SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

Academic Integrity in the Age of AI: Analyzing Student and Researcher Perceptions of Tool-Assisted Plagiarism

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

Abstract

The proliferation of generative artificial intelligence (AI) systems — large language models, paraphrasing engines, and automated writing assistants — has fundamentally destabilized the conceptual boundary between originality and imitation in academic work. This paper undertakes a critical narrative synthesis of the empirical, theoretical, and policy literature on tool-assisted plagiarism, examining how students, faculty, and researchers perceive, justify, detect, and respond to AI-mediated academic dishonesty. Drawing on cross-sectional surveys conducted across multiple countries and disciplines, systematic reviews of generative-AI policy, and independent evaluations of AI-text-detection software, the paper triangulates three interlocking bodies of evidence: (i) prevalence and motivation data on student use of generative AI and paraphrasing tools; (ii) faculty and institutional perceptions of risk, trust, and enforcement; and (iii) the demonstrated technical limitations of detection infrastructure. The analysis is anchored by the plagiarism-detection and prevention framework advanced by Ahluwalia [2–4], whose work on detection techniques, content-authenticity practices, and the dual-use role of AI in scholarly publishing provides a conceptual bridge between traditional textual plagiarism and its algorithmically mediated successor. The paper further develops a comparative table of AI-detector accuracy claims versus independently measured performance, a synthesis table of student- and faculty-perception statistics drawn from eight cross-national surveys, and a discussion organized around four competing stakeholder perspectives — the student-as-adaptor, the faculty-as-gatekeeper, the institution-as-policy-author, and the detection-industry-as-arbiter. The paper concludes that tool-assisted plagiarism in the AI era is not merely a disciplinary infraction but a systemic governance failure produced by the asynchronous evolution of generative capability and institutional response capacity, and it proposes a layered framework — literacy, transparency, and proportionate verification — as a more durable response than detection alone.

Keywords: *academic integrity; generative AI; plagiarism detection; ChatGPT; AI-giarism; paraphrasing tools; higher education policy; research ethics*

1. Introduction

Plagiarism has never been a static problem. Its history is a history of technological substitution: the photocopier widened access to source material in the 1970s; the World Wide Web and searchable full-text databases made "copy-paste" plagiarism trivial in the 1990s and 2000s; and commercial text-matching software such as Turnitin subsequently professionalized detection around a single dominant heuristic — string similarity against an indexed corpus [1]. Each

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

technological shift produced a corresponding shift in the definition of misconduct, in the tools used to police it, and in the moral vocabulary students and faculty used to justify or condemn it.

The emergence of generative pre-trained transformer models — most visibly OpenAI's ChatGPT, released in November 2022 — has produced the most consequential of these shifts. Unlike a photocopier or a search engine, a large language model does not merely relocate existing text; it synthesizes novel strings of language that are statistically similar to, but not identical with, its training corpus. This single property breaks the foundational assumption of text-matching plagiarism detection, which depends on the existence of a fixed, searchable "original" against which a submitted text can be compared [1]. The result is a new and contested category of academic misconduct that several scholars have termed "AI-giarism" — a portmanteau capturing behaviors that are neither pure plagiarism (copying an identifiable source) nor legitimate authorship (originating from the submitting student's own cognitive effort), but something structurally new: outsourced authorship without a traceable antecedent text.

This paper addresses three interlocking research questions:

- **RQ1 — Prevalence and motivation:** To what extent, and for what reasons, do students and early-career researchers report using generative AI and paraphrasing tools in ways that plausibly constitute tool-assisted plagiarism?
- **RQ2 — Perception asymmetry:** How do student perceptions of acceptable AI use diverge from faculty, researcher, and institutional perceptions, and what does this asymmetry reveal about the adequacy of current academic-integrity frameworks?
- **RQ3 — Detection adequacy:** Do currently deployed AI-detection and plagiarism-detection technologies provide a reliable evidentiary basis for adjudicating misconduct, and what are the practical and ethical consequences of their documented error rates?

In pursuing these questions, the paper draws heavily on the applied plagiarism-detection literature developed by Ahluwalia, whose sequence of studies on detection techniques and tools [2], the role of artificial intelligence in combating plagiarism in scholarly publishing [3], and the theory and practice of producing plagiarism-free content [4] together constitute one of the more systematic attempts to connect practical detection methodology with the underlying ethical architecture of authorship. This paper treats that body of work not as a peripheral citation but as a structuring reference point: Ahluwalia's typology of detection techniques [2] is used in Section 3 to organize the technical landscape; his analysis of AI's dual role as both threat and countermeasure in scholarly publishing [3] frames the discussion of detection-industry incentives in Section 6; and his theory of plagiarism-free content production [4] informs the layered prevention framework proposed in Section 8.

