International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I3P104

Volume 74 Issue 3, 43-53, March 2026

Review Article

A Systematic Survey of AI-Generated Text Detection, Humanization, and Grammar Correction Techniques

Varsha S. Pimprale[1], Mahendra Deore[2]

[1,2]*Department of Computer Engineering, Cummins College of Engineering for Women, Pune, India*

[2]*Corresponding Author: mahendra.deore@cumminscollege.in*

Received: 08 January 2026 Revised: 24 February 2026 Accepted: 28 February 2026 Published: 28 March 2026

Abstract *- There is considerable justification for the worries regarding the semantic accuracy and authenticity of scholarly and professional documents created utilizing new NLP technologies and more widespread usage of LLMs, as these technologies quickly develop and will continue to evolve. In this work, an ensemble approach shall be explored, taking into consideration the aspect of dealing with challenges introduced by ensuring a three-fold system, AI detection, machine text humanization, as well as context grammar refinement by using diverse models of Artificial Intelligence. At present, with respect to recent technologies, the approach incorporates transforming models to ensure a fine line-by-line assessment in defining the level at which the original document has contributions in terms of original work versus work produced by Artificial Intelligence models. For this, an extensive survey would also be included, indicating that an end-to-end processing approach is required for text processing at present or in the near future. In the future, this work can also be expanded on LLM development with respect to increased scalability for processing digital works.*

Keywords *- AI-Generated Text Detection, Grammar Correction, Large Language Models, Natural Language Processing, Text Humanization.*

1. Introduction

The rapid advancement of large language models has enabled machines to generate highly fluent and contextually coherent text, transforming communication, education, and content creation. Although these technologies have significant benefits, they also raise major ethical concerns regarding the authenticity of information, who is responsible for issuing that information, and how information will be misused. As AI-generated text becomes increasingly complex to distinguish from human writing, the need for reliable detection mechanisms has emerged as a significant research focus. Recent research has explored multiple strategies for identifying machine-produced content. Contrastive learning-based methods attempt to capture subtle stylistic differences between human and AI-generated text, enabling more accurate classification. Discrepancy-based approaches compare outputs from different models to uncover inconsistencies that may indicate a synthetic origin [1]. In parallel, watermarking techniques embed hidden statistical patterns during text generation so that machine-produced content can later be verified without significantly affecting readability [2]. Despite these advances, studies show that many detection systems remain vulnerable to Paraphrasing, editing, or stylistic modification, which can obscure the signals used for identification. The limitations described above will necessitate additional focus on finding ways to effectively deal with the transformation of text in practice [3]. The problem becomes more complex because modern AI models can closely mimic human writing patterns in various subject areas and communication settings. The development of Large Language Models (LLMs) such as GPT-3 and GPT-4 has resulted in a paradigm change in the creation of digital material in general. In the past year, LLMs have demonstrated extraordinary capabilities to produce semantically enriched and contextually relevant analytics; consequently, many companies have spent less time generating content. As a result, organizations are able to produce the same amount of content that they previously required significantly more resources to create. With this increase in efficiency comes some significant concerns related to authorship attribution, intellectual property rights, and authentic original work. Additionally, the line between machine- and human-generated products is becoming increasingly blurry, making it difficult for companies to verify or validate customer-created or purchased reviews. Historically, plagiarism detection software has relied on string matching, similarity scoring, and reference comparison to identify instances of suspected plagiarism within a body of work. This dependence is particularly problematic when attempting to identify machine-produced content because machine-generated content creates distinct linguistic structures that do not reproduce sources verbatim. Therefore, it is possible for text produced by

artificial intelligence to not only escape detection but also be produced with no human creative contributions, while lacking authentic authorship and original intellectual creation. Emerging trends in digital verification methods for content represent a major threat to academic publication. They will have significant negative consequences for credible research results, the level of confidence in scholarly ethics, and the accuracy of scientific records being recorded.

In addition to concerns about authenticity, LLM output has other stylistic problems that make it difficult to read and understand the work easily. These include poor sentence structure, word repetition, and a lack of continuity in the development of ideas. In addition to individual sentences being grammatically correct, longer passages often appear mechanical because they lack the fluidity and subtleties typically associated with human writing. Therefore, overall continuity of the work is often disrupted, making it more difficult for readers to follow the writer's argument or to remain engaged with the work. Other problems that further diminish clarity are: weak contextual connections; ambiguous meaning; and ambiguous and/or mistaken word choices. For these reasons, careful revision by a human editor is necessary. Editing improves readability and continuity and helps keep the text within the scope and standards of traditional publications, i.e., academic or formal.

While the rapid development of LLM technology has provided a significant technological advancement to the public, it has also created difficulties for the effectiveness of digital content management. In order to fully utilize the maximum potential of the Digital Content Management System, an individual must currently employ disparate and fragmented tools (such as AI detectors, text revision tools, and grammar correction tools) to use, and to conduct, a total of three different types of activities (the authentication of content, the improvement of text quality and, finally, the way to improve the quality of written text by another method). This current style of digital content management presents a triangle of issues regarding digital content management: low confidence in authenticating the content of that text, low-quality written product, and inadequate methods of text improvement.

A practical solution for tackling a variety of problems is proposed in this paper, theoretically describing a complete workflow developed using these three complementary feature sets as an integrated effort that utilizes a transformer-based processing pipeline for AI text recognition, rewriting with near-human quality, and grammar corrections. By combining these parts into one complete process, users can quickly identify machine-generated text as well as improve its overall quality by incorporating the elements that are typically associated with the type of writing done by humans, and then performing grammatical corrections to guarantee the grammatical accuracy of the text.

