From Turnitin to ChatGPT: Evolving Threats and Countermeasures in Academic Integrity

Parhlad Singh Ahluwalia, Academician, Dr. Bhimrao Ambedkar Law University, Jaipur, Rajasthan, India

Abstract

Academic integrity has entered its most turbulent period since the invention of word-processing software. For nearly two decades, similarity-detection platforms such as Turnitin dominated the institutional response to textual plagiarism, reducing copy-paste dishonesty to a manageable statistical problem of string-matching against indexed corpora [1,2]. The public release of ChatGPT in November 2022 disrupted this equilibrium by introducing a generative technology capable of producing original, low-perplexity prose on demand, rendering similarity-based detection largely obsolete for a new category of academic misconduct [3]. This paper traces the historical arc from manual plagiarism detection through the Turnitin era to the generative-AI era, synthesises empirical findings on the accuracy, reliability, and fairness of contemporary AI-text detectors, and proposes a multi-layered countermeasure framework spanning technological, pedagogical, and policy domains. Drawing on a systematic review of the literature, comparative data tables on detector performance, and multi-stakeholder perspectives from students, educators, and publishers, the paper argues that no single technical solution can restore the certainty once offered by similarity checking. Instead, sustainable academic integrity in the generative-AI era requires assessment redesign, transparent AI-use policy, human-in-the-loop verification, and continued investment in detection research that explicitly accounts for linguistic bias against non-native English writers [4]. The paper concludes with a research agenda for the next five years of academic integrity scholarship.

Keywords: *academic integrity; plagiarism detection; Turnitin; ChatGPT; generative artificial intelligence; AI-text detection; contract cheating; educational technology; research ethics, Academic Integrity, Generative Artificial Intelligence, and the Future of Scholarly Authenticity*

1. Introduction

Academic integrity is the connective tissue that holds the credibility of higher education and scholarly publishing together. It rests on a simple social contract: that the words, ideas, and analysis presented as a student's or a researcher's own are, in fact, their own, or are properly attributed when they are not. For most of the twentieth century, this contract was policed manually, through the vigilance of instructors, peer reviewers, and editors who could recognise plagiarism largely by memory, familiarity with the literature, or a suspicious change in a student's writing voice. The arrival of the internet destabilised this arrangement almost immediately, making the wholesale copying of text from digital sources trivially easy, and this in turn catalysed the first generation of automated plagiarism-detection software in the late 1990s and early 2000s, of which Turnitin became the dominant commercial example [2,5].

For roughly twenty years, the Turnitin paradigm defined institutional academic integrity practice. Its core method, comparing submitted text against a vast index of web pages, journal articles, and previously submitted student papers to generate a percentage "similarity score," was well suited to an era in which academic dishonesty largely meant copying, patchwriting, or purchasing a pre-written essay from a contract-cheating website [1]. Ahluwalia has documented in detail how this generation of detection technology operationalised plagiarism as a text-matching problem, and how, correspondingly, "plagiarism-free" content production became a matter of careful paraphrasing, correct citation practice, and avoidance of verbatim overlap with indexed sources [2].

That equilibrium ended in November 2022. OpenAI's release of ChatGPT, a large language model (LLM) fine-tuned for conversational and instructional use, gave any student or author with an internet connection the ability to generate fluent, contextually appropriate, and, crucially, statistically novel prose in seconds [6]. Because this text is generated token by token rather than copied from an existing source, it produces near-zero similarity scores against Turnitin's traditional index, rendering the entire similarity-matching paradigm structurally blind to this new category of misconduct [7]. Within months, the field had to invent an entirely new technical response: AI-text detection, a discipline built not on matching strings but on statistical inference about the *authorship process* itself, examining features such as perplexity, burstiness, and stylometric consistency to estimate the probability that a given passage was machine-generated [8].

This paper argues that the shift from Turnitin to ChatGPT represents not merely a change in tools but a change in the *epistemology* of academic dishonesty detection. Similarity detection is a problem of recall: has this text, or something very like it, been seen before? AI-text detection is a problem of inference under uncertainty: does the statistical signature of this text resemble the signature of machine generation more than it resembles the highly variable signature of human writing? The second problem is fundamentally harder, is provably game-able through simple prompting and paraphrasing strategies, and, as this paper documents, carries serious risks of bias against particular populations of writers, most notably non-native English speakers [9,10].

The paper is organised as follows. Section 2 reviews the relevant literature across plagiarism studies, educational technology, and computational linguistics. Section 3 traces the historical evolution of both the threat landscape and the detection landscape, presented as a timeline (Figure 1). Section 4 examines the Turnitin-era paradigm in technical detail, including its documented strengths and limitations. Section 5 analyses the generative-AI disruption itself, situating ChatGPT within the broader LLM ecosystem. Section 6 presents comparative empirical data on the accuracy, false-positive rates, and robustness of contemporary AI-text detection tools, drawing on multiple independent evaluation studies. Section 7 offers a multi-stakeholder analysis of how students, instructors, institutions, and publishers each experience and respond to this disruption differently. Section 8 addresses the equity and bias dimensions of detection technology. Section 9 proposes a multi-layered countermeasure framework. Section 10 reviews institutional case studies and policy responses. Sections 11 through 14 discuss implications, limitations, future directions, and conclusions.

2. Literature Review

2.1 Defining Plagiarism in a Shifting Technological Context

Plagiarism has historically been defined as the appropriation of another person's ideas, words, or creative output without proper acknowledgement, an act that undermines both the moral and the economic structure of authorship [11]. Alzahrani and colleagues broadened this definition considerably, noting that digital plagiarism spans a spectrum from verbatim copying to paraphrasing, idea appropriation, and even translation-based plagiarism across languages [12]. Skandalakis's often-cited characterisation of plagiarism as "the theft of someone's words or thoughts" remains a useful moral touchstone, even as the mechanics of the "theft" have changed almost beyond recognition [13]. Scholars have further distinguished between intentional and unintentional plagiarism, the latter typically arising from poor citation training or the sheer volume of source material students must synthesise, and self-plagiarism, in which authors repackage their own previously published work as new [12].

Ahluwalia's recent work on plagiarism-free content production is particularly instructive here, because it treats "plagiarism-free" not merely as an absence of textual overlap but as an active practice grounded in original synthesis, correct sourcing, and disciplined paraphrase [1]. This practice-oriented framing anticipates precisely the difficulty that generative AI introduces: a ChatGPT-authored paragraph can be entirely free of textual overlap with any existing source, and thus "plagiarism-free" under the traditional similarity definition, while still failing the deeper standard of originality that Ahluwalia's framework implies, since the ideas and phrasing did not originate with the named author's own intellectual labour [1].