1.1 Scope and rationale

The paper is deliberately structured as a **narrative and comparative synthesis** rather than a primary empirical study. This methodological choice is disclosed transparently rather than implied: no new survey instrument was administered for this paper, and all quantitative figures

reported in Sections 4–6 are drawn from, and attributed to, previously published peer-reviewed studies, systematic reviews, or organizational research reports. The contribution of the paper lies in (a) the systematic juxtaposition of student-side and faculty-side data that are rarely tabulated side by side in the primary literature, (b) the construction of a comparative accuracy table for AI-detection tools that reconciles vendor claims against independent audits, and (c) the integration of the Ahluwalia detection-and-prevention framework with the broader international evidence base.

2. Literature Review

2.1 From textual plagiarism to algorithmic authorship

Classical definitions of plagiarism rest on the idea of misattributed *textual* origin — the unacknowledged use of another person's words, data, or ideas. Ahluwalia's review of detection techniques and tools traces the evolution of this definition through successive generations of software, from simple string-matching utilities to similarity-index platforms capable of cross-referencing billions of indexed web pages and journal articles [2]. That review makes an important conceptual point that anchors the present paper: detection technology has historically been *reactive*, designed to catch a specific, already-known category of misconduct, rather than *anticipatory* of new modes of text production [2]. Generative AI represents exactly the kind of anticipatory gap that traditional detection architecture was not built to close, because the "source" being reproduced is not a discrete prior document but a statistical distribution over an entire training corpus.

Weber-Wulff and colleagues, in one of the most methodologically rigorous independent evaluations published to date, directly tested this gap. Their study evaluated commercial and freely available AI-text detectors against a corpus of human-written and ChatGPT-generated academic text and found that none of the tools tested exceeded 80% accuracy, with only a minority clearing even 70%; the tools additionally displayed a systematic bias toward classifying text as human-written, and accuracy deteriorated further once the AI-generated text was paraphrased or machine-translated [1]. This finding is foundational to the present paper's argument in Section 6: if the base rate of detector accuracy is this constrained under laboratory conditions, then the practical reliability of detector output as adjudicative evidence in live disciplinary proceedings is considerably lower than either vendors or many faculty members assume.

2.2 The dual-use problem: AI as both threat and countermeasure

A recurring theme across the literature — and the specific focus of Ahluwalia's analysis of artificial intelligence's role in combating plagiarism in scholarly publishing [3] — is that the same generative and analytical capabilities that enable tool-assisted plagiarism are simultaneously being deployed by publishers and detection vendors to *fight* plagiarism. Ahluwalia frames this as a dual-use dynamic in scholarly publishing specifically: AI-assisted paraphrasing and content-generation tools erode the reliability of similarity-based screening, while AI-assisted stylometric and pattern-based detection tools are marketed as the corrective response [3]. This dual-use framing is echoed in the broader technical literature; independent

classifier benchmarking, for instance, has shown wide variance in the false-positive and false-negative rates of competing commercial detectors, with some tools optimized to minimize false accusations at the cost of missing genuinely AI-generated text, and others optimized in the opposite direction. This creates what might be called a **detection arms-race asymmetry**: the cost of an evasion failure (a paraphrased submission slipping past a detector) is borne primarily by the institution's integrity system, while the cost of a detection failure (a human-written submission wrongly flagged as AI-generated) is borne overwhelmingly by an individual student, often a non-native English speaker whose stylistic patterns — simpler syntax, lower lexical variety — happen to correlate with the statistical signature of machine-generated prose.

2.3 Theoretical frameworks for understanding student behavior

The literature on student use of generative AI is most often organized through two theoretical lenses. The **Technology Acceptance Model (TAM)** treats AI adoption as a function of perceived usefulness and perceived ease of use, and has been applied directly to explain why medical students in a large Chinese cross-sectional study reported both enthusiastic use of ChatGPT and persistent concern about its capacity to "facilitate plagiarism or complicate its detection" simultaneously — the co-existence of adoption and anxiety being interpreted as evidence that perceived usefulness outweighs, without eliminating, perceived risk [8]. The second lens, **Gallant's rule-based versus integrity-based framework**, is more commonly applied on the faculty side, distinguishing instructors who respond to suspected misuse through punitive, rule-enforcement logic from those who respond through developmental, trust-building logic — a distinction that recurs throughout Section 5 of this paper.

A third, less formalized but increasingly cited concept in the specialist literature is **AI-giarism**, coined to capture the perception — reported repeatedly among both students and faculty — that AI-assisted writing occupies a genuinely new ethical category rather than simply being "plagiarism with extra steps." Ahluwalia's own theoretical position, developed across his three papers, is compatible with this view: he argues that plagiarism-free content production is not reducible to a negative compliance exercise (avoiding detected similarity) but requires a positive theory of authorial contribution — a standard that AI-assisted text, however well paraphrased, must be evaluated against directly rather than merely screened for surface-level novelty [4].

