

Evaluating AI Text Detection Accuracy: A Structured Review of False Positives, False Negatives, and Implications for Academic Integrity Policy in Gulf Region Universities

Husain Ali Merza Shamlooh

202309709@stu.uob.edu.bh

University of Bahrain

Systematic Review

Keywords: structured review, academic integrity, false positives, false negatives, non-native English speakers, GCC universities, Arabic-English bilingual, perplexity analysis, Turnitin, GPTZero, ZeroGPT, Originality.AI, Copyleaks, academic due process, student rights, Gulf Cooperation Council

Posted Date: October 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9581862/v2>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Evaluating AI Text Detection Accuracy: A Structured Review of False Positives, False Negatives, and Implications for Academic Integrity Policy in Gulf Region Universities

Hussain Ali Merza Shamlooh
Department of Computer Engineering
University of Bahrain
Sakhir, Kingdom of Bahrain

202309709@stu.uob.edu.bh; hussainshamlooh06@gmail.com

Abstract Universities across the Gulf Cooperation Council (GCC) have rapidly adopted AI text detection tools for academic integrity enforcement, yet no review of published empirical evidence exists for this regional context. This paper presents a structured narrative review of published independent evidence on the accuracy, bias, and policy implications of five widely deployed AI detection platforms—GPTZero, Turnitin AI Detection, Originality.AI, Copyleaks, and ZeroGPT—with sustained attention to non-native English speaker (NNS) populations and the Arabic-English bilingual writers who make up most of the GCC student body. Drawing on peer-reviewed studies, preprints, and institutional statements, the review finds that: (1) in the largest independent multi-tool evaluation, no detection tool reached 80% accuracy [1]; (2) in the most widely cited bias study, detectors flagged an average of 61.2% of authentic essays by non-native English writers as AI-generated [2], although subsequent evidence from another language found no such bias [3]; (3) no validation study exists for Arabic-English bilingual writers, the most critical unaddressed gap in the regional literature; and (4) simple modification of AI-generated text sharply reduces detection, with accuracy on machine-paraphrased text falling to 26% [1]. Five contributions are made: a structured synthesis of AI detection evidence for the GCC context; formal identification of the Arabic-English bilingual research gap; a four-tier evidence-based policy framework; a structured regional research agenda; and implementation guidance for institutional decision-makers. The evidence indicates that using AI detection scores for disciplinary purposes in Gulf universities carries unacceptable false positive risk, particularly for NNS students, and warrants immediate policy reform.

Keywords AI text detection; structured review; academic integrity; false positives; false negatives; non-native English speakers; GCC universities; Arabic-English bilingual; perplexity analysis; Turnitin; GPTZero; ZeroGPT; Originality.AI; Copyleaks; academic due process; student rights; Gulf Cooperation Council

I. INTRODUCTION

The public release of ChatGPT in November 2022 represented a watershed moment for higher education worldwide. Within weeks of launch, the model demonstrated a capability that fundamentally challenged established assessment practices: the ability to produce fluent, grammatically correct, and structurally coherent academic prose on demand, at zero marginal cost, on virtually any topic [4]. Universities responded quickly, and AI text detection tools were deployed at scale before the empirical evidence required to justify such deployment had been produced. Within a year of launching its AI writing detector in April 2023, Turnitin reported having reviewed more than 200 million student papers [5]. The speed and scale of this institutional response was remarkable; its evidentiary basis was not.

The fundamental reliability problem with AI text detection was illustrated by a widely reported case: AI detectors including GPTZero and ZeroGPT classified the United States Constitution—drafted in 1787—as likely AI-generated [6]. More systematically, Weber-Wulff et al. [1] conducted one of the most comprehensive independent evaluations to date, testing 14 detection tools on a standardised corpus and finding that all scored below 80% accuracy and only five exceeded 70%. Their conclusion was unambiguous: available tools are “neither accurate nor reliable.” Independent evaluations have consistently found a substantial gap between vendor accuracy claims—often 98–99% or higher [7], [8]—and measured performance [1], [9], [10].

The consequences of this accuracy gap are not symmetric and not distributed equally across student populations. False positive errors—human text incorrectly classified as AI-generated—expose students to academic misconduct proceedings for work they completed honestly, with potential consequences including grade failure, academic probation, and expulsion. False negative errors—AI text incorrectly classified as human—allow dishonest submissions to pass undetected, undermining academic integrity. Both error types carry significant institutional and individual consequences, yet they have rarely been studied simultaneously, and their differential implications for policy have not been systematically analysed. The combination of both error types in the same detection system creates a perverse situation in which the system simultaneously fails to catch sophisticated AI misuse while wrongly accusing honest students.

The equity dimension of detection errors is particularly acute in the Gulf Cooperation Council (GCC) higher education context. Research by Liang et al. [2] at Stanford University found that seven widely used AI detectors incorrectly flagged an average of 61.2% of authentic TOEFL essays written by non-native English speakers as AI-generated, while classifying essays by US native speakers with near-perfect accuracy. The proposed mechanism—NNS writers’ tendency toward lower lexical diversity, more common vocabulary, and simpler syntactic constructions, which lowers text perplexity—is plausible, but the finding has not replicated uniformly: a later study of Czech found no systematic bias against non-native writers [3]. What is not known is how detectors behave for Arabic-English bilingual writers, who constitute the majority of the student body at GCC universities and whose first language is typologically distant from English; the essays studied by Liang et al. [2] came from Chinese-background writers. This represents the most urgent unaddressed empirical question in the regional policy context, and it is a gap that this paper formally identifies and documents.

This paper makes five original contributions that are absent from existing literature. First, it provides the first structured synthesis of AI detection evidence specifically for the GCC higher education context. Second, it formally identifies and documents the Arabic-English bilingual research gap as a critical priority for future investigation. Third, it provides a four-tier, evidence-based policy framework designed for Gulf region institutional governance that goes substantially beyond the brief recommendations offered by primary empirical studies. Fourth, it develops a structured regional research agenda prioritised for the GCC context. Fifth, it provides detailed implementation guidance—including procurement criteria, score threshold frameworks, assessment redesign principles, and staff training recommendations—that can be adopted directly by GCC institutions. These contributions are genuinely original even though the underlying empirical data derives from published literature, because their synthesis, application, framing, and institutional translation are entirely new.

A. Research Questions

This review addresses five research questions. RQ1 asks what published independent studies report regarding the accuracy, false positive rate, and false negative rate of the five most widely-deployed AI text detection tools: GPTZero, Turnitin AI Detection, Originality.AI, Copyleaks, and ZeroGPT. RQ2 investigates the documented magnitude of false positive bias against NNS writers and the linguistic and computational mechanisms that explain this bias. RQ3 examines the contextual factors—including academic discipline, text length, degree of AI modification, and writing proficiency—that moderate AI detection accuracy, and the implications of these moderators for GCC assessment contexts. RQ4

considers the legal, ethical, and procedural fairness implications of deploying AI detection tools for academic integrity enforcement in GCC universities, given the documented error rates and their differential distribution across student populations. RQ5 asks what evidence-based policy framework and research agenda should govern the responsible use of AI detection tools in GCC universities.

B. Scope and Structure

This review focuses on five tools selected for their documented institutional prevalence: GPTZero, Turnitin AI Detection, Originality.AI, Copyleaks, and ZeroGPT. The review period covers publications from November 2022 through March 2026, corresponding to the post-ChatGPT detection landscape. The review is limited to English-language academic text detection; detection of code, mathematical content, or non-English text is outside scope. The review synthesises published empirical evidence and does not include unpublished data, proprietary vendor documentation, or speculation beyond what the evidence supports. The absence of Arabic-English bilingual validation data is identified as a limitation of the existing literature rather than this review. Section II provides background on the GCC higher education context and the emergence of AI writing and detection tools. Section III presents the literature review covering technical foundations, empirical accuracy evidence, NNS bias, adversarial evasion, and legal challenges. Section IV describes the review methodology. Section V presents the synthesised findings organised by research question. Section VI discusses the broader implications. Section VII presents the policy framework. Section VIII addresses limitations and future research. Section IX concludes.

II. BACKGROUND AND CONTEXT

A. Higher Education in the Gulf Cooperation Council

The Gulf Cooperation Council, comprising Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and the United Arab Emirates, has undergone a dramatic transformation of its higher education landscape over the past three decades. GCC governments have invested substantially in expanding university capacity, driven by national development strategies that identify knowledge economies and skilled human capital as essential to post-petroleum economic diversification. The author’s institution, the University of Bahrain, exemplifies regional patterns: a primarily Arabic-speaking student body pursuing degree programmes conducted largely in English, with academic writing in English as a second or third language forming a core component of assessment across disciplines including engineering, business, and the natural sciences.

The linguistic profile of GCC university students is directly and critically relevant to AI detection reliability. The majority of students at GCC institutions are Arabic-first-language speakers for whom English is the language of academic instruction but not of daily communication, social interaction, or early educational experience. A substantial proportion of students, particularly in STEM programmes, conduct their studies entirely in English while thinking and communicating primarily in Arabic outside the academic context. This produces a characteristic pattern of English academic writing that differs systematically from native speaker writing in lexical diversity, syntactic complexity, and discourse organisation—differences that, as detailed in Section V, overlap substantially with the statistical signatures that AI detection tools associate with machine-generated text. This overlap is not incidental; it is a structural consequence of how both perplexity-based detection tools and second-language writers relate to probability distributions over English vocabulary.

