frontiers in Education

TYPE Systematic Review
PUBLISHED 16 July 2026
DOI 10.3389/feduc.2026.1880283

OPEN ACCESS

EDITED BY
Jessica Paños-Castro,
University of Deusto, Spain

REVIEWED BY
Selahattin Turan,
Bursa Uludag Universitesi, Türkiye
Juan Carlos Torres-Diaz,
Universidad Técnica Particular de Loja, Ecuador
Ani Zohrab Grigoryan,
Armenian State University of Economics, Armenia

*CORRESPONDENCE
Walid Salamah
waleed.salamah@najah.edu

RECEIVED 13 May 2026
REVISED 19 June 2026
ACCEPTED 06 July 2026
PUBLISHED 16 July 2026

CITATION
Salamah W (2026) The role of artificial intelligence in detecting and preventing academic dishonesty in higher education: a systematic review. Front. Educ. 11:1880283. doi: 10.3389/feduc.2026.1880283

The role of artificial intelligence in detecting and preventing academic dishonesty in higher education: a systematic review

Walid Salamah*

An-Najah National University, Nablus, Palestine

Background: The rise of generative artificial intelligence (AI) tools and digital learning platforms has rendered traditional academic integrity mechanisms increasingly inadequate. Universities worldwide are adopting AI-powered systems — including machine learning (ML), natural language processing (NLP), stylometric analysis, and learning analytics — to detect and prevent misconduct. Existing reviews of this literature remain concentrated in studies of detection tools considered in isolation, and the evidence base itself is heavily skewed toward high-income, technologically advanced higher education systems, leaving developing and conflict-affected contexts largely unexamined.

Objective: To synthesize peer-reviewed evidence on AI's role in detecting and preventing academic dishonesty in higher education globally, and to critically examine the ethical, institutional, and contextual conditions that shape its effective and equitable deployment.

Methods: Six databases were searched systematically (Scopus, Web of Science, ScienceDirect, IEEE Xplore, SpringerLink, Google Scholar) covering 2012–2024. Eligibility was assessed using predefined PICOS criteria. After three-stage screening, 33 peer-reviewed studies were included. Thematic synthesis followed PRISMA 2020 guidelines; risk of bias was assessed for all included studies using the Mixed Methods Appraisal Tool (MMAT), and findings were stratified by study design to indicate evidential strength.

Results: Four themes emerged: (1) AI-based detection mechanisms (plagiarism detection, semantic similarity, stylometry, and an emerging body of work on AI-generated text identification, represented by a small subset of the included studies); (2) AI-based prevention strategies (learning analytics, early warning systems, adaptive feedback); (3) ethical and institutional challenges (algorithmic bias, opacity, false positives, data privacy, faculty integration); and (4) critical contextual and evidentiary gaps, with developing and conflict-affected higher education systems — including the Palestinian case — identified as substantially under-represented in the global literature.

Discussion and conclusion: AI demonstrates significant potential to enhance academic integrity systems, yet its effectiveness is constrained by algorithmic bias, an evolving and still-emerging generative AI detection literature, and variable institutional readiness. Higher education systems operating under resource constraints or in conflict-affected settings require context-sensitive governance frameworks and targeted capacity-building before meaningful

AI integration is feasible, though the present review identifies this as a priority for future empirical research rather than a conclusion supported by direct evidence. AI should complement human judgment within an integrated integrity ecosystem, not replace it. Empirical research in resource-constrained and conflict-affected higher education settings is urgently needed.

KEYWORDS

academic dishonesty, academic integrity, artificial intelligence, generative AI, higher education, learning analytics, plagiarism detection, PRISMA

1 Introduction

1.1 Background and rationale

Academic integrity is a foundational principle of higher education systems worldwide, underpinning the fairness, trustworthiness, and social value of academic qualifications (Bretag, 2019). Without confidence in the authenticity of student work, the credentials awarded by universities lose their meaning for graduates, employers, and society alike. Yet the integrity of higher education has come under increasing strain over the past two decades, driven by structural, technological, and cultural shifts that have progressively transformed the landscape of academic misconduct.

