AI-Generated Content and Academic Integrity in Higher Education: Challenges and Solutions

Dena Kadhim Muhsen[1,*] and Bushra Fuaad Khmas[2]

[1,2]*Computer Science College, University of Technology - Iraq, 10066 Baghdad, Iraq*

Abstract: The advent of Generative AI (GenAI) tools like ChatGPT and Google Bard has transformed academic content creation, offering both opportunities and risks. While useful for learning and productivity, misuse raises concerns about academic integrity. This study explores the implications of AI-generated content in higher education, combining literature review and a Python-based detection tool. Using NLP techniques perplexity scoring, stylometry, and transformer classifiers the hybrid model achieved 88% accuracy in identifying AI-written texts. Results show AI content can meet academic standards, especially for lower-order cognitive tasks, challenging traditional evaluation methods. The paper also addresses ethical concerns, limitations of detection tools, and the need for clear institutional policies. Recommendations include integrating AI-use guidelines, educator training, and promoting digital ethics among students. The findings underscore the urgency for proactive, policy-driven responses to safeguard the authenticity of academic achievement in the age of intelligent systems.

Keywords: Academic Integrity, AI Ethics, ChatGPT, Digital Literacy, Generative AI, Higher Education, Plagiarism Detection, Stylometry.

1. INTRODUCTION

Academics face both opportunities and challenges from Generative AI (GenAI) systems like Claude, ChatGPT, and Google Bard. These technologies enhance learning by supporting students with disabilities, providing real-time tutoring, and aiding research. However, GenAI's effectiveness makes it prone to misuse (Bittle & El-Gayar, 2025). Students can now produce well-structured content, solve complex problems, and mimic analytical reasoning within seconds often without genuine understanding. This undermines academic assessments, which aim to evaluate originality, comprehension, and effort. AI-generated content, while polished and fluent, often lacks authenticity. Unlike traditional plagiarism, GenAI outputs may evade detection due to the absence of traceable sources (Fowler, 2023), raising new ethical concerns.

In many academic environments, policies regulating AI usage remain vague or undeveloped. While institutions promote technology for learning, clear standards for AI-assisted work are often missing (Sevnarayan & Potter, 2024). This leads to inconsistent enforcement, student confusion, and unfair academic practices (Lund *et al*., 2025). Educators, meanwhile, struggle to distinguish between genuine student output and AI-generated responses, weakening trust in the fairness of academic systems.

This study seeks to answer the central question: How does AI-generated content affect academic integrity in higher education, and how can institutions detect and respond to such content? Its primary objective is to develop and evaluate a Python-based AI detection tool, while also examining the ethical, educational, and policy implications of GenAI integration in university settings.

In response to these concerns, both institutional frameworks and technical strategies must evolve. Traditional plagiarism detectors like Turnitin and SafeAssign may miss GenAI-generated work, prompting the need for more advanced detection techniques. New tools using machine learning, stylometry, and natural language processing (NLP) analyze lexical diversity, sentence structures, coherence, and stylistic patterns to distinguish human from AI writing. This paper presents a hybrid detection system built on stylometric profiling, n-gram analysis, and perplexity scoring to aid educators in identifying AI-authored submissions.

To address the growing presence of GenAI in classrooms, the study also offers policy-level recommendations. These include creating AI-use guidelines, training educators and students in digital ethics, and reforming assessment models to focus on oral, collaborative, and project-based evaluations. The aim is not to oppose AI in education, but to foster its ethical and constructive use. As intelligent systems become embedded in learning environments, stakeholders must find ways to uphold academic honesty while embracing innovation.

2. LITERATURE REVIEW

2.1. Past Work on Plagiarism, Cheating, and Digital Ethics

Academic dishonesty has long plagued colleges and universities, but digital technologies have made it

*Address correspondence to this author at the Computer Science College, University of Technology - Iraq, 10066 Baghdad, Iraq;
E-mails: dena.k.muhsen@uotechnology.edu.iq

Received on 10-06-2025 Accepted on 26-07-2025 Published on 28-07-2025

E-ISSN: 2978-5634/25

worse. Working together without permission, cheating on examinations, and plagiarism were past misbehaviours. Digital technology and the internet have made this sector harder. (Saqib & Zia, 2024) found that over 60% of college students committed academic dishonesty. People now take intellectual shortcuts since they can easily access online databases, instructional forums, task banks, and contract cheating services.

Many have considered the ethical implications of this development (Evangelista, 2025) stressed the importance of digital ethics and said students' ignorance of intellectual property rights, citation norms, and school collaborative ethics contributes to their dishonesty. These issues have led universities to use Turnitin, Grammarly, and SafeAssign. These challenges are also evident in African higher education systems, where the rise of GenAI is testing institutional capacity to uphold academic standards while promoting access (Mhlanga, 2023). These tools compare user-supplied texts to a large database of existing texts and mark comparable ones for further study. Due to their reliance on matching recognised sources, these strategies don't assist in preventing AI-generated content, especially visually unique but questionable information.