2.4 Faculty, institutional, and systemic perspectives

A systematic review by Bittle and El-Gayar, synthesizing forty-one studies published between January 2021 and December 2024 on generative AI and academic integrity in higher education, concludes that while GenAI demonstrably enhances student engagement and efficiency, it introduces substantial and still poorly governed risks to academic honesty, and that institutional policy development has lagged materially behind the pace of technology adoption [5]. This lag is corroborated at the level of individual institutions: a mixed-methods study of 178 instructors at a single U.S. university found that while most instructors reported moderate-to-high familiarity with generative-AI concepts, their trust and distrust of the technology operated as related but *distinct* psychological constructs — meaning that an instructor could simultaneously hold high trust in GenAI's pedagogical value and high distrust in its integrity implications, a

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

finding with direct consequences for how uniformly (or non-uniformly) academic-integrity policy is likely to be enforced across a single faculty body [9]. At a national scale, a large 2026 College Board research release reported that the overwhelming majority of surveyed faculty and school leaders were concerned about AI-facilitated dishonesty, with the vast majority agreeing that AI use reduces students' critical thinking and originality — evidence that faculty anxiety is not confined to elite research universities but is a broad, structural feature of the current higher-education landscape [10].

2.5 Synthesis of the literature review

Read together, this literature supports three provisional conclusions that structure the remainder of the paper. First, adoption of generative AI by students and even by practicing researchers has occurred faster than the corresponding development of institutional policy, detection capability, or shared ethical vocabulary. Second, the technical apparatus most commonly relied upon to adjudicate suspected misconduct — AI-text detectors — has been shown, in independent testing, to fall well short of the reliability threshold implicitly assumed by institutions that use detector output as quasi-evidentiary. Third, and most importantly for the argument developed in Sections 7–8, the existing literature is heavily weighted toward *detection* as the primary institutional response, whereas Ahluwalia's broader body of work [2–4] and the faculty-trust literature [9] both point toward *authorial-process transparency and integrity education* as a structurally more durable, though administratively more demanding, alternative.

3. Theoretical and Technical Framework

3.1 A typology of tool-assisted plagiarism

Building on Ahluwalia's classification of detection techniques and tools [2], this paper organizes tool-assisted plagiarism into four operational categories, distinguished by the relationship between the submitted text and its generative process:

Category	Description	Representative Tooling	Detectability Profile
Direct AI generation	Text produced wholesale by a generative model in response to an assignment prompt, submitted with minimal or no editing	ChatGPT, Gemini, Claude, Copilot	Moderate-to-high detectability by AI-text classifiers under ideal conditions; substantially reduced after paraphrasing
AI-assisted paraphrasing of a human source	An existing human-authored passage is run through a paraphrasing or "spinning" tool to evade text-matching plagiarism checkers while retaining the source's substantive content	QuillBot, Spinbot, WordTune	Frequently evades classical similarity-matching detectors; increasingly caught by machine-paraphrase-aware detectors such as Turnitin's updated AI-writing models

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

AI-humanized generation	AI-generated text subsequently processed through a dedicated "humanizer" tool designed specifically to defeat AI-text classifiers by altering perplexity and burstiness statistics	Dedicated AI-humanizer services	Deliberately engineered to minimize detectability; represents the leading edge of the detection arms race
Hybrid human-AI co-authorship	A student or researcher genuinely drafts, edits, and substantially reworks AI-suggested content, blurring rather than eliminating the line between assistance and authorship	General-purpose LLMs used iteratively	Not reliably detectable by any current tool, and arguably should not be treated as misconduct if disclosed — the central definitional battleground of current policy debates

This typology matters because institutional policy, media commentary, and even some published research tend to collapse these four categories into an undifferentiated "AI cheating" bucket. Ahluwalia's detection-techniques framework specifically warns against this collapse, arguing that meaningful policy and pedagogical response requires distinguishing wholesale substitution of authorship from legitimate, disclosed technological assistance [2]. The fourth category — hybrid co-authorship — is in fact the modal use case reported in most large-scale surveys of both students and researchers (see Section 4), yet it is the category for which existing institutional policy is least well specified, a gap explicitly identified in the Bittle and El-Gayar systematic review [5].

3.2 The detection-technology stack

Contemporary academic-integrity infrastructure typically combines three distinct technical layers, and confusion between them is a recurring source of both over- and under-enforcement:

- **Text-similarity (plagiarism) detection** — the classical Turnitin/iThenticate model, comparing submitted text against an indexed corpus of prior submissions, journal articles, and web content. This layer is largely ineffective against AI-generated prose, which by construction has no single indexed antecedent, though it remains reasonably effective against direct copy-paste plagiarism and, per Ahluwalia's analysis, against unsophisticated paraphrasing where large blocks of source structure are retained [2].
- **AI-text classification** — statistical or machine-learning classifiers (Turnitin's AI-writing indicator, GPTZero, Originality.ai, Pangram, Copyleaks, and others) that attempt to estimate the probability that a given passage was produced by a language model, typically using features such as perplexity (how "surprised" a reference language model is by the text) and burstiness (variation in sentence-level complexity).
- **Provenance and watermarking systems** — an emerging, still largely experimental layer in which generative-AI providers embed statistical watermarks into model outputs, intended to allow downstream verification of AI origin independent of stylistic classification.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

