AI vs AI: Evaluating the Cross-Domain Robustness of RoBERTa-Based Detection Across Q&A and Academic Essay Domains

Haleemah Amisu
Department of Computer Sciences
Towson University
Towson, USA
hamisu1@students.towson.edu

Ayandayo Adeleke
Department of Computer Sciences
Towson University
Towson, USA
aadelek7@students.towson.edu

Gabriella Akenn Musa
Department of Computer Sciences
Towson University
Towson, USA
gakennm1@students.towson.edu

***Abstract*—Large Language Models (LLMs) have become increasingly ubiquitous in educational and professional settings, which raises the need for reliable tools to distinguish between human-authored and machine-generated content. While numerous detection systems exist, their reliability across diverse writing formats and generator architectures remains under-explored. This study evaluates the cross-domain robustness of a standard RoBERTa-based detector, which was originally trained on short-form Question-Answer (Q&A) data, when applied to long-form academic writing. We employ a dual-dataset comparative approach. The HC3 dataset (Q&A pairs) serves as a control baseline, while the DAIGT V2 dataset (argumentative student essays) serves as a stress test for the detection model. The results reveal a critical lack of generalization. While the detector performs accurately within its native Q&A domain, its accuracy degrades significantly when applied to academic essays, with a False Positive Rate of 8.4% on human writing. Furthermore, a stratified analysis of the essay dataset exposes a model-specific bias. The detector achieves high sensitivity against open-source models like Mistral (98%), but struggles against Llama-2 (68%) and fails completely against advanced proprietary models like GPT-4 and Claude (0–2% detection rate). These findings suggest that current general-purpose detectors are heavily overfitted to specific text structures and older model signatures, rendering them unreliable for academic integrity enforcement in an evolving AI landscape.**

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

I. Introduction

The rapid growth of Generative AI has caused a major rise in AI-generated text. As a result, AI-generated writing has become common across education, industry, and digital communication. In academic settings, tasks such as researching a topic and writing scholarly papers are essential skills. However, LLMs like ChatGPT, have become widely used in academic article production due to their ability to quickly and effortlessly generate clear, contextually relevant, and coherent text, making them an attractive option for writers [4]. As AI models become capable of generating sophisticated outputs, the ability to reliably distinguish between human-authored and machine-generated content has become a fundamental requirement for maintaining integrity. The core challenge addressed in this study is the capacity of current systems to accurately perform binary classification of the origins as Human or Machine generated texts.

Although AI-detectors can be helpful for identifying possible AI involvement in submitted work, many of them are trained on narrow, highly structured datasets, most commonly short-form Question and Answer text. While such detectors tend to perform well within their training domain, real-world writing is far more diverse. Academic essays, reflective pieces, argumentative writing, and creative compositions differ greatly in structure, length, tone, and complexity. These variations pose challenges for existing detectors because they were not designed to handle such a wide range of writing formats.

Furthermore, there is the rapid evolution of LLMs such as GPT-4, Claude, Llama, and Mistral that generate human-like texts, making detection even more difficult. Current detectors often fail to recognize the writing style of these models and as a result, a detector may incorrectly flag human essays as AI-generated (false positives) or AI-generated texts may go undetected (false negatives), thereby undermining trust in AI-detection tools. A detector that works well on short Q&A text may perform poorly on long-form academic writing, and one that identifies ChatGPT-style outputs may struggle with content produced by Llama or Mistral. This limitation has significant implications in domains where accurate text attribution is essential.

While previous research has examined the generalizability of detectors across formats and AI sources, gaps remain particularly in evaluating how specific detectors perform across multiple domains and against a range of modern AI models. To address this gap, our project focuses on assessing the cross-domain robustness of a representative AI detection system. Specifically, we investigate whether a standard RoBERTa-based detector, trained exclusively on Q&A data, maintains its reliability when applied to academic essays. The findings of this research aim to inform educators, researchers, and policymakers about the limitations of current AI-detection systems and guide the development of more reliable detection approaches in an evolving AI landscape. Our study addresses

the following research questions:

- Is the RoBERTa-based detector accurate within its native training domain (Q&A)?
- Is this accuracy affected when the detector is applied to long-form academic essays?
- Within the essay domain, does the detector's accuracy vary significantly depending on the specific AI model used to generate the text (e.g., GPT-4 vs. Llama)?