The scale of NNS student populations in GCC universities makes the bias documented by Liang et al. [2] potentially consequential. If a GCC university with 10,000 students deployed detectors that behaved as those tested by Liang et al. did—an average false positive rate of 61.2% for NNS writers—and the majority of its students were NNS writers submitting genuine human-authored work, the expected number of false AI flags per semester, each potentially initiating a misconduct investigation, would run into the thousands. No published study has examined the institutional consequences of false positive rates at this scale in a GCC context.

GCC universities have been active adopters of educational technology, including AI detection tools. However, published evidence on how GCC faculty use and interpret detection outputs is scarce, and regionally authored studies have focused on tool accuracy rather than institutional practice [11]. Widespread tool availability without corresponding evidence literacy, policy infrastructure, or local validation is the institutional context that makes this review urgent and directly applicable to current practice across the region.

B. The Emergence of AI Writing Tools in Education

The challenge of AI-generated academic writing emerged rapidly following the public release of ChatGPT in November 2022. Within weeks of launch, commentators were warning that ChatGPT could produce credible student essays [12]. A US survey conducted in early 2023 found that 43% of college students had experience with AI tools such as ChatGPT and that roughly one in five had used them for schoolwork [13], while a self-selected poll at Stanford University found that about 17% of respondents had used ChatGPT for fall-quarter assignments or exams [14]. These

figures varied by institution, sampling method, and definition of ‘AI assistance’, but consistently indicated that AI writing had become widespread before most institutions had developed coherent policies.

Building on discussions of academic integrity in the ChatGPT era [15], three broad categories of AI writing use relevant to the detection challenge can be distinguished: full AI generation, in which an AI system produces the entire submitted text with minimal human input; AI-assisted drafting, in which an AI system produces an initial draft that the student revises and elaborates; and AI-edited writing, in which the student produces an initial draft that an AI system improves. Each category represents a different degree of intellectual engagement and a different relationship to the learning objectives of the assignment. This taxonomy has profound implications for detection: tools optimised to identify fully AI-generated text may perform poorly on hybrid or AI-assisted texts, which are arguably more pedagogically concerning because they involve deliberate deception while appearing superficially human. The categorical problem is compounded by the GCC context, in which students may use AI tools primarily to overcome language barriers rather than to avoid intellectual engagement with content, raising complex questions about the nature of the academic dishonesty involved.

The AI writing tools available to students have also diversified substantially since ChatGPT’s release. By 2024, students could access multiple competing LLMs—including GPT-4, Claude, Gemini, and numerous open-source alternatives—each with different statistical signatures that may confound detection tools trained primarily on ChatGPT outputs. The diversity of available AI writing tools means that detection tools calibrated against one LLM’s outputs may not generalise to others, creating an additional source of accuracy variation that is rarely acknowledged in vendor documentation.

C. The AI Detection Tool Market

The commercial AI text detection market emerged rapidly following ChatGPT’s release. GPTZero, developed by Edward Tian at Princeton University, was among the first widely-used tools, launched in January 2023. Within months, established educational technology companies added AI detection capabilities to existing products; Turnitin, whose services are used by more than 16,000 academic institutions, publishers, and corporations [5], did so in April 2023, adding AI detection capabilities to their existing platforms. The Turnitin case is particularly significant for institutional governance: many universities adopted AI detection capability without a separate procurement decision, simply by enabling a feature within an existing contract. This means that the contractual due diligence that should accompany a new technology procurement—including evaluation of accuracy claims, false positive rates, and population-specific performance—was often not conducted.

The five tools examined in this review represent a cross-section of the detection tool market, varying in technical approach, access model, pricing, output format, and institutional uptake. GPTZero pursued a direct-to-educator model, offering free access for individual teachers alongside institutional licensing. Originality.AI targeted content-industry professionals alongside educational institutions. Copyleaks leveraged existing plagiarism detection relationships to expand into AI detection. ZeroGPT gained widespread casual adoption primarily because it is free and requires no registration, despite having no publicly available methodology or benchmark documentation. This market diversity means that the tools being used across different GCC institutions vary substantially in their technical approaches, accuracy profiles, institutional support, and the vendor accountability that institutions can reasonably expect.

D. Academic Integrity Frameworks in GCC Universities

Academic integrity policy in GCC universities has developed largely within frameworks inherited from Western higher education systems, adapted to local cultural and institutional contexts. Most GCC universities have formal academic integrity policies that specify prohibited behaviours, investigation procedures, and sanction scales ranging from grade reduction to expulsion depending on the severity and frequency of violations. The academic integrity framework of the author’s institution, the University of Bahrain, defines academic misconduct to include plagiarism, fabrication, cheating, and the unauthorised use of AI tools, with sanctions calibrated to the degree of intentionality and impact.

The integration of AI detection tools into these frameworks has typically occurred through ad hoc faculty adoption rather than formal policy revision. Faculty members use detection tools to screen submissions and, where tools indicate high AI probability, refer cases to academic integrity committees. The procedural problem, as this review documents extensively, is that these referral decisions are based on probabilistic scores from tools whose false positive rates have not been validated for the student populations being evaluated. Academic integrity proceedings that affect student academic records and career prospects are being initiated on the basis of evidence that is, in the relevant population, unreliable. This is a failure of institutional governance that this paper seeks to address through an evidence-based policy framework.

III. LITERATURE REVIEW

A. Technical Foundations of AI Text Detection

1) Perplexity-Based Detection

The dominant technical approach in commercial AI text detection is perplexity analysis. Perplexity is a measure derived from language model probability distributions: given a sequence of tokens, perplexity quantifies how ‘surprised’ a reference language model is by each successive token. Mathematically, perplexity is defined as the exponentiated average negative log-likelihood of a sequence, such that lower perplexity indicates that the sequence is more predictable given the reference model; zero-shot detectors such as DetectGPT build on these model log-probabilities [16]. AI-generated text tends to exhibit lower perplexity than human-authored text because LLMs systematically favour high-probability token continuations—the statistical equivalent of always choosing the most expected next word. Human writers, by contrast, make less predictable lexical and syntactic choices driven by creative intent, idiomatic expression, personal voice, and the natural variability of human cognition.

The perplexity-based approach has several well-documented limitations that are directly relevant to the NNS bias problem. Most critically, as LLMs improve, their outputs become more fluent, coherent, and human-like, and therefore harder to distinguish from human writing. Solaiman et al. [17] observed in early GPT-2 experiments that outputs from larger models were harder to detect, and later evaluations found that detectors identified GPT-3.5 text more reliably than GPT-4 text [9], [11]. A detector calibrated against GPT-3.5 outputs may therefore perform substantially worse against GPT-4 or later models. This creates a dynamic accuracy problem that vendors rarely disclose and institutions rarely monitor.

The second critical limitation of perplexity-based detection is its susceptibility to the NNS false positive problem. Because NNS academic writers characteristically produce text with lower perplexity than native speakers—using higher-frequency vocabulary, simpler sentence structures, and more formulaic academic phrases—their writing occupies the same region of the perplexity distribution that detection tools associate with AI generation. This is not a correctable calibration error; it is a structural property of using perplexity as a proxy for AI authorship in any population where language proficiency constrains linguistic variability. The implication is that perplexity-based detection is inherently unsuited to evaluation contexts where NNS writers are assessed.

2) Burstiness Analysis

Burstiness analysis was popularised as a complement to perplexity-based detection, most prominently by GPTZero [6]. Burstiness captures variance in sentence-level complexity: human writing characteristically alternates between simple and complex sentences, creating an irregular, ‘bursty’ pattern of linguistic complexity variation. This reflects the natural rhythm of human writing, in which complex ideas expressed in long, heavily embedded sentences alternate with shorter, simpler sentences that maintain readability and coherence. AI-generated text, by contrast, tends toward more uniform complexity—each sentence is constructed to roughly similar standards of grammatical and lexical sophistication—producing a lower-variance, smoother complexity profile.

GPTZero explicitly combines perplexity and burstiness signals in its detection approach [6]. The tool reports both an overall AI probability score and sentence-level highlighting indicating which individual sentences appear AI-generated. This granular output has interpretability advantages over tools that produce only a single document-level score, as it allows instructors to identify specific portions of a submission that may warrant investigation. However, the burstiness signal may be confounded by academic writing conventions: structured academic prose, particularly in STEM disciplines with standardised methods sections and constrained reporting formats, tends toward more uniform complexity than creative or personal writing, reducing the burstiness signal regardless of authorship. Whether this produces systematically higher false positive rates in STEM disciplines has not yet been tested directly.

3) Neural Classifier Approaches

Several tools employ discriminative classifiers trained on labelled corpora of human and AI-generated text. Unlike perplexity-based methods, which use a reference language model as an indirect statistical signal, such classifiers learn directly to distinguish the two classes; an early example is the fine-tuned RoBERTa detector released alongside GPT-2 [17]. Commercial vendors generally do not disclose their architectures: institutions that declined to enable Turnitin’s detector cited the company’s inability to explain how the tool works or how its accuracy figures were calculated [18], [19].