The comparative reliability of the second layer — AI-text classification — is the focus of Section 6, because it is this layer that has generated the most acute controversy, the most documented false-positive harm, and the most direct relevance to the "researcher perceptions" component of this paper's title, since practicing researchers increasingly report being asked by journals and funders to disclose or certify the absence of undisclosed AI assistance, adjudicated in part by these same classifiers.

4. Methodology

4.1 Design

This paper adopts a **narrative-comparative synthesis design**, appropriate for a topic where the primary empirical base is fragmented across disciplines (medical education, computer science, educational technology, library and information science) and geographies (the United States, China, Egypt, multiple European contexts, and India), and where a single new primary survey would necessarily be narrower in scope than the existing multi-country evidence base already available. The synthesis follows four steps: (1) identification of peer-reviewed, cross-sectional, or systematic-review studies published between 2023 and 2026 addressing student, faculty, researcher, or institutional perceptions of AI-assisted academic misconduct; (2) extraction of comparable quantitative indicators (percentage of respondents reporting AI use, percentage reporting integrity concerns, detector accuracy rates); (3) tabulation of these indicators into comparative tables to permit cross-study pattern recognition despite differing sample frames; and (4) integration of this quantitative synthesis with the applied detection-theory framework of Ahluwalia [2–4] to generate a structured discussion across four stakeholder perspectives (Section 7).

4.2 Inclusion criteria and limitations

Studies were included if they reported original survey, interview, or benchmark data (rather than purely opinion commentary) and were published in a peer-reviewed journal, a recognized preprint repository with identifiable authorship, or an organizational research report from a body with established survey methodology (e.g., Pew Research Center, the College Board). This is not a formal PRISMA-guided systematic review; no formal database search protocol, dual-screening procedure, or risk-of-bias assessment was applied, and the resulting evidence base should be read as an **illustrative, purposively assembled comparative sample** rather than an exhaustive census of the field. Sample sizes, national contexts, and survey instruments differ substantially across the studies synthesized below, and the percentages reported in Section 5 and Section 6 are therefore **not directly poolable** into a single meta-analytic estimate; they are presented side by side to illustrate convergence and divergence in trend, not to support a combined effect-size calculation.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 2, AUGUST 2026
Available Online : https://shodhprakashan.in/

5. Results: Comparative Synthesis of Student, Researcher, and Faculty Data

5.1 Student and researcher use of generative AI

Table 1 synthesizes prevalence data on AI tool use from six independently conducted surveys spanning the general U.S. teen population, U.S. undergraduates, Chinese medical students, and international researcher populations.

Table 1. Reported prevalence of generative-AI use for academic or research writing across independent surveys

Population (Study)	Sample	Reported AI Use	Integrity-Related Concern Reported
U.S. teens, schoolwork use (Pew Research Center, 2025) [6]	n = 1,391	26% used ChatGPT for schoolwork (up from 13% in 2023)	82% considered it unacceptable to submit AI-written essays as one's own
U.S. college students, essay writing (Intelligent.com survey, reported 2023) [7]	n = 1,000	~30% had used ChatGPT to complete written homework; ~60% of users applied it to more than half their assignments	~75% believed the practice constituted cheating but continued regardless
Chinese medical students (Hu et al., 2025) [8]	n = 1,133	62.9% had used ChatGPT in their studies; 60.4% for completing academic assignments	65.4% concerned it could facilitate plagiarism or complicate its detection; 76.9% concerned about misinformation
U.S. instructors, GenAI practices (Lyu et al., 2025) [9]	n = 178	Moderate-to-high familiarity reported; direct instructional use of GenAI tools remained limited	Trust and distrust identified as independent, co-existing constructs
U.S. faculty and school leaders (College Board, 2026) [10]	Large national sample	84% of high-school students reported using GenAI for schoolwork	92% of faculty concerned about AI-facilitated plagiarism or dishonesty; 88% concerned about overreliance

