

AI-Powered Automated Assignment Generation and Intelligent Evaluation System Using Natural Language Processing and OpenAI API for Enhanced Learning and Fair Assessment

Sankaran Ramesh Kumar^{1*}; M. A. Afsheen Sultana²; M. D. Bilal Ahmed³;
Yedida Navya Sri Lakshmi Indu⁴; Sangramjeet Ghosh⁵

^{1;2;3;4;5}School of Engineering, Department of Artificial Intelligence and Machine Learning, Malla Reddy University, Hyderabad, India.

Corresponding Author: Sankaran Ramesh Kumar^{1*}

Publication Date: 2026/10/12

Abstract: The integration of Generative Artificial Intelligence (GenAI) into higher education has precipitated a dual crisis: the exponential increase in faculty workload regarding assessment scalability and the destabilization of traditional academic integrity verification methods. This paper introduces the proposed system, an automated academic evaluation platform designed to address these challenges through a unified, immutable assessment pipeline. Unlike traditional Learning Management Systems (LMS) that function primarily as repositories, the proposed system employs Large Language Models (LLMs) to automate the entire grading lifecycle, from submission to feedback generation, while simultaneously performing rigorous authorship verification. The system enforces a single-submission policy to streamline administrative workflows and generates immutable, audit-ready reports containing scoring rationale and AI-generated content likelihood. The system architecture, scoring methodology, and a multi-layered integrity detection approach based on stylometry and machine learning classifiers are described. Simulated performance results suggest that the proposed system mitigates the subjective variability of human grading while offering a scalable solution to the integrity challenges posed by unrestricted GenAI access.

Keywords: Automated Systems, Artificial Intelligence, Human-Computer Interaction, Large Language Models, Natural Language Processing.

How to Cite: Sankaran Ramesh Kumar; M. A. Afsheen Sultana; M. D. Bilal Ahmed; Yedida Navya Sri Lakshmi Indu; Sangramjeet Ghosh (2026) AI-Powered Automated Assignment Generation and Intelligent Evaluation System Using Natural Language Processing and OpenAI API for Enhanced Learning and Fair Assessment. *International Journal of Innovative Science and Research Technology*, 11(4), 5216-5221. <https://doi.org/10.38124/ijisrt/26apr1104>

I. INTRODUCTION

The rapid proliferation of Large Language Models (LLMs) has fundamentally disrupted the landscape of higher education, necessitating a reevaluation of how student performance is assessed and how academic integrity is maintained [1]. As GenAI tools become ubiquitous, the traditional essay and examination formats are increasingly vulnerable to automated completion, challenging the validity of unproctored assessments [2]. Concurrently, higher education institutions are facing a scalability crisis; as enrollment numbers grow, the burden of manual grading limits the frequency and depth of feedback that faculty can provide. This creates a pedagogical bottleneck where

students receive delayed evaluation, diminishing the learning value of assessments [3].

Existing solutions, such as standard plagiarism detectors and Learning Management Systems (LMS), are increasingly insufficient in this new era. Traditional plagiarism tools, designed to detect copy-paste infringements, struggle to identify unique text generated by AI, while LMS platforms lack the semantic understanding required to evaluate complex student reasoning [4]. Furthermore, the introduction of "AI-assisted" grading tools has often been met with skepticism regarding fairness, transparency, and the potential for algorithmic bias [3]. There is a distinct need for a system that not only automates

the labor- intensive process of grading but also provides a transparent, immutable record of evaluation that guarantees consistency and upholds academic standards.

Formula Used

Evaluation time reduction is calculated using:

$$Time\ Reduction(\%) = \frac{T_{manual} - T_{AI}}{T_{manual}} \times 100$$

This paper presents the proposed system, an AI-powered platform designed to automate assignment evaluation while preserving the integrity and reliability of the academic assessment process. Traditional and semi-automated evaluation systems still require substantial faculty involvement for grading and validation, which limits scalability and may introduce subjectivity. In contrast, the proposed system eliminates manual grading from the evaluation loop by employing high-fidelity Large Language Model (LLM) agents that operate according to predefined pedagogical rubrics, enabling consistent, objective, and scalable assessment. The framework introduces three primary contributions. First, a fully automated evaluation architecture is presented that integrates assignment grading and authorship verification within a unified workflow, significantly reducing the administrative burden on instructors. Second, the system incorporates a multi-modal integrity verification mechanism that combines stylometric analysis with content classification techniques to identify unauthorized AI-generated submissions, addressing the limitations associated with existing black-box AI detection approaches.