The primary limitation of neural classifier approaches is domain generalisation: classifiers trained on corpora from specific domains, writing styles, or AI systems may not generalise to out-of-distribution inputs. A classifier trained predominantly on GPT-3.5 outputs may struggle to identify Claude or Gemini outputs that have different statistical signatures. This generalisation problem is compounded by the rapid pace of LLM development—training data that accurately represents current LLM outputs becomes stale within months as new model generations are released. Neural

classifiers also inherit the NNS bias problem: if the training corpus for the human-text class is drawn predominantly from native English speaker writing, the classifier will learn to associate NNS linguistic features with the AI class.

4) Watermarking Approaches

Kirchenbauer et al. [20] proposed a fundamentally different approach to AI detection: rather than attempting to identify AI-generated text post-hoc through statistical analysis, watermarking embeds imperceptible statistical patterns into text at generation time that can subsequently be detected with high reliability and low false positive rates. The proposed scheme partitions the vocabulary into ‘green’ and ‘red’ tokens using a pseudorandom function keyed to the preceding context, and biases the generation process toward green tokens. A detector with knowledge of the green/red partition function can then identify watermarked text with high accuracy, as the proportion of green tokens in a passage will substantially exceed chance for watermarked text even after moderate editing.

The watermarking approach has significant theoretical advantages over post-hoc detection: it achieves high accuracy with low false positive rates and is robust to the NNS bias problem, since watermark detection does not use perplexity as a signal. However, it faces substantial practical barriers to deployment in educational contexts. It requires LLM providers to implement watermarking at the model level, creating a coordination problem: students who use non-watermarked models, open-source LLMs, or locally-run models would not produce watermarked text, meaning the approach only detects AI use by students employing specific commercial providers that have implemented the scheme. Additionally, the approach is vulnerable to paraphrasing attacks that can disrupt the watermark signal without substantially degrading text quality, and to intentional green-token removal by adversarially informed users. For these reasons, watermarking remains a research approach rather than a deployed institutional solution.

B. Empirical Evidence on Detection Accuracy

1) Comprehensive Multi-Tool Evaluations

Weber-Wulff et al. [1] conducted one of the most comprehensive independent evaluations of AI detection tools to date, testing 14 tools on a corpus of human-written texts, human texts machine-translated into English from other languages, AI-generated texts, and AI-generated texts that had been manually edited or machine-paraphrased. All tested tools scored below 80% overall accuracy, with only five exceeding 70%. The tools showed a systematic bias toward classifying texts as human-written rather than detecting AI-generated text. Obfuscation was highly effective: average accuracy fell to approximately 42% for manually edited AI text and 26% for machine-paraphrased AI text, meaning that most paraphrased AI content went undetected. Machine-translated human texts were also classified less accurately than original human texts. No single tool performed consistently well across all text types and modification conditions.

Walters [9] evaluated 16 publicly available AI detection tools using 126 documents: 42 generated by GPT-3.5, 42 by GPT-4, and 42 written by students. Copyleaks, Turnitin, and Originality.ai achieved high accuracy across all three document sets, whereas most other detectors struggled to distinguish GPT-4 output from human writing. Perkins et al. [10] evaluated six major detectors in 805 tests of AI-generated and human-written samples and found a mean baseline accuracy of only 39.5%, with a mean reduction in accuracy of 17.4% when adversarial techniques such as editing and paraphrasing were applied; the authors conclude that such tools “cannot currently be recommended for determining whether violations of academic integrity have occurred.” This finding has direct implications for institutional enforcement policy.

2) Individual Tool Evaluations

Turnitin has received particular empirical attention given its institutional dominance. At launch, Turnitin stated that its detector was 98% accurate, with false positives occurring less than 1% of the time in its own tests [7]. Independent results are mixed. Weber-Wulff et al. [1] found Turnitin the most accurate of 14 tools and the only one to correctly classify all unmodified AI-generated documents in their corpus, yet its accuracy still fell below 80% overall and degraded substantially on obfuscated text; Perkins et al. [10] likewise observed a large reduction in Turnitin’s accuracy when adversarial techniques were applied. In a small test of 16 samples, The Washington Post found that Turnitin got more than half at least partly wrong, including flagging part of a student’s original essay [7]. Vanderbilt University calculated that even the vendor-claimed 1% false positive rate would have meant roughly 750 wrongly flagged papers among the approximately 75,000 papers it submitted to Turnitin in a year [19].

GPTZero was among the seven detectors tested by Liang et al. [2] and among the 14 tools tested by Weber-Wulff et al. [1], where it did not reach 70% accuracy. In a Bloomberg Businessweek test of 500 college application essays written in 2022, before ChatGPT’s release, GPTZero and Copyleaks falsely flagged 1–2% as likely AI-written [21]. The RAID benchmark—over 6 million generations spanning 11 models, 8 domains, 11 adversarial attacks, and 4 decoding strategies—found that current detectors are easily fooled by adversarial attacks, variations in sampling strategy, and unseen generative models, despite claims by many detectors of 99% accuracy or more [8]. Vendor-published benchmark results, including GPTZero’s own reports of its RAID performance, were excluded from this review under the criteria in Section IV.

ZeroGPT was among the detectors tested by Liang et al. [2] and Weber-Wulff et al. [1], where it did not exceed 70% accuracy. Large-scale false positive figures that circulate online for ZeroGPT originate from vendor blogs, including those of competing ‘humanizer’ services, and were excluded from this review. CopyLeaks achieved high accuracy across GPT-3.5, GPT-4, and human-written documents in Walters [9], achieved the highest accuracy of the tested detectors on adversarially manipulated content (58.7%) in Perkins et al. [10], and falsely flagged 1–2% of pre-ChatGPT essays in Bloomberg’s test [21]; however, it misclassified half of the small human-written control set in Perkins et al. [10]. Originality.AI was among the detectors tested by Liang et al. [2] and performed well in Walters [9], but peer-reviewed, tool-specific false positive data for it remain limited.

C. The Non-Native English Speaker Bias

1) The Stanford Study

Liang et al. [2] conducted the most influential and policy-relevant study on NNS bias in AI detection, testing seven widely used detectors (Originality.AI, Quill.org, Sapling, OpenAI, Crossplag, GPTZero, and ZeroGPT) on 91 TOEFL (Test of English as a Foreign Language) essays written by non-native speakers and 88 US eighth-grade essays written by native speakers. The TOEFL essays, dated 2020 or earlier, were sourced from a Chinese educational forum and therefore predate widely available AI writing tools. The results were stark: while detectors classified the native speaker essays with near-perfect accuracy, they incorrectly flagged an average of 61.2% of TOEFL essays as AI-generated. All seven detectors unanimously misidentified 18 of the 91 TOEFL essays (19.8%) as AI-authored, and 89 of the 91 (97.8%) were flagged by at least one detector. The unanimously flagged essays had significantly lower perplexity than the others, providing direct mechanistic evidence for the perplexity-bias hypothesis.

The practical implications of these findings are severe. A university using any of the seven tested detectors—including widely used tools such as GPTZero, ZeroGPT, and Originality.AI—to screen essays by comparable NNS students would trigger false AI flags for a majority of submissions. If the institution uses multiple tools, the probability of at least one false flag approaches certainty: Liang et al. [2] found that 97.8% of the authentic TOEFL essays received at least one false flag from the seven-detector ensemble. In a GCC university context where NNS writers constitute the majority of the student body, a detection regime behaving in this way would generate a volume of false misconduct signals that would overwhelm investigation capacity, leading either to mass unjust accusations or to the informal abandonment of enforcement—neither of which serves academic integrity.

2) Evidence from Other NNS Populations

The Liang et al. [2] study used essays by Chinese-background writers, raising the question of whether the findings generalise to other NNS populations. The evidence to date is mixed. Institutions have treated the finding as sufficiently credible to inform policy: Vanderbilt University cited it among its reasons for disabling Turnitin’s AI detector [19]. However, a more recent study of Czech by Al Ali et al. [3] found that texts by non-native writers of Czech did not have lower perplexity than native texts and identified no systematic bias against non-native writers across three detector families. The bias therefore appears to depend on language, writer proficiency, and detector design, and its magnitude for any given population cannot be assumed without local testing.

3) The Arabic-English Bilingual Gap

Despite the NNS bias documented by Liang et al. [2], no published study has tested any AI detection tool specifically against Arabic-English bilingual writing corpora. This absence is significant given structural characteristics of Arabic that could affect the perplexity of Arabic-influenced English. Arabic is typologically distant from English in several relevant dimensions: Verb-Subject-Object is a common default word order in Modern Standard Arabic, compared to English’s Subject-Verb-Object; Arabic has an extensive morphological system with rich inflectional and derivational patterns, fundamentally different from English’s predominantly analytic grammar; its root-and-pattern morphology derives words from consonantal roots through vowel patterns, creating derivational relationships absent from English; and Arabic discourse conventions differ from Western academic writing norms in argumentation structure, paragraph organisation, and hedging practices.