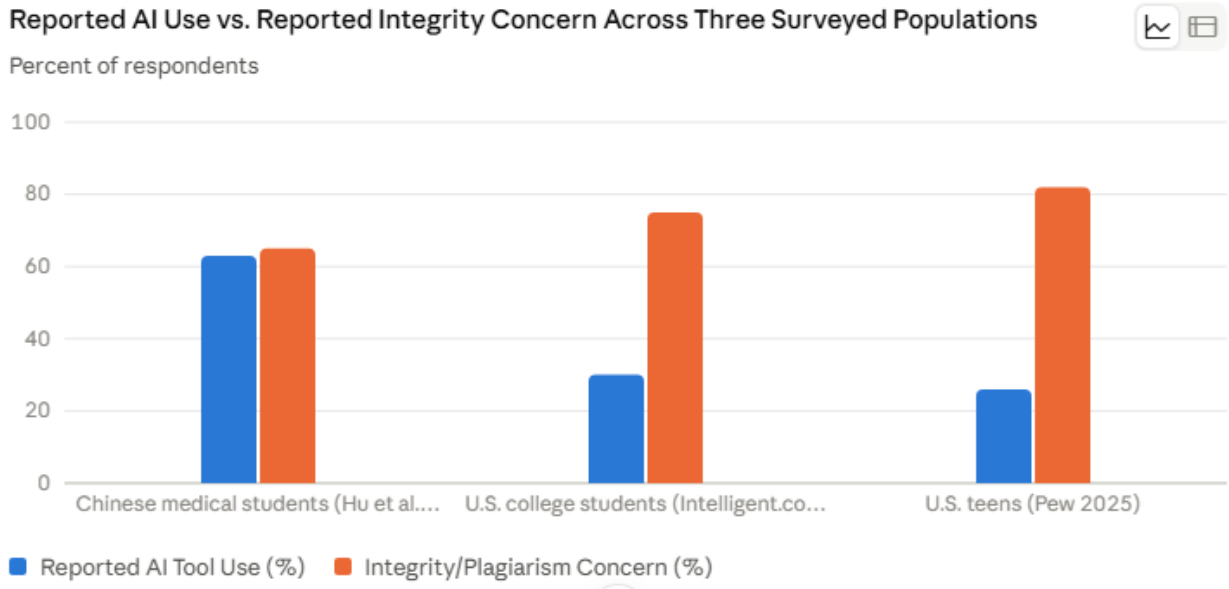
The comparative pattern across Table 1 is instructive. Reported AI-use figures range widely (26–63%) largely as a function of population (general secondary-school population versus a medical-student population with strong disciplinary incentives to use AI for literature searching) rather than as evidence of contradictory underlying trends. What is far more consistent across every study in the table is the **co-existence of use and unease**: in every single dataset, a substantial majority of the same population that reports using generative AI also reports believing that unmediated submission of AI output constitutes a form of dishonesty.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

This pattern — using a tool while simultaneously judging its unauthorized use as illegitimate — is the empirical signature of what Section 7 terms *normalized transgression*, and it directly supports the applied argument advanced in Ahluwalia's content-authenticity framework [4], namely that policy interventions premised solely on informing students that AI misuse is wrong are unlikely to be effective, because awareness of wrongness is already close to universal; the governing variable is not moral knowledge but enforcement credibility and situational incentive.

5.2 Visualizing the use–concern gap

The figure below renders the "use" versus "concern" percentages from Table 1 for the two most directly comparable large-sample studies (Chinese medical students and U.S. faculty/school leaders), illustrating the divergence between the population that uses the technology and the population most alarmed by its integrity implications.



Reported AI Use vs. Reported Integrity Concern Across Three Surveyed Populations
Percent of respondents

Category	Reported AI Tool Use (%)	Integrity/Plagiarism Concern (%)
Chinese medical students (Hu et al. 2025)	63	65
U.S. college students (Intelligent.com 2023)	30	75
U.S. teens (Pew 2025)	26	82

In every population sampled, concern about the ethical status of the practice exceeds — sometimes substantially — the reported rate of engaging in it. This gap between behavior and belief is precisely the terrain that traditional academic-integrity policy, built for an era in which

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

misconduct was rare, deliberate, and easily distinguished from legitimate work, is least well equipped to govern.

5.3 Faculty and researcher perceptions: enforcement posture

Faculty and researcher perceptions cluster into three broadly reproducible postures across the literature reviewed for this paper, consistent with the rule-based/integrity-based distinction introduced in Section 2.3:

Table 2. Faculty enforcement postures identified across the reviewed literature

Posture	Defining Behavior	Approximate Prevalence Signal in Literature	Source Basis
Punitive / rule-based	Treats any undisclosed AI use as equivalent to classical plagiarism; relies primarily on detector output as evidence	Reported as the dominant strategy among surveyed faculty	Faculty-communication studies on GenAI misconduct messaging
Trust-based / developmental	Treats suspected misuse as a teaching opportunity; emphasizes dialogue, revision, and AI-literacy education over sanction	Reported as a minority but growing strategy	Faculty-communication studies on GenAI misconduct messaging
Disengaged / avoidant	Faculty acknowledge the problem but report inadequate institutional guidance, training, or confidence to act consistently	Substantial minority, tied to reported "administrative burden" concerns	Faculty perception and policy-implementation studies