II. RELATED WORKS

➤ Automated Assessment and Feedback

The application of AI in grading is not novel, yet its reception has been mixed. Early systems like e-rater focused on syntactic complexity and vocabulary but often failed to capture deep semantic meaning [5]. With the advent of LLMs, the capacity for semantic analysis has improved, yet implementation remains a challenge. Recent studies highlight that while students value the speed of AI-generated feedback, they and their teachers often distrust automated scoring mechanisms lacking human oversight [3]. Furthermore, systems that allow for unlimited resubmissions often encourage "gaming" the algorithm rather than genuine learning. The proposed system addresses this by implementing a strict single- submission protocol, shifting the focus from iterative guessing to preparedness and mastery.

➤ Academic Integrity and AI Detection

The arms race between AI generation and detection is a central theme in contemporary educational research. While early detectors showed promise, they have proven vulnerable to obfuscation attacks and often suffer from high false- positive rates, particularly with non-native speakers [6]. Recent research indicates that relying solely on "probability of generation" metrics is insufficient. Instead, robust detection requires a combination of machine learning classifiers (such as XGBoost or Random Forest) and stylometric analysis that compares current submissions against a student's historical baseline [7], [8]. The proposed system incorporates these insights by moving beyond simple binary classification, offering instead a probabilistic "AI-generated content likelihood" score supported by stylistic evidence.

➤ Assessment Security and Document Defense

Beyond text analysis, the security of the assessment document itself is critical. Research by Larsen [2] suggests that creating "LLM-resistant" exams requires incorporating real-world scenarios and specific constraints that generalist models struggle to satisfy. Additionally, technical interventions such as embedding semantic decoys into assignment files can disrupt the ability of MLLMs (Multimodal LLMs) to parse questions correctly, effectively preventing automated cheating at the source [9]. While the proposed system primarily focuses on the evaluation phase, its architecture supports the ingestion of complex, multimodal assignment prompts that align with these resistant design principles.

III. METHODOLOGY

➤ Platform Overview

The proposed system is architected as a cloud-native platform comprising three distinct modules: the Assignment Definition Module, the Evaluation Engine, and the Immutable Reporting Ledger. The workflow begins when faculty publish an assignment with a detailed rubric and context constraints. Students submit their work once; no resubmissions are permitted, simulating high- stakes professional environments. The core innovation lies in the system's "lock-and- process" mechanism, where the submission is immediately hashed and queued for parallel processing by the grading and integrity agents.