To answer these questions, we conduct a comparative stress test using the HC3 dataset as a baseline and the DAIGT V2 dataset as an out-of-distribution test case. It is important to note that this study focuses on evaluating the robustness of a representative open-source detector (RoBERTa) which has been trained on Q&A style data. A direct comparison with a detector trained specifically on essay data remains a subject for future research. By investigating limitations across these formats and exploring failure rates among different AI models, this study offers solid theoretical insights and practical guidance for the reliable deployment of detection tools.

II. Related Work

Existing academic literature regarding AI detection has predominantly focused on evaluating the baseline efficacy of available tools, driven by the urgent need to address the rapidly increasing utilization of Large Language Models (LLMs) such as GPT-3 and GPT-4. Given the difficulty in verifying the authenticity of digital text, numerous studies have sought to quantify the reliability of these detectors across standard benchmarks.

Elkhatat et al. [3] 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. This finding is corroborated by Weber et al. [17] who evaluated 12 publicly available AI-detector tools and found that widely used detectors are neither accurate nor reliable, exhibiting a tendency to classify outputs as human-written rather than accurately detecting AI text. Similarly, Chaka [2] evaluated the accuracy of five AI-detection tools and concluded that they were not ready to accurately and consistently detect AI-generated content across different contexts.

Beyond baseline inaccuracy, detectors have been shown to be inaccurate when text is modified. Halaweh et al. [5] reported that detectors such as Turnitin and ZeroGPT failed to consistently identify ChatGPT-generated text, particularly after paraphrasing. This vulnerability suggests that detectors may rely on superficial statistical cues that are easily obfuscated by simple editing or rewriting strategies, rather than detecting semantic signals of AI generation.

Additional concerns have been raised regarding the fairness of AI detector tools, particularly in the context of non-native English writers. A study by Liang et al. [12] found that detectors exhibited high false-positive rates, misclassifying non-native English writing as AI-generated, while native English samples were more reliably identified. This bias poses a significant ethical risk in educational settings, where international students may be disproportionately and wrongfully accused of academic misconduct.

While many studies emphasize the inaccuracy of AI-detection tools, a few have reported promising results. Kar et al. [10] 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. Collectively, these studies suggest that while AI detectors show promise, they face serious limitations, including false positives, inaccuracy, vulnerability to paraphrasing, and language bias.

Recent studies on detecting AI-generated text have shown that encoder-based models, such as RoBERTa, outperform older methods. These models are good at understanding context and are stable. Tests show that they are highly accurate when applied to the same type of data on which they were trained, such as short Q&A content (Mobin & Islam, 2025). However, research has also shown that training models on specific types of data can make them seem better than they are. They often struggle with various types of writing. Tulchinskii et al. [16] found that AI-generated text appears differently in various fields. Therefore, a RoBERTa model trained only on Q&A might not perform well with long academic essays, which have different writing styles and vocabulary.

Another area of study examines how the AI model that creates the text affects detection. Hassan et al. [6] and Jin et al. [8] demonstrated that the performance of a detector is influenced by both the detector itself and the AI model used to create the text. Models like GPT-4 and Llama generate text in distinctive styles. This means that the effectiveness of detectors across different types of writing and AI models is a complex issue that requires further research.

Critically, most existing evaluations focus on a single domain or a limited set of models. There remains a gap in systematically evaluating how these detectors perform when the text format shifts (e.g., from Q&A to Essays) and when the source model changes (e.g., from GPT-3.5 to Llama 2 or Mistral). This study aims to address this gap by conducting a cross-domain stress test of a standard detector against diverse writing styles and generator architectures.

III. Description of the Dataset

To evaluate the robustness of AI detection across distinct domains and model architectures, we selected two publicly available datasets that provide labeled examples of both human and AI-generated text. These datasets were chosen to represent two distinct writing formats: General Q&A(Control), and Academic Writing(Stress Test).

A. Control Dataset: Human–ChatGPT Comparison Corpus (HC3)

We selected the Human–ChatGPT Comparison Corpus (HC3), hosted on Hugging Face, as the baseline control.

This dataset contains approximately 37,000 Question-Answer (QA) pairs characterized by short-form, informational, and factual text. The dataset aggregates questions and human answers from five distinct domains: Finance (financial advice and definitions), Medicine (medical queries and health information), Law (legal questions and explanations), Open-Domain Knowledge (general encyclopedic knowledge), Reddit ELI5(Casual, simplified explanations).