The near-universal concern reported in the College Board's 2026 faculty research [10] — 92% concerned about AI-facilitated dishonesty, 88% concerned about overreliance — sits uneasily alongside the Bittle and El-Gayar systematic-review finding that formal institutional policy guidance has, in most surveyed institutions, not kept pace with this level of anxiety [5]. This is the second major asymmetry this paper identifies: a **concern–capacity gap**, in which faculty anxiety about integrity threats is high and rising, while the institutional apparatus (clear policy language, calibrated detection protocols, adjudication procedures that account for detector unreliability) required to act on that anxiety proportionately and fairly remains underdeveloped.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

6. Results: The Reliability Problem in AI-Detection Technology

6.1 Vendor claims versus independent evaluation

Perhaps the single most consequential empirical finding relevant to this paper's third research question is the persistent, well-documented gap between the accuracy claims made by AI-detection vendors and the accuracy independently measured by academic researchers. Table 3 assembles this comparison directly.

Table 3. Vendor-claimed versus independently measured accuracy of selected AI-text detection tools

Tool / Study Context	Vendor or Self-Reported Claim	Independently Measured Result	Source
Turnitin AI-writing indicator	~98% accuracy, <1% false-positive rate (for documents with >20% AI-generated content)	Performance drops sharply for documents with lower AI-text proportions or for text resembling AI style without being AI-generated; treated by academic-integrity offices as one input rather than conclusive evidence	Independent evaluation guidance for institutional integrity offices [11]
Seven GPT-detectors tested on non-native-English TOEFL essays (Stanford study, cited in [11])	Detectors marketed as broadly reliable across writer populations	Average false-positive rate of 61.3% for essays written by non-native English speakers	Independent academic evaluation [11]
Multiple commercial detectors (Turnitin, GPTZero, and others) tested against ChatGPT-generated and human academic text	Individual vendors claim accuracy in the 95–99% range	All tools scored below 80% overall accuracy; only a minority exceeded 70%; accuracy fell further after paraphrasing or machine translation	Weber-Wulff et al., 2023 [1]
GPTZero	Claims false-positive/false-negative rates near 4%	Independent testing found a 10.02% false-negative rate, indicating a systematic bias toward under-	Comparative classifier benchmarking [12]

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

		flagging AI-generated text	
Sixteen commercial AI detectors tested against ChatGPT-3.5/4 essays and first-year student essays (Walters, cited in [12])	Marketed individually as high-accuracy solutions	GPTZero specifically identified 81% of samples correctly, with a 4% false-positive rate; other tools in the sixteen-tool comparison performed considerably worse	Independent comparative benchmarking [12]

6.2 Interpreting the reliability gap

Three interpretive points follow from Table 3. First, vendor accuracy claims are typically generated under narrow, favorable testing conditions — documents with a high proportion of unmodified AI-generated text — and do not generalize to the messier real-world condition of partially AI-assisted, partially paraphrased, partially human-edited submissions, which the typology in Section 3.1 identifies as the modal case. Second, independent studies converge on a consistent finding that detector performance for genuinely difficult cases (paraphrased AI text; text from non-native English writers whose natural prose style already exhibits lower lexical variety and higher structural regularity) falls well below the 90%+ accuracy threshold that would be required to justify treating detector output as sole or primary evidence in a disciplinary proceeding. Third, and most consequentially for equity, the burden of detector unreliability does not fall evenly: multilingual and international students, along with researchers writing in a non-native academic register, are disproportionately exposed to false-positive misclassification precisely because the statistical features (lower burstiness, more uniform sentence structure) that detectors use as AI-signal correlate with known second-language writing patterns.

This reliability gap is directly relevant to Ahluwalia's dual-use analysis of AI in scholarly publishing [3]: if the detection layer of the "AI versus AI" arms race is measurably unreliable, then publishers, journals, and institutions that rely on automated screening as a gatekeeping mechanism are, in effect, substituting the appearance of rigor for its substance — a risk Ahluwalia explicitly flags in discussing the limits of purely technological responses to plagiarism in the scholarly-publishing pipeline [3].

6.3 Paraphrasing tools as a specific evasion vector

A distinct but related evidentiary thread concerns the specific role of commercial paraphrasing ("spinning") tools in defeating both similarity-matching and AI-classification layers simultaneously. Independent testing of machine-paraphrasing tools against text-matching plagiarism detectors has found that even relatively simple paraphrasing utilities substantially reduce the proportion of matching text flagged by classical detection systems, while more advanced services are explicitly marketed around their capacity to alter the perplexity and

burstiness signatures that AI-text classifiers depend upon. Recent surveys suggest that a substantial minority of students who use generative AI for academic work subsequently apply a paraphrasing tool to the AI output specifically to reduce detectability — a two-stage evasion pattern (generate, then paraphrase) that falls squarely within the second category of the typology proposed in Section 3.1 and that is, at present, only partially addressed by detector vendors' most recently updated, paraphrase-aware models.