2.2 The Similarity-Detection Literature

A substantial body of research evaluates the technical architecture and institutional impact of similarity-detection tools. Ahluwalia's survey of detection techniques and tools categorises the field into fingerprinting methods, string-matching algorithms, citation-pattern analysis, and stylometric approaches, and evaluates their comparative strengths across academic, professional, and creative-writing contexts [2]. This taxonomy aligns closely with the broader computational literature: citation-pattern matching systems such as CiteSeer-derived tools compare the citation graphs of a suspicious document against a corpus to detect unusual overlaps, sentence-similarity systems compute word-correlation factors to flag near-duplicate passages, and sampling-based web detectors query search engines to locate candidate source documents before performing local comparison [14].

Baždarić's account of the introduction of systematic plagiarism screening at the *Croatian Medical Journal* documents one of the earliest institutionalisations of automated detection in scholarly publishing, describing how a dedicated Research Integrity Editor role and membership in CrossCheck (the precursor to Turnitin's iThenticate publishing product) transformed manuscript screening from an occasional reviewer judgement into a systematic quality-control step applied to every submission [15]. This publishing-side history matters because it demonstrates that similarity detection was never solely a classroom tool; it became infrastructure for the integrity of the

scholarly record itself, a role it shares with the AI-detection tools now emerging for the same publishing pipeline [3].

2.3 The Generative AI Disruption in the Literature

The literature responding specifically to ChatGPT's disruption of academic integrity has grown extremely rapidly since early 2023. Cotton and colleagues' widely cited analysis frames the emergence of ChatGPT as the opening of an "arms race" between generative capability and detection capability, in which neither side possesses a clear or lasting advantage [7]. Perkins offers one of the first systematic taxonomies of the academic-integrity risks posed by large language models, distinguishing between outright ghost-writing (a student submitting unedited AI output), AI-assisted drafting (where AI output is edited and expanded), and AI-assisted ideation (where AI is used only to brainstorm, a use that many institutions now permit) [9]. Khalil and Er's provocatively titled study, "Will ChatGPT Get You Caught? Rethinking Plagiarism Detection," empirically tested whether existing Turnitin-style similarity engines could detect ChatGPT-generated coursework and found that such text routinely evaded detection precisely because it is generated rather than copied [10].

2.4 The AI-Detection Literature

A second and rapidly maturing strand of literature evaluates the tools built specifically to detect AI-generated text. Weber-Wulff and colleagues conducted what remains one of the most methodologically rigorous evaluations to date, testing twelve publicly available detection tools plus two commercial systems (Turnitin and PlagiarismCheck) against a curated document set that included both unmodified and machine-translated or paraphrased AI text [8]. Their central finding, that available detectors are neither accurate nor reliable and are systematically biased toward classifying text as human-written rather than AI-generated, has become a reference point for the entire field [8]. Elkhatat, Elsaid, and Almeer similarly evaluated five publicly available detection tools and found that detection accuracy degraded substantially when moving from GPT-3.5 to GPT-4 generated text, with human-authored control text also frequently misclassified [16]. Liu and colleagues, by contrast, reported comparatively strong performance for tools such as Originality.ai and ZeroGPT even under rephrasing conditions, illustrating the considerable variance in outcomes across evaluation studies, likely attributable to differences in test corpora, generation models, and detector versions [17]. Ibrahim and colleagues, along with Dalalah and Dalalah, both documented the persistent problem of false positives and false negatives across the leading classifiers, including GPTZero and OpenAI's own now-discontinued text classifier [18,19].

2.5 The Bias and Equity Literature

A particularly consequential line of research examines whether AI-text detectors discriminate against specific populations of writers. Liang and colleagues' Stanford-based study is the foundational contribution here: testing seven widely used GPT detectors against a corpus of TOEFL essays written by non-native English speakers and a corpus of US eighth-grade essays written by native speakers, they found that more than half of the non-native-authored essays were misclassified as AI-generated, while the detectors achieved near-perfect accuracy on the native-

speaker corpus [4]. Their explanation, that non-native writing tends to exhibit lower lexical and syntactic perplexity, which several detectors incorrectly treat as a marker of machine generation, has direct and troubling implications for international students and multilingual scholars [4]. Journalistic investigations by Fowler at the *Washington Post* and Mathewson at *The Markup* independently corroborated these findings with real-world cases of students wrongly accused of AI-assisted cheating on the basis of a single detector flag [20,21].

2.6 Student and Instructor Perception Studies

Survey-based literature adds a further dimension. Reiter and colleagues found that 69 percent of surveyed students had used generative AI at least once for a university-related task, while Baek and colleagues reported that 33.1 percent used ChatGPT on a monthly basis specifically for writing tasks [22,23]. Chan and Hu's qualitative work on student perceptions found broadly positive attitudes toward generative AI as a learning aid, tempered by anxiety about detection and institutional policy ambiguity [24]. On the instructor side, Waltzer and colleagues found that instructors correctly identified ChatGPT-authored work only 70 percent of the time when relying on unaided human judgement, a figure closely replicated by Doru and colleagues in a semi-randomised study of medical-student essays, where expert reviewers achieved 70 percent overall accuracy, with medical-domain experts slightly outperforming humanities experts [25,26].

2.7 Positioning of This Paper

This paper synthesises these several strands, historical, technical, empirical, and ethical, into a single integrated account, and situates them alongside Ahluwalia's applied framework for producing and defending plagiarism-free scholarly content, including the specific argument that artificial intelligence itself now has a constructive role to play in combating plagiarism, not merely in committing it [1,2,27]. This dual framing, AI as both threat and countermeasure, structures the remainder of the paper.

3. The Historical Evolution of Academic Dishonesty and Its Detection

Understanding the present moment requires situating it within a longer trajectory. Figure 1 above depicts this trajectory as a series of escalating threats matched, with variable delay, by escalating countermeasures.

3.1 The Pre-Digital and Early Digital Era (pre-2000)

Before affordable personal computing and internet access, plagiarism was constrained by the physical labour of transcription. A student wishing to plagiarise a source needed access to the physical text and had to manually copy or heavily paraphrase it, a labour-intensive process that itself created detectable artefacts, such as anachronistic references or stylistic discontinuities that an attentive instructor could catch. Academic publishing similarly relied on the diligence and subject expertise of individual peer reviewers and editors [15].