Kaplan’s early work in contrastive rhetoric [22] characterised English paragraphs written by Arabic-background students as relying heavily on parallel, coordinate constructions—a generalisation that later scholarship has criticised but that continues to inform research on Arabic-influenced English writing. If Arabic-first-language students with intermediate to advanced English proficiency produce more predictable vocabulary choices, simpler clause structures, and more formulaic phrase-level patterns than native speakers, their texts would have lower perplexity and could be at elevated risk of false positives. Equally, the Czech evidence [3] shows that such effects cannot be assumed. The GCC research gap is therefore not merely an absence of regional data: whether existing findings over- or underestimate the risk for the populations most affected by GCC institutional deployment is unknown.

D. Adversarial Evasion and Detection Robustness

A consistent and policy-critical finding across the detection literature is the ease with which AI detection can be evaded through text modification. Sadasivan et al. [23] demonstrated both theoretically and empirically that watermarking, zero-shot, neural-network-based, and retrieval-based detectors can be defeated by adversarial paraphrasing: their recursive paraphrasing attack, in which AI-generated text is repeatedly rephrased by a second language model, substantially reduced detection rates while largely preserving text quality. Weber-Wulff et al. [1] found that machine paraphrasing reduced average detection accuracy to 26%, and Perkins et al. [10] found that adversarial techniques reduced the accuracy of six detectors by a mean of 17.4% from an already low baseline of 39.5%, confirming that evasion is accessible to any student willing to apply minimal effort.

This evasion evidence has a deeply asymmetric implication for academic integrity enforcement. Detection tools are most likely to flag unsophisticated AI use—students who paste AI-generated text directly into submission systems without any modification—while systematically failing to detect the more sophisticated and academically concerning uses, where AI assistance is blended with human writing or where simple paraphrasing is applied. The student population most likely to be caught is therefore not the most sophisticated or academically concerning user of AI tools, but rather the most naive. Combined with the documented NNS false positive bias, the result is a detection regime that catches naive users and NNS students at high rates while missing sophisticated academic dishonesty entirely—precisely the inverse of what a fair and effective academic integrity system should produce.

E. Academic Integrity, Due Process, and Legal Challenges

Procedural fairness in academic misconduct proceedings requires, at minimum, that students be told the evidence against them and given a meaningful opportunity to respond—principles long established in US case law on student discipline [24], [25]. It follows that disciplinary evidence must be reliable, interpretable, and open to meaningful challenge. AI detection tools present challenges on all three dimensions. Their outputs are probabilistic scores that cannot be traced to specific AI-generated source material and cannot be independently verified; their internal reasoning is not explainable to students or adjudicators in human-interpretable terms; and their reliability varies across student populations in ways that institutions typically do not disclose to students, faculty, or academic integrity committees.

Documented disputes illustrate the concrete risks of detection-based enforcement. In February 2025, an executive MBA student filed a federal lawsuit challenging a one-year suspension by Yale School of Management that followed a GPTZero flag; the complaint argued that Yale’s own policies discouraged such tools because of their high false positive rates, particularly for non-native English speakers [26]. Institutional responses have also shifted. The University of Manchester opted out of the first release of Turnitin’s AI detector in April 2023, citing the vendor’s inability to explain how the tool works or how its quoted accuracy had been calculated [18], and Vanderbilt University disabled the feature in August 2023, citing false positive risk, bias against non-native writers, and lack of transparency [19]. Perkins et al. [10] conclude that current detectors cannot be recommended for determining whether academic integrity violations have occurred, and TEQSA-commissioned guidance argues for reorienting institutional effort toward assessment reform [27].

F. Regional and GCC-Specific Literature

Published research specifically addressing AI detection in GCC or broader Arab-majority higher education contexts is limited. Elkhatat et al. [11], writing from institutions in Qatar, evaluated five detectors (OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag) on ChatGPT-3.5, ChatGPT-4, and human-written texts, finding that the tools were more accurate for GPT-3.5 than GPT-4 output and produced false positives and uncertain classifications for human-written text. Dalalah and Dalalah [28] examined the detection of ChatGPT-generated academic text and found that genuine literature-review passages were particularly prone to being misclassified as AI-generated. Neither study examined Arabic-English bilingual student writing as a distinct category, and evidence on how GCC faculty use and interpret detection outputs remains scarce—gaps that motivate the research agenda in Section VIII.

IV. REVIEW METHODOLOGY

A. Review Design and Rationale

This paper employs a structured narrative review. This approach was selected over a full systematic review for three reasons. First, the relevant evidence base is small and spans peer-reviewed studies, preprints, and institutional statements; the primary gap is the absence of its synthesis and contextualised application to the GCC setting, not the absence of primary evidence. Second, heterogeneity in corpora, tools, thresholds, and reported metrics precludes meta-analytic pooling. Third, the policy claims advanced in Section VII require an evidence base in which every quantitative claim can be traced to a primary source; accordingly, all figures reported in this review were checked against the original publications. Structured narrative reviews are an established contribution in applied fields where evidence must be synthesised and contextualised to guide practice [29].

B. Literature Search Protocol

A comprehensive database search was conducted across five sources: Google Scholar, Semantic Scholar, arXiv, PubMed, and the ACM Digital Library in March 2026. Multiple search strategies were employed at three levels. The primary search combined terms for AI text detection with accuracy, error rate, and educational setting. The search string was adapted for each database’s syntax and query language, with the core terms: (“AI text detection” OR “AI content detection” OR “GPT detection” OR “ChatGPT detection”) AND (“accuracy” OR “false positive” OR “false negative” OR “bias” OR “reliability”) AND (“education” OR “academic integrity” OR “university” OR “student”). Secondary searches used tool-specific terms for each of the five target platforms: GPTZero, Turnitin AI detection, Originality.AI accuracy, CopyLeaks AI detection, and ZeroGPT accuracy. Tertiary searches targeted population-specific terms including non-native English AI detection, ESL writing AI detector, Arabic English academic writing AI, and GCC higher education AI integrity. Reference lists of all included studies were manually screened for additional relevant sources through both forward citation tracking (who cited this paper) and backward citation chaining (who was cited by this paper).

C. Inclusion and Exclusion Criteria

Studies were included if they met all of the following criteria: publication between November 2022 and March 2026, corresponding to the post-ChatGPT AI detection landscape; empirical measurement of at least one of accuracy, false positive rate, false negative rate, or precision-recall metrics for AI text detection in academic contexts; evaluation of at least one of the five target tools or evidence directly and specifically relevant to NNS bias in AI detection; publication in a peer-reviewed journal, indexed conference proceedings, credible institutional report, or verified pre-print repository such as arXiv or SSRN (institutional statements and reputable investigative journalism were consulted only for vendor claims, documented cases, institutional responses, and clearly labelled illustrative tests, not as primary accuracy evidence); use of a corpus of at least 20 text samples with sufficient methodological detail to allow assessment of result reliability; and sufficient description of the evaluation methodology to assess the conditions under which reported metrics apply. Studies were excluded if they reported only vendor-provided benchmark data without independent verification; focused exclusively on plagiarism detection without AI-specific evaluation components; were commentary, opinion, or editorial pieces without original empirical data; evaluated AI detection tools not among the five target platforms and did not address NNS bias; or predated the ChatGPT release period and therefore evaluated an earlier generation of detection challenge. Foundational technical studies (for example, on perplexity, watermarking, and adversarial paraphrasing) and broader benchmarks such as RAID [8] were consulted as background to explain detector behaviour, but were not used as tool-specific accuracy evidence.

D. Study Selection and Data Extraction

Records identified by the search were screened by title and abstract and then by full text against the inclusion and exclusion criteria. Because this is a structured narrative rather than a systematic review, record counts are not reported as a PRISMA flow diagram. The principal studies forming the evidence base are summarised in Table I.

Data were extracted from each included study using a standardised extraction template capturing: study design and corpus characteristics; tools evaluated and evaluation methodology; accuracy metrics reported including overall accuracy, FP rate, FN rate, precision, recall, and F1 where available; population characteristics including NES versus NNS status, academic discipline, and text length; and key findings and conclusions. Where a study reported multiple metrics across different conditions—for example, accuracy for edited versus unedited AI text—all condition-specific metrics were extracted separately to enable condition-specific analysis. Key extracted metrics are summarised in Table I.