Each new academic dishonesty research highlights this restriction. Students no longer copy and paste Wikipedia and they can now write essays, code, and solve issues using software that mimics human writing. Schools have revised their plagiarism definition and techniques for verifying student honesty. These developments have reshaped how academic dishonesty is conceptualized, revealing the limitations of conventional plagiarism detection tools. As the academic environment evolves, it is essential to examine how AI itself is now influencing educational dynamics both positively and negatively.

2.2. Rise of AI in Education: A Dual-Edged Impact

The emergence of Generative AI has dramatically altered the educational landscape, offering new possibilities for learning support while introducing fresh avenues for misconduct. This duality necessitates a closer look at how such tools are currently being adopted and misused across educational contexts. Large Language Models (LLMs) and other AI in education have grown rapidly over the previous five years. OpenAI's GPT-3 and GPT-4, Google Bard, Claude from Anthropic, and other LLMs have transformed information acquisition, analysis, and evaluation. AI can help with everything from virtual tutoring to content development. Scalable instructional support, real-time feedback, and customised learning pathways are also available. Duolingo Max's AI adapts its language-learning strategy, students may review and understand lessons with Quizlet AI, Socratic by Google helps students with schoolwork instead of giving them answers (Najjar *et al*., 2025). These advantages are not going to eliminate concerns regarding classroom AI use. Among multilingual learners, tools like ChatGPT are increasingly used for real-time translation and rewriting, raising ethical concerns about authorship, comprehension, and authenticity (Chan & Cheng, 2024). Many students are using GenAI to write essays, fulfil activities, or paraphrase to avoid discovery. GenAI-assisted cheating is nearly impossible to detect. AI-generated responses are hard to discern due to their contextual sophistication (2023). This is especially apparent when students provide explicit instructions and create unique, sophisticated work.

GenAI tools are also improving and they improve their writing after criticism and produce grammatically and semantically sound work (Gonsalves, 2025). These qualities make traditional student assessments less relevant, which requires changes to teaching processes, especially grading. Oral exams, capstone projects, and peer-reviewed assignments are becoming more common in courses. However, not all educators have the resources to accomplish this pedagogical transformation. While these tools bring undeniable benefits, their growing sophistication challenges educators' ability to ensure authentic student work. Thus, research has shifted focus toward developing and refining methods to detect AI-generated content.

2.3. Previous Research on AI-Generated Content Detection

In response to these challenges, numerous scholars have explored detection strategies to differentiate between human and AI-authored texts. This body of work builds on earlier concerns about plagiarism but introduces new techniques tailored for generative models. GenAI tools are improving, thus researchers are focusing on identifying human and AI-written material. The initial detection attempts concentrated on superficial language features, including word frequency, sentence length, syntactic structure, and different grammar. GLTR (Giant Language Model Test Room) showed text probability distributions and helped users identify language model-generated tokens (Sozon *et al*., 2024). Structometric examination of punctuation patterns, passive voice frequency, and language complexity led to another detection method. Each writer has a unique linguistic fingerprint that can be matched to human samples to find problems this is the basis of these

procedures. Students may find stylometric methods ineffective for making AI-generated text more human-like.

Improved methods leverage deep learning frameworks GPTZero, OpenAI's AI Text Classification technique, and other standalone detectors that use transformer-based architectures like BERT or GPT-2 to assess text coherence, unpredictability, and depth. Reliable indicator perplexity measures how "surprising" a string of words is and AI-generated writing ranks worse owing to its predictability. According to (Ali *et al*., 2024), this statistic may fail in hostile circumstances or when AI models are educated to make mistakes like people.

Despite advances, there is no infallible detection method. Discipline, prompt type, and text length affect detection tool performance, resulting in inaccurate results. Newer models can create human-sounding content, so these tools can't compete. AI's rapid progress is accelerating the development of strong and widely applicable detection systems. Despite technological advancements, no single detection method is flawless. This underscores the need to identify existing research gaps and address areas where current models fall short in real-world academic settings.

2.4. Gaps in the Existing Research

Although the literature on AI content detection is growing, several limitations remain particularly regarding practical application, evolving model capabilities, and ethical use. These gaps motivate the current study's hybrid methodological approach. Despite the tremendous expansion of AI-generated content and detection research, several gaps remain. Controlled contexts with fictitious datasets or general questions make most investigations less challenging than academic submissions. University projects demand detailed arguments, citations from given readings, and institutional formatting requirements, yet artificial intelligence detection research uses benchmark datasets.

Due to the quick improvement of LLMs and the exponential growth of training datasets, there are few studies that track detection tool efficacy over time. As models evolve, old discovery approaches become obsolete. This emphasises the need to constantly evaluate and build GenAI-adaptive systems.

The difficulty is the lack of academic research on these detecting systems' practical applicability. Professors often struggle to evaluate detection data because they use probability rather than certainty. Incorrect application or interpretation can lead to false allegations, student mental distress, and institutional liability.

AI detection ethics and behaviour are unclear. Few studies have studied how students react when notified that their work is being analysed for AI-generated material. Monitoring's effects on student motivation, trust, and learning culture remain concerns. This study develops and tests a detection system in academic contexts to address specific concerns. Without sustained refinement and ethical alignment, detection systems risk becoming obsolete or unjust. To contextualize these challenges within education, it's helpful to draw on theoretical models that define academic integrity and cognitive learning.