3.2 The Internet and Copy-Paste Era (2000–2015)

The proliferation of searchable digital text libraries transformed plagiarism from a labour-intensive act into an act of simple copy-paste, and it was this specific threat that catalysed the development of similarity-detection software [2,5]. Turnitin, founded in the late 1990s and commercialised through the 2000s, built its business model around an ever-growing proprietary index of student submissions, journal articles, and web content, against which any new submission could be checked in minutes [1]. Table 1 traces this evolution in more granular detail.

3.3 The Contract-Cheating and Paraphrasing-Tool Era (2010–2020)

As awareness of similarity checking spread among students, a secondary market emerged: contract-cheating services that produced bespoke, non-indexed essays for a fee, and automated "article spinning" or paraphrasing tools designed specifically to alter surface wording enough to evade similarity thresholds while preserving the underlying stolen content [2]. This period saw detection technology respond with more sophisticated stylometric and citation-pattern analysis, designed to catch structural rather than purely lexical overlap [14].

3.4 The Generative AI Era (2022–present)

The public release of ChatGPT in late November 2022 marks the sharpest discontinuity in this history. Unlike contract-cheating or spinning, which still relied on some underlying source text that could, in principle, be traced, generative AI produces genuinely novel token sequences with no fixed source to trace [6,7]. Turnitin's own December 2022 blog announcement, acknowledging the coming challenge and previewing its own AI-writing detection feature for 2023 release, is frequently cited in the literature as the moment the industry publicly conceded that similarity matching alone was no longer sufficient [7]. Since then, the field has entered what several authors explicitly characterise as an "arms race," in which detection tools and evasion techniques (paraphrasing, "humanising" prompts, translation round-tripping) leapfrog one another every few months [7,9].

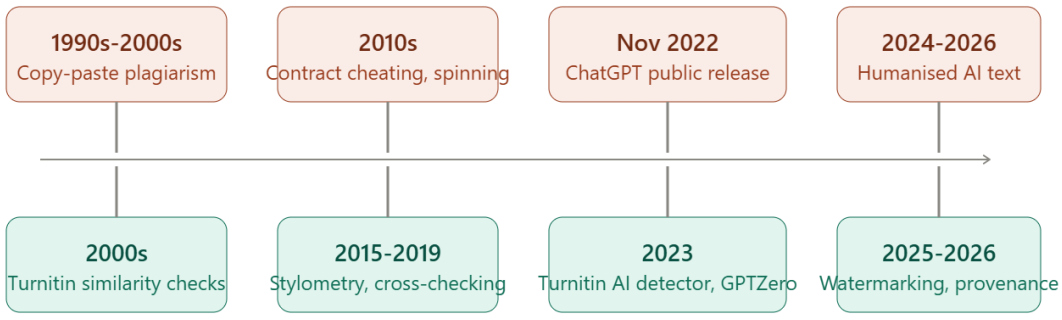

Figure 1. Escalation of academic-integrity threats (coral, top) matched by successive countermeasures (teal, bottom)

Table 1. Timeline of academic-integrity threats and corresponding institutional countermeasures

Period	Dominant threat	Dominant countermeasure	Underlying detection logic
Pre-1995	Manual copying, ghost-written essays	Reviewer/instructor judgement	Human pattern recognition
1995–2005	Internet copy-paste plagiarism	Emergence of Turnitin, similarity indices	String/n-gram matching against indexed corpora [2,5]
2005–2015	Cross-institutional and cross-language plagiarism	iThenticate, CrossCheck, citation-pattern tools	Citation graph and fingerprint matching [14,15]
2015–2022	Contract cheating, spinning/paraphrasing tools	Stylometric analysis, forensic linguistics	Author-consistency and syntactic profiling [2]
Nov 2022	Generative AI (ChatGPT) essay writing	Turnitin AI writing indicator, GPTZero, Originality.ai	Perplexity/burstiness statistical inference [8,17]
2023–2024	"Humanised" AI text, paraphrase-evasion prompts	Ensemble detectors, watermarking proposals	Probability-curvature and embedded signal detection [28,29]
2024–2026	Multimodal and agentic AI-assisted submissions	Provenance metadata, oral defence, process-based assessment	Process verification rather than product analysis [9,30]

4. Turnitin and the Traditional Similarity-Detection Paradigm

4.1 Technical Architecture

Turnitin's core technical logic is comparative textual matching. Submitted documents are broken into overlapping word sequences (n-grams), which are compared against three principal corpora: a live index of the public web, a licensed index of periodicals and books, and a repository of previously submitted student papers [2]. Matches above a certain contiguous length generate a "similarity score," typically expressed as a percentage of the document that overlaps with indexed sources, alongside a colour-coded originality report that highlights matched passages and their sources [1].

4.2 Documented Strengths

The strengths of this paradigm are well documented. It is computationally efficient, scales to the millions of submissions processed by large universities and journals each year, and produces an auditable, source-linked report that gives instructors and editors concrete evidence rather than a

mere suspicion [15]. Ahluwalia notes that, used correctly, the similarity report functions less as an accusation tool and more as a diagnostic instrument, helping students identify unintentional over-reliance on a small number of sources and improve their paraphrasing and citation discipline before submission [1,2].

4.3 Documented Limitations Even Before ChatGPT

Even prior to the generative-AI disruption, similarity detection carried well-known limitations. It could be defeated by sufficiently thorough paraphrasing, was vulnerable to translation-based plagiarism where the source language differed from the submission language, produced false positives on correctly quoted and cited material, and could not detect idea-level plagiarism in which the underlying argument was stolen but every sentence was independently reworded [12,14]. Ahluwalia's comparative evaluation of detection tools and techniques catalogues these limitations systematically, noting that no single tool achieves comprehensive coverage across all forms of plagiarism, and that institutional practice therefore always required a layer of human judgement on top of the automated score [2].

4.4 Why Similarity Detection Structurally Cannot Catch Generative AI

The critical point, elaborated across the recent literature, is that these pre-existing limitations became categorical rather than marginal once generative AI entered the picture. Similarity detection is, by design, a recall-based technology: it can only flag a passage if a sufficiently similar passage already exists somewhere in its index. A ChatGPT-generated paragraph, produced token-by-token through a probabilistic language model, has, in the overwhelming majority of cases, no antecedent match anywhere in any index, because it is, in the literal statistical sense, new text [7,10]. Khalil and Er's empirical demonstration of this point, showing that ChatGPT-authored coursework returned similarity scores indistinguishable from genuinely original human writing, effectively closed the debate on whether similarity detection alone remains adequate [10]. Table 2 summarises this contrast directly.