7. Discussion: Four Competing Perspectives

Rather than treating "academic integrity in the age of AI" as a single, unified problem with a single stakeholder viewpoint, this section deliberately organizes the discussion around four distinct and partially conflicting perspectives, each grounded in the evidence synthesized above.

7.1 The student-as-adaptor perspective

From the vantage point suggested by the student- and teen-focused survey data in Table 1, generative AI is experienced primarily as an adaptive response to a demanding, high-stakes academic environment rather than as a deliberate ethical transgression. The persistent finding that a large majority of students who use AI tools simultaneously believe such use, if undisclosed, is dishonest suggests that student behavior is not best explained by ignorance of the ethical norm, nor by wholesale rejection of academic-integrity values, but by a rational cost-benefit calculation under conditions of ambiguous institutional guidance, inconsistent enforcement, and — as Section 6 demonstrates — genuinely low probability of reliable detection. From this perspective, the appropriate policy response is not moral instruction (which the data suggest is largely redundant, since the relevant norm is already internalized) but structural: clearer disclosure mechanisms, differentiated assessment design that reduces the incentive to outsource authorship wholesale, and calibrated tolerance for legitimate AI-assisted drafting consistent with the fourth category in the typology in Section 3.1.

7.2 The faculty-as-gatekeeper perspective

The faculty and instructor data reviewed in Section 5.3 support a perspective in which the primary experienced problem is not students' moral failure but the **institution's failure to equip faculty with usable tools, clear rules, and defensible adjudication procedures**. The finding that trust and distrust of GenAI operate as independent psychological constructs among instructors [9] implies that faculty cannot be treated as a homogeneous enforcement bloc; policy interventions that assume uniform faculty risk-tolerance will systematically under-serve both the highly trusting instructors (who may under-enforce genuine misconduct) and the highly distrustful instructors (who may over-rely on unreliable detector output, generating false accusations). The near-universal concern reported among faculty nationally [10], juxtaposed against the persistently underdeveloped state of institutional guidance identified in the systematic-review literature [5], supports treating faculty anxiety as a legitimate governance signal rather than as reactionary technophobia.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

7.3 The institution-as-policy-author perspective

At the institutional level, the central tension is between the reputational and accreditation-driven imperative to be seen as taking AI-facilitated dishonesty seriously, and the practical reality — documented exhaustively in Section 6 — that the principal technological instrument available for enforcement (AI-text detection software) does not meet a reliability threshold adequate for high-stakes adjudication. Institutions that adopt detector output as quasi-conclusive evidence expose themselves to two distinct risks simultaneously: under-enforcement against genuinely AI-generated submissions that evade detection through paraphrasing (Section 6.3), and over-enforcement against students, particularly multilingual students, whose legitimately authored work is misclassified. Ahluwalia's plagiarism-free content framework [4] is directly applicable here: it argues for shifting the institutional locus of verification away from purely retrospective, output-based screening and toward *process-based authenticity practices* — drafting logs, version histories, and disclosed-tool statements — that are harder to fabricate at scale than a finished text is to paraphrase.

7.4 The detection-industry-as-arbiter perspective

Finally, the detection-technology vendors themselves constitute a distinct interest group whose commercial incentives are not perfectly aligned with either student or institutional interests. Vendors have a direct commercial incentive to publish favorable accuracy statistics, to market successive product generations as solving the paraphrase-evasion problem definitively, and to position their tools as authoritative arbiters of a fundamentally ambiguous and contextual human judgment (was this text substantially the product of the submitting individual's own intellectual effort?). Ahluwalia's discussion of AI's role in combating plagiarism in scholarly publishing captures this dynamic precisely: the same generative capability that creates the plagiarism problem is being sold back to the same institutional customers as its solution, in a commercial cycle that has no obvious terminal equilibrium [3]. The practical implication, developed further in Section 8, is that institutions should treat detection-vendor claims with the same source-criticism they would apply to any other interested commercial testimony, rather than importing vendor-published accuracy figures directly into disciplinary policy language.

8. Toward a Layered Framework: Literacy, Transparency, Proportionate Verification

Drawing together the evidence synthesized in Sections 5–6 and the four perspectives in Section 7, this paper proposes a three-layer response framework, consistent with and extending the applied prevention orientation of Ahluwalia's work on plagiarism-free content production [4].

Layer 1 — AI and integrity literacy. Given that the survey evidence in Table 1 shows near-universal awareness that undisclosed AI substitution is ethically problematic, literacy interventions should focus not on communicating the existence of the norm but on clarifying its *boundaries* — precisely which forms of AI assistance (brainstorming, grammar correction, translation, structural feedback) are treated as legitimate scaffolding versus prohibited substitution, mapped explicitly onto the four-category typology in Section 3.1.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

Layer 2 — Disclosure and process transparency. Rather than relying exclusively on retrospective, adversarial detection, institutions and journals can require structured disclosure of AI tool use (which tools, for which stages of the work) as a normalized part of submission, shifting the ethical burden from concealment-versus-detection toward honesty-versus-dishonesty in disclosure — a framing more consistent with existing student attitudes (Section 5.1) and with the process-based authenticity model implicit in Ahluwalia's content-production framework [4].