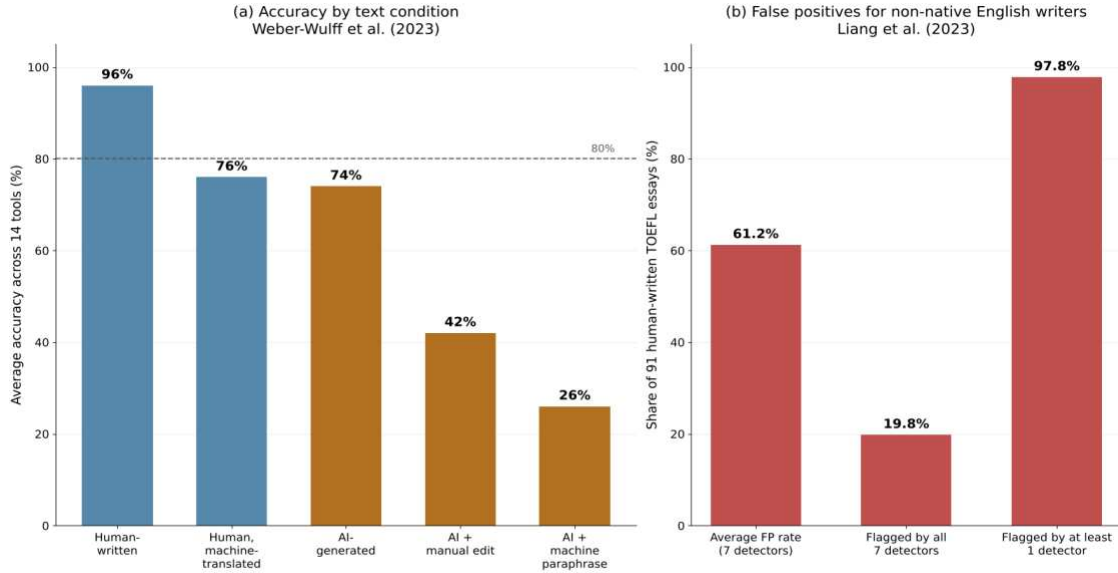
E. Quality Assessment and Synthesis Approach

Each included study was appraised qualitatively for corpus representativeness and size, ground-truth reliability, completeness of reported metrics, and generalisability to deployment contexts. Evidence synthesis followed the narrative synthesis guidance of Popay et al. [30] rather than meta-analysis, because methodological heterogeneity across studies precluded quantitative pooling. Confidence in synthesised findings is graded as High, Moderate, or Low based on consistency across studies, methodological quality, and directness of evidence to the GCC context.

V. SYNTHESISED FINDINGS

A. RQ1: Tool-Level Accuracy

Figure 1 presents two of the most policy-relevant quantitative findings in the reviewed literature: average detection accuracy by text condition in the 14-tool evaluation of Weber-Wulff et al. [1], and the false positive outcomes for non-native English writers reported by Liang et al. [2]. Table I summarises the verified independent evidence for each of the five target tools, with sources and confidence grades.



Values taken directly from the cited studies. Panel (a): blue = human-written text (correct = classified human); orange = AI-generated text (correct = classified AI).

Fig. 1. (a) Average detection accuracy across 14 tools by text condition, from Weber-Wulff et al. [1]. (b) False positive outcomes for 91 human-written TOEFL essays by non-native English writers across seven detectors, from Liang et al. [2]. All values are taken directly from the cited studies.

Table I
VERIFIED INDEPENDENT EVIDENCE FOR FIVE AI DETECTION TOOLS

Tool	Vendor Claim	Independent Accuracy Evidence	FP on Human Text	FP on NNS Text	Under Modification	Confidence	Key Sources
GPTZero	Very high (often $\geq 99\%$) ^a	Below 70% overall in 14-tool test [1]	1–2% of 500 pre-2022 essays [21]	One of 7 detectors averaging 61.2% [2]	Paraphrase cut average accuracy to 26% (all tools) [1]	Moderate	[1], [2], [21]
Turnitin AI	98% accuracy; <1% FP [7]	Most accurate of 14 tools but <80%; detected all unmodified AI texts [1]; high accuracy on all document sets [9]	None in controlled corpus [1]; partly misflagged an original student essay [7]	Not tested; no published NNS data	Large accuracy loss under adversarial techniques [10]	High	[1], [9], [10], [7]
Originality.AI	Very high (often $\geq 99\%$) ^a	High accuracy on all document sets [9]	No peer-reviewed estimate located	One of 7 detectors averaging 61.2% [2]	No peer-reviewed data located	Low	[9], [2]
Copyleaks	Very high (often $\geq 99\%$) ^a	High accuracy on all document sets [9]; tested in [11]	1–2% of 500 pre-2022 essays [21]; half of a small control set [10]	Not tested	Highest accuracy on manipulated text (58.7%) [10]	Moderate	[9], [10], [21], [11]
ZeroGPT	Very high (often $\geq 99\%$) ^a	Below 70% overall in 14-tool test [1]	No peer-reviewed estimate located	One of 7 detectors averaging 61.2% [2]	Paraphrase cut average accuracy to 26% (all tools) [1]	Low	[1], [2]

^a RAID [8] notes that many detectors claim accuracy of 99% or more; tool-specific marketing claims change frequently and should be checked at the time of reading. FP = false positive; NNS = non-native English speaker. Figures from Bloomberg Businessweek [21] and The Washington

Post [7] are journalistic tests reported for illustration. 'Not tested' means the tool was not among the detectors in the cited NNS study. Confidence grades reflect the consistency and quality of the supporting evidence.

As shown in Figure 1 and Table I, three cross-cutting patterns characterise the verified evidence. First, vendor accuracy claims—98% for Turnitin and commonly 99% or more across the market [7], [8]—substantially exceed independently measured performance: in the 14-tool evaluation of Weber-Wulff et al. [1], even the best-performing tool (Turnitin) scored below 80%, and most tools scored below 70%. Second, false positive rates on general-population human writing vary widely, from none in some controlled corpora [1] and 1–2% in a large journalistic test [21] to half of a small control set for one tool [10]; even a 1% rate translates into hundreds of false flags at institutional scale [19]. Third, for NNS writers the most widely cited evidence indicates dramatically higher false positive rates [2], but replication across languages is mixed [3] and no data exist for Arabic-English bilingual writers. This three-pattern structure—vendor overclaiming, non-trivial false positive rates, and potentially severe but population-dependent NNS bias—establishes the evidentiary basis for the policy framework in Section VII.

B. RQ2: The NNS Bias

The evidence that AI detectors can be biased against NNS writers is strong in the study that first documented it but has not replicated uniformly (Confidence: Moderate). Liang et al. [2] found a 61.2% average false positive rate for TOEFL essays by Chinese-background writers across seven detectors, compared with near-perfect accuracy for US native-speaker essays, and showed that the unanimously flagged essays had significantly lower perplexity. By contrast, Al Ali et al. [3] found no systematic bias against non-native writers of Czech, whose texts did not have lower perplexity than native texts. Together, these studies suggest that the bias arises where second-language writing is measurably more predictable than native writing, and that its size depends on the language, the writers' proficiency, and the detector. Where it does arise, it cannot be resolved simply by adjusting detection thresholds, because NNS and AI-generated text then overlap in the perplexity space that perplexity-based detectors use as their primary signal.

For GCC universities, the implications are potentially severe. Even at the 1% document-level false positive rate claimed by Turnitin [7], Vanderbilt University estimated that around 750 of the roughly 75,000 papers it submitted in a year could have been wrongly flagged [19]. If detectors behaved for Arabic-English bilingual writers as they did for the TOEFL essays in Liang et al. [2], a department of 500 students each submitting one major essay per semester would see approximately 306 entirely human-authored submissions flagged as AI-generated. These figures are illustrative rather than predictive: the Arabic-English bilingual population has not been studied, and its false positive rate could be higher or lower than in existing studies. That uncertainty is itself the central policy problem, because consequential disciplinary decisions are being made without population-specific validation data.

C. RQ3: Contextual Moderators

Several contextual factors moderate AI detection accuracy, each with implications for GCC assessment practice. Text genre and discipline are plausible moderators: Dalalah and Dalalah [28] found that genuine literature-review passages were especially prone to being misclassified as AI-generated, and the burstiness argument in Section III-A-2 suggests that formulaic STEM prose may be at elevated risk. However, no reviewed study directly compared false positive rates across disciplines, so this remains a hypothesis requiring testing in GCC assessment contexts (Confidence: Low).

Text modification is the most powerful moderator identified in the reviewed literature, with consistent findings across multiple studies. Weber-Wulff et al. [1] found that manual editing reduced average detection accuracy to approximately 42% and machine paraphrasing to 26%. Perkins et al. [10] found that adversarial techniques reduced the accuracy of six detectors by a mean of 17.4% from a baseline of 39.5%. Sadasivan et al. [23] showed that recursive paraphrasing substantially reduces detection rates across detector types. These findings mean that AI detection tools are most likely to flag the simplest and most naive AI use while systematically missing sophisticated use that involves deliberate evasion—an asymmetry with direct equity implications, since students with greater technical knowledge of AI tools and detection methods are better positioned to evade detection.

The model generating the text also matters: detectors identified GPT-3.5 output more reliably than GPT-4 output [9], [11], and RAID found detectors easily fooled by unseen generative models and changes in sampling strategy [8]. Text length is likely to matter as well, since statistical signals are weaker in short texts, but it was not systematically examined in the reviewed studies; this is particularly relevant for GCC assessment contexts where short-answer questions are common. Finally, consistency across tools is limited: Elkhatah et al. [11] reported inconsistent outputs, including false positives and uncertain classifications for human-written text, and Weber-Wulff et al. [1] found that no tool performed consistently across text types. Limited agreement means that the output of any single tool carries reduced evidential value, and that conflicting outputs from two tools applied to the same submission further undermine the basis for any enforcement action.

D. RQ4: Legal, Ethical, and Due Process Implications

The synthesised evidence supports a clear legal and ethical assessment: current AI detection tools do not meet the evidentiary standards required for academic misconduct proceedings in higher education. Academic misconduct proceedings that can result in grade penalties, suspension, or expulsion require evidence that is reliable (produced by a method with documented, validated accuracy in the relevant population), interpretable (explainable to the student and adjudicators in a way that enables meaningful challenge), and specific (pointing to specific acts rather than general statistical probabilities), as argued in Section III-E. AI detection tool outputs fail on all three criteria. Their accuracy is not validated for NNS populations, as established by RQ1 and RQ2. Their internal reasoning is opaque and not explainable in human-interpretable terms. And their outputs are probabilistic scores rather than evidence of specific acts.