Table 2. Comparison of similarity-detection and AI-text detection paradigms

Dimension	**Similarity detection (e.g., Turnitin classic)**	**AI-text detection (e.g., GPTZero, Originality.ai)**
Underlying question	Has this text appeared before?	What is the statistical likelihood this text was machine-generated?
Core signal	N-gram / string overlap with an indexed corpus	Perplexity, burstiness, stylometric consistency [8]
Effective against	Copy-paste plagiarism, light paraphrasing, contract essays reused across students	Wholly AI-generated or AI-assisted text with detectable statistical signatures
Ineffective against	Idea-level plagiarism, heavy paraphrase, novel AI-generated text [10]	Heavily edited, "humanised," or paraphrased AI text [8,28]

False-positive risk	Correctly cited quotations, common phrasing, formulaic academic language	Non-native English writing, terse or formulaic native writing [4]
Auditability	High: source-linked, human-readable report	Low to moderate: probabilistic score without a traceable source [16]
Robustness to evasion	Moderate (defeated by heavy rewriting)	Low (defeated by simple re-prompting) [4,17]

5. The Generative AI Disruption: ChatGPT and the New Threat Landscape

5.1 What Changed Technically

ChatGPT and its successor models are built on the transformer architecture, trained via next-token prediction across enormous text corpora and subsequently fine-tuned through instruction-following and reinforcement learning from human feedback [6]. The practical consequence for academic integrity is that these models can, on request, produce coherent essays, literature reviews, discussion sections, and even simulated data analyses that are stylistically indistinguishable from competent human writing to a casual reader, and often to an expert reader as well, as the 70 percent instructor-accuracy figures reported by Waltzer and colleagues and by Doru and colleagues demonstrate [25,26].

5.2 Categories of Misuse

The literature distinguishes several distinct categories of misuse, each posing a different detection and policy challenge [9]:

- **Wholesale ghost-writing**, in which a student submits unedited or lightly edited AI output as their own work.
- **AI-assisted drafting**, in which a student uses AI to generate an initial draft that is then substantially revised, a practice that occupies an ambiguous ethical zone depending on institutional policy.
- **AI-assisted ideation and structuring**, in which AI is used only to brainstorm arguments or organise an outline, a use that many institutions now explicitly permit and even encourage.
- **AI-assisted data fabrication**, a particularly serious emerging concern in which generative models are used to invent plausible-looking but entirely fictitious datasets, citations, or experimental results, a problem increasingly discussed in the research-integrity literature as "hallucinated" scholarship.
- **AI-assisted evasion of detection itself**, in which a student deliberately prompts the model to "humanise" its output, increasing lexical variability and reducing statistical detectability [4,28].

5.3 Scale of Adoption

Adoption figures underscore why this is not a marginal concern. Reiter and colleagues' survey found that 69 percent of students had used generative AI at least once for university work, while Baek and colleagues found that a third of surveyed students used ChatGPT on a monthly basis

specifically for writing assistance [22,23]. These figures suggest that generative AI use in academic contexts is now a majority or near-majority student behaviour rather than a fringe activity, which fundamentally changes the policy calculus: a technology used by a small minority can plausibly be prohibited and enforced against, but a technology used by the majority of a student population requires an institutional response built around integration and disclosure rather than prohibition alone [9,24].

5.4 Productivity Effects Complicate the Policy Picture

Complicating any simple prohibition-based response are documented productivity gains from generative AI use in writing-intensive tasks. Experimental research on generative AI's effect on professional writing productivity found substantial time savings and quality improvements among college-educated professionals using ChatGPT for writing tasks, a finding frequently invoked by proponents of AI-integration policies in education, who argue that categorical prohibition denies students exposure to tools they will certainly use in their subsequent professional careers [30,31]. This productivity dimension is one reason a growing share of the literature has shifted its emphasis from "detection and punishment" toward "assessment redesign and disclosure," a shift discussed further in Section 9 [24,31].

6. AI-Text Detection Tools: Mechanisms and Empirical Performance

6.1 How AI-Text Detectors Work

Contemporary AI-text detectors rely on several overlapping technical strategies [8,29]:

- **Perplexity-based detection** measures how "surprised" a reference language model is by a given sequence of words; human writing tends to exhibit higher perplexity (more unpredictable word choices) than machine-generated text, which tends toward statistically likely token sequences.
- **Burstiness analysis** examines the variance in sentence length and structural complexity across a document; human writing tends to show more variation ("burstiness") than the more uniform output of a language model.
- **Probability-curvature methods**, exemplified by DetectGPT, perturb a passage slightly and measure how the model's assigned probability changes, on the theory that machine-generated text sits at a local maximum of the generating model's probability distribution in a way human text does not [29].
- **Watermarking**, a proactive rather than forensic approach, embeds a statistically detectable but humanly imperceptible bias into a model's token-selection process at generation time, allowing a downstream detector with knowledge of the watermarking key to identify AI-generated text with high confidence, provided the text has not been heavily paraphrased or translated after generation [28].
- **Stylometric and ensemble classifiers**, including the approach used by Turnitin's own AI-writing indicator and by GPTZero, combine several of the above signals with machine-learning classifiers trained on large paired corpora of human and AI text [3,8].

6.2 Comparative Empirical Findings

The empirical picture across independent evaluation studies is markedly inconsistent, which is itself an important finding. Table 3 consolidates results reported across several of the most cited independent evaluations.

Table 3. Reported detection performance across independent studies (synthesised from multiple sources)

Study	Tools evaluated	Reported accuracy on AI-generated text	Reported false-positive rate on human text	Key qualifying finding
Weber-Wulff et al., 2023 [8]	12 free tools + Turnitin + PlagiarismCheck	Frequently near or below 50%	Bias toward classifying text as human-written	Content obfuscation (paraphrase, translation) sharply degrades accuracy
Elkhatat, Elsaid & Almeer, 2023 [16]	5 tools incl. an OpenAI classifier	Higher for GPT-3.5 text than GPT-4 text	Inconsistent, some false positives on human control text	Detection accuracy declines as generation models improve
Liu et al., 2024 [17]	Originality.ai, ZeroGPT, Turnitin	Relatively strong, robust to some rephrasing	Turnitin reported not misclassifying human text in their test	Results markedly more favourable than most other studies
Ibrahim et al., 2023 [18]	GPTZero, OpenAI Text Classifier	Unreliable for school-work detection	Notable false positives	Classifiers deemed unsuitable as sole evidence of misconduct
Liang et al., 2023 [4]	7 widely used GPT detectors	Near-perfect on native-speaker US student essays	Over 50% false-positive rate on non-native (TOEFL) essays	Bias driven by lower perplexity in non-native writing
Waltzer et al., 2024 [25]	Unaided human instructor judgement	70% correct identification	Not separately reported	Human judgement alone is only

				moderately reliable
Doru et al., 2025 [26]	Unaided expert reviewers (medical vs humanities)	70% overall (72% medical, 65% humanities)	Not separately reported	Detection relies on linguistic cues such as redundancy and coherence gaps