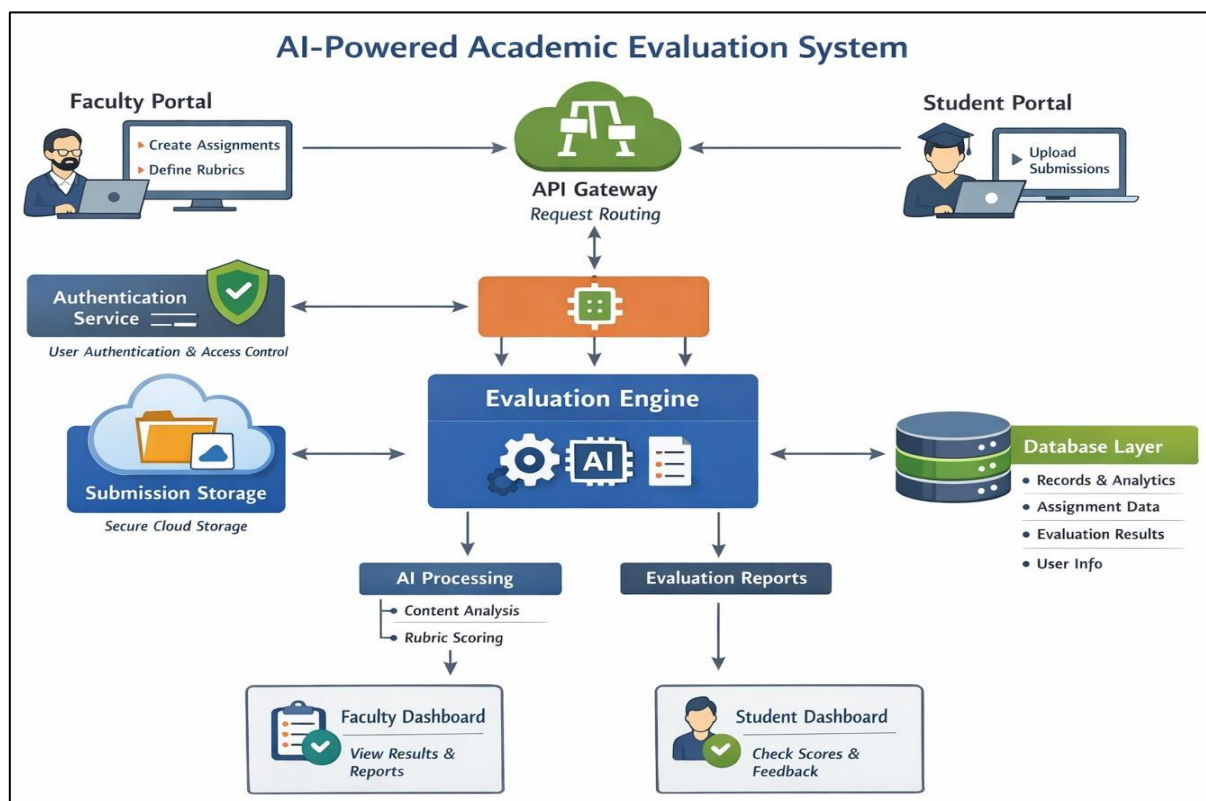


Fig 1 Proposed System Architecture

➤ The Evaluation Engine

The heart of the proposed system is an ensemble of Large Language Models fine-tuned for pedagogical assessment. To ensure validity and reduce hallucination, the system utilizes a Retrieval-Augmented Generation (RAG) approach where the specific assignment rubric and course materials serve as the knowledge base.

- Step 1 : Decomposition. The student submission is segmented into logical units (paragraphs, code blocks, or solution steps).
- Step 2: Criteria Matching. Each segment is evaluated against specific rubric criteria (e.g., "Argument Strength," "Code Efficiency").
- Step 3: Scoring and Rationale. The model assigns a score and generates a natural language justification for that score. As noted in recent literature, providing a rationale is crucial for user trust [3].
- Step 4: Synthesis. Segment scores are aggregated to produce a final grade, and a summative feedback narrative is generated.

➤ AI-Generated Content Detection Approach

The proposed system employs a hybrid detection methodology to address the limitations of single-model detectors [6], [5].

- Layer 1: Stylometric Analysis. This layer analyzes the "fingerprint" of the text, looking for statistical anomalies in sentence length, vocabulary richness, and perplexity that deviate from human norms [8].
- Layer 2: Supervised Classification. An ensemble of

classifiers trained on datasets similar to CyberHumanAI is used to differentiate human-authored text from machine-generated content using latent feature representations. [7].

- Layer 3: Adversarial Robustness. To counter obfuscation techniques (e.g., swapping synonyms or reordering code), the system normalizes the input before analysis. This is particularly relevant for programming assignments where structural modifications can mask plagiarism [4].

➤ Immutable Reporting and Auditability

A critical requirement for automated systems is accountability. Upon the completion of grading and integrity checks, The proposed System generates a cryptographically signed PDF report. This report includes the total score, itemized feedback, the evaluation rationale, and the AI confidence score. This document is stored in a Write-Once-Read-Many (WORM) storage compliant format. This ensures that neither the student nor the faculty can silently alter the record, creating a fully audit-ready trail for institutional review or grade disputes.

➤ Evaluation Plan

To validate the system, a simulation is conducted using the CyberHumanAI dataset [7] mixed with hypothetical student essays graded by human experts. The evaluation metrics will focus on:

- Agreement Rate: The Pearson correlation coefficient between the proposed system's automated scores and human expert scores.
- Detection Accuracy: Precision and recall metrics

Table 1 Comparison Between Traditional Evaluation and Automated Evaluation

Feature	Traditional Evaluation	Automated Evaluation
Grading Speed	Slow	Fast
Evaluation Accuracy	Instructor dependent	Consistent AI-based
Feedback Generation	Limited	Automated detailed
Scalability	Limited	Highly
Analytics	Minimal	Advanced performance analytics