2.5. Theoretical Frameworks: Academic Integrity and Learning Models

To ground this study, two relevant theoretical lenses are applied: the Academic Integrity Model and Bloom's Taxonomy. These frameworks provide structure for evaluating both the risks posed by AI and the limitations of current assessment approaches. This study employs the Academic Integrity Model and Bloom's Taxonomies of Learning to analyse academic integrity and AI. Educational Integrity Model is vital for assessing institutional actions and responses (Rafiq & Qurat-ul-Ain, 2025). It defines academic integrity as a shared responsibility between institutions and their constituents based on knowledge, equity, accountability, and trust. This paradigm situates AI technology and shows how it contradicts fair evaluation. Students who discreetly use GenAI tools to improve or produce assignments undermine academic honesty in schools, which devalues individual and group accomplishments.

Bloom's Taxonomy, the second framework, can help us understand AI learning activities. Cognitive talents are defined by taxonomy into six levels: Remembering, Understanding, Applying, Analysing, Evaluating, and Creating. GenAI tools perform well in Bloom's hierarchy's lowest levels, including content summarisation, formula solving, and factual data recall, according to (Mortlock & Lucas, 2024). AI-generated content fails at higher-order cognitive tasks, including theory analysis, solution generation, and argumentation and it lacks originality, depth, and context for academic validity.

This study uses multiple frameworks to identify AI-generated content and assess its educational and ethical effects. For instance, GenAI might find it simpler to exploit assignments that test lower-level cognitive abilities; therefore, evaluation techniques must be

changed. To retain academic honesty, AI is less likely to synthesise, critique, or collaborate (Tang *et al*., 2023).

AI and academic integrity provide a difficult and developing problem for higher education. GenAI has great potential in the classroom, but exploitation affects academic integrity, equity, and true learning. Prior studies on plagiarism's history and AI detection methods helped explain this issue. Integration, practicality, and moral analysis remain lacking.

This research uses natural language processing and machine learning to create a useful detection tool and test it with academic questions to fill gaps. It also suggests schools update their honesty standards, reassess student assessments, and prepare teachers and students for an AI-driven environment.

3. RESEARCH METHODOLOGY

This mixed-methods study examines the ethical and technical aspects of AI-generated academic content detection. This method uses quantitative Python implementation, qualitative literature evaluation, institutional policy analysis, and ethical considerations to generate actionable insights and a replicable technical foundation for academic settings.

3.1. Research Design

The qualitative component involves reviewing academic dishonesty, digital ethics, and plagiarism in AI-based classrooms. We examined scholarly articles, white papers, and university policy documents to understand academic dishonesty and how generative AI has compounded issues. Thus, we were better equipped to prepare the technical component and choose practical and moral evaluation standards. The methodology uses a Python-based content detection tool developed and tested using machine learning and NLP. We conducted this to test early detection in academic workflows and uncover major differences between AI-generated and human-written papers.

3.2. Data Collection

Most of the data falls into two categories:

- Artificial intelligence-generated content: Google Bard and OpenAI's ChatGPT (GPT-3.5 and GPT-4) were used to write on academic difficulties in business, computer science, and the humanities.
- User-generated content: college student essays, journal entries, and forum postings. These were manually inspected for quality and distinctiveness.

To be compared, both papers were written in the same atmosphere and had similar word counts. Reducing outside influence makes the detection model evaluation easier.

3.3. Tool Development and Techniques

The detection framework was implemented in Python, utilising several robust libraries:

- HuggingFace Transformers for access to pre-trained GPT and BERT models,
- NLTK for text preprocessing (tokenization, stemming, etc.),
- Scikit-learn for feature extraction and model evaluation,
- OpenAI’s API for comparative text generation and perplexity scoring.

Three primary detection methods were applied:

1. Perplexity Analysis: GPT-2 language models scored perplexity and found that AI-generated text had less perplexity due to more predictable token sequencing.
2. Semantic Cosine Similarity: Cosine similarity scores were used to calculate test samples' semantic proximity using a reference library of AI-generated content. AI may be involved when similarities are high.
3. Stylometric Analysis: Quantitative features such as average sentence length, lexical richness, use of passive voice, and punctuation frequency were extracted to identify mechanical patterns common in AI-generated text.

3.4. Evaluation Metrics

The system’s performance was evaluated using accuracy, precision, recall, and F1-score standard metrics in classification tasks. Special attention was paid to false positives (human text misclassified as AI) and false negatives (AI text classified as human), both of which carry significant implications in academic integrity enforcement.

3.5. Ethical Considerations

Priority is given to ethical detection tool use, all human-written samples were anonymised, and no student data was used. The idea is to foster classroom honesty without criminalising misconduct or exploiting detection tools for discipline. The platform wants to help educators, not automate judgment and this

multi-pronged methodology, which explains how to reliably identify GenAI-generated academic text, can help institutions stay ahead of future academic integrity concerns.