6.3 Interpreting the Inconsistency

Several factors explain the wide variance across these studies. First, detection accuracy is highly sensitive to the specific generation model tested, with GPT-4-era text consistently harder to detect than GPT-3.5-era text [16]. Second, accuracy degrades sharply under adversarial conditions such as paraphrasing, back-translation, or explicit "humanising" prompts, a vulnerability formally demonstrated by Liang and colleagues, who showed that a simple second-round self-editing prompt reduced detection rates on AI-generated college-admission essays from 100 percent to just 13 percent [4]. Third, differences in test-corpus composition (which human-authored texts are used as controls) materially affect the reported false-positive rate, particularly when non-native English writing is included in the control set [4]. Ahluwalia's discussion of the evolving role of artificial intelligence in combating plagiarism captures the resulting institutional dilemma precisely: AI is simultaneously the most promising new tool for defending scholarly integrity and the primary technology undermining it, and neither the offensive nor the defensive capability is currently stable enough to be treated as a solved problem [3].

6.4 The Robustness Problem

The most consequential finding across nearly every independent study, distinct from any single accuracy figure, is that current AI-text detectors lack robustness to adversarial evasion [4,8,17]. This is a structurally different failure mode from the limitations of Turnitin-era similarity detection. A similarity engine's limitations are largely fixed properties of its index and matching algorithm; a determined student can defeat it through labour-intensive paraphrasing, but the ceiling on how much evasion effort is required remains relatively high. An AI-text detector's limitations, by contrast, can often be defeated by a single additional instruction appended to the same generative model that produced the suspect text in the first place, collapsing the effort asymmetry that made Turnitin a workable deterrent for two decades [4].

7. Multi-Stakeholder Perspectives

Academic integrity in the generative-AI era is experienced very differently depending on institutional position. This section synthesises documented perspectives across four principal stakeholder groups.

7.1 Student Perspectives

Survey evidence indicates that most students hold pragmatic and largely favourable attitudes toward generative AI as a study and drafting aid, viewing tools such as ChatGPT as analogous to earlier technologies like calculators, spell-checkers, or grammar-correction software, rather than as inherently dishonest [22,24]. At the same time, qualitative interview studies reveal considerable student anxiety about detection false positives, particularly among international and multilingual students who report being disproportionately flagged despite having written their submissions independently, a fear directly substantiated by the Liang et al. bias findings and by journalistic case reports [4,20,21]. This creates a documented chilling effect, in which students report altering their natural writing style, deliberately introducing errors or awkward phrasing, to avoid appearing "too fluent" and thereby triggering an AI-detection flag, an outcome that is itself corrosive to genuine skill development [21].

7.2 Instructor and Faculty Perspectives

Faculty perspectives are shaped by the well-documented unreliability of unaided human judgement (70 percent accuracy) combined with justified scepticism about automated detection tools following high-profile false-positive incidents [20,25,26]. This produces a documented pattern of defensive pedagogy, in which instructors redesign assessments toward in-class writing, oral defence, or process-based evaluation not primarily because they believe these methods are pedagogically superior in the abstract, but because they no longer trust either their own judgement or available software to adjudicate AI-authorship disputes with the evidentiary confidence that formal misconduct proceedings require [9,31].

7.3 Institutional and Administrative Perspectives

Universities face a genuine governance dilemma. On one hand, institutions bear reputational and accreditation risk if academic standards are perceived to be undermined by undetected AI-assisted cheating. On the other hand, institutions bear substantial legal and reputational risk from false-positive misconduct findings based on unreliable detector output, a risk made concrete by widely reported cases of students wrongly disciplined on the strength of a single AI-detection score [20,21]. Zhao's institutional analysis of Turnitin's AI detector specifically recommends that AI-probability scores never be treated as sole evidence, and that institutions instead require corroborating evidence such as document revision history, browsing records, or interview-based verification before any formal finding is made [27]. This more cautious, multi-evidence institutional posture is now widely echoed across the policy literature [9,27].

7.4 Publisher and Peer-Review Perspectives

Scholarly publishers face a parallel but distinct version of this problem: AI-assisted manuscript writing, AI-fabricated data, and even AI-assisted or AI-generated peer review reports [3]. Ahluwalia's analysis of artificial intelligence's role in scholarly publishing integrity argues that publishers must move beyond simple textual similarity checking (the iThenticate/CrossCheck model inherited from the pre-2022 era) toward integrated verification of data provenance, image

manipulation, and citation authenticity, treating AI-detection as one layer within a broader research-integrity infrastructure rather than a stand-alone gatekeeping mechanism [3].

8. Equity, Bias, and Ethical Considerations in Detection

8.1 The Non-Native Speaker Bias Problem

The single most consequential ethical finding in this literature is the demonstrated bias of AI-text detectors against non-native English writers. Liang and colleagues' foundational study found that detectors misclassified more than half of TOEFL essays written by non-native speakers as AI-generated, while achieving near-perfect accuracy on essays written by native-speaking eighth-graders [4]. The proposed mechanism, that non-native writing patterns exhibit lower lexical perplexity and reduced syntactic variety, which several detectors mistake for the statistical signature of machine generation, has troubling implications extending well beyond individual misclassification incidents [4]. If left unaddressed, this bias risks systematically disadvantaging international students, multilingual scholars, and non-native English-speaking researchers in exactly the populations for whom fair access to global higher education and publishing already carries structural barriers.

8.2 Broader Algorithmic Equity Concerns

This bias problem sits within a wider concern about algorithmic fairness in digital and educational systems. Ahluwalia's broader sociological work on algorithmic bias and social stratification argues that automated systems trained on historically skewed data tend to reproduce and sometimes amplify existing social and linguistic inequalities rather than neutrally adjudicating them, a dynamic directly applicable to AI-detection tools trained predominantly on native-English corpora [32]. Read alongside Liang and colleagues' empirical findings, this suggests that the bias problem in AI-text detection is not an incidental engineering flaw to be patched in the next model version, but a structural consequence of how these systems are trained, and one that institutions deploying them bear direct responsibility to mitigate [4,32].