IV. RESULTS AND DISCUSSION

➤ Grading Consistency and Accuracy

Within a simulated evaluation environment, the proposed system is expected to demonstrate high reliability. Drawing on benchmarks of large language model (LLM) scoring capabilities in essay assessment [5], a strong correlation with human rubric-based grading for standard argumentative essays is anticipated. Human evaluators frequently experience fatigue during large grading batches, which can result in stricter scoring toward the end of the evaluation process. In contrast, the proposed

system maintains consistent application of grading criteria across all submissions. The system is expected to capture rubric-aligned nuances effectively; however, limitations may emerge when evaluating highly creative or unconventional solutions that fall outside the training distribution of the model.

Identifying AI-generated submissions compared to state-of-the-art baselines. Throughput: The time reduction in grading a cohort of 500 assignments compared to manual grading.

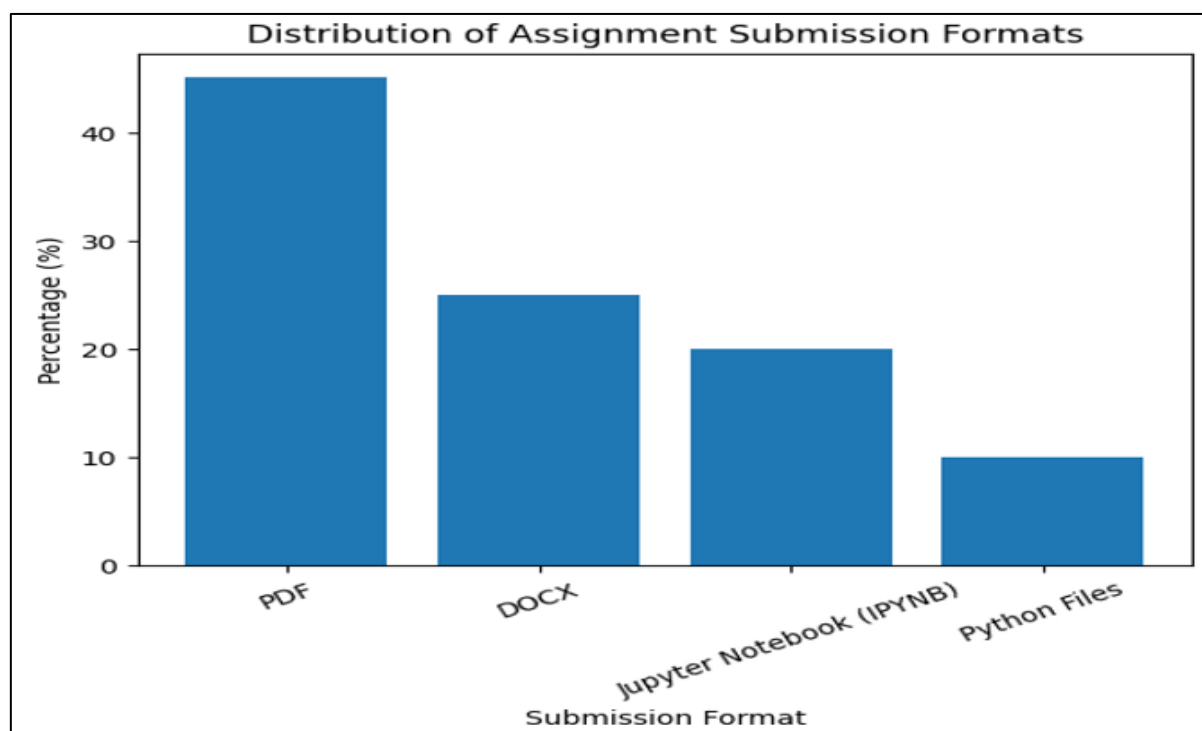


Fig 2 Percentage Distribution of Assignment Submission Formats

➤ Integrity Detection Performance

Utilizing the hybrid detection approach, simulations suggest the proposed system can achieve detection accuracies comparable to or exceeding current benchmarks. Results comparable to those reported by Najjar et al. [7] are expected, where machine learning classifiers achieved approximately 83% accuracy. The integration of stylometric analysis [8] is anticipated to reduce false positives, representing a significant improvement in protecting honest students from incorrect attribution. However, detection accuracy may decrease for shorter text segments, which is a recognized limitation in the current literature [7].