The NNS bias compounds these evidentiary concerns into an equity and potential legal liability issue. If AI detection tools systematically produce higher false positive rates for NNS writers—as Liang et al. [2] found for the population they studied—their use without population-specific corrective measures constitutes a form of indirect discrimination in academic integrity enforcement. A facially neutral policy applying the same detection threshold to all students produces disproportionately adverse outcomes for NNS students, a group whose equitable treatment many educational equity frameworks seek to protect. In legal contexts that apply proportionality analysis to educational policies, this structural inequity may create institutional liability independent of whether any individual misconduct finding is subsequently reversed on appeal.

E. RQ5: Summary of Evidence for Policy

Drawing together the findings from RQ1 through RQ4, the synthesised evidence establishes seven key findings with associated confidence grades that directly ground the policy framework in Section VII. Finding 1 (Confidence: High): In the largest independent multi-tool evaluation, no AI detection tool reached 80% accuracy [1], and vendor-claimed accuracy substantially exceeds independently measured performance. Finding 2 (Confidence: Moderate): Detectors can show severe false positive bias against NNS writers—61.2% on average in the most widely cited study [2]—but the effect has not replicated in every language [3], so it must be measured for each population. Finding 3 (Confidence: High): Simple modification of AI-generated text sharply reduces detection—to 26% average accuracy for machine-paraphrased text in one large evaluation [1]—meaning tools reliably catch only the most naive AI use. Finding 4 (Confidence: High): No AI detection tool has been validated against Arabic-English bilingual writing corpora, meaning GCC institutions are making consequential disciplinary decisions based on tools whose performance in their student population is unknown. Finding 5 (Confidence: Low): Text genre and discipline may moderate false positive rates [28], but this has not been tested directly. Finding 6 (Confidence: Moderate): Detection performance is inconsistent across tools and degrades for newer generative models [11], [9], [8], undermining the evidential value of single-tool outputs. Finding 7 (Confidence: Moderate): Universities have begun disabling or restricting AI detection features on reliability, transparency, and equity grounds [18], [19].

VI. DISCUSSION

A. The Detection Paradox

The synthesised evidence reveals a fundamental detection paradox that has not been adequately acknowledged in institutional deployment decisions. AI detection tools are most likely to catch the least sophisticated AI users—those who paste AI-generated text directly into submission systems without any modification—while systematically failing against sophisticated users who apply basic evasion techniques. Combined with the documented NNS false positive bias, the result is a detection regime that penalises honest NNS students and naive AI users at high rates while allowing sophisticated academic dishonesty to evade detection. This is not merely a measurement error; it is a systematic directional bias that produces exactly the wrong enforcement outcomes. The students most likely to be penalised are those who are either NNS writers submitting genuine human work, or naive AI users who lack awareness of evasion methods. The students most likely to escape detection are those who deliberately incorporate AI assistance while performing enough editing to evade perplexity-based signals. If academic integrity enforcement is understood as a system designed to deter and identify the most academically harmful AI use, current detection technology is not merely inadequate; it is inversely effective.

This paradox should prompt a fundamental reconsideration of the assumption that AI detection tools serve academic integrity. The evidence suggests that the primary effect of current detection tools in NNS-majority institutions is to generate false accusations against honest students, not to deter or identify genuine academic misconduct. Institutional investment in detection tools may therefore be counterproductive: it allocates resources to a technology that is failing at its stated purpose while creating legal and reputational risks through systematic false accusations. The alternative—investing in assessment redesign, AI literacy education, and human-centred investigation processes—may be both more effective and more equitable.

B. The Equity Crisis and Vendor Transparency Problem

The NNS bias in AI detection constitutes a systemic equity crisis, not a minor calibration issue that can be resolved through threshold adjustment or tool improvement. Higher education systems have invested substantially in equity, diversity, and inclusion initiatives designed to ensure that all students, regardless of linguistic background, have equal opportunity to demonstrate their learning. The uncritical deployment of AI detection tools in NNS-majority institutions directly undermines this investment by introducing a technical system that systematically disadvantages students on the basis of their linguistic background. The equity dimension is compounded by the power dynamics of academic misconduct proceedings: students accused of misconduct must typically demonstrate their innocence in a process where the institution holds substantial procedural advantages, and for NNS students facing AI misconduct allegations, this challenge is doubled by the need to overcome both the detection verdict and limited familiarity with institutional processes.

The vendor transparency failure compounds the equity problem by making it difficult for institutions to make informed decisions. Vendors commonly claim accuracy of 98–99% or higher [7], [8]; no tool approached this in the largest independent multi-tool evaluation [1]. To the author’s knowledge, none of the five vendors examined publicly reports false positive rates disaggregated by writer linguistic background, despite the documented risk of NNS bias. Institutions serving predominantly NNS student populations therefore have no way to assess the false positive rates they are imposing on their students without conducting their own independent validation studies—a requirement that most institutions lack the resources and expertise to fulfil. This represents a fundamental failure of vendor accountability that institutional procurement processes must address through contractual requirements for population-specific performance disclosure.

C. The Arms Race Dynamic and Long-Term Implications

Sadasivan et al. [23] show theoretically that reliable detection becomes harder as the distributions of human and AI-generated text converge, implying an arms race between AI generation and AI detection that has significant long-term implications for institutional policy. As LLMs improve, their outputs become harder to detect; detection tools are updated to catch current-generation outputs; LLMs continue to improve. This creates an inherent instability in detection-based enforcement: a tool that achieves acceptable accuracy today may be substantially less accurate within months as new model generations are released. Much of the published evaluation evidence concerns 2022–2023 outputs from GPT-3.5 and early GPT-4; detectors already performed worse on GPT-4 than on GPT-3.5 [9], [11], and their performance against subsequent generations is largely unknown.

This dynamic argues strongly against institutional policies that treat AI detection as a stable, reliable infrastructure component comparable to plagiarism detection through text matching. Text-matching plagiarism detection is inherently stable: the copied text does not change, and the database of sources against which it is matched grows over time, improving detection rather than degrading it. AI detection works through statistical pattern recognition that must continuously adapt to a moving target. For GCC universities, the combination of high NNS false positive rates, absent Arabic-English bilingual validation data, and the dynamic accuracy problem creates a policy situation in which the honest institutional position must be that AI detection cannot currently serve as a reliable enforcement mechanism for the student populations being evaluated. Institutional resources should therefore be redirected toward the stable, equity-neutral alternatives discussed in Section VII: assessment redesign, AI literacy education, and investment in the regional validation research that would provide the evidence base for any future enforcement regime.

VII. EVIDENCE-BASED POLICY FRAMEWORK FOR GCC UNIVERSITIES

This section presents a four-tier policy framework for the responsible governance of AI detection tools in GCC universities, grounded directly in the synthesised evidence from Sections V and VI. Figure 2 illustrates the framework architecture, showing the four tiers and their sequential implementation relationship from foundational governance through to student and staff support. Each tier is grounded in specific evidence findings identified in Section V, with the connection between evidence and recommendation made explicit throughout. The score threshold framework is presented in Table II.

Four-Tier Evidence-Based Policy Framework for GCC Universities

Derived from a structured review of published independent evidence (2022–2026)

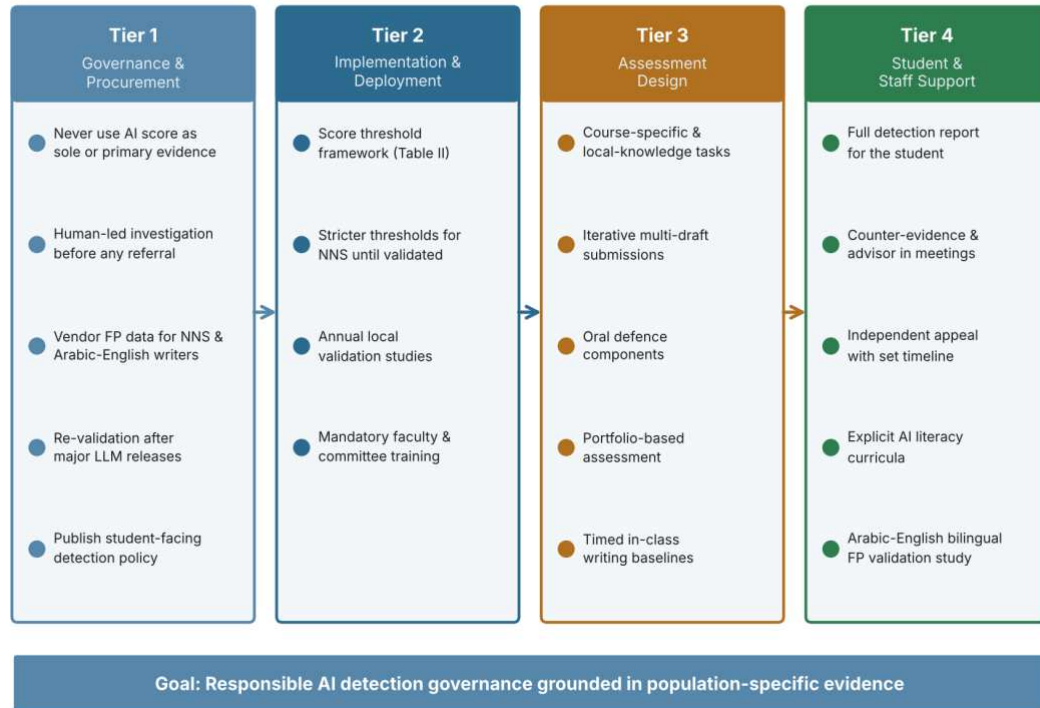


Fig. 2. Four-tier evidence-based policy framework for GCC universities, derived from the structured review of published independent evidence (2022–2026). Arrows indicate the progressive implementation sequence from Tier 1 governance and procurement through to Tier 4 student and staff support.