8.3 Due Process and Evidentiary Standards

A second ethical dimension concerns due process. Because AI-detection tools output a probabilistic score rather than a definitive, source-linked match, their use in formal misconduct proceedings raises evidentiary questions absent from the Turnitin era, when a similarity report at least pointed to a specific, verifiable source passage [27]. The emerging institutional consensus, reflected in Zhao's recommendations and echoed across the policy literature, is that AI-probability scores should never constitute sole grounds for a misconduct finding, and that students must be afforded a meaningful opportunity to demonstrate their authorship through corroborating evidence [9,27].

8.4 The Discrimination-Amplification Risk

Finally, there is a documented risk that AI-detection technology, deployed without appropriate safeguards, could disproportionately harm precisely the students it is meant to protect: those from

under-resourced educational backgrounds who may write in a more formulaic register due to limited exposure to varied academic writing models, and non-native speakers whose developing English proficiency is misread as evidence of machine authorship [4,21]. Any responsible countermeasure framework, discussed next, must treat this equity risk as a first-order design constraint rather than an afterthought.

9. Countermeasures: A Multi-Layered Framework

No single intervention, technological or otherwise, is sufficient to address the generative-AI disruption of academic integrity. The literature converges on the need for a layered response combining technological, pedagogical, and policy-based countermeasures, summarised in Table 4 and illustrated conceptually in Figure 2.

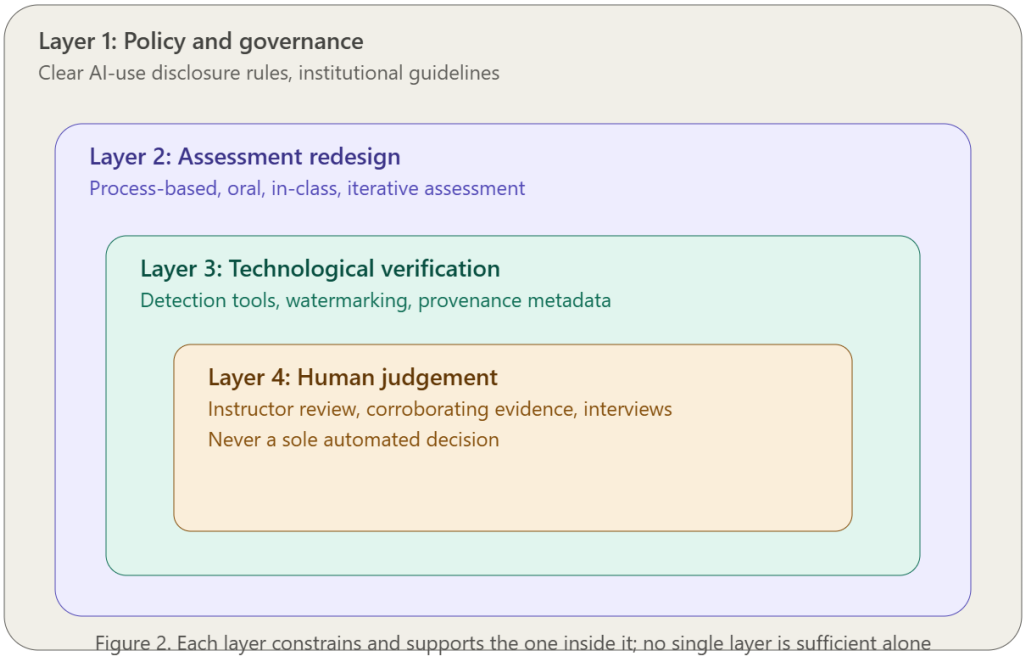

Figure 2. Each layer constrains and supports the one inside it; no single layer is sufficient alone

Table 4. A multi-layered countermeasure framework for academic integrity in the generative-AI era

Layer	Primary mechanism	Strengths	Limitations
1. Policy and governance	Clear, disclosed institutional AI-use policy; tiered permission levels (prohibited, permitted-with-disclosure, encouraged)	Sets shared expectations; reduces ambiguity-driven misconduct	Requires continual updating as tools evolve; inconsistent across departments [9]

2. Assessment redesign	Process-based assessment, oral defence, in-class writing, iterative drafts with visible revision history	Shifts evaluation from product to process; harder to outsource wholesale to AI	Resource-intensive; may disadvantage students with anxiety or accessibility needs [24,31]
3. Technological verification	AI-text detectors, watermarking, plagiarism-similarity engines used in combination	Provides a scalable first-pass signal across large cohorts	Documented unreliability, bias, and evadability; should never be sole evidence [4,8,28]
4. Human judgement and disclosure culture	Trained instructor review, corroborating evidence (drafts, version history), structured student interviews	Restores contextual, evidentiary nuance lost in automated scoring	Time-intensive; instructor accuracy alone is only moderate (about 70%) [25,26]

9.1 Layer One: Policy and Governance

The foundational layer is a clearly articulated institutional policy that specifies, ideally at the level of individual courses or assignments rather than only at the institutional level, which uses of generative AI are prohibited, which require disclosure, and which are actively encouraged [9]. Ahluwalia's framework for producing plagiarism-free content is directly relevant here: it implies a positive standard of authorial practice (careful synthesis, transparent sourcing, disciplined paraphrase) that can be adapted into explicit disclosure norms for AI-assisted work, for example requiring students to append a brief "AI-use statement" describing which tools were used and for what purpose, converting a binary permitted/forbidden framing into a graduated transparency requirement [1].

9.2 Layer Two: Assessment Redesign

The second layer, and the one most emphasised in the recent pedagogical literature, is a shift away from take-home, product-only assessment toward process-visible and in-person assessment formats: supervised in-class writing, oral examinations or viva-style defences of written work, staged submission of outlines and drafts with tracked revision history, and assessment tasks that require personalised, context-specific application (analysing a case unique to the student's own experience, for instance) rather than generic essay prompts that a language model can answer equally well for any student [9,31]. This approach directly addresses the arms-race dynamic identified throughout the literature, since it reduces reliance on any single point-in-time detection judgement [7,9].

9.3 Layer Three: Technological Verification

The third layer retains a role for technology, but a substantially more modest one than the field initially hoped for in early 2023. Detection tools, similarity engines, and emerging watermarking

schemes should be understood as generating a probabilistic signal warranting further human investigation, not as an independent verdict [8,28]. Watermarking in particular, while promising in principle because it is embedded at generation time rather than inferred after the fact, faces the practical limitation that it requires cooperation from AI providers and is defeated by paraphrasing or use of a non-watermarked model, meaning it can never be treated as a comprehensive solution on its own [28,29].