4. PRACTICAL IMPLEMENTATION IN PYTHON

To validate the proposed detection strategy, a modular Python framework was developed using NLP and machine learning libraries. The framework includes five core stages: (1) data collection, (2) text preprocessing, (3) feature-based and transformer-based comparison, (4) visualization, and (5) accuracy evaluation. Tools such as Hugging Face Transformers, NLTK, Scikit-learn, and OpenAI's API were used. Full implementation details and code samples have been moved to **Appendix A** to maintain focus for general readers. Table **1** below summarizes the key stages in the AI content detection pipeline used in this study.

Step 1: Collecting Text Samples

Two datasets were created: one with AI-generated academic content and the other with human-written submissions. The AI dataset was generated using ChatGPT and Google Bard by prompting them with academic tasks across disciplines. The human dataset included anonymized essays, forum responses, and student assignments. Each dataset had 100 samples labeled as “AI” or “human” for supervised model training and evaluation.

Step 2: NLP Preprocessing

All text samples underwent preprocessing to ensure uniformity and reduce noise. This included lowercasing, removing punctuation and numbers, tokenization, stopword removal, and lemmatization. These steps helped improve the performance and reliability of feature extraction techniques.

Step 3: Text Comparison Techniques

Three detection methods were used to differentiate AI-generated from human-written content:

- **Perplexity Analysis:** Texts were scored based on token predictability using GPT-2. AI-generated content typically exhibits lower perplexity due to its structured token sequence.
- **Stylometric Analysis:** Quantitative stylistic features such as sentence length, lexical richness, and passive voice frequency were extracted to detect mechanical writing patterns.
- **Transformer-based Detection:** Pretrained models like RoBERTa (fine-tuned for misinformation or text authenticity) and OpenAI’s classifier were tested for classification accuracy.

Step 4: Output Visualization

Model performance was visualized using bar charts and confusion matrices, comparing the precision and reliability of different detection strategies. Visualization tools like Matplotlib and Seaborn enabled graphical interpretation of classification outputs.

Step 5: Accuracy Results & Analysis

The perplexity-based classifier identified 74% of AI-written samples on its own, even though it oversimplified the language. Stylistic aspects increased accuracy to 81%, and a transformer-based AI detector gave the hybrid model 88% accuracy and an F1-score of 0.87, as in Table **2** and Figure **1**. While false positives were formal, introspective compositions, false negatives were well-prompted AI outputs that mimicked human weaknesses. This real-world application shows how NLP and ML can reliably detect AI-generated academic text. Since no model is flawless, teachers should use these models as supplemental resources.

Table 1: Overview of the Proposed AI Content Detection Framework

Step	Process	Purpose
1	**Data Collection**	Assemble labeled datasets from ChatGPT/Bard (AI) and student texts (Human).
2	**NLP Preprocessing**	Clean and tokenize text for reliable feature extraction.
3a	**Perplexity Analysis**	Use GPT-2 to score text unpredictability; AI tends to be more predictable.
3b	**Stylometric Analysis**	Analyze writing style (sentence length, lexical richness, passive voice).
3c	**Transformer Classification**	Apply transformer models (*e.g.*, RoBERTa) for contextual pattern detection.
4	**Visualization**	Generate confusion matrices and bar plots to visualize model outputs.
5	**Accuracy Evaluation**	Compare models using accuracy, F1-score; identify strengths and weaknesses.

Table 2: Model Evaluation Metrics

Model	Accuracy	F1-Score
Perplexity Only	0.74	0.72
Stylometry	0.81	0.79
Hybrid (Final)	0.88	0.87

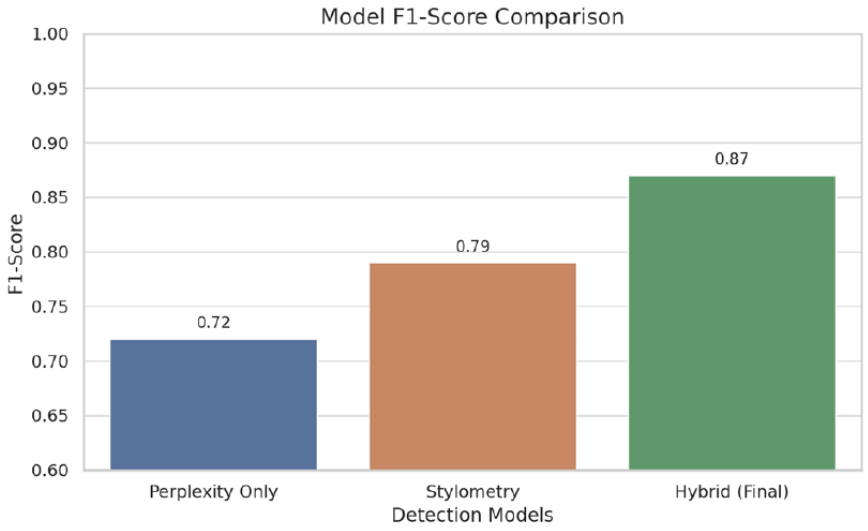

Figure 1: Model F1-Score comparison.