A. Tier 1: Governance and Procurement

The most immediate governance change required is the prohibition of AI detection scores as sole or primary evidence in academic misconduct proceedings. This recommendation is directly grounded in Finding 1 (no tool reached 80% accuracy), Finding 2 (potentially severe, population-dependent NNS bias), and Finding 6 (inconsistent outputs across tools): taken together, these findings establish that no AI detection score, regardless of its magnitude, meets the evidentiary standard required for disciplinary action without corroborating evidence. In practice, this means that a detection score should trigger a human-led investigation involving examination of the student's portfolio of writing for stylistic consistency, comparison with examination scripts or other in-person writing samples where available, a conversation with the student about their understanding of the submission, and review of any contextual factors including the student's language background and academic history. Only after this investigation should a misconduct referral be considered, and the detection score should constitute one input among many rather than the evidentiary foundation of the referral.

Institutions procuring or renewing AI detection tool contracts should require vendors to provide, as contractual conditions: validated false positive rates for NNS populations disaggregated by first language background, measured on corpora of at least 500 texts under conditions representative of the institution's student population; validated false positive rates specifically for Arabic-English bilingual writing for GCC institutions; methodology documentation sufficient to assess the validity of accuracy claims including corpus composition, threshold settings, and evaluation protocol; commitment to performance re-validation within 90 days of each major LLM model release; a process for reporting and investigating individual false positive claims by students; and disclosure of known performance degradation categories. Vendors that cannot provide this documentation should not be selected for institutional deployment. Institutions should also publish clear, student-accessible documentation specifying which tools are used, what score thresholds trigger investigation, what additional evidence is required before formal misconduct charges are filed, how students can request review of a detection finding, and what support is available during investigations.

B. Tier 2: Implementation and Deployment

Table II presents a score threshold framework for institutional use, derived from the accuracy and false positive data in Table I and Figure 1. The framework provides four score ranges with associated interpretations, required actions, and additional evidence requirements. All thresholds should be adjusted downward by at least one category for NNS students until NNS-specific false positive rates have been locally validated, reflecting Finding 2: NNS writers may face dramatically higher false positive rates than the general-population benchmarks on which these thresholds are based.

Table II
PROPOSED AI DETECTION SCORE THRESHOLD FRAMEWORK FOR GCC INSTITUTIONS

Score Range	Interpretation	Required Action	Additional Evidence Required
0–20%	Inconclusive	No action; document and file	None
21–50%	Low suspicion	Faculty review of submission context	Portfolio comparison recommended
51–80%	Moderate suspicion	Mandatory human-led investigation	Portfolio review plus in-person discussion required
81–100%	High suspicion	Full investigation procedure	Portfolio, exam comparison, interview; independent reviewer recommended

All thresholds should be adjusted downward by at least one category for NNS students until NNS-specific false positive rates are locally validated. No score range, including 81–100%, should trigger misconduct charges without corroborating evidence independent of the detection score.

Every institution should conduct periodic local validation studies testing deployed detection tools against locally-produced student writing samples with known authorship. The recommended methodology is to collect 100–200 student essays from previous semesters with confirmed human authorship (established through examination comparison or direct observation), submit these essays to all deployed detection tools, record the false positive rate for the institution’s specific student population, and use this data to calibrate institution-specific thresholds. This validation should be conducted annually and after any major tool update or new LLM model release, given the dynamic accuracy problem documented in Section VI-C. Faculty and academic integrity committee members who use or evaluate AI detection outputs should receive mandatory training covering the technical basis of AI detection and its known limitations, the documented NNS bias and its implications, the institution’s specific threshold policy, the evidentiary standards required for misconduct proceedings, and strategies for gathering corroborating evidence. Such training directly addresses the evidence-literacy gap identified in Section II-A.

C. Tier 3: Assessment Design

The most sustainable institutional response to AI-assisted academic misconduct is not technological enforcement but pedagogical redesign. Assessments that are difficult to complete using AI assistance are inherently more robust than those relying on detection for integrity, because they reduce the incentive for AI misuse rather than attempting to catch it after the fact. Consistent with TEQSA-commissioned guidance that trustworthy judgements about student learning require multiple, inclusive, and contextualised approaches to assessment [27], institutions can design assessments that incorporate course-specific content that AI systems cannot access, require iterative development through multiple drafts that make AI substitution visible through discontinuities in voice and understanding, include oral defence components that ensure students can demonstrate understanding of what they have submitted, and require reflection on personal experiences or local knowledge that AI cannot provide without substantial personalisation. Portfolio-based assessments tracking writing development across multiple submissions are particularly robust, as they make AI substitution visible through stylistic and conceptual discontinuities between drafts. Timed in-class writing assignments provide authenticated writing samples for comparison with take-home submissions, creating a human baseline for stylistic comparison that is more reliable than any detection tool output.

D. Tier 4: Student and Staff Support

Students must have clear, accessible, and documented rights to challenge AI detection findings. The institution’s appeal process should include access to the full detection report including the raw score, tool name, version, and any sentence-level annotations; a written explanation of the tool’s methodology in accessible non-technical language appropriate for students without computational backgrounds; the opportunity to present counter-evidence including portfolio writing samples, examination scripts, research notes, drafts, and other documentation of the writing process; the right to be accompanied by a student union representative or advisor during investigation meetings; and a documented decision timeline with the right of further appeal to an independent reviewer not involved in the original investigation. Institutions should also develop explicit AI literacy curricula that clearly define the distinction between

permitted and prohibited AI use, explain the academic rationale for these distinctions, and provide practical guidance on citation and disclosure of AI assistance where it is permitted. The most urgent research priority for Tier 4 is the first empirical study of Arabic-English bilingual false positive rates at GCC universities, using a corpus of 200–500 authentic student essays tested across all five target tools collaboratively across multiple GCC institutions. Without this study, all enforcement use of AI detection tools in GCC universities rests on extrapolations from studies of different linguistic populations.

VIII. LIMITATIONS AND FUTURE RESEARCH

A. Limitations of This Review

This review has three principal limitations that should inform the interpretation of its findings and recommendations. First, as a literature review rather than a primary empirical study, its findings are bounded by the scope and quality of the reviewed literature. The absence of Arabic-English bilingual validation data identified in Section V-E as Finding 4 with High confidence is the most significant gap, and the GCC-specific policy implications necessarily rely on extrapolation from studies of other NNS populations. This extrapolation is evidence-based and explicitly qualified throughout the paper, but it remains extrapolation: the magnitude of the false positive problem for Arabic-English bilingual writers at GCC universities is unknown until an empirical study is conducted. In addition, as a structured narrative review rather than a full systematic review, it may have missed relevant studies, and its confidence grades reflect the author’s judgement.

Second, the AI detection landscape is evolving rapidly in ways that may affect the generalisability of reported accuracy figures. Studies published in 2022 and 2023 evaluated tools against outputs from GPT-3.5 and early GPT-4; by March 2026, both the LLMs generating student text and the detection tools attempting to identify it had evolved substantially. The accuracy figures reported in Table I and illustrated in Figure 1 represent the state of the evidence as of the review period and may not reflect current performance. The recommendation for periodic local validation is specifically designed to address this instability: an institution that validates its deployed tools annually will have continuously updated performance data rather than relying on dated published benchmarks.

Third, the review focuses on five specific tools selected for their institutional prevalence as of early 2026. Other tools not examined here may have different performance profiles, and the field continues to evolve rapidly with new entrants. The policy recommendations are intentionally tool-agnostic where possible, designed to apply to any AI detection tool through the procurement criteria, threshold framework, and validation methodology described in Section VII. However, the specific accuracy figures cited in support of these recommendations are tool-specific and should not be generalised to tools not included in the review without independent validation.

B. Priority Research Agenda

Six research priorities are identified, ordered by urgency for the GCC context. The first and most urgent priority is an Arabic-English bilingual validation study: a collaborative study across multiple GCC universities using a corpus of 200–500 authentic student essays tested against all five target tools, producing the first validated false positive rates for Arabic-English bilingual writers. This study is the single most important empirical contribution that GCC researchers could make to the evidence base for responsible AI detection policy in the region, and it is the prerequisite for any evidence-based enforcement policy that does not rely on extrapolation from other linguistic contexts.