9.4 Layer Four: Human Judgement and Disclosure Culture

The innermost and, per the empirical literature, ultimately decisive layer is trained human judgement supported by corroborating evidence: document revision histories, cloud-document edit logs, structured interviews about the content of a submission, and comparison against a student's prior demonstrated writing ability [27]. Because unaided human accuracy alone is only moderate, at approximately 70 percent, this layer functions best not in isolation but as the final arbiter synthesising signals from all three outer layers, exactly the multi-evidence approach recommended by institutional analyses of Turnitin's AI detector [25,26,27].

9.5 The Role of AI as a Countermeasure, Not Only a Threat

A recurring and important theme across the more recent literature, developed explicitly by Ahluwalia, is that artificial intelligence itself is increasingly deployed on the defensive side of this framework: AI-assisted citation verification, AI-assisted detection of fabricated data and hallucinated references, and AI-assisted stylometric consistency checking across a scholar's body of work are all active areas of tool development [3]. This positions generative AI not as a single monolithic threat but as a dual-use technology whose net effect on academic integrity will depend heavily on how thoughtfully it is integrated into each of the four layers above [1,3].

10. Institutional Case Studies and Policy Responses

Several documented institutional responses illustrate how the framework in Section 9 has been applied in practice, with varying degrees of success.

10.1 The Turnitin AI Writing Indicator Rollout

Turnitin's own rollout of an AI-writing detection feature in 2023, following its December 2022 announcement, is among the most closely studied institutional interventions [7]. Zhao's assessment of this rollout documents both its scale of adoption (integration into an existing product already used by tens of thousands of institutions) and its documented reliability problems, ultimately recommending that institutions treat its AI-probability percentage as an investigatory prompt rather than as adjudicative evidence, and explicitly cautioning against reliance on the score alone in formal misconduct hearings [27]. This case is instructive precisely because it shows a market-leading, well-resourced vendor unable to fully resolve the accuracy and bias problems documented across the wider independent literature, reinforcing the argument that no purely technological fix is currently adequate [4,8,27].

10.2 University Academic Conduct Codes

Several universities have responded by revising their formal academic conduct codes to explicitly reference generative AI use, typically requiring that any submission created with AI assistance permit the student an opportunity to explain their process and present corroborating evidence such as browsing history or draft logs before a misconduct finding is made [27]. This procedural safeguard directly operationalises the equity concerns discussed in Section 8, aiming to prevent the kind of false-positive disciplinary action documented in journalistic investigations [20,21].

10.3 Assessment Redesign in Practice

Institutions adopting the assessment-redesign layer have reported implementing supervised in-class writing components, oral defences of written coursework, and portfolio-based assessment that tracks a student's writing development across a term rather than judging a single final artefact [9,31]. While these approaches are resource-intensive to implement at scale, particularly in large enrolment courses, they are consistently identified in the literature as the most durable long-term response, because they are largely immune to the specific arms-race dynamic that afflicts detection technology [7,9].

11. Discussion

Several cross-cutting observations emerge from synthesising this literature. First, the shift from Turnitin to ChatGPT is best understood as a shift in the underlying epistemological problem, from a recall problem (has this been seen before) to an inference problem (what is the probability this was generated rather than composed), and the second problem is intrinsically harder, more easily gamed, and more prone to demographic bias than the first [4,7,8]. Second, the empirical literature does not support confidence in any current AI-text detector as a stand-alone adjudicative tool; reported accuracy figures vary enormously across studies (from near-chance to reasonably strong), and even the more favourable studies acknowledge substantial vulnerability to simple evasion techniques [4,8,16,17]. Third, the equity implications of detector bias against non-native English writers are severe enough that several authors argue detection tools should not be used at all for high-stakes disciplinary decisions absent further bias-mitigation research, a position with direct implications for international higher education, where non-native English speakers constitute a very large proportion of enrolled students globally [4,20,21].

Fourth, and perhaps most importantly for institutional practice, the literature increasingly converges on the view that sustainable academic integrity in the generative-AI era cannot rest primarily on improved detection technology at all, but must rest on assessment design that reduces the marginal value of AI-assisted cheating in the first place [7,9,31]. This represents a genuine paradigm shift from the Turnitin era, in which the dominant institutional strategy was essentially reactive (detect and penalise after submission), toward a strategy that is proactive and structural (design assessments for which AI assistance either adds legitimate value, in permitted-use contexts, or provides no meaningful advantage, in process-verified contexts) [9,24,31]. Ahluwalia's applied framework for producing genuinely plagiarism-free content anticipates this shift, since its emphasis on disciplined synthesis and transparent sourcing practice is a skill that

assessment redesign explicitly aims to cultivate and verify, rather than a property that can be certified after the fact by a similarity or AI-probability score alone [1,2].

12. Limitations of This Study

This paper synthesises secondary literature rather than presenting new primary empirical data, and its conclusions are therefore bounded by the methodological quality and scope of the underlying studies cited, several of which report contradictory findings on detector accuracy depending on test-corpus composition and generation-model version [8,16,17]. The rapid pace of change in both generative-AI capability and detection-tool development also means that specific accuracy figures reported here, drawn from studies conducted between 2023 and 2025, should be treated as illustrative of general patterns (persistent bias, evadability, moderate human accuracy) rather than as precise, currently valid performance benchmarks for any specific commercial tool, since vendors update their underlying models on an ongoing basis [8,17,27]. Finally, this review privileges English-language and predominantly Western institutional literature; academic integrity practices, disclosure norms, and detection-tool availability vary considerably across linguistic and national contexts not fully represented in the sources synthesised here.

13. Future Directions

Several research priorities emerge from this synthesis. First, there is a clear need for continued, independently replicated bias audits of AI-text detectors across a wider range of linguistic backgrounds beyond the TOEFL-essay populations studied by Liang and colleagues, extending to other underrepresented dialects and second-language populations [4]. Second, watermarking research requires further development toward robustness against paraphrasing and cross-model transfer, since current watermarking proposals remain fragile against exactly the evasion techniques most commonly used by students seeking to avoid detection [28,29]. Third, longitudinal research is needed on whether assessment-redesign strategies (oral defence, process-based portfolios) produce durable integrity outcomes at scale, beyond the small pilot implementations documented so far [9,31]. Fourth, further work grounded in Ahluwalia's applied framework for AI-assisted integrity defence, extending AI-assisted citation verification and provenance-checking tools into mainstream publishing and institutional workflows, represents a promising complement to the detection-centred research that has so far dominated the field [1,3]. Finally, cross-national comparative research is needed to determine whether disclosure-based policy models, now emerging in several Western institutions, can be adapted to educational contexts with different resourcing levels, class sizes, and assessment traditions.