The machine-generated responses in this dataset were produced primarily using ChatGPT (GPT-3.5). This dataset serves to establish the baseline accuracy of the detector in its native training environment, where the text format (Q&A) and model architecture (GPT-3.5) are consistent with the detector's expected input.

B. Stress Test Dataset: DAIGT V2 (Academic Essays)

To evaluate the detector performance in an educational context, we used the DAIGT V2 (Detect AI Generated Text Version 2) dataset. The original dataset released for the Kaggle LLM Detect AI Generated Text competition contained a critical limitation, it included approximately 1,375 human essays but only 3 AI-generated essays. To rectify this severe imbalance, we utilized the DAIGT V2 dataset, a community-curated expansion that augments the original data with over 20,000 AI-generated essays while preserving the original academic context.

The dataset consists of long-form, structured, argumentative essays written by students (grades 6-12). Both the human and AI essays were written in response to the same set of standardized academic prompts, ensuring that detection results are not skewed by topic. The specific prompts include: "Car-free cities", "Does the electoral college work?", "Distance Learning", "Seeking multiple opinions", and, "Facial action coding systems" Unlike the HC3 dataset, DAIGT V2 contains text generated by a diverse array of Large Language Models, allowing for analysis of model-specific detection bias. The dataset includes essays generated by: proprietary models (GPT-4,GPT-3.5, Claude Anthropic, PaLM, and Cohere Command) and open-source models (Llama-2; 7b and 70b variants, Falcon-180b, and Mistral-7b).

This dual-dataset design allows the study to assess whether detection accuracy remains consistent across different kinds of text (Q&A vs. Essays) and whether it is robust to different generative architectures (GPT vs. Llama/Mistral), strengthening the reliability and applicability of the findings.

IV. Methodology

To rigorously evaluate detector robustness, we developed a multi-faceted experimental framework designed to isolate the variables of text format and generator architecture. We first establish a performance baseline using in-domain data before subjecting the model to an out-of-distribution stress test on academic writing, followed by a stratified analysis of model-specific detection rates.

A. Detector Architecture

We evaluated a pre-trained RoBERTa-base detector model (Hello-SimpleAI/chatgpt-detector-roberta) developed specifically for identifying machine-generated text. RoBERTa (Robustly Optimized BERT Pretraining Approach) is a transformer-based architecture that modifies the key hyperparameters of BERT, including removing the Next Sentence Prediction (NSP) objective and training with larger mini-batches. This specific detector was fine-tuned on the HC3 dataset, making it a native detector for the Q&A format. We selected this model as a representative black box detector to simulate the real-world scenario where educators use tools without access to the internal training data or weights.

B. Evaluation Metrics

To quantify the performance and risk associated with AI detection, we model the problem as a binary classification task where Human = 0 and AI = 1. We define the fundamental confusion matrix components as follows:

- TP (True Positive): AI correctly identified as AI.
- TN (True Negative): Human correctly identified as Human.
- FP (False Positive): Human incorrectly identified as AI (Type I Error).
- FN (False Negative): AI incorrectly identified as Human (Type II Error).

Based on the components, we utilize the following metrics for evaluation:

- **Accuracy:** The overall effectiveness of the classifier across both classes.

  $$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN)$$

- **False Positive Rate (FPR):** The probability of a Type I error, quantifying the rate at which human-authored text is misclassified as machine-generated.

  $$\text{FPR} = (FP)/(FP + TN)$$

- **False Negative Rate (FNR):** The probability of a Type II error, quantifying the rate at which AI-generated text successfully evades detection

  $$\text{FNR} = (FN)/(FN + TP)$$

C. Experimental Design

The study employed a two-phase comparative stress test:

- **Phase I - Baseline Validation (In-Domain)**: We first applied the detector to the HC3 dataset (Q&A pairs) to establish a performance baseline. Since the detector was trained on similar data, we expected optimal performance (High Accuracy, Low FPR). This phase confirms that the detector functions correctly under ideal conditions.
- **Phase II - Cross-Domain Stress Test (Out-of-Distribution)**: We then applied the exact same detector to the DAIGT V2 dataset (Academic Essays). This isolates the variable of Text Format. Any degradation in performance can be attributed to the shift from short Q&A to long-form essays.

