# Responsible AI Use in Nigerian Secondary-School Essay Preparation: A Framework and Preliminary Study of AI Detection Risks in WAEC/NECO-Style Essays

**Shaibu Abdullateef Topa**

**Department of Computer Science, Faculty of Computing, Bayero University, Kano, Nigeria**

**DOI: https://doi.org/10.47772/IJRISS.2026.100500277**

**Received: 01 May 2026; Accepted: 07 May 2026; Published: 29 May 2026**

## ABSTRACT

Generative AI tools are increasingly used by students for writing support, including English essay preparation. In Nigeria, this raises important questions for WAEC/NECO-style examinations, where essay writing remains a major component of assessment. While AI can support learning through outlining, grammar feedback, and vocabulary development, it also creates academic-integrity concerns. At the same time, AI-text detection approaches may be unreliable or unfair, especially for non-native English writers and hybrid human-AI writing. This paper proposes a responsible AI-use framework for Nigerian secondary-school essay preparation and presents a preliminary baseline study of AI-detection risks using WAEC/NECO-style essay prompts.

**Keywords:** Academic Integrity, Detection, Examinations, Generative AI, Responsible AI Framework

## INTRODUCTION

Nigeria operates one of the largest education systems in Africa, structured around a 6-3-3-4 policy framework, which denotes six years of primary, three years of junior secondary, three years of senior secondary, and four years of tertiary education, as established by the National Policy on Education. Secondary education is regulated by the Federal Ministry of Education alongside examination bodies such as the West African Examinations Council (WAEC) and the National Examinations Council (NECO), which set curriculum standards and conduct terminal assessments that determine students' eligibility for higher education. Despite its enormousness, the Nigerian secondary school system is characterized by inadequate facilities, poor staffing, and inequitable access between urban and rural areas. Such structural challenges define the landscape in which teachers and students interact with new educational technologies, including AI-driven writing tools.

English Language holds a uniquely central position in Nigerian secondary education, serving simultaneously as the medium of instruction and as a compulsory examination subject at both the WASSCE and NECO SSCE. A credit pass in English Language is a minimum entry requirement for admission into all Nigerian universities and tertiary institutions. Essay writing forms a core component of these examinations, requiring candidates to produce structured continuous writing across argumentative, expository, narrative, and descriptive genres under timed conditions. Many Nigerian students, for whom English is a second or additional language, struggle with coherence, formal register, and paragraph development, which makes essay writing one of the most demanding pedagogical tasks (Ogunmodimu, 2015). In this examination-driven environment, AI writing tools have gained significant attention from both students and educators.

Artificial intelligence (AI) refers to the capacity of computer systems to simulate human cognitive functions such as reasoning, language comprehension, and problem-solving (Russell & Norvig, 2021). In educational settings, AI applications now encompass intelligent tutoring systems, automated essay scoring, adaptive learning platforms, and natural language processing tools designed to support literacy development (Holmes et al., 2019). These technologies have rapidly transitioned from experimental to mainstream adoption, creating significant challenges for school systems, especially in developing contexts, as they attempt to address technologies that advance more quickly than existing policy frameworks. In Nigerian secondary schools, the central issue is no longer whether students will encounter AI tools, but rather how institutions will effectively guide their use.

Generative AI encompasses a class of artificial intelligence systems designed to produce original content, including text, images, and code, in response to user prompts (Sengar et al., 2025). Tools such as ChatGPT, Google Gemini, and Microsoft Copilot, which are powered by large language models (LLMs), can generate fluent and contextually appropriate essays within seconds. These tools are widely accessible via smartphones, which serve as the primary internet device for most Nigerian students. The rapid adoption of generative AI among secondary-school students worldwide, including those in sub-Saharan Africa, has surpassed institutional preparedness. As a result, teachers and school administrators often lack clear policies or pedagogical frameworks to manage the use of these technologies in writing instruction and assessment.

Generative AI tools present significant opportunities for secondary-school writing instruction. These technologies can provide immediate feedback on grammar, vocabulary, and structure, forms of support that many teachers in large Nigerian classrooms are unable to deliver consistently at scale (Kohnke et al., 2023). For students who write in English as a second language, AI tools may also function as accessible scaffolds for mastering formal register and essay organization (Warschauer et al., 2023). Despite these advantages, substantial risks are evident. The most prominent is academic dishonesty, specifically students submitting AI-generated essays as their own, which directly undermines the integrity of WAEC and NECO assessments (Cotton et al., 2023). Another significant concern is cognitive dependency, in which habitual AI use reduces students' capacity to independently plan, draft, and revise their writing. In a high-stakes examination system where written performance determines access to higher education, both risks have considerable consequences.

In response to AI-assisted dishonesty, many institutions have explored automated AI-writing detection systems to identify machine-generated writing. However, recent research raises concerns regarding their reliability (Turnitin, 2023). Detection accuracy is inconsistent, can be circumvented through paraphrasing, and is notably biased against non-native English writers, whose prose patterns are more likely to be misclassified as AI-generated (Liang et al., 2023). This situation presents a significant fairness risk for Nigerian students, whose English writing may be unjustly flagged. In this context, the present paper offers two contributions. First, it proposes a responsible AI-use framework tailored to Nigerian secondary-school essay preparation, balancing pedagogical benefits with academic integrity. Second, it presents a preliminary empirical study examining whether WAEC/NECO-style student essays are disproportionately misidentified by commonly used AI detection tools or baseline AI-text classification approaches, with the goal of informing fair, evidence-based institutional policy.

## LITERATURE REVIEW

### Generative AI and Student Writing

Scholarly attention to the impact of generative AI on student writing has grown substantially since the public release of large language model (LLM)-based tools in late 2022. The literature reveals a major tension: AI writing tools provide powerful support for students, but educational systems remain uncertain about how to govern their use.

Kohnke, Moorhouse, and Zou (2023) identify this tension in classroom instruction. They argue that AI tools can offer immediate and personalized feedback on grammar, vocabulary, and structure. This is especially important in large or under-resourced classrooms where teachers may not be able to provide detailed feedback to every student.

Warschauer et al. (2023) extend this argument by showing that LLM-based tools can support ESL writers by modelling formal language, coherent paragraph organization, and field-specific vocabulary. These functions align with established research on the importance of feedback in writing development.

However, the literature also recognizes serious risks. Cotton et al. (2023) argue that AI-assisted cheating differs from traditional plagiarism because it does not involve copying an existing source. Instead, it may replace the student’s own cognitive work with machine-generated content.

This concern goes beyond academic dishonesty. Holmes, Bialik, and Fadel (2019) warn that overreliance on AI tools may encourage students to become passive consumers of machine-generated text. Such dependency could weaken students' ability to plan, draft, and revise writing independently.

Institutional responses to AI-assisted writing have also attracted scholarly criticism. Luo (2024) finds that many universities policies frame AI mainly as a threat requiring prohibition or punishment. Similarly, (Weng et al., 2024) show that assessment responses often focus on avoiding AI misuse rather than teaching responsible AI use.

This literature suggests that educational institutions need a balanced approach. Rather than treating all AI use as misconduct, schools should distinguish between AI use that supports learning and AI use that replaces student authorship. This distinction is central to the present study's focus on Nigerian secondary-school essay preparation.

**AI Detection and Academic Integrity**

The rapid adoption of AI detection tools has surpassed the availability of empirical evidence validating their reliability. Tools such as Turnitin's AI detector, GPTZero, ZeroGPT, and DetectGPT are now frequently integrated into academic integrity workflows. Nevertheless, existing literature raises significant concerns regarding the ability of these tools to reliably determine student authorship.

Mitchell et al. (2023) state that many detection approaches rely on statistical properties of text, such as perplexity and probability curvature. These methods aim to determine whether a text exhibits distributional patterns characteristic of machine-generated writing.

However, these statistical signals are unstable. Sadasivan et al., (2023) demonstrate that detection performance varies across large language model (LLM) versions and can be compromised by paraphrasing. Consequently, a text may evade detection not due to human authorship, but as a result of being rewritten or lightly modified.

The vulnerability of AI detectors to paraphrasing has been further demonstrated by Krishna et al. (2023). Their work suggests that even modest rewording can reduce detection scores below institutional thresholds. This has serious implications for schools that may rely on detector outputs as evidence of misconduct.