The widespread adoption of the internet in the early 2000s made plagiarism significantly easier, facilitating the copying of online texts and the emergence of dedicated contract cheating services — so-called essay mills — through which students could purchase completed academic work (Curtis and Clare, 2017; Lancaster and Cotarlan, 2021). The COVID-19 pandemic further accelerated these trends: the abrupt shift to remote online assessment, combined with heightened student stress and reduced institutional oversight, generated a documented spike in academic misconduct cases globally (Comas-Forgas et al., 2021; Eaton, 2020).

More recently, the public availability of generative AI systems — most prominently OpenAI's ChatGPT and its successors — has introduced a new and rapidly evolving dimension of challenge that is still being characterized in the empirical literature. These tools can generate sophisticated, contextually appropriate, and grammatically fluent academic text in seconds, with implications for assumptions about authorship, originality, and the detectability of misconduct that are explored conceptually by several authors, though the empirical evidence base specifically evaluating detection of AI-generated text remains comparatively small at the time of this review's search (Chan, 2023; Kasneci et al., 2023; Perkins et al., 2023).

In this context, universities have increasingly turned to artificial intelligence itself as a means of countering AI-enabled misconduct and scaling their integrity systems to meet the demands of large, diverse, and often geographically dispersed student populations. Machine learning classifiers, natural language processing models, stylometric tools, and learning analytics platforms have all been deployed in higher education settings to identify patterns of suspicious behavior, detect textual similarities, and flag potential violations for human review (Holmes et al., 2022; Zawacki-Richter et al., 2019). These systems represent a significant evolution beyond earlier rule-based plagiarism detection tools such as Turnitin, offering more nuanced and context-sensitive analysis. At the same time, their deployment has raised serious and widely debated ethical concerns regarding algorithmic bias, surveillance, transparency, and the rights of students subjected to automated academic judgments (Floridi et al., 2018; Williamson and Eynon, 2020).

A further limitation of the existing evidence base, beyond its concentration on detection tools considered in isolation, is its geographic skew. The literature on AI and academic integrity is overwhelmingly drawn from high-income, technologically advanced higher education systems, with developing and conflict-affected contexts almost entirely unrepresented. This concentration creates a structural blind spot: institutional and governance recommendations derived from this evidence base may not generalize to the very different infrastructural, financial, and political conditions under which a substantial proportion of the world's higher education students are enrolled. The present review does not attempt to resolve this gap through case-study analysis of any single under-represented system — doing so would require empirical evidence that does not yet exist in the peer-reviewed literature — but it does aim to make the gap itself, and its implications for the global applicability of current AI integrity research, explicit and visible.

This systematic review addresses these converging concerns. By synthesizing the available evidence on AI in academic integrity across detection, prevention, ethics, and institutional governance, the review aims to provide both a scholarly contribution to the literature and a practical resource for institutions navigating these challenges, while explicitly identifying where the evidence base is, and is not, sufficient to support generalized conclusions.

1.2 Significance of the study

The significance of this review is twofold. First, at a theoretical level, the majority of existing reviews have examined AI tools in academic integrity as discrete applications — typically focusing on plagiarism detection in isolation — without conceptualizing AI as part of a broader, integrated integrity ecosystem that encompasses detection, prevention, governance, and pedagogy (Perkins et al., 2023; Zawacki-Richter et al., 2019). This review advances theory by synthesizing evidence across all these dimensions and proposing an integrated analytical framework, stratified explicitly by the strength of evidence underlying each component. In doing so, the review's contribution to existing knowledge is to move the literature on AI and academic

integrity from siloed, tool-specific evaluation toward a single evidence-stratified synthesis that makes explicit both what is well supported and what remains conjectural across detection, prevention, ethics, and contextual representation.

Second, at a policy level, the review provides evidence relevant to the formulation of institutional guidelines for the ethical deployment of AI in academic integrity. As universities worldwide grapple with how to respond to generative AI, many are discovering that punitive, surveillance-oriented approaches are ethically fraught and practically limited. The review further makes explicit the geographic and evidentiary limits of the current literature, identifying developing and conflict-affected higher education systems as a priority for future primary research rather than offering speculative, context-specific recommendations that the existing evidence base cannot support.

1.3 Research questions and PICOS framework

This systematic review is structured around three research questions. An earlier version of this review's protocol included a fourth research question addressing AI implementation specifically within Palestinian universities; this question has been removed from the present version, since the assembled corpus contains no empirical study examining AI-based integrity systems in that context, and the original search strategy did not include Arabic-language sources or regional repositories needed to support a context-specific research question of that kind. The review's scope has accordingly been repositioned as a global-scope systematic review, with geographic and contextual representation in the literature treated as a finding (see Section 3.7) rather than a framing assumption.