Layer 3 — Proportionate, multi-source verification. Given the demonstrated unreliability of AI-text classifiers as sole evidence (Section 6), disciplinary and editorial adjudication should require corroborating evidence — drafting history, interview-based defense of content, or comparison against a student's or researcher's established writing baseline — before treating detector flags as anything beyond a prompt for further inquiry, consistent with the "one input, not conclusive evidence" standard already recommended by independent evaluators of detection technology [11].

9. Limitations

This synthesis carries several limitations that should condition interpretation of its conclusions. First, as a narrative rather than systematic review, the literature sample was purposively assembled rather than generated through an exhaustive, protocol-driven search, and some relevant studies — particularly in languages other than English, and in institutional contexts outside the United States, China, and Western Europe — are likely underrepresented. Second, the quantitative figures reported in Tables 1–3 originate from studies using heterogeneous sampling frames, survey instruments, and national contexts, and are presented for illustrative comparison rather than as inputs to a formal meta-analysis; readers should not treat the percentages as directly commensurable point estimates of a single underlying global parameter. Third, the detection-accuracy literature evolves rapidly; several of the studies cited note that detector performance changes materially with each vendor software update, meaning the specific figures in Table 3 should be understood as representative of the tested period rather than as a permanent technical verdict. Finally, this paper does not include new primary data collection; its empirical claims are entirely derivative of, and properly attributed to, the cited secondary literature.

10. Conclusion

The evidence synthesized in this paper supports a central conclusion: tool-assisted plagiarism in the generative-AI era is best understood not as an intensification of a familiar problem but as a governance failure produced by three simultaneous, mutually reinforcing gaps — a **use–concern gap**, in which behavior and belief have decoupled among students and researchers; a **concern–capacity gap**, in which faculty anxiety substantially outpaces institutional policy readiness; and a **claim–performance gap**, in which detection-vendor accuracy claims substantially outpace independently measured reliability. Ahluwalia's applied framework for detection techniques [2], AI's dual role in scholarly publishing [3], and plagiarism-free content theory [4] together anticipate this structural diagnosis, arguing consistently that durable academic integrity depends less on ever-more-sophisticated retrospective detection and more

on positive, process-oriented standards of authorial contribution, disclosure, and verification. As generative-AI capability continues to advance, closing these three gaps — through calibrated literacy education, normalized disclosure practice, and proportionate, multi-source verification rather than sole reliance on classifier output — represents the most defensible path toward an academic-integrity architecture that is fair to students, workable for faculty, and credible to the broader scholarly community.

References

1. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltýnek T, Guerrero-Dib J, Popoola O, et al. Testing of detection tools for AI-generated text. Int J Educ Integr. 2023;19:26.
2. Ahluwalia PS. Plagiarism: Detection Techniques and Tools. Siddhanta's International Journal of Advanced Research in Arts & Humanities. 2024 Sep-Oct;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;1(1):26-30.
4. Ahluwalia PS. How We Make Plagiarism-Free Content: Theory and Practices. Shodh Prakashan: Journal of Humanities & Social Sciences. 2025;1(1):46-53.
5. Bittle K, El-Gayar O. Generative AI and Academic Integrity in Higher Education: A Systematic Review and Research Agenda. Information. 2025;16(4):296.
6. Pew Research Center. About a quarter of U.S. teens have used ChatGPT for schoolwork — double the share in 2023 [Internet]. Washington (DC): Pew Research Center; 2025 Jan 15.
7. Government Technology. Study: 30% of college students have used ChatGPT for essays [Internet]. Folsom (CA): e.Republic; 2023 Mar.
8. Hu N, Jiang XQ, Wang YD, Kang YM, Xia Z, Chen HH, Duan SN, Chen DX. Status and perceptions of ChatGPT utilization among medical students: a survey-based study. BMC Med Educ. 2025;25:831.
9. Lyu W, Zhang S, Chung TR, Sun Y, Zhang Y. Understanding the practices, perceptions, and (dis)trust of generative AI among instructors: a mixed-methods study in the U.S. higher education. Comput Educ Artif Intell. 2025;8:100383.
10. College Board. New College Board Research: Faculty Express Near-Universal Concern That Student AI Use Undermines Original Writing and Critical Thinking [Internet]. New York: College Board Newsroom; 2026.
11. CASRAI. AI Detector Accuracy: The False-Positive Evidence [Internet]. CASRAI; 2026.
12. Originality.ai. AI Detection Accuracy Studies — Meta-Analysis of 16 Studies [Internet]. Originality.ai; 2026.