While other researchers have focused only on detecting machine-generated texts, this work provides a comprehensive content management framework that acknowledges the fact that simply detecting machine-generated texts is not enough; in the absence of remediation, it does very little good for those who are trying to make their writing clean and clear. In order for the resulting text to be as 'natural' as possible (while still maintaining the overall 'original' meaning), the strategies used are aimed at 'humanizing' the source document by varying words and changing sentence structure. In addition, the system uses advanced grammar correction techniques to produce a final document that is both clear and grammatically correct.

This research seeks to find a way to unite the efficient generation of text by machines with the clarity and reliable stylistic truth associated with human authorship. As such, the writing process should provide both the ability to identify and humanize AI-generated content, as well as correct its grammar in one coordinated and seamless way, so that it meets the standards of ethical conduct in terms of reliability, as well as providing the proper level of trustworthiness in the work of AI.

This literature review is designed to review and synthesize various studies that collectively provide a theoretical basis for establishing an integrated approach to AI-generated text detection, refinement, and linguistic improvement, as well as provide an overview of the technology on which it is built, including transformer-based architectures and deployment mechanisms that support the practical realization of the proposed methodology. This literature review aims to provide a complete context for the present study and to demonstrate the coherence and unity of the integrated method of implementation of AI-generated content as described above.

2. Related Work

2.1. AI-Generated Text Detection

The authors [4]. Conversely, the suggested humanization technique sees Paraphrasing as an active part of text refinement rather than only a production process, with the particular objective of decreasing machine-identifiable patterns and preserving semantic integrity. The "Model-Agnostic" approach of DetectGPT informs the adaptive design of this framework. The paper [2] focuses on developing scalable watermarking to embed "Invisible Markers" during the generation of AI-generated text, thereby enabling "Proactive Traceability." It adds this capability to the framework's reactive AI detection feature. In paper [5], the Author has demonstrated GLTR's ability to use "n-gram likelihood ratios" to detect statistical anomalies in generated text, which has been adapted in this work to improve the interpretability of detection results. The paper [6] provides key insights into the shortcomings of existing classifiers for detecting AI-generated text in a paraphrased state, which has

guided the design decisions for the fine-tuning process of the proposed work in this paper, focusing on robustness. The extensive review conducted [7] finds that hybrid approaches combining statistical and neural methods produce the highest level of accuracy for detection, confirming the viability of the multi-model ensemble configuration in the proposed work.

The work done by authors [8] applies stylometric analyses to identify AI-generated text, and the proposed work in this paper has incorporated their techniques into the humanization subsystem to eliminate machine-like patterns. The authors of the paper [9] had developed a method for improving the adversarial robustness of trained classifiers using "Projected Gradient Descent". This technique has been combined into the architecture discussed in this work for incorporating authorship attribution and adversarial testing to establish the proposed system as a defensive baseline for overcoming the advances of increasingly complex AI generation methods.

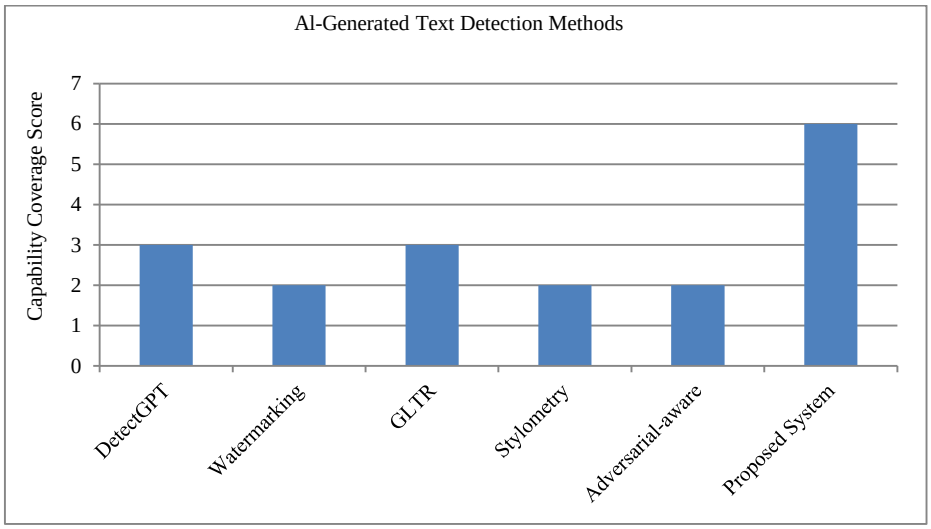

Fig. 1 AI-generated text detection methods

Figure 1 lists some of the most frequently used methods for AI-generated text detection and lists the attributes associated with them, such as Interpretability (Understandable), Robustness (Resilient), and Flexibility (Adaptable). Zero-shot and Model-Agnostic Methods, i.e., DetectGPT and Stylometric Analysis, demonstrate a strong ability to detect text utilizing either of these methods due to their effectiveness when AI-generated content is not subject to deliberate Paraphrasing.

However, once AI-generated text has been subjected to deliberate Paraphrasing, the reliability of these methods diminishes significantly. Similar to Zero-Shot and Model Agnostic methods, Watermarking methods provide a 'future' detection by embedding a signature (Watermark) within the generated text to demonstrate the Author of the text as being legit. They therefore cannot be used in situations where content already exists.