Metric	Value
Overall Accuracy	99.4%
Human Precision	1.0
Human Recall	1.0
AI Precision	0.99
AI Recall	0.99
FPR	0.46%
FNR	0.90%

TABLE I
TABLE 1: BASELINE PERFORMANCE METRICS (HC3 DATASET)

Metric	Value
Overall Accuracy	84.06%
Human Precision	0.84
Human Recall	0.92
AI Precision	0.85
AI Recall	0.72
FPR	8.4%
FNR	27.7%

TABLE II
TABLE 2: STRESS TEST PERFORMANCE METRICS (ACADEMIC ESSAYS)

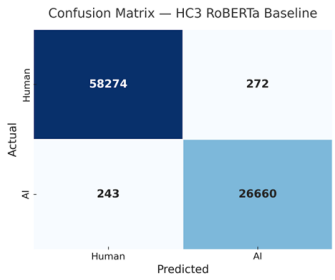

Fig. 1. Confusion Matrix

AI Model/Source	Accuracy (Detection Rate)
mistral7binstruct_v1	98.88%
mistral7binstruct_v2	96.57%
mistralai/Mistral-7B-Instruct-v0.1	96.25%
falcon_180b_v1	92.51%
llama_70b_v1	87.88%
kingki19_palm	83.24%
radek_500 (GPT-3.5)	81.40%
palm-text-bison1	79.08%
llama2_chat	81.40%
NousResearch/Llama-2-7b-chat-hf	63.50%
chat_gpt_moth (GPT-3.5)	63.44%
cohere-command	58.29%
darragh_claude_v6	98.88%
darragh_claude_v7	2.30%
radekgpt4	0.00%

TABLE III
TABLE 3: DETECTION ACCURACY BY SOURCE MODEL (*Sorted By Detector Success Rate*)

Furthermore, we performed a stratified analysis by splitting the DAIGT dataset based on the underlying generator model. By calculating the accuracy for each subset (e.g., Accuracyllama, AccuracyGPT4, etc), we determined whether the detector's reliability is universal or biased toward specific model architectures.

To enable a direct and accurate comparison, we aligned our evaluation metrics with those used in standard open-detector benchmarks. Specifically, we utilized the same confusion matrix derivatives (accuracy, precision, recall, and F1-score) that characterize the original architecture's performance. By keeping these metrics constant while varying the data domain (Q&A vs. Essays) and source models, we isolated the detector's generalization capability without introducing confounding variables in the evaluation method itself.

V. RESULTS

The experimental results of the comparative stress tests are organized into three phases to directly address the research questions. First, we establish the detector's baseline performance on the HC3 (Q&A) dataset to verify its efficacy in its native domain. Second, we evaluate the detector's performance on the DAIGT V2 (Essay) dataset to quantify the impact of text format on reliability. Finally, we present a stratified analysis of the essay dataset, breaking down detection rates by specific AI models to identify potential biases in the detector's architecture.

A. Phase I: Baseline Performance (Q&A Format)

In the control phase, the RoBERTa detector was evaluated on the HC3 dataset, which matches its training distribution (Short Q&A, primarily GPT-3.5). As hypothesized, the model demonstrated robust performance within its native domain. (see Table 1)

The accuracy of 99.40% and the perfect recall score (1.00) for human-authored text indicate that the detector is highly reliable for Q&A content. Within its native domain, the risk of wrongfully accusing a human author is negligible.

B. Phase II: Cross-Domain Stress Test (Essay Format)

When the same detector was applied to the DAIGT V2 dataset (Academic Essays), a significant degradation in performance was observed across all metrics. This confirms the hypothesis that the detector struggles to generalize from Q&A structures to long-form argumentation. (see Table 2)

Comparing the essay accuracy (84.06%) to the Q&A baseline (99.40%), we observe a sharp decline in reliability when the text format shifts to academic writing. The rise in the False Positive Rate to 8.4% is particularly concerning, as it implies that nearly 1 in 12 human-written essays was incorrectly flagged as AI-generated. Additionally, the high False Negative Rate of 27.7% indicates that over a quarter of AI-generated essays successfully evaded detection.