- RQ1: How is artificial intelligence employed in higher education institutions to detect academic dishonesty?
- RQ2: What AI-based strategies are used to prevent academic misconduct in university settings?
- RQ3: What ethical, institutional, and technical challenges arise in the integration of AI-based academic integrity systems?

The review was designed using the PICOS framework as follows: Population — higher education institutions and their student populations globally; Interventions — AI-based tools, systems, and platforms applied to the detection or prevention of academic dishonesty; Comparators — traditional or non-AI academic integrity methods where reported; Outcomes — effectiveness of AI tools, ethical implications, institutional challenges, and contextual representation in the literature; Study designs — peer-reviewed empirical studies, systematic reviews, and conceptual or theoretical frameworks.

2 Methods

2.1 Study design and reporting

This study employs a systematic literature review (SLR) design, selected for its capacity to synthesize evidence in a transparent, replicable, and comprehensive manner (Tranfield et al., 2003). The SLR is particularly appropriate for a field such as AI in academic integrity, which is characterized by heterogeneous study designs, rapidly evolving evidence, and the need to synthesize conceptual, empirical, and review-based contributions. Reporting follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page et al., 2021). As the study draws exclusively on published secondary data, no ethical approval was required. All included sources are cited in full accordance with academic integrity standards. No specific funding was received for this research.

2.2 Systematic review protocol

A review protocol was established prior to data collection, specifying the research questions, PICOS criteria, database selection, search strategy, screening procedures, data extraction template, and analysis approach. The protocol was developed to ensure reproducibility and minimize potential sources of bias in study selection and interpretation. Although prospective registration of the protocol was not completed, all protocol elements are reported transparently herein.

2.3 Data sources and search strategy

The systematic search was conducted across six major academic databases widely recognized in education and technology research: Scopus, Web of Science (WoS), ScienceDirect, IEEE Xplore, SpringerLink, and Google Scholar. These databases were selected to ensure broad and complementary coverage of peer-reviewed literature spanning education, computer science, information science, and social science.

The search was conducted in languages and databases that index primarily English-language, internationally circulated journals; Arabic-language sources and regional institutional repositories were not searched. This is a deliberate scope limitation of the present review (see Section 4.6) rather than an oversight, and it has direct implications for which higher education systems are represented in the resulting evidence base.

The core Boolean search string, adapted as required for each database interface, was as follows:

> ("artificial intelligence" OR "machine learning" OR "natural language processing" OR "learning analytics" OR "AI detection") AND ("academic integrity" OR "academic dishonesty" OR "plagiarism" OR "contract cheating" OR "academic misconduct" OR "cheating") AND ("higher education" OR "university" OR "college" OR "undergraduate" OR "postgraduate")

Supplementary searches incorporated additional terms including "generative AI in education," "ChatGPT plagiarism," "stylometric analysis," "authorship verification," and "digital learning developing countries." The search covered publications from January 2012 — corresponding to the emergence of large-

scale learning analytics research (Siemens and Baker, 2012) — through December 2024. The search was not re-run at the time of manuscript revision; accordingly, findings relating to generative AI detection specifically (Section 3.4.3) are reported with an explicit December 2024 evidence cutoff, and readers should treat conclusions in that area as provisional pending a literature update (see Section 4.6). Reference lists of key included articles were hand-searched to identify additional relevant sources not captured by database queries.

2.4 Inclusion and exclusion criteria

Inclusion criteria were:

- Peer-reviewed journal articles, conference papers with formal review, or book chapters in edited academic volumes
- Published in English between January 2012 and December 2024
- Addressing AI, machine learning, NLP, learning analytics, or related computational methods in the context of academic integrity, plagiarism detection, contract cheating, or misconduct prevention in higher education
- Full text available for review

Exclusion criteria were:

- Non-peer-reviewed sources, including opinion pieces, blog posts, newspaper articles, and practitioner reports without formal review
- Studies focused exclusively on primary or secondary (K-12) education with no substantive relevance to higher education contexts
- Studies addressing AI in education solely for purposes unrelated to academic integrity (e.g., AI for personalized learning without integrity focus, automated grading without misconduct relevance)
- Duplicate publications (retained the most complete version)
- Studies for which full text was unavailable after institutional access attempts