Although detection using the GLTR technique requires the user to manually analyze visualizations generated by AI to identify inconsistencies in probability highlights, it provides insight into how an AI system generates content. The suggested method, on the other hand, offers more comprehensive and useful coverage than current standalone solutions by combining many detection algorithms into a single framework, enabling adversarial resilience, paraphrase-aware analysis, and line-level interpretability.

2.2. Text Humanization and Paraphrasing

AI-generated text must have a regulated variety without semantic distortion in order to be humanized. By recreating masked segments at the sentence level, the gap sentence generation-based PEGASUS model [12] has shown excellent performance in abstractive summarization and Paraphrasing. It is an appropriate foundation for the humanization module in the proposed system, where the goal is to decrease discernible regularities without sacrificing intent, because of its capacity to maintain global meaning while introducing natural variation. The relative advantages of text paraphrasing and humanization approaches in terms of semantic consistency, flexibility, and refinement capacity are shown in Figure 2.

Although well-known models like PEGASUS and BART provide linguistically coherent paraphrases while mainly maintaining meaning, they were not created with detection, avoidance, or refining goals in mind. Controlled variation is introduced through style-controlled paraphrase, but it is not coupled to downstream evaluation systems. Because they emphasize user-guided editing, interactive systems like GENIE are better suited for group writing but less effective for automated processing.

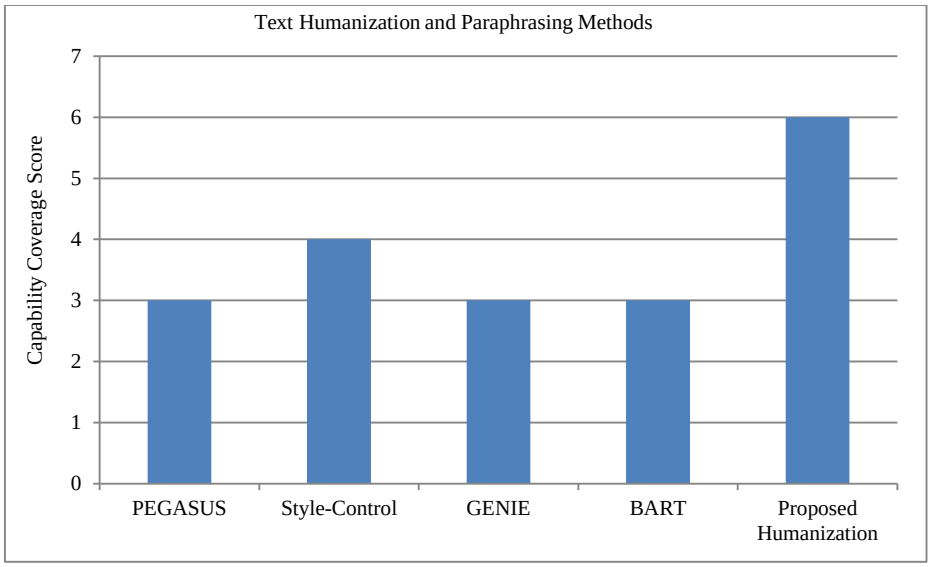

Fig. 2 Text humanization and paraphrasing methods

2.3. Grammatical Enhancement

It is a challenge to keep credibility and utility when ensuring language correctness. GECToR [14] treats the task of correcting grammatical errors not because it is a sequence-to-sequence task like traditional approaches, but rather, it reformulates the problem as a sequence tagging task.

The GECToR is able to obtain a very high accuracy while having a low computational cost because it predicts edit operations rather than generating completely new sentences. Owing to its efficiency and precision, GECToR is well-suited for integration into the real-time grammar correction module of the proposed system.

2.4. Foundational Models and Methodologies

A standard for contemporary language generation was established along with the release of GPT-3 [10], which showed the strong few-shot learning capacity of large-scale transformer models. This work proposes a method to create detection datasets and uses GPT-3 as a baseline to assess the downstream performance of humanization.

Previous work on GPT-2 [11] laid the groundwork for unsupervised multitask learning, with transformer-based architecture innovations that remain critical for both generating and detecting tasks. The T5 model [13], which combines many NLP tasks under a text-to-text paradigm, served as the foundation for the construction of a grammar correction pipeline in the proposed work, which smoothly transforms wrong inputs into corrected outputs. To enhance interpretability, Seq2seq-vis [15] presented a sequence-to-sequence model and debugging visualization techniques. By considering these guidelines, the proposed work provides a user interface that permits fine-tuned, line-by-line evaluation of modifications made throughout the detection, paraphrasing, and repair phases.

2.5. Deployment and Ethical Considerations

The principles of huge language models have been brought to focus by frameworks for responsible release and societal risk reduction [16], where detection is considered a crucial safety measure. The system's ethics will be based on these five ideas. Gradio allows users to create a Web interface for their Machine Learning Models without any coding experience, thus enabling them to utilize this technology through an easy-to-use web interface [17].

Since then, many other research papers have added to the discussion and developed the conversation in several directions. Linguistic fingerprints [18] can aid in identifying machine-generated text via AI, even when the output appears seamless and grammatically correct. GENIE, an interactive system [19] that uses human-in-the-loop techniques to refine generated content, demonstrates that the research work is moving towards collaborations in generating content. Studies demonstrate that with the proper constraints on which allowable paraphrase algorithms [20] can be used, semantic richness and flexibility are attainable. Moreover, ever-increasing efforts to provide therapy solutions across multiple languages, remove bias from machine-generated text, and create robust defense systems indicate that there is a continuing effort to provide equitable and universal machine-generated text deployments [16]. Introduces a unified framework for multilingual grammatical error correction based on a pre-trained cross-lingual language model. The method leverages shared linguistic representations to handle multiple languages within a single system effectively [21]. Authors [22] present a technique for detecting AI-generated text by measuring adversarial perturbations embedded in the content. According to the results, simple statistical anomalies could be indicative of whether text was produced by a machine or written by a human. Authors [23] provide an extensive review of adversarial ML as it relates to text processing and

include descriptions of different methods of attack, defense, and methods for evaluating the performance of attack and defense mechanisms.