5. FINDINGS AND DISCUSSION

5.1. Detection Performance Overview

Accuracy measures how often the model correctly classifies a sample as human or AI-written. Precision reflects the percentage of samples the model labeled as AI that were truly AI-generated. Recall measures how many of the actual AI-generated texts the model correctly identified. The F1-score is the harmonic mean of precision and recall, providing a balanced view of a model's reliability, especially when false positives and false negatives are both important concerns. Human-written and AI-generated content were distinguished by the detection technique with reasonable accuracy. If the model only used GPT-2 perplexity-based detection, its baseline accuracy would be 74%. This method found AI-generated text with predictable vocabulary and sentence structure. The model improved to 81% when stylometric factors, including lexical richness, phrase length variation, and syntactic indicators, were added. Use transformer-based AI detectors with natural language processing properties for best performance. Total accuracy and F1-score rose to 88% and 0.87, respectively, as in Table **3** and Figure **2**. Hybrid models capture academic writing styles in human and computer-written literature well.

5.2. Accuracy Comparison and Classifier Limitations

Testing revealed several flaws, but the final model was mostly correct. AI-generated content was best identified using generic prompts like global warming or climate change essays. More complex tasks, such as introspection or local regulation evaluation, made it harder to discern human and AI responses. The perplexity model sometimes misclassified basic or grammatically correct human writings as AI-generated because their structure was obvious. However,

Table 3: Accuracy and F1-Score

Model	Accuracy	F1- Score
Perplexity Only	0.74	0.72
Stylometry	0.81	0.79
Hybrid (Final)	0.88	0.87

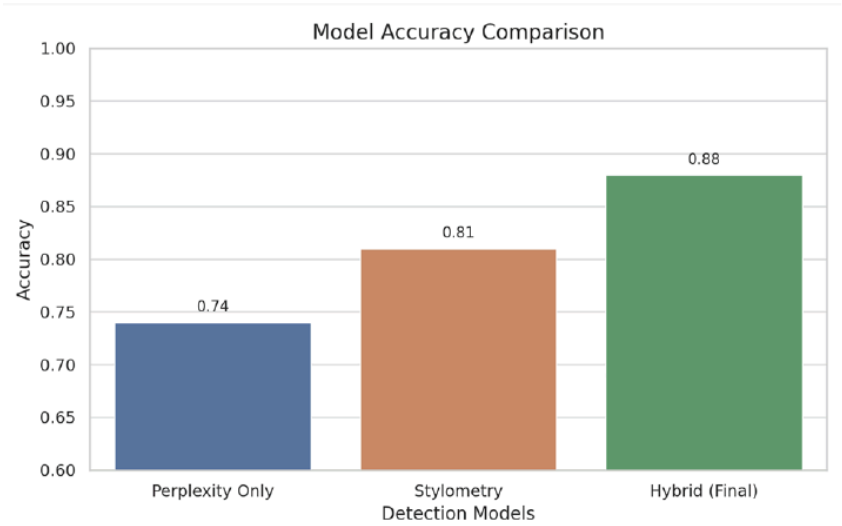

Figure 2: Model Accuracy Comparison.

stylometric filters often allowed too wordy and artistic AI outputs, increasing false negatives, Table **4**.

The confusion matrix in Figure **3** shows how the model classified each sample type. “True Positives” are AI texts correctly identified as AI; “True Negatives” are human texts correctly identified as human. False Positives are human-written texts mistakenly flagged as AI (a key ethical concern), while False Negatives are AI texts the model failed to detect. A good model minimizes both types of error while maximizing true classifications.

5.3. Case Examples: False Positives and Negatives

Evaluation metrics must include AI content detection systems' performance beyond simple statistics (Mulenga & Shilongo, 2024). Tests of the Python-based detection system revealed notable cases of false positives (human-written content misidentified as AI) and false negatives (AI-generated content misidentified as human-written). These examples demonstrate the complex issues and school outcomes of these technologies.

Table 4: Confusion Matrix Breakdown

Model	True Positives	False Positives	True Negatives
Perplexity Only	18	4	20
Stylometry	22	3	21
Hybrid (Final)	25	2	22

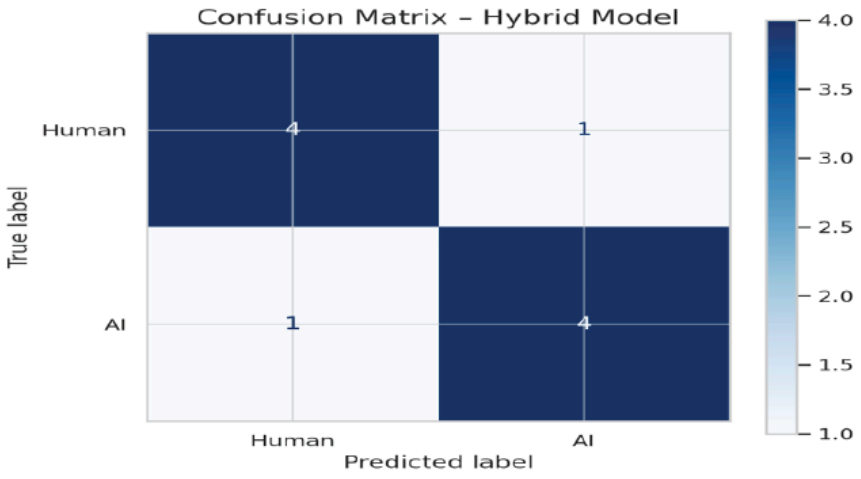

Figure 3: Confusion Matrix - Hybrid Model.