Other studies also show inconsistency across detection tools. San Miguel et al. (2025), for example, found that different detectors produced conflicting judgments on the same student essays. Such inconsistency weakens the evidential value of AI detection in academic integrity cases.

Consequently, several scholars contend that AI detection should serve only as one indicator within a comprehensive review process. Peterson's (2025) Academic Integrity Reporting framework emphasizes that detection outputs alone cannot establish student intent. This distinction is critical, as academic misconduct requires evidence that a student knowingly misrepresented authorship, not merely the presence of suspicious writing patterns.

In summary, the literature indicates that detection-only approaches are insufficient. Effective AI governance should integrate technical screening with human review, student explanations, analysis of writing history, and well-defined institutional policies.

**Bias Against Non-Native English Writers**

The fairness implications of AI detection have received increasing attention, especially in relation to writers for whom English is a second or additional language. This issue is directly relevant to Nigerian secondary-school students, many of whom write English as an additional language.

Liang et al. (2023) provide one of the most important empirical demonstrations of this problem. They found that GPT detectors assigned higher AI-probability scores to essays written by non-native English writers, even when those essays were confirmed to be human-written.

The proposed explanation is that many detectors associate predictable language with AI-generated text. However, second-language writing may also be more predictable because it often uses simpler vocabulary, more regular sentence structures, and less syntactic variation. This creates a risk that authentic non-native writing may be misread as machine-generated.

Khalil and Er (2023) support this concern by showing that false identification rates may increase when writing differs from the standard American or British English patterns on which many tools are likely trained. This is significant for Nigerian students because Nigerian English contains recognized regional features that may differ from those dominant varieties.

San Miguel et al. (2025) report a similar pattern in a Philippine tertiary context. Students who wrote in simpler or more predictable English structures were more likely to be flagged by detection tools, regardless of whether AI assistance was actually used.

The institutional consequences of such bias are substantial. False accusations of AI use can harm a student's reputation and may result in disciplinary action. In high-stakes examination systems, such as WAEC and NECO, these outcomes could impede access to higher education.

Therefore, AI detection systems should be applied with caution in English as a Second Language (ESL) contexts. Detection outputs should not serve as automatic evidence of misconduct, particularly when the student population writes in a recognized non-native variety of English.

**Nigerian English and WAEC/NECO Essay Writing**

Nigerian English occupies a distinctive position within the World Englishes framework. It is widely recognized as a nativized variety of English with its own phonological, lexical, syntactic, and discourse features. These features distinguish it from the standard British English norms often used in formal assessment.

This situation generates a significant pedagogical challenge. Nigerian secondary-school students are typically required to write in formal standard English, although their daily linguistic environment includes Nigerian English, indigenous languages, and multilingual code-switching.

Ogunmodimu (2015) identifies several recurring challenges in Nigerian student writing, such as coherence management, use of formal register, paragraph development, and clarity of expression. These challenges do not indicate a lack of writing ability. Instead, they reflect the reality that students compose texts within a complex multilingual environment while being evaluated according to formal examination standards.

This linguistic context has direct implications for AI-generated text detection. Characteristics such as simplified syntax, predictable essay introductions, restricted lexical range, and formulaic organization are prevalent in Nigerian examination writing. These features closely resemble those that AI detectors often associate with machine-generated content.

The associated risk is therefore not purely theoretical. Nigerian students may produce original essays that superficially resemble AI-generated or AI-edited writing. This resemblance often results from examination preparation practices that emphasize memorized structures, standardized introductions, and conventional transitional phrases.

The structure of Nigerian secondary education is also a significant factor. Large class sizes, insufficient access to qualified English teachers, and a predominant focus on summative assessment result in limited individualized feedback for students. In these circumstances, AI writing tools may be appealing due to their provision of immediate grammar correction, vocabulary enhancement, and structural guidance.

This situation presents a policy dilemma. Treating all AI use as academic dishonesty may deprive students of legitimate writing support, whereas permitting unrestricted AI use could compromise the validity of essay

assessments. An effective policy framework must therefore differentiate between AI as a tool for learning support and AI as a replacement for student authorship.

The response of the West African Examinations Council (WAEC) to AI use during the 2023 West African Senior School Certificate Examination (WASSCE) cycle underscores the urgency of this issue. According to John Kapi (2023), some candidates were identified after copying AI-generated refusal messages verbatim into their answer booklets. This type of evidence is compelling because it constitutes direct behavioral proof of AI-assisted copying.

However, such cases represent only the most obvious form of AI misuse. They do not address more subtle cases, such as AI-assisted revision or essays that resemble AI output because of formulaic examination writing. This limitation reinforces the need for careful, context-sensitive AI governance in Nigerian secondary education.

**Research Gap**

The literature reviewed above shows that research on generative AI, student writing, and AI detection is expanding rapidly. However, important geographical and institutional gaps remain.

Most existing empirical research focuses on higher education settings in North America, Europe, East Asia, and Southeast Asia. Comparatively little attention has been given to secondary education in Sub-Saharan Africa, despite the region's linguistic diversity and large student population.

This gap is significant for Nigeria. WAEC and NECO examinations are high-stakes assessments, and English essay writing plays an important role in determining access to higher education. Yet there appears to be limited empirical work examining how AI detection methods perform on WAEC/NECO-style essays.

A related gap concerns Nigerian English. Existing research shows that AI detectors may misclassify non-native English writing, but little work has examined how this risk applies to Nigerian English and examination-oriented student writing.

There is also a policy gap. Existing AI governance frameworks are often designed for universities and assume access to formal academic integrity offices, trained investigators, and established appeals procedures. These assumptions may not hold in Nigerian secondary schools.

The present study addresses these gaps by combining a preliminary empirical evaluation with a responsible AI-use framework. It compares human-style, AI-generated, and AI-edited WAEC/NECO-style essays using baseline text-classification methods and uses the findings to propose a context-sensitive framework for Nigerian secondary-school essay preparation.

## METHODOLOGY

**Research Design**

This study adopts a mixed conceptual and empirical research design. The conceptual component proposes a responsible AI-use framework tailored to Nigerian secondary-school essay preparation, while the empirical component evaluates the reliability and fairness risks of AI-generated text detection applied to WAEC/NECO-style essays. The study is exploratory and preliminary in scope, aiming not to develop a definitive detection system, but to examine whether common detection approaches can distinguish among human-style, AI-generated, and AI-edited essays in a Nigerian English examination context. This design is appropriate because the study simultaneously addresses a policy question and a technical fairness question, both of which are best served by combining conceptual framework development with preliminary empirical evidence.

The study is guided by three research questions:

1. How reliably can AI-generated text detection methods distinguish human-style, AI-generated, and AI-edited WAEC/NECO-style essays?

2. Are human-style Nigerian English essays at elevated risk of being falsely classified as AI-generated by current detection tools?

3. What responsible AI-use framework can guide students, teachers, and schools in Nigerian secondary-school essay preparation?

The first two research questions are addressed through the empirical component of the study, which uses a constructed dataset of WAEC/NECO-style essays and transparent machine-learning baseline classifiers. The third research question is addressed through the conceptual component, which synthesizes findings from the literature review and empirical results into an actionable framework. The integrated design enables the empirical findings to both test and inform the framework, making the overall contribution stronger than either component would be in isolation.

**Dataset Creation**

A small experimental dataset was constructed using WAEC/NECO-style English essay prompts. The prompts were selected to reflect the principal essay genres assessed in Nigerian senior secondary school examinations, specifically informal letters, formal letters, argumentative essays, narrative essays, descriptive essays, speeches, articles, and expository essays. These genres were chosen because they collectively represent the full range of essay types that appear in the WAEC English Language paper, and because they exhibit meaningful differences in register, discourse structure, and stylistic expectation, which makes them a useful test of whether detection approaches generalise across essay types or perform inconsistently across genres.