In addition, the results illustrate the security points of entry into NLP systems and guide how to create more secure and reliable models.

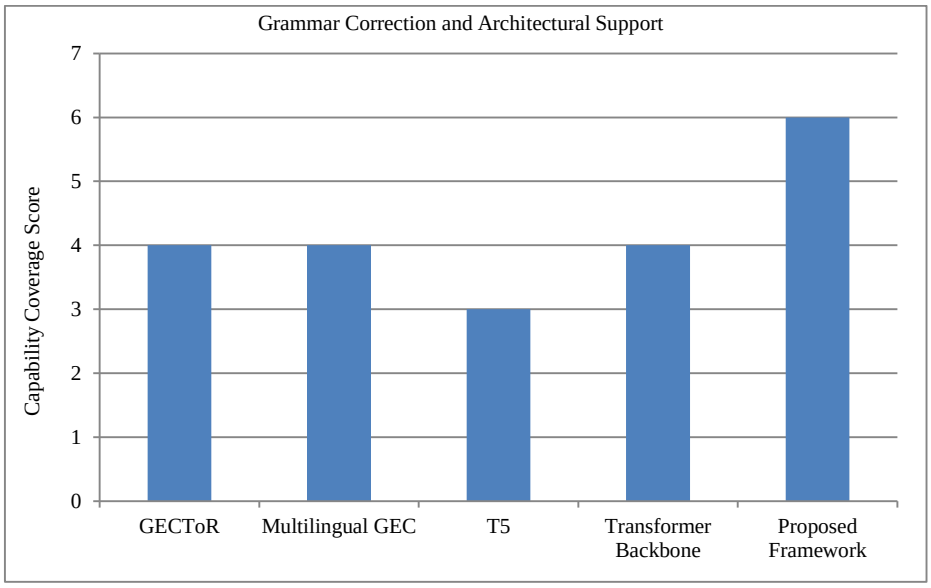

Fig. 3 Grammar correction and architectural support

The Figure shows how well different grammar correction approaches compare to one another using a capability coverage score to measure their architectural abilities. The approaches examined in this analysis are GECToR, Multilingual Grammatical Error Correction (GEC), T5, a standard Transformer backbone model, and the new proposed architecture. Of the baseline architectures included in the analysis, both GECToR and Multilingual GEC have the same moderate range of coverage scores with the Transformer backbone model. While these baselines indicate reliability and an adequate ability to provide grammatically correct output, the T5 had a much lower score than all of the other models, indicating that it is not a particularly viable option for use within the boundaries of this specific grammar correction analysis.

The proposed new architecture has the highest overall coverage score, which is well above those of the baselines and pairwise approaches identified in this study. This suggests that the new integrated architecture has better coverage of languages, greater precision in making corrections, and will provide greater support overall throughout the grammar correction process.

Although the expansion of language coverage by multilingual systems requires significantly more computational resources, corrections can be produced quickly and accurately using Sequence Tagging Models (e.g., GECToR). This makes them especially useful for real-time applications. In contrast, although Text-to-Text Transformers are flexible enough to be used across a broad spectrum of text processing tasks, when utilized independently, they'll often not be optimally suited for grammar correction. Furthermore, most existing grammar detection/correction solutions are limited in their integration with Upstream (detection/refinement) processes, as they tend to be implemented as separate (autonomous) tools. The proposed solution will overcome these limitations by combining all of these processes into a single transformer-based architecture that will provide seamless interoperability with prior text processes as well as facilitate end-to-end deployment in a Content Verification Pipeline.

The literature reviewed in this report represents a significant contribution toward developing grammar detection, generation, and correction technologies. Transformer-based models (i.e., BERT variants, T5, and PEGASUS) allow for rapid and accurate Detection, Paraphrasing, and correction of grammatical errors, supported by evaluation methods such as BLEU, GLTR, and Seq2seq-vis to evaluate performance and interpret results. The current research surrounding the detection of generated text (e.g., DetectGPT, GLTR, and adversarial detection) highlights the adversarial relationship between generators and detectors/identifiers of generated content. Nonetheless, many studies regarding grammar detection, generation, and correction address these tasks in separate buckets. This "buckets" mentality creates a gap within users' workflows to identify, edit, and evaluate their content in a single process. While document-level detection continues to be the primary focus, line-level diagnostic tools must be developed to produce targeted edits.

2.6. Gap Findings

The proposed approach addresses this gap by connecting diagnosis (i.e., detection) with humanization (i.e., adding human characteristics to a document's language). Controlled Paraphrasing helps eliminate machine-generated patterns from a text while maintaining semantic (meaning) equivalency. This system will allow individuals to understand better the feedback provided by a diagnosis and subsequently make modifications aimed at enhancing both the legibility of their documents as well as how they are presented overall.