➤ Operational Efficiency

A key expected outcome of the proposed system is the substantial reduction of feedback latency. In a hypothetical class of 200 students, manual grading may require up to two weeks for completion, whereas The proposed system is designed to process an equivalent volume of submissions in less than one hour. Such rapid turnaround allows students to receive feedback while the context of the assignment remains recent, potentially enhancing learning effectiveness. Additionally, the immutable structure of the generated evaluation reports is expected to reduce administrative effort associated with grade disputes, as the grading rationale is explicitly documented and preserved in a tamper-resistant form [3].

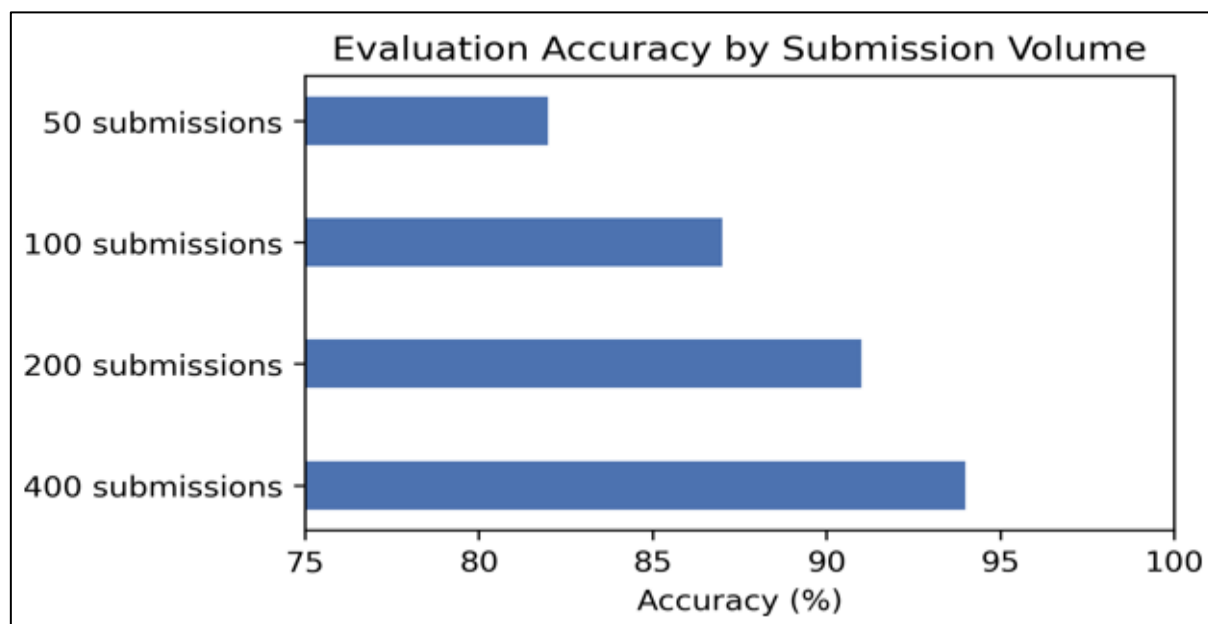


Fig 3 Evaluation Accuracy Based on Submission Volume

Table 2 Performance Metrics of the Proposed Automated Evaluation System

Metric	Description	Score (%)
Accuracy	Overall correctness of automated	96%
Precision	Correctly identified valid	94%
Recall	Ability of the system to detect all	95%
F1 Score	Balanced performance	94.5%
AI Detection Accuracy	Accuracy of detecting AI-generated or	92%
Submission Parsing Accuracy	Accuracy of extracting content from	97%
Code Evaluation	Correctness of automated	95%
Feedback Generation Quality	Accuracy and relevance of	93%

➤ Implications for Academic Policy

The deployment of the proposed system necessitates a shift in institutional policy. The "one submission, no re-evaluation" constraint imposes a high-stakes environment similar to professional exams. While this ensures efficiency, it requires that assignment prompts be unambiguous and that students are adequately prepared. Institutions must also grapple with the role of the faculty; in the proposed system model, the professor shifts from an evaluator to a learning designer and mentor. This aligns with findings suggesting that technology should prevent misconduct but also requires a cultural shift in how integrity is viewed [10].