For each prompt, three versions of essay responses were produced, corresponding to the three analytical categories of the study. Human-style essays were manually written to imitate the linguistic features characteristic of Nigerian secondary-school English writing, including straightforward vocabulary, examination-oriented organisational patterns, and expression features associated with second-language writing in a high-stakes assessment context. AI-generated essays were produced by submitting the same prompts to a generative AI tool and recording its complete output without human editing. AI-edited essays were created by taking the human-style essays and using AI to improve grammar, clarity, and fluency while explicitly instructing the model to preserve the original ideas and structural organisation of the source essay.

Each entry in the dataset contains the following fields: essay ID, prompt, essay type, source label, word count, essay text, detector result, model prediction, and researcher notes. The source label field identifies the category to which each essay belongs and serves as the ground-truth label for classification evaluation. The result and model prediction fields are populated during the detection and classification stage. Further details on prompt selection, essay construction, AI-generation prompts, AI-editing prompts, and matched-triplet cleaning are provided in Appendices A–F.

**Table 3.1. Dataset Composition by Category**

| Category | Target Size | Source |
|---|---|---|
| Human-style | 70 essays | Manually composed to imitate Nigerian secondary-school essay writing |
| AI-generated | 70 essays | Produced by generative AI from WAEC/NECO-style prompts |
| AI-edited | 70 essays | Human-style essays revised by AI for grammar, fluency, and clarity |
| Total | 210 essays | — |

An important methodological limitation must be acknowledged at this point. Because the human-style essays are manually produced rather than collected from actual examination candidates, the dataset should be treated as a preliminary simulation rather than a representative sample of the population of WAEC or NECO examination essays. The findings produced from this dataset are intended to provide indicative rather than conclusive evidence regarding detection risks. Future studies should validate and extend these findings using

real student essays collected under appropriate ethical approval and with the informed consent of students, guardians, and relevant examination authorities.

**Essay Categories**

The dataset is divided into three analytically distinct categories, each representing a different mode of student-AI interaction. These categories were operationalised as follows.

**Human-style:** The human-style category represents essays written to resemble authentic Nigerian secondary-school student examination responses. These essays are characterised by features that the literature identifies as typical of second-language examination writing in the Nigerian context, including relatively straightforward vocabulary choices, predictable essay openings, conventional paragraph structures, and occasional expression patterns that reflect L1 transfer or non-standard usage. This category is used principally to test the false-positive risk: the extent to which AI detection tools or classifiers erroneously flag authentic student writing as machine-generated.

**AI-generated:** The AI-generated category consists of essays produced entirely by a generative AI tool in response to the essay prompt, without any subsequent human editing. These essays typically exhibit more polished grammatical accuracy, smoother discourse transitions, more consistent formal register, and a broader vocabulary range than the human-style essays. This category tests the capacity of detection approaches to identify unambiguously machine-generated content and provides the clearest contrast with the human-style category for classification purposes.

**AI-edited:** The AI-edited category consists of essays created by submitting human-style essays to a generative AI tool with an instruction to correct grammar, improve clarity, and enhance fluency while preserving the original content and organisational structure. This category is analytically significant because it models the most likely pattern of student AI use in practice. As Cotton et al. (2023) observe, students who use AI for academic writing rarely submit wholly AI-generated text; more commonly, they use AI as a revision and correction tool. The AI-edited essays therefore represent hybrid human-AI writing, which is expected to be the most challenging category for detection approaches to classify correctly.

**Table 3.2. Essay Category Definitions and Analytical Purpose**

| Category | Description | Analytical Purpose |
|---|---|---|
| **Human-style** | Manually written to resemble Nigerian secondary-school examination essays | Tests false-positive misclassification risk |
| **AI-generated** | Fully produced by generative AI from the essay prompt | Tests detector sensitivity to machine-generated content |
| **AI-edited** | Human-style essay revised by AI for grammar and fluency | Tests detection reliability for hybrid human-AI writing |

The inclusion of the AI-edited category distinguishes this study from much of the existing detection literature, which tends to evaluate binary human-versus-AI classification tasks. In the Nigerian secondary-school context, where AI use is most likely to take the form of grammar assistance, vocabulary improvement, and structural revision, a binary classification framing does not adequately capture the actual risk landscape that schools and examination bodies face. The three-category design therefore provides a more ecologically valid evaluation of detection reliability.

**Detection and Classification Method**

The detection and classification stage was carried out using transparent machine-learning baseline classifiers developed from textual features. The purpose was to examine whether human-style, AI-generated, and AI-edited essays could be distinguished using simple, reproducible text-classification methods.

**Machine-learning baseline:** A simple machine-learning baseline was developed to assess whether basic textual features can separate the three essay categories with reasonable accuracy. The baseline uses Term Frequency-Inverse Document Frequency (TF-IDF) vectorisation to represent essay text as numerical feature vectors. These vectors were used to train a Logistic Regression classifier, with a Linear Support Vector Machine (SVM) evaluated as an alternative model for comparison. The purpose of this baseline is not to produce a state-of-the-art detection system, but to examine whether the three essay categories are linguistically distinguishable using simple, transparent, and reproducible methods.

The classification task was framed as a three-class problem: human-style, AI-generated, and AI-edited. The dataset was partitioned into training and testing sets using an 80:20 split. The classifier was trained on the training partition and evaluated on the held-out test partition. Given the small size of the dataset, results from the train-test split are supplemented by five-fold cross-validation to produce more stable estimates of generalisation performance. All analysis was conducted in Python using the scikit-learn library, with pandas used for dataset management.

The results are interpreted in relation to the broader literature on AI-text detection reliability documented in this section. Given the exploratory and preliminary character of the study, the findings are presented as indicative evidence rather than definitive conclusions regarding the capacity of baseline classifiers to identify AI-assisted writing in Nigerian secondary-school essay contexts.

**Table 3.3: Machine-Learning Baseline: Technical Specifications**

| Component | Specification |
|---|---|
| Text representation | TF-IDF vectorisation (unigrams and bigrams) |
| Primary classifier | Logistic Regression (C = 1.0, max_iter = 1000) |
| Alternative classifier | Support Vector Machine (linear kernel) |
| Classification task | Three-class: Human-style, AI-generated, AI-edited |
| Data split | 80% training, 20% testing |
| Validation strategy | Five-fold cross-validation |
| Programming language | Python (scikit-learn, pandas) |

Given the exploratory and preliminary character of the study, the findings are presented as indicative evidence rather than as definitive conclusions regarding the capacity of any particular detection tool or classifier to identify AI-assisted writing in Nigerian secondary-school essay contexts.

**Evaluation Metrics**

The performance of the detection and classification approaches was evaluated using a standard set of classification metrics: accuracy, precision, recall, F1-score, and confusion matrix analysis. In addition, the false positive rate for the human-style category was treated as a key fairness-oriented evaluation measure. The rationale for this emphasis is grounded in the ethical stakes of the study: in a high-stakes examination context where a false detection finding may result in disqualification or other serious consequences, the cost of incorrectly classifying an authentic student essay as AI-generated is substantially higher than the cost of failing to detect an AI-generated essay. A detection approach that achieves high overall accuracy by correctly classifying AI-generated essays while also generating frequent false positives for human-style essays is not acceptable from a fairness standpoint, even if its aggregate performance metrics appear satisfactory.

The metrics are defined as follows:

1. Accuracy measures the proportion of all essays correctly classified across all three categories. It provides an overall summary of classifier performance but may mask category-level imbalances in a small dataset.

2. Precision measures, for a given predicted category, what proportion of essays assigned to that category actually belong to it. High precision for the AI-generated category means that essays predicted as AI-generated are indeed AI-generated.
3. Recall measures, for a given true category, what proportion of essays belonging to that category were correctly identified. High recall for the human-style category means that authentic essays are not frequently missed and misclassified elsewhere.
4. F1-score is the harmonic mean of precision and recall for a given category. It provides a balanced measure that is informative when precision and recall are in tension, which is expected for the AI-edited category.
5. Confusion matrix analysis identifies the specific types of classification errors made by the classifier and the AI-detection tools, including which category pairs are most frequently confused. This analysis is particularly important for understanding whether human-style essays are misclassified as AI-generated, or whether AI-edited essays are misclassified as human-style.
6. False positive rate for the human-style category measures the proportion of human-style essays incorrectly classified as AI-generated. This metric is given priority in the interpretation of results because of its direct relevance to the fairness argument developed in the previous sections.