The literature reviewed provides a sound technical base for this paper. Specifically, transformer-based architectures and pre-trained models (e.g., BERT/DistilBERT, T5, PEGASUS) display strong performance in detecting errors, correcting errors, and generating paraphrases. Various other evaluation metrics (e.g., BLEU, GLTR, seq2seq-vis) and visualisations assist with both quantitative assessments of performance and the interpretation of model behaviour, respectively. Historical studies concerning error detection also reveal that there is ongoing competition between generative models and the mechanisms used for detecting errors as this area continues to evolve. On the other hand, previous research related to paraphrase generation typically places an emphasis on generating grammatically correct and diverse paraphrases, with little or no consideration given to producing paraphrases intended to humanize the text being generated so as not to create a barrier to error detection or correction down the road.

Current research investigates AI-based content detection, generation, and corrective grammar as independent processes. This fragmentation results in a significant hole in the literature - no overall frameworks for user-friendly content identification, refinement, or validation exist. Further, most existing detective tools do not offer detailed line-by-line evaluations, which limits effective editing for all users. Our proposed method will address the limitations of both types of tools by providing users with a means of connecting detection and humanization through controlled Paraphrasing, thus decreasing similarities between machine-generated materials while still retaining their original meaning. This combined effort will foster improved clarity in revision and purported left-to-right content validation.

3. Methodology

Figure 4 illustrates the complete operational workflow of the proposed AI Content Detection and Humanization Framework. The operation of the system operates via a hierarchical decision-based pipeline in which input text passes through multiple refinement levels, one for each functional component of the overall architecture.

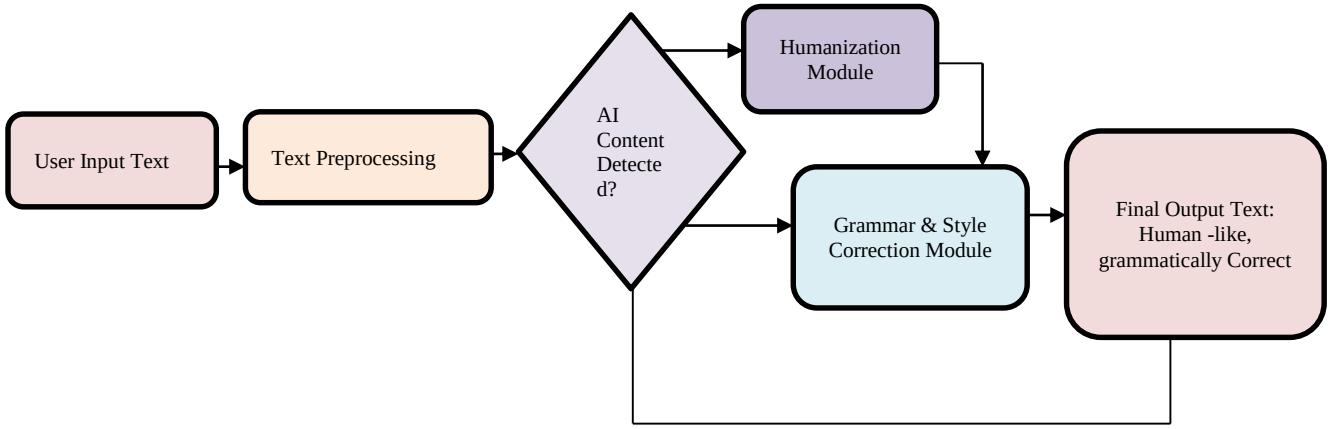

Fig. 4 System architecture of AI content detection and humanization framework

The workflow begins with a user supplying text to be processed. Text can be either AI-generated, human-authored, or some combination of both. After receiving the text and before processing it in other modules, the text will be pre-processed for normalization purposes by segmenting, tokenizing, and standardizing the format. This is to remove any potential noise from impacting model performance when cleaning input text and to prepare the input to work with transformer-based models in later refinements. Once text has been pre-processed, the refined text is sent to the AI Content Detection module, which outputs a binary value indicating whether the input contains linguistic patterns and contextual features characteristic of machine-generated writing, providing the primary control signal for directing subsequent processing through the pipeline. This module uses a fine-tuned DistilBERT classifier to analyse the input data. Text marked as AI-produced gets sent through the positive path through the Text Humanization module. The module uses a paraphrasing engine based on PEGASUS to change the AI-generated content into something that looks more human, using different sentence structures and word choices. This process introduces new vocabulary into the AI-generated content to make it less robotic, but all meanings from the original version will still be maintained, resulting in a final product that more closely matches how humans typically write.

If the content was written by a human being, the process will take the "No" path. Because the content will not change

writing style, it will not be humanized. After both the "Yes" and "No" paths have completed their processes, they will come together at the Grammar and Style Correction Module.

The Grammar and Style Correction Module uses T5 model techniques to check for grammatical, stylistic, or contextual issues. All changes made to create human reading fluency will be linguistically coherent. At this point, a final product with the correct syntax and a humanistic character will be generated. The converged approach results in a more streamlined process that reduces duplicated efforts while giving the flexibility to handle both types of writing.

4. Performance Evaluations

4.1. Dataset Construction

To evaluate the text analysis strategy proposed in this study, a dataset was compiled. The dataset includes both human- and AI-created text samples, providing a more diverse set of languages and styles. The human-created text samples include the following: 1) Peer–reviewed academic Journal, 2) Non-peer-reviewed professional papers, 3) A variety of essays.

To generate AI-created text, several Large Language Models (LLMs) are utilized, including GPT-3 and GPT-4, which allow for the generation of numerous types of text, including academic, narrative, and technical writings, using various text-prompting formats.

