partitioning in the Kaggle dataset to prevent leakage between training and test sets across the essay prompts [1]. Models will be trained on the training set, hyperparameters tuned using the validation set, and final evaluation conducted on the test set.

The evaluation of the fine-tuned transformers will use the same metrics as the existing open detectors, enabling direct comparison. Accuracy, precision, recall, and F1-score will be reported, and confusion matrices will be used to analyze the error patterns. In addition to quantitative results, explainability techniques such as SHAP or LIME will be applied to a subset of predictions, highlighting which tokens or textual elements most influenced the models' decisions. This interpretability analysis will allow for discussion of not only how accurately models perform, but also why they arrive at particular classifications.

IV. ANTICIPATED OUTCOMES

At the conclusion of this study, we expect to deliver a comprehensive assessment of the current state and potential of AI text detection. Specifically, we anticipate three primary outcomes.

First, we aim to have a clear analysis of the performance of existing detectors, establishing their strengths and weaknesses when applied to both long-form essays and domain-specific Q&A responses. This will clarify how well current tools function in practice and where they are most likely to fail.

Second, by fine-tuning transformer-based models on the Kaggle and HC3 datasets, we expect to demonstrate whether modern representation-learning approaches can surpass existing detectors in accuracy, precision, and robustness.

Finally, through the application of explainability techniques such as SHAP and LIME, we expect to gain insight into the specific linguistic and stylistic factors that influence classification decisions. This interpretability analysis will help explain why detectors succeed or fail, particularly in cases of false positives or false negatives. Together, these outcomes will address the central question of the study: how accurately can AI-generated text be distinguished from human-authored writing, and what limitations remain in current approaches.

REFERENCES

[1] https://www.kaggle.com/competitions/llm-detect-ai-generated-text/data.

[2] https://huggingface.co/datasets/Hello-SimpleAI/HC3/tree/main.

[3] Elkhatat, A.M., Elsaid, K., & Almeer, S.(2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. Int J Educ Integr 19, 17 https://doi.org/10.1007/s40979-023-00140-5.

[4] Halaweh, M. & Refae, G.E. (2024). Examining the Accuracy of AI Detection Software Tools in Education. 2024 Fifth International Conference on Intelligent Data Science Technologies and Applications (IDSTA), 2024, pp. 186-190, doi:10.1109/IDSTA62194.2024.10747004.

[5] Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S. et al. (2023). Testing of detection tools for AI-generated text. Int J Educ Integr 19, 26 (2023). https://doi.org/10.1007/s40979-023-00146-z.

[6] Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. Journal of Applied Learning & Teaching, 2023, 6(2). https://doi.org/10.37074/jalt.2023.6.2.

[7] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. https://doi.org/10.48550/arXiv.2304.02819.

[8] Kar, S.K., Bansal, T., Modi, S., & Singh, A. (2025). How Sensitive Are the Free AI-detector Tools in Detecting AI-generated Texts? A Comparison of Popular AI-detector Tools. Indian J Psychol Med. 2025;47(3):275–278. DOI: 10.1177/02537176241247934.

[9] Mindner, L., Schlipe, T., & Schaaf, K. (2023). Classification of human- and ai-generated texts: Investigating features for chatGPT. Lecture Notes on Data Engineering and Communications Technologies, 152–170. https://doi.org/10.1007/978-981-99-7947-9_12.

[10] Khalil, M., Er, E. (2023). Will ChatGPT get you caught? Rethinking of plagiarism detection. In: Zaphiris, P., Ioannou, A. (eds.) HCII 2023. LNCS, vol. 14040, pp. 475–487. Springer, Cham. https://doi.org/10.1007/978-3-031-34411-4_32