**Table 3.4. Evaluation Metrics and Their Role in the Study**

| Metric | Definition | Role in This Study |
|---|---|---|
| Accuracy | Proportion of all essays correctly classified | Overall performance summary |
| Precision | Proportion of predicted-positive essays that are true positives | Measures prediction quality per category |
| Recall | Proportion of true-positive essays correctly identified | Measures coverage per category |
| F1-score | Harmonic mean of precision and recall | Balanced per-category performance measure |
| Confusion matrix | Cross-tabulation of predicted versus actual categories | Identifies specific misclassification patterns |
| False positive rate (human-style) | Proportion of human-style essays classified as AI-generated | Primary fairness metric |

Note. False positive rate for the human-style category is treated as the primary fairness metric because false accusations of AI use in high-stakes examination contexts may carry serious academic consequences for students.

The results of the evaluation are interpreted in relation to the broader goal of responsible AI governance in Nigerian secondary-school essay assessment. Rather than treating detection scores as determinative evidence of academic dishonesty, the study examines whether the performance and error profiles of current detection approaches are sufficiently reliable and fair to justify their use in a high-stakes Nigerian examination context. A detection approach that generates frequent false positives for human-style Nigerian English essays would, on the basis of this evaluation, fail to meet the minimum threshold of fairness required for use as punitive evidence under any institutional academic integrity framework.

**Proposed Responsible AI Use Framework**

The empirical findings of this study, alongside an extensive review of the literature on generative AI in education, non-native English writing, and academic integrity, motivate the development of a structured policy instrument for Nigerian secondary schools. Therefore, this section presents the RAISE Framework, an evidenced five-

component model that offers a guide concerning the ethical and practical challenges of AI-assisted essay writing in the context of WAEC/NECO assessment for students, teachers, and school administrators.

**Rationale for a Framework Approach**

Existing literature consistently demonstrates that binary approaches to AI in education, such as blanket prohibition or unrestricted permissiveness, are both ineffective and inequitable (UNESCO, 2023). Prohibition is ineffective because it is technically unenforceable and restricts students' access to legitimate learning tools. Conversely, permissiveness undermines assessment validity and disadvantages students who do not utilize AI assistance. A framework approach is preferable, as it recognizes the spectrum of AI involvement in writing, ranging from minimal support to full substitution, and provides normative guidance for each level.

In the Nigerian context, this gradient is especially relevant. English is used as a second or additional language by the majority of WAEC/NECO candidates (Jekayinfa & Aburime, 2021), and many students legitimately benefit from grammar-correction tools, vocabulary aids, and structured feedback. A framework that equates all AI assistance with academic dishonesty would deny these educational benefits and may increase the risk of false-positive accusations.

**The RAISE Framework**

The RAISE Framework is organised around five core principles, each addressing a distinct dimension of responsible AI use in essay preparation. The acronym is intended as a mnemonic for students, teachers, and policymakers:

**Table 4.1. The RAISE Framework**

| Code | Principle | Description | Status |
|---|---|---|---|
| R | Role Clarity | AI tools should serve as a learning support mechanism and never as a replacement for the student's own cognitive engagement. | Acceptable |
| A | Academic Honesty | Students must not submit AI-generated content as their own original work. Disclosure of AI assistance is required in all assessed tasks. | Mandatory |
| I | Improvement Focus | AI use is most ethical when directed at improvement activities: grammar feedback, vocabulary enrichment, structural revision, and coherence checking. | Acceptable with disclosure |
| S | Student Ownership | The final essay must authentically represent the student's own ideas, argument, and voice. AI refinement of structure is permissible; AI substitution of thought is not. | Conditional |
| E | Ethical Assessment | Schools and examination bodies must not rely solely on automated AI detection tools for disciplinary decisions. Detection output should serve as a low-confidence alert requiring human review. | Mandatory |

Each principle is elaborated below.

**R — Role Clarity**

AI tools must be positioned as supporting instruments rather than as authorial agents. In pedagogical terms, this principle corresponds to Vygotsky's zone of proximal development: AI assistance is most educationally beneficial when it scaffolds a student's existing capacity rather than replacing the cognitive effort that develops writing competence. Teachers should make explicit to students that AI-generated content does not belong to the student's intellectual record and that submitting it as such is a form of academic misrepresentation.

**A — Academic Honesty**

Students who submit fully or substantially AI-generated essays as their own original work violate the integrity

of the assessment process. This principle aligns with the WAEC/NECO examination regulations that prohibit examination malpractice (West African Examinations Council, 2024). Where schools permit AI use for revision or feedback, a disclosure requirement should accompany submission, analogous to existing norms for citing secondary sources, so that assessors can contextualise the work appropriately.

**I — Improvement Focus**

AI use is most ethically defensible when it is directed at improvement rather than generation. Asking an AI tool to correct grammar, improve sentence clarity, or suggest more precise vocabulary in a student-written draft constitutes a form of automated tutoring that is broadly comparable to teacher feedback. This kind of AI assistance actively develops student writing competence and should be distinguished in policy from AI use that generates the substantive content of the essay.

**S — Student Ownership**

Regardless of the degree of AI assistance involved in editing, the argumentative content, central ideas, narrative perspective, and evaluative judgements in the final essay must originate with the student. This principle is especially significant for WAEC essay genres such as narrative writing, argumentative essays, and personal letters, where the examinable skill is precisely the student's ability to generate and organise their own ideas in standard English. AI polishing of student-authored text is permissible within the limits of this principle; AI replacement of student-authored text is not.

**E — Ethical Assessment**

Schools and examination bodies must recognise that AI detection tools do not constitute reliable evidence of academic misconduct. As the results of this study demonstrate, both TF-IDF-based models tested in this preliminary study produce a broad AI-assisted flag rate of 64.3%, meaning that approximately two thirds of authentic student essays would be incorrectly identified as AI-assisted. Using detection output as sufficient grounds for disciplinary action against students is therefore ethically indefensible. Detection tools may serve as a low-confidence alert that triggers human review, but the review process must include oral examination, consultation of the student's writing history, and opportunity for the student to respond.

**AI Use Levels**

The RAISE Framework is operationalised through four discrete AI use levels that map different modes of AI involvement in essay writing to corresponding ethical designations. These levels provide students with concrete guidance about acceptable and unacceptable behaviour and give teachers a vocabulary for discussing AI use in the classroom.

**Table 4.2. AI Use Levels for Secondary-School Essay Preparation**

| Level | Category | Example | Ethical Status |
|---|---|---|---|
| Level 1 | Practice Support | Generating essay topics, WAEC/NECO-style prompts, outlines, and vocabulary lists. | Acceptable |
| Level 2 | Feedback Support | Grammar correction, structural feedback, coherence and style review of a student-authored draft. | Acceptable with disclosure |
| Level 3 | Heavy Rewriting | AI substantially rewrites most of the student's original draft, altering ideas and voice. | Ethically risky |
| Level 4 | Full Substitution | Student submits a fully AI-generated essay as their own original examination work. | Unacceptable |

Levels 1 and 2 are endorsed as legitimate uses of AI in learning. Level 3 requires disclosure and teacher consultation. Level 4 constitutes examination malpractice under existing WAEC/NECO regulations and should be treated accordingly, subject to the due-process protections.

### Implications for Nigerian Schools and Examination Bodies

It appears that effective implementation of the RAISE Framework would need to occur at three levels. At the institutional level, schools might consider embedding AI literacy within their information and communication technology curricula, ideally with clear instruction on how Levels 1 through 4 differ in practice. In the classroom, English teachers could arguably benefit from designing essay assessments that require students to document their process, for example through draft submissions, peer feedback, and reflective commentary, in addition to the final essay. This approach may help to reduce the appeal of AI-assisted shortcuts by making the process itself more visible. At the policy level, it seems important for WAEC and NECO to develop formal guidelines on AI use in exam preparation that acknowledge the distinctions outlined in the framework, rather than adopting a blanket approach that equates all forms of AI assistance with examination malpractice.