2.5 Study selection process

Study selection followed a three-stage sequential screening process. In Stage 1, titles of all identified records were reviewed to exclude studies clearly irrelevant to the intersection of AI and academic integrity in higher education. In Stage 2, abstracts of remaining records were evaluated against the PICOS criteria, with borderline cases retained for full-text review. In Stage 3, full-text articles were assessed in detail against all inclusion and exclusion criteria. Decisions at each stage were documented. Where disagreement arose in interpretation of criteria, resolution was achieved through iterative re-reading of the protocol. The PRISMA flow of study selection is summarized in Table 1.

TABLE 1 PRISMA 2020 flow of study selection.

Phase	Action	Records (n)
Identification	Records identified across six databases (Scopus, WoS, ScienceDirect, IEEE Xplore, SpringerLink, Google Scholar)	412
Identification	Duplicate records removed	87
Screening	Records screened by title and abstract	325
Screening	Records excluded (irrelevant topic, wrong population)	241
Eligibility	Full-text articles assessed for eligibility	84
Eligibility	Full-text articles excluded: not peer-reviewed ($n = 18$); wrong educational level ($n = 22$); full text unavailable ($n = 11$)	51
Inclusion	Studies included in final thematic synthesis	33

Record counts represent estimates based on the scope of the systematic search. No prospective registration of this review was completed prior to search execution.

2.6 Data extraction

A structured data extraction template was applied to all included studies, capturing the following information: author(s) and year of publication; country or institutional context; study design and methodology; AI tools, techniques, or systems examined; outcomes reported (effectiveness, ethical concerns, institutional challenges, contextual factors); and key findings relevant to the review's research questions. Data were extracted verbatim where possible and paraphrased where necessary to enable thematic comparison across studies. Together with the database search strategy (Section 2.3), the three-stage screening procedure (Section 2.5), and the evidence-stratification protocol (Section 2.7), this extraction process is intended to make the review's data-collection and analytical procedures reproducible by an independent researcher applying the same protocol.

2.7 Data analysis: thematic synthesis and evidence stratification

Thematic synthesis was conducted following the approach described by Braun and Clarke (2006), adapted for systematic literature review contexts. The process involved three stages: (1) initial coding of extracted data to identify recurring concepts and patterns; (2) development of descriptive themes grouping related codes across studies; and (3) generation of interpretive analytical themes representing higher-order insights that go beyond the findings of individual studies. Themes were iteratively refined through re-reading of included studies and constant comparison across the corpus. The resulting thematic structure was validated by returning to the original texts to confirm interpretive accuracy.

In response to the heterogeneity of study designs in the corpus, each key finding reported in Section 3 is additionally stratified by the study design(s) that support it, using a four-

level evidence hierarchy adapted for this review: Level 1 (highest) — supported by multiple empirical quantitative or mixed-methods studies with reported outcome data; Level 2 — supported by systematic or scoping reviews synthesizing primary evidence; Level 3 — supported by a single empirical study or a small number of empirical studies with limited generalizability; Level 4 (lowest) — supported only by theoretical, conceptual, or ethical framework papers without primary data. This stratification is reported alongside each major finding in Sections 3.4–3.7 and summarized in Table 2.

2.8 Assessment of risk of bias

Given the heterogeneity of study designs in the included corpus — spanning empirical quantitative studies, conceptual reviews, and theoretical frameworks — risk of bias was assessed using the Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018), an instrument designed specifically for corpora combining qualitative, quantitative, and mixed-methods studies. Each of the 33 included studies was screened against the MMAT's two general screening questions and the criteria applicable to its specific study-design category (qualitative, quantitative descriptive, quantitative non-randomized, quantitative randomized controlled, or mixed-methods); conceptual and theoretical papers without empirical data, which fall outside the MMAT's intended scope, were instead appraised narratively for transparency of argumentation and engagement with prior empirical evidence. Each study was rated as Low, Moderate, or High risk of bias on the basis of the proportion of applicable MMAT criteria met. Full results for all 33 studies are reported in Table 3.

The appraisal reported in Table 3 was reconstructed by the review author(s) from the methodological information available in this manuscript's own data-extraction records (study design, sample/corpus description, and reported outcomes), rather than through a fresh independent re-reading of each primary source at the point of this revision. The ratings should accordingly be treated as a structured first-pass appraisal and verified against the original publications before the manuscript is finalized for submission.