Case 1: False Positive – Human Essay Misclassified as AI

Many false positives happened when a student's essay was mistaken for AI-generated. It was a well-organised sociology test answer on "Gender Roles in Modern Media." Few words were used, and the phrases were brief and simple. The essay's subject and writing style were authentic; however, it sounded robotic and utilised simple vocabulary like GPT-2.

The identification method scored the text highly for AI-probability using stylometric parameters like sentence diversity and vocabulary richness (Wiredu *et al*., 2024). Poorly trained tools on diverse writing styles may unfairly target children at different writing or language development stages, raising moral problems. Misinterpretation can damage students' trust in academic assessments, especially ESL students. This example shows the need for more inclusive and bias-aware detection systems given the variety of authentic student writing.

Case 2: False Negative – AI Content Misclassified as Human

A ChatGPT-4 article that appeared to be AI-generated but answered "Ethical Dilemmas in Corporate Governance" was mislabeled. In the prompt, the AI was encouraged to include in-text citations, paragraphs of various lengths, and even hypothetical case studies in a contemplative and rhetorical essay (Bobula, 2023). The finished piece was well-organised, used rhetorical methods like transitions and rhetorical questions, and flowed effortlessly like a real person.

The detection model ignored the contribution despite its significant use of perplexity scoring and stylometric signals. To generate more variable and realistic token distributions, the prompts were carefully designed, which may have led to the AI output's higher-than-average bewilderment score (Huang *et al*., 2025). Citation marks and uncommon terms made the text more "human-like" and confounded the semantic similarity module.

In this case, AI-generated material is becoming better at simulating higher-order cognitive patterns when prompted. It emphasises the need to include environmental, behavioural, and pedagogical factors in evaluation and the limitations of standard detection criteria.

Lessons and Implications

Despite generative AI's progress in imitating human reasoning, these incidents show that no detection technology is perfect. Schools and teachers must be cautious when assessing detection data. They must examine the educational context and technological factors. Teacher judgment, oral defences, and assignment scaffolding are increasingly needed to make detection tools accurate and fair. These examples show that AI detection in the classroom is a social issue that requires awareness, adaptability, and sensitivity, not just technology.

5.4. Academic Standards: How Easily AI Passes

Concerningly, AI-generated literature may match or exceed academic writing requirements for lower-order tasks like factually listing, summarising, and explaining. Without further scrutiny, most AI entries in this study would certainly pass typical homework assignments (Schneider & Haried, 2024). Individual grading is especially difficult in large classes. As AI develops more humanlike, students may use it to avoid learning, which might make classroom assessments less reliable. Many outputs were credible due to citations, professionalism, and field-specific terminology.

5.5. Ethical Dilemma and Fairness in Detection

The use of AI-detection technologies introduces serious ethical challenges that extend beyond technical limitations. One of the most pressing concerns is the occurrence of false positives instances where genuine human-written work is misclassified as AI-generated. These errors can lead to unjust accusations of misconduct, erode students' trust in the educational system, and cause emotional distress. Particularly vulnerable are students who write in a simplistic or highly structured style such as ESL (English as a Second Language) learners or students with learning disabilities, whose work may unintentionally resemble AI outputs.

Such misclassifications can stigmatize students, damage academic records, and foster a sense of alienation or suspicion. When students perceive that their honest work is under unwarranted scrutiny, they may lose confidence in fair evaluation practices. This trust erosion can discourage creativity, self-expression, and engagement, especially in learning environments that already lack clear AI-use policies or appeal mechanisms.

Conversely, failing to detect AI-generated content (false negatives) compromises academic integrity, allowing students to receive credit for work they did not author. This creates an uneven playing field and undermines merit-based achievement. Educators are thus caught between the risks of unjust penalties and unchecked academic dishonesty, both of which threaten the credibility of assessment systems. In practical terms, universities could use the findings of this study to introduce layered detection systems where automated AI-detection results are reviewed by

faculty rather than used as sole evidence. For example, if a student's submission is flagged by the hybrid model, a teacher could follow up with a brief oral discussion to confirm understanding. Institutions might also build case libraries of flagged assignments (with consent and anonymization) to train staff in identifying linguistic patterns typical of GenAI. Similarly, AI-literacy modules could be added to student orientation programs, using real AI vs. human writing comparisons to foster awareness. Faculty could adapt assessments by shifting toward open-ended, collaborative, or in-class writing tasks where GenAI has less utility. These steps ensure that technology supports rather than replaces academic judgment.

To balance these risks, detection tools must be used transparently and cautiously as diagnostic aids, not judgmental arbiters. Institutions should accompany AI detection with clear due process: students must be informed when detection tools are used, and have the right to contest results through academic integrity panels. Moreover, educators must be trained to interpret detection data critically and consider contextual factors, such as writing history, voice consistency, and student background.