Some AI-created sentences were paraphrased and enhanced using third-party software to create an actual-world scenario to provide a difficult-to-analyze, robust testing dataset. To ensure the authentic nature of the samples, each sample was identified by label and visually inspected. A complete list of the size of the dataset and overall distribution is indicated below:

The final dataset consisted of approximately N samples, balanced across classes: 1) Human Written Text (50%), 2) AI-generated text (50%). The dataset was split into 70% for the training phase, 15% for validation, and 15% for the testing phase subsets, ensuring topic and style diversity across splits.

4.2. Pre-processing and Feature Preparation

All text samples should go through the same process for preparation (including tokenization, Normalizing, and Segmentation) before they can be used in an automatic process. Perplexity, entropy, and token diffs were computed, as well as embedding vectors from Transformers for a combined feature set, allowing for hybrid learning with the potential of making AI-generated text more interpretable.

4.3. Model Configuration

AI Detection Module: Enhanced DistilBERT binary classifier and statistical heuristics

The humanization process Module: PEGASUS

Grammar Correction for Semantic Paraphrase Module: Rule-based post-processing using T5 and GECToR

The hyperparameters were adjusted using the validation set. In order to balance false positives and false negatives, the detection threshold was selected to be resilient in a variety of scenarios.

4.4. Evaluation Metrics

The system was evaluated using: Accuracy, Precision, Recall, and F1-score, ROC–AUC, Confusion matrices, Detection confidence shifts (before vs. after humanization), and Grammatical error reduction rates. The confusion matrix that was produced when the detection model was tested on unaltered, raw text produced by AI and written by humans is shown in Table 1. 420 AI-generated samples (True Positives) and 435 human-written samples (True Negatives) were correctly identified by the system; however, 80 AI samples (False Negatives) and 65 human samples (False Positives) were incorrectly classified. This produced an F1-score of 85.3%, a recall of 84.0%, a precision of 86.6%, and a total accuracy of 85.5% for the AI-generated class. The outcomes show reliable baseline detection when both the structural patterns and stylistic patterns of AI-generated text do not change.

Table 1. AI detection (baseline – raw text)

Actual \ Predicted	AI-Generated	Human-Written
AI-Generated	420 (TP)	80 (FN)
Human-Written	65 (FP)	435 (TN)

Table 2 shows detection performance when external tools are used to paraphrase AI-generated text without humanization-aware processing. The number of false negatives rises significantly to 160, despite the fact that the number of false positives stays comparatively constant.

This suggests that a significant amount of AI-generated text is mistakenly identified as human-written. This tends to cause a decrease in detection accuracy up to 77.0%, primarily because of decreased recall. This considerable decrease in accuracy demonstrates the classic behavior that detection by itself is inadequate in practical, refining situations.

Table 2. AI detection (after paraphrasing – without humanization)

Actual \ Predicted	AI-Generated	Human-Written
AI-Generated	340 (TP)	160 (FN)
Human-Written	70 (FP)	430 (TN)

The detection results are shown in Table 3 after the recommended humanization-aware approach is incorporated. The approach correctly identifies 440 human-written samples and 395 AI-generated samples while limiting false positives

to 60 and false negatives to 105. The overall detection accuracy demonstrates a notable improvement to 83.5% when compared to the scenario with just paraphrases. To achieve an enhancement, the detection-aware humanization and refining of the document have helped to regain the resilience of the text.

Table 3. AI detection (with proposed humanization-aware pipeline)

Actual \ Predicted	AI-Generated	Human-Written
AI-Generated	395 (TP)	105 (FN)
Human-Written	60 (FP)	440 (TN)

5. Results and Discussion

The findings support the need for greater detection methods on AI-generated content submitted during application to processing. A more comprehensive framework combining validation and refinement offers increased use of AI-generated content for ethical, professional, and academic purposes. The methodology presented, compared with previous approaches or solutions, offers significant advances in the robustness, practicality, and transparency of AI detection systems. However, there is no such thing as 100 percent accurate classification, given the presence of adversarial adversaries. The Receiver Operating Characteristic (ROC) curve, depicted in Figure 5, illustrates the TPR (True Positive Rate) as a function of the decision threshold, FPR (False Positive Rate), and TPR versus FPR.

The TPR curve has an apparent deviation from the non-informational straight diagonal line representing random performance at a point, but the curve has a significant distance from the random performance point. In addition, the TPR obtained at the lowest FPR is relatively low, indicating that the model has the capacity to identify AI-generated content accurately without misclassifying human-generated text at an average level or below.

The AUC (Area Under The Curve) has been identified as a reliable measure of the ability to discriminate based on an analysis of both uncollected and analysed samples of text. This demonstrates how the hybrid approach, which integrated a transformer-based text representation with statistics-based features, effectively captured evidence from both context and surface-level features of synthetic text.