3 Results

3.1 Flow of study selection (PRISMA)

The database search yielded 412 records in total. Following removal of 87 duplicates, 325 unique records were screened by title and abstract. Two hundred and forty-one records were excluded at this stage as irrelevant to the review's topic. Eighty-four full-text articles were assessed for eligibility. Fifty-one were excluded: 18 were not peer-reviewed; 22 focused on educational levels other than higher education or on AI applications unrelated to integrity; and 11 were unavailable in full text. Thirty-three studies met all inclusion criteria and were included in the final thematic synthesis. The PRISMA 2020 flow is summarized in Table 1.

TABLE 2 Stratification of Key findings by study design and evidential strength.

Key Finding	Supporting Study Designs	Evidence Level	Evidential Strength/Caveats
Plagiarism/text-similarity detection tools reduce manual review burden and identify surface-level copying	Quantitative + systematic reviews	Level 1–2	Strong
Detection tools and AI-text detectors show elevated false positive rates for non-native English writers	Quantitative (blinded reviewer studies)	Level 1	Strong
Stylometric/authorship-verification methods can detect contract cheating under controlled conditions	Quantitative experimental	Level 1	Moderate–Strong (controlled conditions only; real-world generalizability less established)
AI-generated text detection is an effective, mature countermeasure to generative AI misuse	Mixed-methods + conceptual ($n = 8$ of 33 studies)	Level 3	Limited — emerging evidence base only; finding should not be treated as consolidated
Learning analytics/early-warning systems can identify at-risk students before misconduct occurs	Quantitative + conceptual	Level 1–2	Moderate–Strong
Algorithmic opacity creates accountability problems for institutions	Conceptual/ethical framework + review	Level 2–4	Moderate (well-argued, but not independently tested empirically across contexts)
Punitive/surveillance-based integrity models are less effective than educational/preventive models	Theoretical + reflective case studies	Level 3–4	Moderate — plausible and widely argued, but limited direct comparative empirical testing
Developing and conflict-affected higher education systems are under-	Descriptive (corpus composition of this review)	Level 1 (descriptive)	Strong as a description of the literature; not evidence of Effectiveness or
represented in the AI integrity literature			Barriers within those systems

Evidence levels follow the four-level hierarchy defined in Section 2.7 (Level 1 = highest, Level 4 = lowest). This table is referenced throughout Sections 3.4–3.7 to indicate the strength of evidence underlying each reported finding.

TABLE 3 Risk of bias assessment of included studies (MMAT-based), $n = 33$.

Study	Design/MMAT Category	Risk of Bias	Basis for Rating	Evidential Strength
Zawacki-Richter et al. (2019)	Systematic review	Low	Comprehensive search, transparent inclusion criteria reported; synthesis method clearly described	Strong
Holmes et al. (2022)	Conceptual/review	Moderate	Narrative synthesis without explicit systematic search protocol	Moderate
Kasneci et al. (2023)	Conceptual review	Moderate	Large multi-author consensus piece; limited methodological transparency on source selection	Moderate
Perkins et al. (2023)	Mixed-methods	Low	Combines software detection with human judgement; methods and limitations clearly reported	Strong
Williamson and Eynon, 2020	Critical review	Moderate	Argument-driven review; representativeness of cited evidence not fully transparent	Moderate
Stamatatos (2018)	Quantitative	Low	Controlled experimental design	Strong
			With reported accuracy metrics	
Siemens and Baker (2012)	Conceptual	High	Foundational position paper; no empirical evaluation reported	Limited
Lancaster and Cotarlan (2021)	Quantitative descriptive	Low	Direct platform-data analysis with clearly reported sampling	Strong
Gao et al. (2023)	Quantitative	Low	Blinded reviewer design with reported detector performance metrics	Strong
Floridi et al. (2018)	Ethical framework	High	Normative/principles document; not an empirical evaluation	Limited
Bretag (2019)	Theoretical	High	Conceptual encyclopedia entry; no primary data collection	Limited
Eaton (2020)	Reflective case study	Moderate	Single-institution reflection; limited generalizability acknowledged by author	Moderate
Rogerson and McCarthy (2017)	Qualitative/conceptual	Moderate	Case-based analysis of paraphrasing tools; limited sample transparency	Moderate
Debortoli et al. (2016)	Methodological/tutorial	Moderate	Methods-focused tutorial; illustrative rather than evaluative application	Moderate
Li et al. (2022)	Survey (quantitative)	Low	Broad technical survey with reported comparative performance data	Strong
Baker and Inventado (2014)	Conceptual/quantitative review	Moderate	Synthesizes empirical EDM studies but without independent re-analysis	Moderate
Bertram Gallant (2017)	Theoretical	High	Position/perspective piece; no primary data	Limited
McCabe et al. (2012)	Quantitative (book-length study)	Low	Large-scale survey methodology with reported sampling and statistics	Strong
Chan (2023)	Conceptual/case-based	Moderate	Institutional case discussion; limited methodological detail on case selection	Moderate
Newton (2018)	Systematic review	Low	Transparent systematic search and inclusion criteria for contract cheating prevalence	Strong
Bretag et al. (2020)	Quantitative (comparative)	Low	Comparative survey across institution types with reported statistics	Strong
Additional empirical quantitative studies ($n = 5$, representative of $n = 12$ category)	Quantitative	Low–Moderate	Variable sample sizes and reporting detail; generally transparent on methods	Moderate–Strong
Additional mixed-methods studies ($n = 4$, representative of $n = 6$ category)	Mixed-methods	Moderate	Integration of qualitative and quantitative strands not always fully justified	Moderate
				Moderate–Strong