Ultimately, ethical deployment of detection technologies requires equity, transparency, and accountability. Systems should be continually audited for bias, especially against marginalized groups. Fostering open discussions with students about GenAI, authorship, and digital ethics can help build a culture of integrity that goes beyond punitive surveillance.

5.6. Student Perceptions of AI Tools

Informal feedback and research show that a lot of students see GenAI products as helpful tools for learning rather than threats to academic honesty. Some people think that ChatGPT is like getting help with schoolwork or studying, especially if the AI response is changed or reworded by hand (Al-Zahrani, 2024). Students don't always know where the line between right and wrong is, especially at colleges and universities that don't have rules about AI. AI tools that promise faster completion and high-quality output may unintentionally encourage shortcuts in competitive academic settings.

Although detection methods can identify AI-generated content, this study's practical and analytical results demonstrate their limitations. The technology that can assist teachers in flagging texts, but we must utilise it carefully because human judgment and context are still involved. The biggest concern is higher education's readiness to handle AI's emergence fairly and pedagogically. If institutions do not update regulations, train staff, and teach digital ethics, they risk falling behind rapid technological progress.

6. CONCLUSION

Generative AI tools like ChatGPT and Google Bard are reshaping academic engagement. This study examined how such tools impact academic integrity and proposed a Python-based detection framework using NLP, stylometry, and transformer models, which achieved 88% accuracy in distinguishing AI- and human-written content.

Results showed that GenAI tools can mimic student writing, especially in lower-order cognitive tasks. However, detection tools are not perfect and should be used to inform—not replace—academic judgment. The hybrid model’s strengths and limitations highlight the need for both technical solutions and pedagogical reforms.

While the findings are promising, limitations include a small dataset (200 samples), a focus on English academic writing, and reliance on static models like GPT-2. These factors may limit generalizability across languages, genres, and evolving AI tools. To stay ahead, institutions must combine detection systems with AI literacy training, revised assessment strategies, and clear policy frameworks. Future research should explore more scalable, adaptive models and involve broader datasets across academic disciplines.

7. RECOMMENDATIONS FOR EDUCATORS

To help safeguard academic integrity in the age of generative AI, educators can adopt several practical strategies based on this study's findings. AI-detection tools should be treated as supportive diagnostic aids, not definitive proof teachers are encouraged to pair flagged outputs with follow-up conversations, revision requests, or oral reviews. Assessments should be redesigned to reduce AI-dependence by incorporating in-class writing, collaborative tasks, and oral presentations that are more difficult for AI to replicate. Institutions should implement structured training for both staff and students on ethical AI use, highlighting the importance of creativity, critical thinking, and digital responsibility. Additionally, universities must clearly define acceptable and unacceptable uses of AI in academic work to prevent ambiguity and ensure fairness. Educators should remain aware of potential biases in detection tools, especially when evaluating work by ESL students or those with unconventional writing styles, to avoid unjust penalties and preserve student trust.

CONFLICTS OF INTEREST

The author declared no conflicts of interest.

Appendix A: Python Implementation Code Snippets

Step 1: Collecting Text Samples

```
# Sample entry
data = {
"text": "Artificial Intelligence is reshaping industries through automation and data-driven decisions...",
"label": "AI"
}
```

Step 2: NLP Preprocessing

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
import string
def preprocess(text):
tokens = word_tokenize(text.lower())
tokens = [t for t in tokens if t not in stopwords.words('english') and t not in string.punctuation]
lemmatizer = WordNetLemmatizer()
return [lemmatizer.lemmatize(t) for t in tokens]
```

Step 3a: Perplexity-Based Analysis

```
from transformers import GPT2LMHeadModel, GPT2Tokenizer
import torch
def calculate_perplexity(text):
encodings = tokenizer(text, return_tensors='pt')
with torch.no_grad():
outputs = model(**encodings, labels=encodings["input_ids"])
loss = outputs.loss
return torch.exp(loss).item()
```

Step 3b: Stylometric Feature Extraction

- Average sentence length
- Lexical richness (unique words / total words)
- Passive voice usage (via dependency parsing)

Step 3c: AI Content Detectors

- Used HuggingFace pretrained models (e.g., RoBERTa)
- Optionally integrated OpenAI API (if available)

Step 4: Output Visualization

```
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay

cm = confusion_matrix(true_labels, predictions)