## RESULTS

This section reports the results of the machine learning classification experiments. Two baseline pipelines, TF-IDF with Logistic Regression and TF-IDF with Linear Support Vector Machine (SVM), were evaluated using a matched-triplet dataset of 198 essays (66 Human-style, 66 AI-generated, and 66 AI-edited). The evaluation metrics comprised test-set accuracy, per-class precision, recall, and F1 scores, five-fold stratified cross-validation scores, and two fairness metrics relevant to the risk of false accusation in educational contexts.

### Dataset Summary

Following the data cleaning procedure, the final dataset comprised 198 essays drawn from 66 matched prompt triplets. Each triplet consisted of exactly one Human-style essay, one AI-generated essay, and one AI-edited essay written in response to the same WAEC/NECO-style prompt. The dataset was stratified into a training set of 158 essays and a test set of 40 essays (approximately 80:20 split), maintaining equal class proportions in both partitions. Eight essay types were represented, with Informal Letter (n = 45), Article (n = 39), and Speech (n = 39) being the most frequent. All essay type frequencies were perfectly balanced across the three source categories, confirming that essay genre is not a confounding variable in the classification task.

An exploratory analysis of word count distributions revealed a significant asymmetry: AI-generated essays demonstrated a substantially narrower interquartile range (approximately 450 to 564 words) compared to Human-style (158 to 908) and AI-edited (162 to 946) essays. This observation suggests that text-length features, rather than exclusively stylistic or vocabulary-based indicators, may influence the detection of AI-generated essays.

### Classification Performance

Table 5.1 presents the overall classification performance of both models on the held-out test set and across five-fold cross-validation.

**Table 5.1. Overall Classification Performance of Baseline Models**

| Model | Test Acc. | CV Acc. (5-fold) | CV Macro F1 | Strict FPR | Broad Flag Rate |
|---|---|---|---|---|---|
| Logistic Regression | 0.550 | 0.621 ± 0.048 | 0.606 ± 0.046 | 0.000 | 0.643 |
| Linear SVM | 0.575 | 0.702 ± 0.043 | 0.689 ± 0.048 | 0.000 | 0.643 |

The Linear SVM outperformed Logistic Regression on all primary performance metrics. Cross-validation accuracy for the SVM was 0.702 (± 0.043) compared to 0.621 (± 0.048) for Logistic Regression: a difference of approximately 8 percentage points. Cross-validation macro F1 scores showed a similar pattern: 0.689 (± 0.048) for the SVM versus 0.606 (± 0.046) for Logistic Regression. Both models exhibited lower test-set accuracy than cross-validation accuracy, a discrepancy attributable to the small test set size (n = 40), which introduces substantial sampling variance.

**Per-Class Performance**

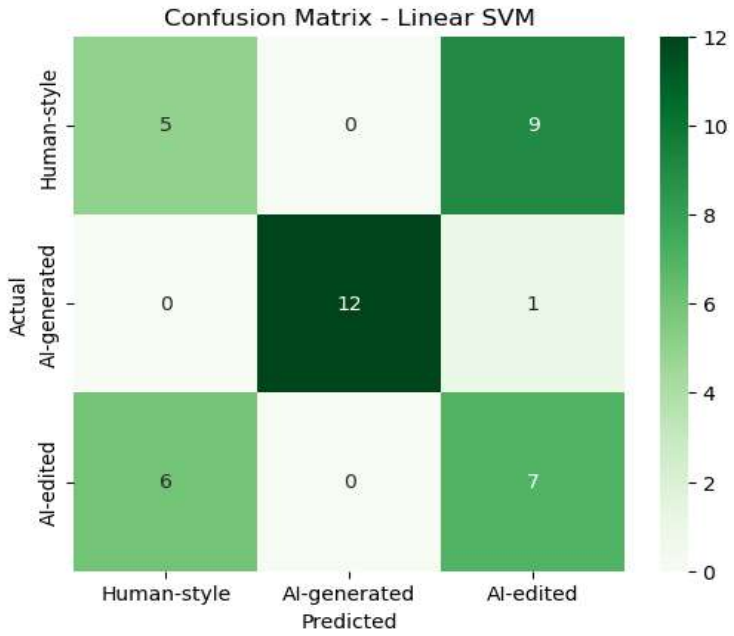
Table 5.2 disaggregates classification performance by source category for the test set.

**Table 5.2. Per-Class F1 Scores on the Test Set (n = 40)**

| Model | Human-style F1 | AI-generated F1 | AI-edited F1 | Macro F1 | Accuracy |
|---|---|---|---|---|---|
| **Logistic Regression** | 0.36 | 1.00 | 0.31 | 0.55 | 0.550 |
| **Linear SVM** | 0.38 | 0.96 | 0.41 | 0.59 | 0.575 |

Both models demonstrated near-perfect performance in detecting AI-generated essays, achieving F1 scores of 1.00 for Logistic Regression and 0.96 for Linear SVM. These findings align with previous research, which indicates that fully synthetic text produced by large language models exhibits strong statistical regularities, such as characteristic vocabulary, syntactic patterns, and consistent length, that TF-IDF features can effectively capture (Jawahar et al., 2020).

**Fig 5.1. Confusion Matrix: Linear SVM**




By contrast, performance on the Human-style and AI-edited categories was substantially weaker. Human-style essays attained F1 scores of 0.36 and 0.38 for Logistic Regression and Linear SVM respectively, while AI-edited essays achieved F1 scores of 0.31 and 0.41. The confusion matrices shown above reveal that the dominant error pattern in both models involves Human-style essays being misclassified as AI-edited: Logistic Regression misclassified 9 of 14 Human-style essays as AI-edited, and the SVM misclassified 9 of 14 in the same direction. Neither model misclassified any Human-style essay as AI-generated.

**Fairness Metrics**

Two fairness metrics were established to evaluate the risk of false accusation within an educational deployment context. The strict false positive rate (Strict FPR) measures the proportion of Human-style essays misclassified specifically as AI-generated, which represents the most serious error type because it suggests that a student submitted entirely machine-written work. The broad AI-assisted flag rate measures the proportion of Human-style essays misclassified as AI-generated or AI-edited, thereby reflecting the risk of AI-related accusations.

Both models achieved a Strict FPR of 0.000, indicating that no Human-style essay was labeled as AI-generated in the test set. This result is significant for contexts where the most serious accusation, namely full AI authorship,

constitutes the primary disciplinary concern. However, the broad AI-assisted flag rate for both models was 0.643, meaning that 9 out of 14 Human-style test essays were predicted as AI-edited. In educational settings where any AI-related flag initiates a formal investigation, this rate suggests that approximately six out of ten genuine student essays would be incorrectly subjected to scrutiny.

# DISCUSSION

The results of this preliminary study yield three interlocking findings that are both technically informative and practically significant for the Nigerian secondary-school context. This section discusses each finding in turn, draws connections to the broader literature, and identifies the limitations that constrain the generalisability of these conclusions.

### AI-Generated Essays Are Reliably Detectable by Baseline Models

A notable finding of this study is the near-perfect F1 performance of both baseline models in identifying AI-generated essays. Despite utilizing a relatively small dataset of 198 essays and a feature representation limited to TF-IDF bigrams, both classifiers achieved F1 scores exceeding 0.96 for fully synthetic essays. This outcome aligns with existing literature on AI text detection, which indicates that large language models generate text with distinctive distributional properties, such as consistent sentence length, reduced lexical diversity, and characteristic transitional vocabulary, which classical machine learning methods can effectively leverage (Gehrmann et al., 2019; Mitchell et al., 2023).

In practical terms, this finding suggests that if a student were to submit a completely AI-generated essay with no human editing, a simple TF-IDF classifier would likely flag it correctly. However, the educational significance of this result must be contextualised carefully: as AI models evolve and as students learn to request outputs in specific styles or to introduce deliberate imperfections, the distinctiveness of AI-generated text is likely to diminish. The near-perfect detection of AI-generated essays in this study may therefore represent an upper bound on detection performance with current models, rather than a stable property of the classification task.

### Human-Style and AI-Edited Essays Are Difficult to Distinguish

The principal challenge identified by this study is the classification boundary between Human-style and AI-edited essays. Both models systematically misclassified Human-style essays as AI-edited, producing per-class F1 scores below 0.40 for both of these categories. This finding has a straightforward linguistic explanation: AI editing, when applied to a student-written draft, typically preserves the original ideas, narrative structure, and cultural references of the source text while improving grammatical accuracy, sentence fluency, and lexical precision. The resulting AI-edited essay is therefore lexically and grammatically closer to standard written English than the original Human-style essay, but it retains the argumentative structure, topic-specific vocabulary, and discourse patterns of the student's own writing.