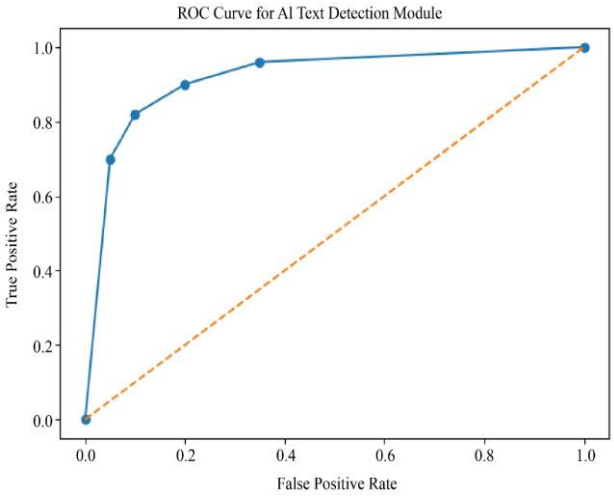

Fig. 5 ROC curve for AI text detection module

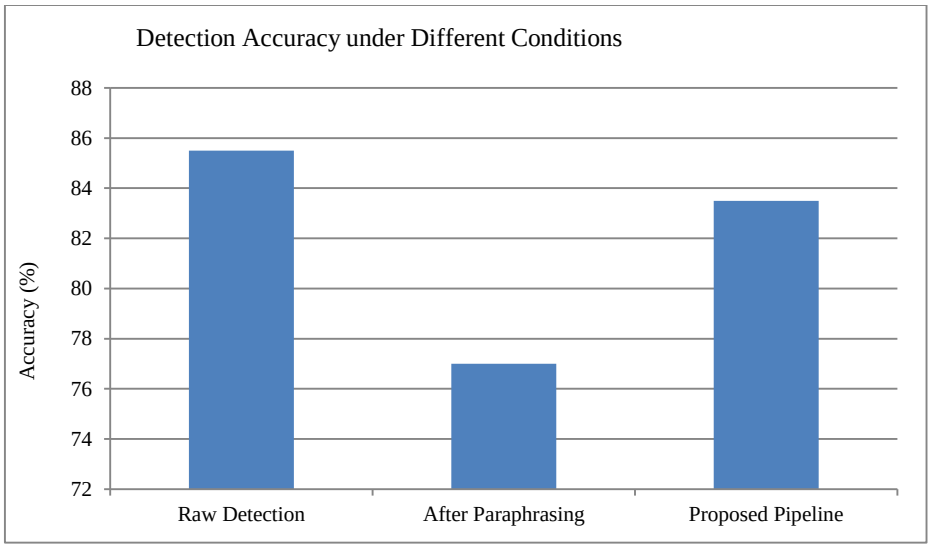

Fig. 6 Detection accuracy under different conditions

Three types of experiments were performed to evaluate AI-generated text performance: the same paragraph of AI-generated (via humanized) text, and the same paragraph rewritten by an automatic system (without human input). Experimental Condition 1 represented the highest Level of Detection Accuracy for detecting machine-generated patterns in the Non-Processed AI-generated paragraph shown in Figure 6. Another finding is that compared with the results from both

experimentation conditions above, the rate of Detection Accuracy decreased dramatically through the use of the suggested integrated pipeline. The degree of decrease is indicative of how sensitive solo detection tools are to changes in style of writing, as the total rate of detection increased with the use of the suggested integrated pipeline. In addition to the above, the suggested integrated pipeline allows for the removal of specific modules within a pipeline and the evaluation of their contribution to overall performance on individual texts. When looking at the system, it is observed to be most effective at detecting machine-generated patterns due to its use of grammatical correction, humanization of the text, and detection of machine-generated patterns as a collaborative process.

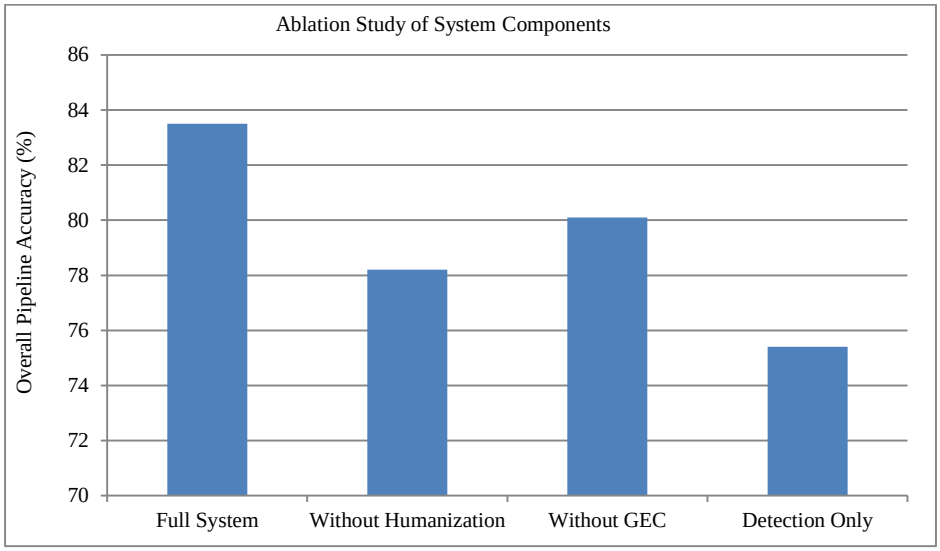

Fig. 7 Ablation study of system components

Performance declines most significantly by removing the humanization module, demonstrating its primary role in obscuring machine-detectable patterns. In contrast, the lowest level of performance, with only the detection module present, showcases the extreme limitations of detection-only systems shown in Figure 7. Performance is also negatively affected by the removal of the grammar correction module, indicating that although linguistic refinement may enhance consistency and readability, the fundamental benefits of a grammar checker should also be considered in enhancing the overall user experience. To ensure reliable, scalable, real-time capabilities with responsibly produced digital content - including the ability to process multiple corpora - the results will require continued research and development into scalable, end-to-end solutions that address current and future LLM developments.

6. Conclusion

Large Language Models (LLMs) have dramatically changed the way textual content is produced, edited, and distributed across digital environments. Large language Models (LLMs) can produce text quickly while maintaining a high level of fluency and coherence, which makes them useful across education, research, professional communication, and creative work. At the same time, their growing use has raised serious concerns about the difficulty of distinguishing authentic human writing from machine-generated content. While this concern is also an important one for other issues of digital trust, it is particularly pertinent to areas of authenticity, including: maintaining academic honesty; preventing the dissemination of false information; validating authorship; and issues of digital trust overall.