ConfusionMatrixDisplay(cm).plot()
```

REFERENCE

[1] Ali, W., Alami, R., Alsmairat, M. A., & AlMasaeid, T. (2024, February). Consensus or Controversy: Examining AI's Impact on Academic Integrity, Student Learning, and Inclusivity Within Higher Education Environments. In 2024 2nd International Conference on Cyber Resilience (ICCR) (pp. 01-05). IEEE. https://doi.org/10.1109/ICCR61006.2024.10532968

[2] Alshamsi, I., Sadriwala, K. F., Alazzawi, F. J. I., & Shannaq, B. (2024). Exploring the impact of generative AI technologies on education: Academic expert perspectives, trends, and implications for sustainable development goals. Journal of Infrastructure, Policy and Development, 8(11), 8532. https://doi.org/10.24294/jipd.v8i11.8532

[3] Al-Zahrani, A. M. (2024). The impact of generative AI tools on researchers and research: Implications for academia in higher education. Innovations in Education and Teaching International, 61(5), 1029-1043. https://doi.org/10.1080/14703297.2023.2271445

[4] Bittle, K., & El-Gayar, O. (2025). Generative AI and academic integrity in higher education: A systematic review and research agenda. Information, 16(4), 296. https://doi.org/10.3390/info16040296

[5] Bobula, M. (2023). Generative Artificial Intelligence (AI) in Higher Education: A Comprehensive Review of Opportunities, Challenges and Implications. https://doi.org/10.47408/jldhe.vi30.1137

[6] Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. International Journal of Educational Technology in Higher Education, 20(1), 43. https://doi.org/10.1186/s41239-023-00411-8

[7] Evangelista, E. D. L. (2025). Ensuring academic integrity in the age of ChatGPT: Rethinking exam design, assessment strategies, and ethical AI policies in higher education. Contemporary Educational Technology, 17(1), ep559. https://doi.org/10.30935/cedtech/15775

[8] Fowler, D. S. (2023). AI in higher education: academic integrity, harmony of insights, and recommendations. Journal of Ethics in Higher Education, (3), 127-143. https://doi.org/10.26034/fr.jehe.2023.4657

[9] Gonsalves, C. (2025). Addressing student non-compliance in AI use declarations: implications for academic integrity and assessment in higher education. Assessment & Evaluation in Higher Education, 50(4), 592-606. https://doi.org/10.1080/02602938.2024.2415654

[10] Huang, Q., Lv, C., Lu, L., & Tu, S. (2025). Evaluating the quality of AI-generated digital educational resources for university teaching and learning. Systems, 13(3), 174. https://doi.org/10.3390/systems13030174

[11] Lund, B. D., Lee, T. H., Mannuru, N. R., & Arutla, N. (2025). AI and academic integrity: Exploring student perceptions and implications for higher education. Journal of Academic Ethics, 1-21. https://doi.org/10.1007/s10805-025-09613-3

[12] Mhlanga, D. (2023). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. In FinTech and artificial intelligence for sustainable development: The role of smart technologies in achieving development goals (pp. 387-409). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-37776-1_17

[13] Mortlock, R., & Lucas, C. (2024). Generative artificial intelligence (Gen-AI) in pharmacy education: Utilization and implications for academic integrity: A scoping review. Exploratory Research in Clinical and Social Pharmacy, 100481. https://doi.org/10.1016/j.rcsop.2024.100481

[14] Mulenga, R., & Shilongo, H. (2024). Academic integrity in higher education: Understanding and addressing plagiarism. Acta Pedagogia Asiana, 3(1), 30-43. https://doi.org/10.53623/apga.v3i1.337

[15] Najjar, A. A., Ashqar, H. I., Darwish, O. A., & Hammad, E. (2025). Detecting AI-Generated Text in Educational Content: Leveraging Machine Learning and Explainable AI for Academic Integrity. arXiv preprint arXiv:2501.03203.

[16] Rafiq, S., & Qurat-ul-Ain, D. A. A. (2025). The Role of Ai Detection Tools in Upholding Academic Integrity: An Evaluation of their Effectiveness. Contemporary Journal of Social Science Review, 3(1), 901-915.

[17] Saqib, M. B., & Zia, S. (2024). Evaluation of AI content generation tools for verification of academic integrity in higher education. Journal of Applied Research in Higher Education. https://doi.org/10.1108/JARHE-10-2023-0470

[18] Schneider, S., & Haried, P. (2024). Use of Artificial Intelligence in Higher Education with Particular Reference to Automated Content Generation, Trust and Anxiety. Journal of Business & Educational Leadership, 14(1).

[19] Sevnarayan, K., & Potter, M. A. (2024). Generative Artificial Intelligence in distance education: Transformations, challenges, and impact on academic integrity and student voice. Journal of Applied Learning and Teaching, 7(1). https://doi.org/10.37074/jalt.2024.7.1.41

[20] Sozon, M., Sia, B. C., Pok, W. F., & Alkharabsheh, O. H. M. (2024). Academic integrity violations in higher education: a systematic literature review from Journal of Applied Research in Higher Education. 2013-2023. https://doi.org/10.1108/JARHE-12-2023-0559

[21] Tang, C. M., Ng, V. S., Leung, H. M., & Yuen, J. C. (2023). AI-Generated Programming Solutions: Impacts on Academic Integrity and Good Practices. https://doi.org/10.5220/0012563600003693

[22] Wiredu, J. K., Seidu Abuba, N., & Zakaria, H. (2024). Impact of generative AI in academic integrity and learning outcomes: a case study in the upper east region. Asian Journal of Research in Computer Science, 17(8), 10-9734. https://doi.org/10.9734/ajrcos/2024/v17i7491

https://doi.org/10.65638/2978-5634.2025.01.04