C. Phase III: Stratified Analysis by Generator (Model Agnosticism)

To detect any disparities in how the detector handles different AI models, we further broke down the detection rate by the specific source model. The results of this analysis reveal a clear bias towards older models, and a complete failure against newer architectures. (see Table 3)

The stratified analysis reveals that the detector demonstrated high efficacy against specific open-source models, achieving

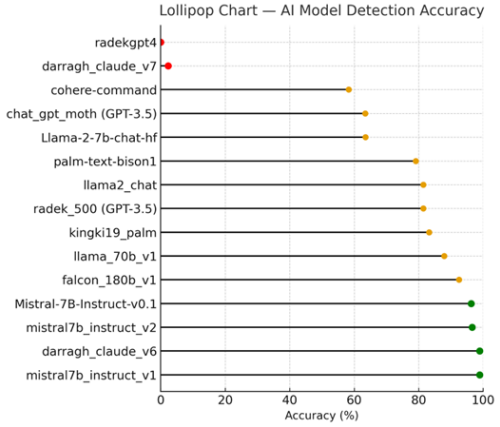

Fig. 2. Lopllipop Chart of AI Model Detection Accuracy

near-perfect accuracy on Mistral (98.9%) and strong performance on Falcon (92.5%), suggesting these models share statistical characteristics with the detector's training data. In contrast, performance degraded significantly for Llama-2 based models, with detection rates dropping to approximately 68%. Most critically, the detector exhibited a complete failure in identifying text generated by advanced proprietary models; detection rates for GPT-4 and Claude were negligible (0.0% and 2.3% respectively), indicating that current state-of-the-art generation capabilities have effectively outpaced the detection logic of RoBERTa-based systems.

VI. Discussion

The results indicate that the reliability of the RoBERTa-based detector is contingent on the specific model used to generate the text. While the detector successfully identified content from models such as Mistral and Falcon, it failed to generalize to more advanced architectures like GPT-4 and Claude. This performance disparity suggests that the detector relies on specific linguistic patterns or artifacts present in earlier model generations (such as GPT-3.5) that are absent or refined in newer models. Consequently, the tool does not function as a universal AI detector but rather as a classifier for specific subsets of model outputs.

The study also shows that a RoBERTa-based AI text detector works well only with certain types of texts. It performs best with short Q&A data, where it is nearly perfect. This confirms that these detectors are very effective when used with data similar to that on which they were trained. In the HC3 domain, the detector rarely made mistakes, indicating that it worked well for its intended purpose. This finding aligns with other studies that have reported precision for models trained on structured short text [13], [15].

However, the detector struggles with long academic writing. The drop from 99.4% accuracy in Q&A to 84.06% in essays reveals a significant limitation in cross-domain generalization. This indicates that it strives to work with various types of texts. The detector made more mistakes, with 8.4% false positives and 27.7% false negatives. This means that it incorrectly labels many human-written essays and misses many AI-generated essays. This supports earlier findings by (Tulchinskii et al., 2023) that skills learned in one area do not transfer well to others [16]. Academic writing is more complex and different from Q&A, which explains the lower reliability of the detector.

The analysis also reveals that the detector's weaknesses vary with different AI models. It detects Mistral and Falcon well because they are related to GPT-3.5, on which the detector was trained. However, it fails with Llama-2, GPT-4, and Claude, where the detection rates range from 0% to 2%. This highlights a significant issue: new AI models do not exhibit the same patterns that older detectors rely on. As AI models become increasingly adept at mimicking human writing, existing detectors become less effective. This trend aligns with earlier research by (Ippolito et al., 2020; Mitchell et al., 2023) indicating that improved AI models render older detectors less effective [9], [14].

These findings raise ethical concerns. The 8.4% false-positive rate means that human-written essays might be wrongly flagged as AI-generated. This raises fairness concerns, particularly for students whose writing differs from the typical training data, such as multilingual or non-native English speakers. Simultaneously, the high false-negative rate means that students using advanced AI might not be detected, leading to unfair academic integrity.

These results show three important things; doing well in one area does not mean success in another, detection tools need to keep up with new model abilities, and bias in models is a key issue that needs more attention. The study suggests that future detection tools should utilize training methods that adapt to diverse areas, incorporate data from various fields, and incorporate enhanced language features to function effectively in real-world education.