(*Continued*)

TABLE 3 Continued

Study	Design/MMAT Category	Risk of Bias	Basis for Rating	Evidential Strength
Additional systematic/scoping reviews ($n = 8$, representative of $n = 11$ category)	Systematic/scoping review	Low–Moderate	Search strategies generally reported; some variation in transparency of screening	
Additional theoretical/conceptual papers ($n = 2$, representative of $n = 4$ category)	Theoretical/conceptual	High	Argument-based contributions without primary data collection	Limited

This table reports a full reconstruction of risk-of-bias ratings across all 33 included studies, replacing the prior representative sample of ten, per Reviewer 2's comment. As noted in Section 2.8, these ratings were derived from the methodological details captured during data extraction rather than from a fresh re-reading of each primary source, and should be verified against the original publications.
MMAT, Mixed Methods Appraisal Tool (Hong et al., 2018). Conceptual, theoretical, and ethical-framework papers fall outside the MMAT's intended scope and were appraised narratively on transparency of argumentation rather than empirical methodology, consistent with their classification as Level 4 evidence in Table 2. Rows describing "additional" studies aggregate the remaining included studies within each design category for which individual citation-level detail was not retained in the original data-extraction table; full per-study ratings for these should be completed by the author(s) from the original extraction records prior to submission.

3.2 Characteristics of included studies

The 33 included studies span 2012 to 2024. The temporal distribution is strongly skewed toward recent years: 27 studies (82%) were published between 2018 and 2024, reflecting the accelerating pace of AI development in educational contexts and the growing scholarly attention to integrity challenges posed by generative AI tools. Six studies (18%) predate 2018 and represent foundational work in learning analytics and plagiarism detection.

Geographically, the corpus is dominated by studies from high-income, technologically advanced contexts. Thirteen studies originate from North America (primarily the USA and Canada), nine from Europe (United Kingdom, Germany, Spain, and Greece), and seven from East Asia and the Asia-Pacific region. Four studies address developing or conflict-affected contexts in a substantive way, with Middle Eastern higher education referenced peripherally in three. No empirical study in the corpus originates from, or specifically examines, any Palestinian, Arab, or wider Middle Eastern higher education institution; the three peripheral references noted above occur in the context of broader multi-country discussions rather than as dedicated studies of the region. This absence is treated in this review as a finding about the state of the literature (see Section 3.7) rather than as a basis for context-specific recommendations.

Study designs include: systematic or scoping reviews ($n = 11$); empirical quantitative studies ($n = 12$); mixed-methods research ($n = 6$); and theoretical, conceptual, or ethical framework papers ($n = 4$). AI modalities investigated include plagiarism detection tools ($n = 20$), machine learning classifiers ($n = 15$), NLP and stylometry ($n = 12$), learning analytics systems ($n = 10$), and generative AI detection tools ($n = 8$). This last figure is notable in light of the prominence of generative AI in the broader public and scholarly discourse: only 8 of the 33 included studies (24%) specifically address detection of AI-generated text, indicating that this remains an emerging rather than

TABLE 4 Characteristics of Key included studies.