TF-IDF representations are limited in their ability to capture deeper structural similarities, as they primarily reflect surface word frequency rather than semantic coherence or stylistic consistency. For example, a Human-style essay written by a competent Nigerian secondary-school student with accurate grammar and a developed vocabulary may yield a TF-IDF signature similar to that of an AI-edited essay in the same genre. This similarity arises because both texts employ standard English vocabulary at comparable frequencies. Such overlap accounts for the high broad flag rate observed in this study.

The challenge of distinguishing AI-edited from Human-style essays extends beyond a technical limitation of feature representation. Instead, it reflects an inherently ambiguous phenomenon in which AI editing produces hybrid authorship that exists along a continuum between fully human and fully AI writing. Researchers employing advanced detection methods, such as perplexity-based approaches, stylometric fingerprinting, and transformer-based classifiers, consistently find that AI-edited text is the most difficult category to classify reliably (Ippolito et al., 2020; Verma et al., 2024). The present study corroborates this trend at the baseline level.

### The Broad Flag Rate Represents a Practical Barrier to Deployment

The primary practical implication of this study is that both baseline classifiers yield broad AI-assisted flag rates of 64.3% on Human-style essays. Such rates would be ethically unacceptable in any real-world examination context. If an educational institution were to employ either classifier as the basis for formal investigations of AI-assisted cheating, the majority of authentic student essays would be incorrectly referred for disciplinary review. The resulting harm to students, including reputational damage, increased anxiety, loss of examination time, and potential grade penalties during appeals, would be disproportionate. This impact would be especially pronounced for students whose natural writing style resembles AI-edited output, such as proficient second-language writers.

This finding aligns with previous research documenting systematic bias in AI detection tools against non-native English writers. Liang et al. (2023) reported that GPT-detector tools incorrectly classified a significant proportion of TOEFL essays written by non-native speakers as AI-generated. This bias was attributed to the tendency of both AI-generated text and non-native academic writing to employ simpler, more standardized syntactic structures compared to native academic prose. Nigerian secondary-school students, who write in English as a second or additional language, face a similar risk. The linguistic features characteristic of developing L2 writing may overlap with those that detectors associate with AI authorship.

The zero Strict False Positive Rate (FPR) result, which indicates that neither model misclassified any Human-style essay as fully AI-generated on the test set, offers some reassurance that the most severe form of false accusation can be avoided. However, this finding should be interpreted with caution due to the small test set size ($n = 40$). Furthermore, it does not address the broader issue of high flag rates in scenarios where any AI-related flag prompts formal scrutiny.

### Limitations

Several limitations affect the conclusions of this study. First, the dataset was constructed from simulated Human-style essays designed to approximate the style of Nigerian secondary-school student writing, rather than authentic student examination scripts. Although significant effort was made to represent the vocabulary, grammatical patterns, and discourse features of Nigerian secondary-school writing, simulated data cannot fully capture the range of individual variation found in real examination scripts. Applying these models to authentic WAEC or NECO examination data may result in different performance characteristics.

Second, the feature representation employed in this study, namely TF-IDF with unigrams and bigrams, is a well-established yet relatively shallow approach to text representation. More advanced methods, such as transformer-based language models like BERT or DistilBERT fine-tuned on Nigerian English text, may provide significantly improved discrimination between Human-style and AI-edited essays. This study establishes a baseline for comparison with future, more sophisticated approaches.

Third, the dataset size of 198 essays, although suitable for a baseline study and balanced across categories and prompts, is small by contemporary NLP research standards. Consequently, cross-validation results are estimates with considerable uncertainty, as indicated by the standard deviations reported in Table 3. Additionally, single-split test-set results on 40 essays are subject to substantial sampling variance.

It is also important to note that the AI-edited essays in this dataset were generated by instructing a generative AI tool to enhance the grammar and clarity of the Human-style essays. This approach reflects just one possible way in which AI might be used for editing. In reality, students may interact with AI tools in a variety of ways, such as requesting sentence-level rewrites, seeking alternative word choices, or reorganizing paragraphs. These different forms of AI assistance could plausibly result in essays with distinct statistical characteristics. Future research would benefit from incorporating a broader range of AI-editing procedures during dataset construction.

## ETHICAL IMPLICATIONS

The findings of this study carry substantial ethical implications for Nigerian secondary-school education, examination governance, and the broader field of AI deployment in high-stakes assessment contexts. This section

examines these implications across four dimensions: the rights of students, the responsibilities of educators and institutions, the obligations of AI tool developers, and the considerations specific to the Nigerian social and linguistic context.

**Student Rights and Due Process**

A primary ethical implication of this study is that students accused of AI-assisted cheating based on automated detection tools face a significant risk of wrongful accusation. Given a broad flag rate of 64.3% on authentic student essays, the classifiers evaluated in this study, if used without human oversight, represent an unreliable basis for academic discipline. In contexts where disciplinary actions may influence examination grades, academic records, or university admissions, the principle of due process necessitates that evidence of misconduct achieves a higher standard of reliability than that demonstrated by the tools assessed in this study.

Schools and examination bodies must therefore establish formal safeguards before any AI detection output is used in a disciplinary process. At a minimum, these safeguards should include: (a) the right of the accused student to view and respond to the detection evidence; (b) mandatory human review of any AI-flagged essay by a qualified English teacher familiar with the student's prior writing; (c) the option of an oral examination or written explanation in which the student can demonstrate knowledge of and authorship over the submitted essay; and (d) an independent appeals process. These protections are analogous to existing academic integrity procedures in higher education institutions and should be extended to the secondary-school level in anticipation of increasing AI use.

**Institutional Responsibilities**

Schools, teachers, and examination bodies bear a shared responsibility not to deploy unproven AI detection tools in contexts where the consequences of error are serious. This responsibility is not merely pedagogical; it reflects a broader ethical obligation to avoid harm. The Nigerian Constitution and the Child Rights Act of 2003 recognise the rights of young persons to education free from arbitrary interference (Nigerian Child Rights Act, 2003). Wrongful accusations of examination malpractice, if not handled with appropriate procedural care, may constitute a form of such interference.

Beyond avoiding harm, institutions have a positive responsibility to prepare students and teachers for an educational environment in which AI tools are widely available. Prohibition-only policies that do not educate students about the distinction between legitimate and illegitimate AI use are unlikely to be effective and may increase the risk that students engage in AI-assisted work covertly, without the benefit of teacher guidance. The RAISE Framework proposed provides a policy instrument for meeting this positive responsibility.

**Language Fairness and the Nigerian English Context**

The ethical concerns surrounding AI detection bias seem particularly pronounced in the context of Nigerian secondary schools, largely due to the linguistic realities faced by students. Most WAEC and NECO candidates are using English as a second or even third language, and their writing typically reflects what is known as Nigerian English. This variety is recognised in sociolinguistics for its distinct phonological, grammatical, and lexical characteristics. It appears that recent research on AI detection tools suggests a tendency to misclassify non-native English writing as AI-generated, likely because both types of text often display features such as simpler syntax, a more limited vocabulary, and lower textual perplexity.

For Nigerian secondary-school students, this situation arguably means they are at a greater risk of being falsely flagged by AI detection systems compared to their peers who write in native English varieties. This is not a matter of increased likelihood of academic dishonesty, but rather a consequence of their genuine writing sharing certain surface-level features with AI-generated text. If such detection tools are introduced into WAEC or NECO settings without careful consideration of these biases, the result could be not only technical inaccuracies but also a pattern of systematic unfairness. In effect, students from communities where English proficiency is still developing may be disproportionately penalised. This raises questions of educational equity that go beyond the technical limitations of the tools themselves.