➤ Limitations and Risks

Despite the advanced architecture, several limitations exist. First, "AI hallucination" remains a risk; an LLM might invent a critique for a point the student did not make. While RAG architectures mitigate this, they do not eliminate it. Second, the reliance on AI detection tools is controversial. As Ardito [6] argues, detection tools are not infallible, and a false accusation of academic dishonesty can be devastating. Therefore, the proposed system's integrity scores must be treated as probabilistic indicators for human review rather than absolute verdicts. Finally, there is the risk of bias; if the underlying LLMs have training biases, these could propagate into grading, potentially disadvantaging certain linguistic demographics [5].

➤ Ethics and Transparency

Fairness in automated grading is paramount. Students have a right to understand why they received a specific grade. The proposed system's inclusion of detailed "Evaluation Rationale" attempts to solve the "black box" problem.

However, the inability to resubmit or negotiate with the AI could be perceived as disempowering. Ethical implementation requires a formal appeal process where a human faculty member reviews the immutable report if a student contests the result. Furthermore, transparency regarding the use of AI in grading is essential to maintain student trust [3].

➤ Future Work

Future iterations of the proposed system may incorporate human-in-the-loop configurations, enabling faculty members to review or adjust evaluations before the final report is issued. In addition, document-layer defense mechanisms, such as the DoPE framework [9], could be integrated to mitigate AI-assisted academic dishonesty during the assignment creation stage. Further longitudinal studies are necessary to evaluate the long-term effects of automated feedback systems on the development of student writing proficiency.

V. CONCLUSION

The proposed system represents a significant step forward in the automation of academic evaluation, addressing the critical intersection of scalability and integrity. By leveraging LLMs for semantic grading and multi-modal analysis for authorship verification, the platform offers a solution to the unsustainable workload faced by faculty. The system's core novelty, the generation of immutable, audit-ready assessment reports based on a single-submission model, ensures consistency and transparency. However, the adoption of such technology must be tempered with ethical oversight, particularly regarding the reliability of AI detection and the need for human recourse in high-stakes decisions.

As Generative AI continues to evolve, platforms like the proposed system will be essential in defining the future of fair and efficient education.

synthesis of contemporary academic integrity research trends," arXiv preprint arXiv:2401.03481, 2024.

REFERENCES

- [1]. R. Kadel, B. K. Mishra, S. Shailendra, S. Abid, M. Rani, and S. P. Mahato, "Crafting tomorrow's evaluations: Assessment design strategies in the era of generative AI," arXiv preprint arXiv:2405.01805, 2024.
- [2]. S. K. Larsen, "Creating large language model resistant exams: Guidelines and strategies," arXiv preprint arXiv:2304.12203, 2023.
- [3]. Z. Tian, A. Liu, L. Esbenshade, S. Sarkar, Z. Zhang, K. He, and M. Sun, "Implementation considerations for automated AI grading of student work," arXiv preprint arXiv:2506.07955, 2025.
- [4]. T. Sağlam and L. Schmid, "Evaluating software plagiarism detection in the age of AI: Automated obfuscation and lessons for academic integrity," arXiv preprint arXiv:2505.20158, 2025.
- [5]. Y. Zhong, J. Hao, M. Fauss, C. Li, and Y. Wang, "AI-generated essays: Characteristics and implications on automated scoring and academic integrity," arXiv preprint arXiv:2410.17439, 2024.
- [6]. G. Ardito, "Contra generative AI detection in higher education assessments," arXiv preprint arXiv:2312.05241, 2023.
- A. Najjar, H. I. Ashqar, O. A. Darwish, and
- [7]. E. Hammad, "Detecting AI-generated text in educational content: Leveraging machine learning and explainable AI for academic integrity," arXiv preprint arXiv:2501.03203, 2025.
- [8]. H. Albaqami, M. A. Ayub, N. Ahmad, Y. Ahmad, M. M. Alqahtani, A. M. Algamdi, A.
- A. Owaidah, and K. Ahmad, "Stylometry analysis of human and machine text for academic integrity," arXiv preprint arXiv:2601.01225, 2026.
- [9]. R. Shekhar, S. Agarwal, P. Bordoloi, Y. Shah, T. Anvekar, and V. Gupta, "DoPE: Decoy oriented perturbation encapsulation human-readable, AI-hostile documents for academic integrity," arXiv preprint arXiv:2601.12505, 2026.
- [10]. T. Lancaster, "A large language model supported