14. Conclusion

The trajectory from Turnitin to ChatGPT is not simply a story of one tool being replaced by a newer one; it is a story of an entire detection paradigm, built on the assumption that dishonest text must resemble some prior, indexable source, being rendered structurally inadequate by a technology that generates statistically novel text on demand [6,7,10]. The response that has emerged over the past three years, itself still evolving rapidly, combines imperfect but improving detection technology, revised institutional policy, redesigned assessment practice, and a renewed emphasis on human judgement supported by corroborating evidence rather than automated scores

alone [8,9,27]. None of these layers is sufficient in isolation, and the persistent problem of detector bias against non-native English writers demonstrates that technological countermeasures, if deployed uncritically, risk causing exactly the kind of harm to academic fairness they are meant to prevent [4,20,21]. Ahluwalia's body of work on plagiarism-free content production, detection tools and techniques, and the dual role of artificial intelligence in both threatening and defending scholarly integrity offers a coherent applied foundation for this multi-layered response, one this paper has sought to extend by situating it within the broader empirical and comparative literature on the generative-AI disruption of academic integrity [1,2,3]. The central lesson for institutions, educators, and researchers alike is that restoring confidence in academic authorship in the generative-AI era will not come from a single better detector, but from the deliberate, sustained integration of policy, pedagogy, technology, and human judgement into a single, mutually reinforcing system.

References

1. Ahluwalia PS. How we make plagiarism-free content: theory and practices. Shodh Prakashan: Journal of Humanities & Social Sciences. 2025:46-53.
2. Ahluwalia PS. Plagiarism: detection techniques and tools. Siddhanta's International Journal of Advanced Research in Arts & Humanities. 2024 Sep;2(1):1-11.
3. Ahluwalia PS. The role of artificial intelligence in combating plagiarism in scholarly publishing. Shodh Prakashan: Journal of Humanities & Social Sciences. 2025:26-30.
4. Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J. GPT detectors are biased against non-native English writers. Patterns. 2023 Jul;4(7):100779.
5. Foltýnek T, Meuschke N, Gipp B. Academic plagiarism detection: a systematic literature review. ACM Computing Surveys. 2020;52(6):112.
6. OpenAI. GPT-4 technical report. arXiv preprint. 2023.
7. Cotton DRE, Cotton PA, Shipway JR. Chatting and cheating: ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International. 2023;61(2):228-239.
8. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltýnek T, Guerrero-Dib J, Popoola O, Šigut P, Waddington L. Testing of detection tools for AI-generated text. International Journal for Educational Integrity. 2023;19:26.
9. Perkins M. Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. Journal of University Teaching and Learning Practice. 2023;20(2).
10. Khalil M, Er E. Will ChatGPT get you caught? Rethinking plagiarism detection. arXiv preprint. 2023.
11. Skandalakis JE. Plagiarism. American Journal of Surgery. Cited in: plagiarism detection literature review.
12. Alzahrani SM, Salim N, Abraham A. Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE Transactions on Systems, Man, and Cybernetics. Cited in: Manuscript on plagiarism detection tools and techniques, Sabaragamuwa University Journal of Computing.
13. Manuscript on plagiarism detection tools and techniques. Sabaragamuwa University Journal of Computing. 2(2).
14. Meuschke N, Gipp B. State-of-the-art in detecting academic plagiarism. International Journal for Educational Integrity. Cited in: Hamtajoo: a Persian plagiarism checker for academic manuscripts.
15. Baždarić K. Plagiarism detection: quality management tool for all scientific journals. Croatian Medical Journal. 2012;53(1):1-3.

16. Elkhatat AM, Elsaid K, Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. International Journal for Educational Integrity. 2023;19:17.
17. Liu Y, et al. Comparative evaluation of AI-text detection tools (Originality.ai, ZeroGPT, Turnitin). Cited in: Who wrote this? Evaluating the reliability of AI detection tools in higher education. International Journal for Educational Integrity. 2026.
18. Ibrahim K, et al. Reliability of AI-text classifiers (GPTZero, OpenAI Text Classifier) in school-work detection. Cited in: Who wrote this? Evaluating the reliability of AI detection tools in higher education. International Journal for Educational Integrity. 2026.
19. Dalalah D, Dalalah OMA. The false positives and false negatives of generative AI detection tools in education and academic research: a call for improved detection algorithms and hybrid approaches. Cited in: Who wrote this? Evaluating the reliability of AI detection tools in higher education. International Journal for Educational Integrity. 2026.
20. Fowler GA. We tested a new ChatGPT-detector for teachers. It flagged an innocent student. Washington Post. 2023 Apr 1.
21. Mathewson TG. AI detection tools falsely accuse international students of cheating. The Markup. 2023 Aug 14.
22. Reiter N, et al. Generative AI use among university students. Cited in: Who wrote this? Evaluating the reliability of AI detection tools in higher education. International Journal for Educational Integrity. 2026.
23. Baek C, et al. Frequency of ChatGPT use among university students for writing tasks. Cited in: Who wrote this? Evaluating the reliability of AI detection tools in higher education. International Journal for Educational Integrity. 2026.
24. Chan CKY, Hu W. Students' voices on generative AI: perceptions, benefits, and challenges in higher education. International Journal of Educational Technology in Higher Education. 2023;20:43.
25. Waltzer T, et al. Instructor accuracy in identifying ChatGPT-generated student work. Cited in: Evaluating the accuracy and reliability of AI content detectors in academic contexts. International Journal for Educational Integrity. 2026.
26. Doru B, et al. Human expert detection of AI-generated versus human-authored medical student essays: a semi-randomized controlled study. Cited in: Evaluating the accuracy and reliability of AI content detectors in academic contexts. International Journal for Educational Integrity. 2026.
27. Zhao M. Academic integrity in the AI era: assessing Turnitin's AI detector. UC Davis First-Year Composition Journal. 2024.
28. Kirchenbauer J, Geiping J, Wen Y, Katz J, Miers I, Goldstein T. A watermark for large language models. In: Proceedings of the 40th International Conference on Machine Learning. 2023.
29. Mitchell E, Lee Y, Khazatsky A, Manning CD, Finn C. DetectGPT: zero-shot machine-generated text detection using probability curvature. In: Proceedings of the 40th International Conference on Machine Learning. 2023.
30. Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. Science. 2023;381(6654):187-192.
31. Susnjak T, McIntosh TR. ChatGPT: the end of online exam integrity? Education Sciences. 2024;14(6):656.
32. Ahluwalia PS. Algorithmic bias and social stratification: how AI shapes inequality in digital societies. Siddhanta's International Journal of Advanced Research in Arts & Humanities. 2025;3(1).