**Responsibilities of AI Tool Developers**

The ethical responsibilities of those who develop AI text generation and detection tools arguably extend well beyond the immediate contexts in which these products are marketed. When AI writing assistants are introduced into environments that include Nigerian secondary-school students, as is clearly the case with tools such as ChatGPT, Gemini, and Claude, it appears that their developers must also consider the broader educational consequences that may arise from their use. In particular, there is a foreseeable risk that students will turn to these tools to produce examination essays, and this risk cannot be easily dismissed.

Similarly, developers of AI detection tools seem to bear a comparable responsibility to communicate the limitations of their products with accuracy and to take care that these tools are not used in assessment settings where the false positive rates observed in this and related studies could potentially harm students. The findings of the present study suggest that even relatively simple TF-IDF classifiers, which are much less complex than the neural network models behind most commercial detection tools, can flag nearly two thirds of genuine student essays. It follows that commercial detection tools which have not undergone rigorous validation on Nigerian English examination writing arguably should not be recommended to Nigerian schools or examination bodies unless these limitations are made explicit.

**The Broader Principle: Detection Should Support, Not Substitute, Human Judgement**

The ethical responsibilities of developers working on AI text generation and detection tools seem to reach beyond the immediate settings in which these products are promoted. Once AI writing assistants become accessible in contexts involving Nigerian secondary-school students, as is now the case with tools like ChatGPT, Gemini, and Claude, it becomes necessary for developers to reflect on the wider educational implications that might follow from their use. There is, for instance, a plausible risk that students will rely on these tools to generate examination essays. This possibility, in my view, cannot be overlooked.

The main ethical point that emerges here is that, given their current stage of development, AI detection tools do not appear suitable to function as autonomous decision-makers in academic integrity processes. While this observation arguably holds in a general sense, it seems particularly pressing in situations such as Nigerian secondary-school examinations. In these settings, students may be exposed to serious consequences if found guilty of malpractice, linguistic diversity may heighten the risk of bias, and institutional resources for ensuring due process are often limited.

For the foreseeable future, it seems that AI detection tools should be regarded as just one element within a broader, human-led review process. The output from such tools might reasonably prompt a teacher to examine an essay more closely, to compare it with a student's earlier work, or to invite the student to explain their writing process. However, it would not be appropriate to use detection results as the sole basis for grade penalties, examination disqualification, or disciplinary action. The underlying principle here is that detection ought to inform and support, rather than replace, human judgement. This ethical stance aligns with the RAISE Framework's fifth component (E for Ethical Assessment) and arguably should be adopted as a guiding standard by Nigerian schools and examination authorities.

# CONCLUSION

This study set out to address a question of increasing urgency in Nigerian secondary-school education: how reliably can AI-generated and AI-edited WAEC/NECO-style English essays be detected, and what are the associated risks of false accusation for genuine student writers? The answer, on the evidence of this preliminary study, is unambiguous: automated detection approaches are not yet sufficiently reliable for unsupervised use in Nigerian secondary-school essay assessment.

The experimental results reveal a fundamental asymmetry in detection performance. Fully AI-generated essays are readily identifiable, with both baseline classifiers achieving near-perfect F1 scores in this category. The consistent length, characteristic vocabulary, and limited stylistic variation typical of large language model output provide TF-IDF-based classifiers with reliable discriminating features. In contrast, AI-edited essays, where

human-authored content has been enhanced by AI, are extremely challenging to distinguish from authentic student writing. The primary error pattern in both models involved misclassifying Human-style essays as AI-edited, resulting in an overall AI-assisted flag rate of 64.3%. In educational settings where any AI-related flag prompts disciplinary investigation, this error rate suggests that approximately six out of ten genuine student essays would be subject to unwarranted scrutiny.

These findings align with the broader international literature on AI text detection, which consistently identifies hybrid human-AI authorship as the hardest category to classify (Ippolito et al., 2020; Verma et al., 2023), and with specific concerns about the bias of detection tools against non-native English writers (Liang et al., 2023). In the Nigerian context, where the majority of WAEC/NECO candidates are second-language English writers, the risk of biased false flagging is structurally elevated. AI detection tools trained predominantly on native English academic writing may encode implicit assumptions about what constitutes authentic student prose that are systematically violated by the features of Nigerian English, even when that writing is entirely genuine.

In response to these findings, the RAISE Framework is proposed as a five-component model of Responsible AI Support for Essay preparation, serving as a practical policy instrument for Nigerian secondary schools, teachers, and examination bodies. The framework delineates a clear typology of AI use, ranging from acceptable practice support and feedback assistance at Levels 1 and 2, to ethically problematic heavy rewriting at Level 3, and culminating in unacceptable full substitution at Level 4. Its five principles (Role Clarity, Academic Honesty, Improvement Focus, Student Ownership, and Ethical Assessment) offer normative guidance grounded in both educational theory and the empirical realities of detection performance.

Three broad conclusions follow from this study for different stakeholder groups. For students, the study demonstrates that AI editing of essays is technically detectable to some degree, and that its use without disclosure creates both ethical risks and reputational vulnerabilities; it also demonstrates, however, that legitimate use of AI for grammar correction and feedback support is less likely to attract detection and is consistent with responsible academic practice under the RAISE Framework. For teachers and school administrators, the study demonstrates that automated detection tools must not be used as the primary basis for academic integrity decisions; human review, student consultation, and due-process safeguards are essential complements to any technological screening. For WAEC, NECO, and education policymakers, the study makes a case for the urgent development of formal AI use policies that reflect the educational and linguistic realities of Nigerian secondary-school students, rather than importing wholesale the detection-centred approaches developed for higher-education contexts in English-speaking countries.

Future research should address the principal limitations of the present study through several approaches. First, the dataset should be expanded to include authentic student examination scripts, obtained under appropriate ethical safeguards, to determine whether detection patterns observed with simulated data generalise to real student writing. Second, more advanced feature representations, such as stylometric features, perplexity scores computed with pre-trained language models, and transformer-based classifiers, should be evaluated against the TF-IDF baselines established in this study. Third, the research should be extended to include the bias analysis for which this paper has provided the conceptual groundwork, specifically by testing whether detection tools systematically misclassify Nigerian English essays at a higher rate than essays written in other varieties. Fourth, the RAISE Framework should be piloted in selected Nigerian secondary schools and evaluated for its effectiveness in promoting responsible AI use and reducing misconduct, with the goal of establishing an evidence base for its formal adoption as educational policy.

The central message of this study is both cautionary and constructive. Caution is necessary because, in its current state, AI detection is not a reliable tool for penalizing students who may be innocent. Constructive action is required because the educational challenges posed by AI, such as fostering authentic writing competence in an era of automated writing assistance, can be addressed through robust frameworks, effective pedagogy, and thoughtful policy. The objective should shift from detecting student use of AI to cultivating educational environments where students recognize the importance of developing their own writing abilities and where AI supports, rather than replaces, learning.

**Data and Code Availability**

The dataset and analysis notebook used in this study are available from the author upon reasonable request. The dataset consists of simulated WAEC/NECO-style essay samples created for research purposes. No real student examination scripts, personal student records, or identifiable participant data were collected or used in this study.