VII. Limitations

While this study provides important insights into the cross-domain reliability of AI detection, particularly the RoBERTa-based detection, it is important to note certain limitation to contextualize the findings:

- **Asymmetrical Datasets:** While the study compares detection across formats (Q&A vs. Essays), the baseline Q&A dataset (HC3) consists primarily of ChatGPT outputs. As a result, we could not evaluate whether the performance gaps observed for Llama and GPT-4 persist in short-form Q&A contexts or if they are specific to the essay format.
- **Temporal Scope:** This study evaluates a detector and generative models representative of the 2023–2024 landscape. The specific 0% detection rate for GPT-4 is a finding relevant to this specific detector version. As generative models and detection methods continue to evolve, these specific accuracy figures may shift, although the underlying challenge of model generalization remains.
- **Sample size variance:** Sample sizes in the DAIGT V2 dataset vary across generative models due to the resource constraints associated with generating proprietary model

data. While the sample size for GPT-4 (N=200) is smaller than for open-source models (N=2421), the consistent 0.0% detection rate across this subset provides a strong indication of detection failure that is unlikely to be an artifact of sample size alone.

VIII. Conclusion

This project evaluates the cross-domain robustness of a standard RoBERTa-based AI detector by examining its performance across two fundamentally different writing formats and multiple generator architectures. By benchmarking the detector on the HC3 Q&A dataset and then subjecting it to an out-of-distribution stress test using the DAIGT V2 academic essay dataset, we provide a comprehensive assessment of its generalization capabilities.

We identified a critical limitation; although the detector performs well within its native Q&A domain, its accuracy drops sharply when applied to long-form academic writing. This decline raises concerns about relying on such systems for high-stakes academic integrity decisions. In addition, the detector's performance varies widely across different AI models. These inconsistencies suggest that current detectors are overfitted to older model patterns and simpler text structures rather than genuinely capable of distinguishing human writing from AI-generated content.

Overall, this study demonstrates that the RoBERTa-based detector cannot be considered a general-purpose solution for AI-content attribution. Instead, it behaves as a narrow classifier optimized for a limited subset of models and formats. As AI text generation continues to advance, the gap between detector capabilities and real-world demands will only widen. These findings highlight the urgent need for improved detection systems that are explicitly designed to handle diverse writing styles, evolving model architectures, and dynamic linguistic patterns.

Future research should explore detectors trained on varied datasets, hybrid approaches that incorporate stylistic, semantic, and metadata-based signals, and evaluation frameworks that reflect real-world academic and professional contexts. Until such advances are realized, institutions should be cautious in using AI detectors as definitive evidence and consider them only as supportive tools within broader human-centered evaluation processes.

References

[1] Ahmad, P. N., Shah, A. M., & Lee, K. (2025). Enhanced Propaganda Detection in Public Social Media Discussions Using a Fine-Tuned Deep Learning Model: A Diffusion of Innovation Perspective. *Future Internet*, *17*(5), 212.

[2] 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).

[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.

[4] Erol, G., Ergen, A., Gülşen Erol, B. *et al.* Can we trust academic AI detective? Accuracy and limitations of AI-output detectors. *Acta Neurochir* **167**, 214 (2025).

[5] 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

[6] Hassan, E., Talaat, A. S., & Elsabagh, M. A. (2025). Intelligent text similarity assessment using Roberta with integrated chaotic perturbation optimization techniques. *Journal of Big Data*, *12*(1), 164.

[7] Jawahar, G., Sagot, B., & Seddah, D. (2019, July). What does BERT learn about the structure of language?. In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

[8] Jin, H., Ashrafi, N., Abdollahi, A., Liu, W., Wang, J., Gui, G., ... & Feng, H. (2025). LLM Encoder vs. Decoder: Robust Detection of Chinese AI-Generated Text with LoRA. *arXiv preprint arXiv:2509.00731*.

[9] Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020, July). Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1808-1822).

[10] 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

[11] 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.

[12] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers.

[13] Mindner, L., Schlippe, T., & Schaaff, K. (2023). Classification of human- and ai-generated texts: Investigating features for chatGPT. Lecture Notes on Data Engineering and Communications Technologies, 152–170.

[14] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023, July). Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning* (pp. 24950-24962). PMLR.

[15] Mobin, M. K., & Islam, M. S. (2025). LuxVeri at GenAI Detection Task 3: Cross-Domain Detection of AI-Generated Text Using Inverse Perplexity-Weighted Ensemble of Fine-Tuned Transformer Models. *arXiv preprint arXiv:2501.11918*.

[16] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Nikolenko, S., Burnaev, E., ... & Piontkovskaya, I. (2023). Intrinsic dimension estimation for robust detection of ai-generated texts. *Advances in Neural Information Processing Systems*, *36*, 39257-39276.

[17] 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).