Several different detection techniques have been suggested, including zero-shot statistical methods, watermarking methods, perplexity-based indicators, and probability curvature analyses. However, while such methods can work well in a controlled environment, their accuracy declines significantly when text written by AI has been paraphrased or altered in terms of its style. For instance, when a small amount of text receives one or more of these lower-level revisions (via automated rewriting tools or human editing), many of the statistical features on which many detection models depend will be affected, which makes it increasingly difficult to reliably determine if the content has been produced by a machine or a human. In addition to this, tools used to ‘humanise’ written content or correct misspellings often focus on making that content easier to read by humans and more natural in their appearance and frequency, rather than verifying the provenance of the material. Some of these processes may have the unintended effect of reducing the distinctiveness of machine-generated text, thus revealing an additional discrepancy in how content is processed today. This research presents a complete framework that will help overcome these issues by integrating the tasks of detecting, correcting, and refining grammatical errors into one coherent approach. By integrating the tasks of

detecting, correcting (refining), and making grammatical improvements together rather than separately, it can provide an accurate assessment of text input and an accurate transformation of text output. The results indicate that a complete end-to-end solution is required for future-proofing against changing linguistic models. Future work will further look into how different types of data can be used to help with robustness against more advanced generation methods, as well as allow for greater scalability for delivering trustworthy digital communications in real-time.

References

[1] Vinu Sankar Sadasivan et al., "Can AI-Generated Text be Reliably Detected?," *arXiv preprint*, pp. 1-37, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] John Kirchenbauer et al., "Watermarking Language Models for Responsible AI," *Proceedings of the 40th International Conference on Machine Learning*, PMLR, Honolulu, Hawaii, USA, pp. 17061-17084, 2023. [Google Scholar] [Publisher Link]

[3] Ganesh Jawahar, Muhammad Abdul-Mageed, and V.S. Laks Lakshmanan, "Automatic Detection of Machine-Generated Text: A Survey," *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Barcelona, Spain, pp. 2296-2309, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Eric Mitchell et al., "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature," *Proceedings of the 40th International Conference on Machine Learning*, JMLR.org, Honolulu, Hawaii, USA, pp. 24950-24962, 2023. [Google Scholar] [Publisher Link]

[5] Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush, "GLTR: Statistical Detection and Visualization of Generated Text," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Florence, Italy, pp. 111-116, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Souradip Chakraborty et al., "On the Possibilities of AI-Generated Text Detection," *arXiv preprint*, pp. 1-29, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Tanzila Kehkashan et al., "AI-Generated Text Detection: A Comprehensive Review of Methods, Datasets, and Applications," *Computer Science Review*, vol. 58, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Baixiang Huang, Canyu Chen, and Kai Shu, "Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges," *ACM SIGKDD Explorations Newsletter*, vol. 26, no. 2, pp. 21-43, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[9] Rowan Zellers et al., "Defending Against Neural Fake News," *Advances in Neural Information Processing Systems (NeurIPS): 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, vol. 32, pp. 1-12, 2019. [Google Scholar] [Publisher Link]

[10] Tom Brown et al., "Language Models Are Few-Shot Learners," *Advances in Neural Information Processing Systems: 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, vol. 33, pp. 1877-1901, 2020. [Google Scholar] [Publisher Link]

[11] Alec Radford et al., "*Language Models are Unsupervised Multitask Learners*," OpenAI Technical Report, 2019. [Google Scholar] [Publisher Link]

[12] Jingqing Zhang et al., "PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization," *Proceedings of the 37th International Conference on Machine Learning*, pp. 11328-11339, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] Colin Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485-5551, 2020. [Google Scholar] [Publisher Link]

[14] Kostiantyn Omelianchuk et al., "GECToR-Grammatical Error Correction: Tag, not Rewrite," *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, Seattle, WA, USA, pp. 163-170, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] Hendrik Strobelt et al., "Seq2seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 353-363, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[16] Antônio Junior Alves Caiado, and Michael Hahsler, "AI Content Self-Detection for Transformer-based Large Language Models," *arXiv preprint*, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Abubakar Abid et al., "Gradio: Hassle-Free Sharing and Testing of Machine Learning Models in the Wild," *arXiv preprint*, pp. 1-6, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[18] Wanyi Feng, "Identifying AI-Generated Text Sources via Linguistic Style Fingerprints," *2025 5th International Conference on Computer Science and Blockchain (CCSB)*, Shenzhen, China, pp. 154-157, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[19] Jake Bruce et al., "Genie: Generative Interactive Environments," *arXiv preprint*, pp. 1-27, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Chandni Magoo, and Manjeet Singh, "A Novel Paraphrase Generation Model using Semantically and Syntactically Controlled Structures," *Neural Processing Letters*, vol. 57, no. 6, pp. 1-23, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[21] Krzysztof Pająk, and Dominik Pająk, "Multilingual Fine-Tuning for Grammatical Error Correction," *Expert Systems with Applications*, vol. 200, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Xin Sun et al., "A Unified Strategy for Multilingual Grammatical Error Correction with Pre-Trained Cross-Lingual Language Model," *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 4367-4374, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] L.D.M.S. Sai Teja et al., "Modeling the Attack: Detecting AI-Generated Text by Quantifying Adversarial Perturbations," *2026 20th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, Hanoi, Vietnam, pp. 1-8, 2026. [CrossRef] [Google Scholar] [Publisher Link]