## REFERENCES

1. Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228–239. https://doi.org/10.35542/osf.io/mrz8h
2. Gehrmann, S., Strobelt, H., & Rush, A. M. (2019, June 10). *GLTR: Statistical Detection and Visualization of Generated Text*. arXiv.Org. https://arxiv.org/abs/1906.04043v1
3. Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial Intelligence in Education. Promise and Implications for Teaching and Learning.
4. Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). *Automatic Detection of Generated Text is Easiest when Humans are Fooled* (arXiv:1911.00650). arXiv. https://doi.org/10.48550/arXiv.1911.00650
5. Jawahar, G., Abdul-Mageed, M., & Lakshmanan, L. V. S. (2020). *Automatic Detection of Machine Generated Text: A Critical Survey* (arXiv:2011.01314). arXiv. https://doi.org/10.48550/arXiv.2011.01314
6. Jekayinfa, O. J., & Aburime, A. O. (2021). Comparative Analysis of WAEC and NECO SSCE English Language Results in Ilorin, Kwara State, Nigeria. *EDUCARE*, *14*(1), 49–60. https://doi.org/10.2121/edu-ijes.v14i1.1449
7. John Kapi. (2023, December 19). *WAEC reveals techniques used to identify WASSCE candidates used AI in cheating*. GhanaWeb. https://www.ghanaweb.com/GhanaHomePage/NewsArchive/WAEC-reveals-techniques-used-to-identify-WASSCE-candidates-used-AI-in-cheating-1901240
8. Khalil, M., & Er, E. (2023). *Will ChatGPT get you caught? Rethinking of Plagiarism Detection* (arXiv:2302.04335). arXiv. https://doi.org/10.48550/arXiv.2302.04335
9. Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for Language Teaching and Learning. *RELC Journal*, *54*(2), 537–550. https://doi.org/10.1177/00336882231162868
10. Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). *Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense* (arXiv:2303.13408). arXiv. https://doi.org/10.48550/arXiv.2303.13408
11. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023, April 6). *GPT detectors are biased against non-native English writers*. Patterns, 4(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779
12. Luo (Jess), J. (2024). A critical review of GenAI policies in higher education assessment: A call to reconsider the "originality" of students' work. *Assessment & Evaluation in Higher Education*, *49*(5), 651–664. https://doi.org/10.1080/02602938.2024.2309963
13. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023, January 26). *DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature*. arXiv.Org. https://arxiv.org/abs/2301.11305v2
14. Nigerian Child Rights Act (2003). Laws of the Federation of Nigeria. - References—Scientific Research Publishing. (2003). https://www.scirp.org/reference/referencespapers?referenceid=3054259
15. Ogunmodimu, M. (2015). Language Policy in Nigeria: Problems, Prospects and Perspectives. 5(9).
16. Peterson, S. (2025). Addressing student use of generative AI in schools and universities through academic integrity reporting. *Frontiers in Education*, *10*. https://doi.org/10.3389/feduc.2025.1610836
17. Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach, Global Edition. Pearson Higher Ed.
18. Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023, March 17). *Can AI-Generated Text be Reliably Detected?* arXiv.Org. https://arxiv.org/abs/2303.11156v4
19. San Miguel, E., Inoncillo, F., & Malto, J. (2025). Analysing Students' Use of AI in Essay Outputs: A Case Study in Purposive Communication. *International Journal of Research and Scientific Innovation*, *XII*, 1620–1627. https://doi.org/10.51244/IJRSI.2025.12040130

20. Sengar, S. S., Hasan, A. B., Kumar, S., & Carroll, F. (2025). Generative artificial intelligence: A systematic review and applications. *Multimedia Tools and Applications*, *84*(21), 23661–23700. https://doi.org/10.1007/s11042-024-20016-1
21. Turnitin. (2023). *AI writing detection capabilities and limitations*. Turnitin. https://www.turnitin.com
22. UNESCO. (2023). Guidance for generative AI in education and research. https://doi.org/10.54675/PCSP7350
23. Verma, V., Fleisig, E., Tomlin, N., & Klein, D. (2024). *Ghostbuster: Detecting Text Ghostwritten by Large Language Models* (arXiv:2305.15047). arXiv. https://doi.org/10.48550/arXiv.2305.15047
24. Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). *The Affordances and Contradictions of AI-Generated Text for Second Language Writers*. https://escholarship.org/uc/item/337773bk
25. Weng, X., Xia, Q., Gu, M., Rajaram, K., & Chiu, T. K. F. (2024). Assessment and learning outcomes for generative AI in higher education: A scoping review on current research status and trends. *Australasian Journal of Educational Technology*, *40*(6), 37–55. https://doi.org/10.14742/ajet.9540
26. West African Examinations Council. (2024). *West African Examinations Council*. WAEC. https://www.waec.org.ng

## Appendix A: Dataset Construction Procedure

A.1 Prompt Selection

The essay prompts used in this study were designed to reflect common WAEC/NECO-style English Language continuous writing tasks. The prompts were selected to represent major essay genres commonly encountered in Nigerian senior secondary-school English examinations, including informal letters, formal letters, speeches, articles, debates, argumentative essays, narrative essays, descriptive essays, and reports.

The aim of prompt selection was not to reproduce official WAEC or NECO examination questions verbatim, but to construct examination-style prompts that reflect the structure, tone, and writing expectations of Nigerian secondary-school essay preparation. Each prompt was designed to support the creation of three comparable essay versions: one Human-style essay, one AI-generated essay, and one AI-edited essay.

After initial inspection of the dataset, prompts were standardized by removing extra spacing, converting text to lowercase, and matching prompts across the three categories. Only prompts that appeared exactly once in each category were retained in the final matched-triplet dataset. This produced a final dataset of 198 essays, consisting of 66 Human-style essays, 66 AI-generated essays, and 66 AI-edited essays.

## Appendix B: Creation of Human-Style Essays

The Human-style essays were manually written to simulate Nigerian secondary-school English essay responses. These essays were not collected from real students and should therefore be interpreted as simulated student-style writing rather than authentic examination scripts.

The Human-style essays were written to reflect features commonly associated with Nigerian secondary-school essay writing, including straightforward vocabulary, examination-oriented structure, conventional openings and closings, simple paragraph organization, and occasional expression patterns associated with second-language English writing. The aim was to create plausible WAEC/NECO-style essays that could serve as a preliminary baseline for evaluating AI-detection risks.

To avoid ethical concerns, no real student essays, personal records, or examination scripts were used. The simulated nature of the Human-style essays is recognized as a methodological limitation of the study, and future research should validate the findings using authentic student writing collected with informed consent and appropriate ethical approval.

**Appendix C: AI-Generated Essay Prompt Template**

AI-generated essays were produced by submitting WAEC/NECO-style prompts to a generative AI tool. The AI was instructed to produce complete essay responses suitable for a Nigerian senior secondary-school examination context.

The following prompt template was used: "Write a WAEC/NECO-style English essay on the following topic. The essay should be suitable for a Nigerian senior secondary-school student. Write not less than 350 words. Use clear paragraphing, formal examination-style English, and remain relevant to the topic.

Topic: [insert essay prompt]"

The AI-generated essays were recorded as complete outputs without manual editing. This allowed the study to evaluate fully AI-generated writing as a separate category from Human-style and AI-edited essays.

**Appendix D: AI-Edited Essay Prompt Template**

AI-edited essays were created by submitting Human-style essays to a generative AI tool for revision. The AI was instructed to improve grammar, clarity, fluency, and sentence structure while preserving the original ideas and organization of the Human-style essay.

The following prompt template was used:

"Edit the following WAEC/NECO-style essay for grammar, clarity, fluency, and sentence structure. Preserve the original ideas, meaning, tone, and overall organization. Do not turn it into a completely new essay. Keep the essay close to the original length and maintain a Nigerian secondary-school examination style.

Essay: [insert Human-style essay]"

The resulting AI-edited essays represent hybrid human-AI writing. This category was included because students may use AI tools not only to generate full essays, but also to polish or revise drafts they have already written.

**Appendix E: Dataset Fields**

Each essay record in the dataset contains the following fields:

1. essay_id: A unique identifier for each essay.

2. prompt: The essay question or writing task.

3. essay_type: The essay genre, such as article, speech, debate, informal letter, formal letter, narrative essay, descriptive essay, argumentative essay, or report.

4. source_label: The category of the essay: Human-style, AI-generated, or AI-edited.

5. word_count: The recorded number of words in the essay.

6. text: The full essay text.

7. notes: Additional researcher notes about the essay source or construction process.

**Appendix F: Matched-Triplet Cleaning Procedure**

The initial dataset contained 210 essays, with 70 essays in each of the three categories. During data inspection, some prompt-level inconsistencies were identified, including duplicated prompts and slight wording or spacing variations. To improve comparability across categories, a matched-triplet cleaning procedure was applied.

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue V May 2026

First, prompts were standardized by converting all prompt text to lowercase, removing extra whitespace, and trimming leading or trailing spaces. Next, the dataset was grouped by standardized prompt and source label. Only prompts that appeared exactly once in each of the three categories were retained. A valid matched triplet therefore consisted of one Human-style essay, one AI-generated essay, and one AI-edited essay responding to the same prompt.

After this cleaning procedure, the final dataset contained 198 essays: 66 Human-style essays, 66 AI-generated essays, and 66 AI-edited essays. This matched-triplet dataset was used for the final modelling analysis.