The second priority is a longitudinal accuracy tracking programme systematically monitoring AI detection performance as LLM capabilities evolve, with regular benchmarking against current-generation LLM outputs at intervals of no more than six months. The third priority is empirical research on GCC student AI use patterns—what tools students use, for what purposes, in what disciplines, and with what degree of post-generation modification—to inform more targeted and proportionate academic integrity responses that address the actual phenomenon rather than a hypothetical version of it. The fourth priority is extension of NNS validation studies to French-English, Urdu-English, Hindi-English, and Tagalog-English bilingual populations, representing other major student language groups in GCC universities. The fifth priority is longitudinal policy impact assessment evaluating academic integrity outcomes associated with different institutional approaches to AI governance. The sixth priority is participatory policy design research involving students as active participants in developing and evaluating AI use policies, improving both policy legitimacy and institutional understanding of student perspectives on fairness and due process.

IX. CONCLUSION

This paper has presented a structured review of published independent evidence on AI text detection accuracy and its implications for academic integrity policy in Gulf region universities. The evidence on accuracy is consistent: in the largest independent multi-tool evaluation, no detection tool reached 80% accuracy [1], and simple modification of AI-generated text sharply reduces detection [1], [10]. On bias, the most widely cited study found that detectors flagged an

average of 61.2% of authentic essays by non-native English writers as AI-generated [2], although later evidence from another language did not find such bias [3]. The risk to NNS students is therefore real but population-dependent—and for the Arabic-English bilingual writers who make up most of the GCC student body, it is entirely unmeasured.

For GCC universities, where Arabic-English bilingual students constitute the majority of the student body and where no validation study has been published for this specific population, these figures represent an acute and largely unacknowledged procedural fairness crisis. Universities across the region are making consequential disciplinary decisions that affect student academic records and career prospects using tools whose performance in their specific student population is entirely unknown. This is not a minor technical concern to be managed through policy caveats; it is a fundamental failure of evidence-based institutional governance that carries both legal and ethical risks. The evidence is clear that deploying these tools without population-specific validation for disciplinary purposes is unjustifiable given what is currently known and not known about their false positive rates for NNS populations.

This paper has addressed this gap through five original contributions. It provides the first structured synthesis of AI detection evidence for the GCC context, assembling and contextualising the evidence base that institutional decision-makers need but currently lack. It formally identifies and documents the Arabic-English bilingual research gap as the most urgent empirical priority for regional researchers. It develops a four-tier, evidence-grounded policy framework illustrated in Figure 2 and detailed in Section VII, covering governance and procurement, implementation and deployment, assessment design, and student and staff support. It provides a structured six-priority research agenda ordered by urgency for the GCC context. And it translates the synthesised evidence into practical implementation tools—the score threshold framework in Table II, vendor procurement criteria, and a local validation and staff training protocol, all set out in Section VII—that institutions can adopt directly.

The policy position of this paper is unambiguous. AI detection tools should not be used as primary or sole evidence in academic integrity proceedings at GCC universities until their performance against Arabic-English bilingual writing has been independently validated and found sufficient. This is not a position of institutional passivity in the face of AI misuse—it is a position of institutional responsibility in the face of documented tool unreliability. The appropriate response to the challenge of AI-assisted academic misconduct is to redesign assessments that make AI assistance less effective, educate students clearly about expectations and consequences, invest in the regional validation research that responsible governance requires, and develop human-centred investigation processes that treat students as people rather than probability scores. Accurate evidence is the foundation on which fair and effective policy must be built.

ACKNOWLEDGMENT

The author acknowledges the published researchers whose empirical work forms the evidence base of this review.

Funding: This research received no external funding.

Ethics declaration: Not applicable

Disclosure of interest: The author reports there are no competing interests to declare.

Data availability statement: No new data were generated or analysed in support of this research. This study is a review of previously published literature.

REFERENCES

- [1] D. Weber-Wulff et al., “Testing of detection tools for AI-generated text,” *Int. J. Educ. Integr.*, vol. 19, no. 1, art. no. 26, 2023, doi: 10.1007/s40979-023-00146-z.
- [2] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou, “GPT detectors are biased against non-native English writers,” *Patterns*, vol. 4, no. 7, art. no. 100779, 2023, doi: 10.1016/j.patter.2023.100779.
- [3] A. Al Ali, J. Helcl, and J. Libovický, “Different time, different language: Revisiting the bias against non-native speakers in GPT detectors,” in *Proc. EACL Student Research Workshop*, 2026. [Online]. Available: <https://arxiv.org/abs/2602.05769>
- [4] OpenAI, “Introducing ChatGPT,” *OpenAI Blog*, Nov. 2022. [Online]. Available: <https://openai.com/blog/chatgpt>
- [5] Turnitin, “Turnitin marks one year anniversary of its AI writing detector with millions of papers reviewed globally,” press release, Apr. 9, 2024. [Online]. Available: <https://www.turnitin.co.uk/press/turnitin-first-anniversary-ai-writing-detector>
- [6] B. Edwards, “Why AI detectors think the US Constitution was written by AI,” *Ars Technica*, Jul. 2023.
- [7] G. A. Fowler, “We tested a new ChatGPT-detector for teachers. It flagged an innocent student,” *The Washington Post*, Apr. 2023.
- [8] L. Dugan et al., “RAID: A shared benchmark for robust evaluation of machine-generated text detectors,” in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2024. [Online]. Available: <https://arxiv.org/abs/2405.07940>
- [9] W. H. Walters, “The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors,” *Open Inf. Sci.*, vol. 7, no. 1, art. no. 20220158, 2023, doi: 10.1515/opis-2022-0158.
- [10] M. Perkins, J. Roe, B. H. Vu, D. Postma, D. Hickerson, J. McGaughan, and H. Q. Khuat, “Simple techniques to bypass GenAI text detectors: Implications for inclusive education,” *Int. J. Educ. Technol. High. Educ.*, vol. 21, art. no. 53, 2024, doi: 10.1186/s41239-024-00487-w.

- [11] A. M. Elkhatat, K. Elsaid, and S. Almeer, "Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text," *Int. J. Educ. Integr.*, vol. 19, art. no. 17, 2023, doi: 10.1007/s40979-023-00140-5.
- [12] C. Stokel-Walker, "AI bot ChatGPT writes smart essays—should professors worry?," *Nature*, Dec. 2022, doi: 10.1038/d41586-022-04397-7.
- [13] L. Welding, "Half of college students say using AI on schoolwork is cheating or plagiarism," *BestColleges*, Mar. 17, 2023.
- [14] M. A. Cu and S. Hochman, "Scores of Stanford students used ChatGPT on final exams, survey suggests," *The Stanford Daily*, Jan. 22, 2023.
- [15] D. R. E. Cotton, P. A. Cotton, and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT," *Innov. Educ. Teach. Int.*, vol. 61, no. 2, pp. 228–239, 2024, doi: 10.1080/14703297.2023.2190148.
- [16] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-shot machine-generated text detection using probability curvature," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023.
- [17] I. Solaiman et al., "Release strategies and the social impacts of language models," *arXiv:1908.09203*, 2019.
- [18] University of Manchester, "Use of AI (artificial intelligence) detection tools in Turnitin," *Student News*, Apr. 2023. [Online]. Available: <https://studentnews.manchester.ac.uk/2023/04/05/use-of-ai-artificial-intelligence-detection-tools-in-turnitin/>
- [19] Vanderbilt University, "Guidance on AI detection and why we're disabling Turnitin's AI detector," *Brightspace Support*, Aug. 16, 2023. [Online]. Available: <https://www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector/>
- [20] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023.
- [21] J. Davalos and L. Yin, "AI detectors falsely accuse students of cheating with big consequences," *Bloomberg Businessweek*, Oct. 2024.
- [22] R. B. Kaplan, "Cultural thought patterns in inter-cultural education," *Lang. Learn.*, vol. 16, nos. 1–2, pp. 1–20, 1966.
- [23] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, "Can AI-generated text be reliably detected?," *arXiv:2303.11156*, 2023.
- [24] *Dixon v. Alabama State Board of Education*, 294 F.2d 150 (5th Cir. 1961).
- [25] *Goss v. Lopez*, 419 U.S. 565 (1975).
- [26] Poets & Quants, "Denying alleged AI use, EMBA student sues Yale SOM over year-long suspension," Feb. 25, 2025. [Online]. Available: <https://poetsandquants.com/2025/02/25/denying-alleged-ai-use-emba-student-sues-yale-som-over-year-long-suspension/>
- [27] J. M. Lodge, S. Howard, M. Bearman, P. Dawson, and Associates, *Assessment Reform for the Age of Artificial Intelligence*. Melbourne, VIC, Australia: Tertiary Education Quality and Standards Agency (TEQSA), 2023.
- [28] D. Dalalah and O. M. A. Dalalah, "The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT," *Int. J. Manag. Educ.*, vol. 21, no. 2, art. no. 100822, 2023.
- [29] H. M. Cooper, *Synthesizing Research: A Guide for Literature Reviews*, 3rd ed. Thousand Oaks, CA, USA: Sage, 1998.
- [30] J. Popay et al., "Guidance on the conduct of narrative synthesis in systematic reviews," *ESRC Methods Programme*, Lancaster University, 2006.