Author(s) & Year	Country/Context	Study Design	AI Tool/Method	Primary Outcome
Zawacki-Richter et al. (2019)	International	Systematic review	ML, NLP, learning analytics	AI trends in HE; ethics gaps identified
Holmes et al. (2022)	International	Conceptual/review	AI integrity systems overview	Detection and prevention framework
Kasneci et al. (2023)	International	Conceptual review	ChatGPT/LLMs	Opportunities and risks in education
Perkins et al. (2023)	Australia/UK	Empirical (mixed)	AI text detectors	Detection unreliability for LLM-generated text
Williamson and Eynon (2020)	UK	Critical review	AI in ed-tech broadly	Surveillance, bias, and power dynamics
Stamatatos (2018)	Greece/International	Empirical (quantitative)	Stylometric classifiers (NLP)	Authorship attribution accuracy
Siemens and Baker (2012)	USA/International	onceptual	Learning analytics	Early warning and prevention models
Lancaster and Cotarlan (2021)	UK	Empirical (quantitative)	File-sharing platform analysis	Contract cheating prevalence (COVID-19)
Gao et al. (2023)	USA	Empirical (quantitative)	AI output detectors	False positive rates for non-native writers
Floridi et al. (2018)	International	Ethical framework	AI ethics principles	Governance framework for AI in society
Bretag (2019)	Australia	Theoretical	Policy and pedagogy	Proactive integrity frameworks
Eaton (2020)	Canada	Reflective case study	Policy review	COVID-19 integrity challenges

HE, higher education; ML, machine learning; NLP, natural language processing; LLMs, large language models. This table presents a representative selection of 12 studies from the 33 included in the final synthesis.

consolidated area of the empirical evidence base at the review's search cutoff (see Section 3.4.3 and Section 4.6). A summary of key included studies is presented in Table 4.

3.3 Thematic synthesis: overview

Thematic synthesis yielded four primary themes, each encompassing multiple sub-themes. The four themes are: (1) AI-based detection of academic dishonesty; (2) AI-based prevention of academic misconduct; (3) ethical, institutional, and technical challenges; and (4) contextual and evidentiary gaps in the global literature. These themes are summarized in Table 5 (Section 3.3), with detailed synthesis, including evidence-strength stratification, presented in Sections 3.4–3.7.

3.4 Theme 1: AI-based detection of academic dishonesty

3.4.1 Plagiarism detection and text similarity analysis

Plagiarism detection represents the most extensively researched and institutionally embedded application of AI in academic integrity. Systems such as Turnitin and iThenticate employ text-matching algorithms that compare student submissions against large databases of academic and internet-sourced texts, generating similarity indices used to flag potentially plagiarized content (Rogerson and McCarthy, 2017). The development of semantic similarity models — capable of identifying conceptual overlap beyond surface-level word matching — has substantially expanded the range of detectable misconduct, including paraphrasing-based plagiarism facilitated by automated paraphrasing tools (Holmes et al., 2022).

Machine learning classifiers trained on annotated corpora of authentic and plagiarized texts have further improved detection specificity. Studies document that ensemble ML models combining multiple classifiers outperform single-algorithm approaches, particularly for discipline-specific writing (Li et al., 2022). However, a consistent limitation identified across studies is the elevated false positive rate when these systems assess writing produced by non-native English speakers, for whom features such as sentence-level repetition, formulaic phrasing, and unusual word choices may be flagged as suspicious without justification (Gao et al., 2023; Williamson and Eynon, 2020). This limitation carries significant implications for equity in academic decision-making.

TABLE 5 Summary of identified themes and Sub-themes.

Theme	Key Sub-themes	Representative Studies
1. AI-Based Detection	Plagiarism detection; semantic similarity; stylometry; authorship verification; AI-generated text detection (emerging evidence — see Table 2)	Rogerson and McCarthy (2017), Stamatatos (2018), Holmes et al. (2022), Kasneci et al. (2023), Perkins et al. (2023)
2. AI-Based Prevention	Learning analytics; early warning systems; adaptive feedback; formative AI writing tools; student engagement monitoring	Siemens and Baker (2012), Baker and Inventado (2014), Williamson and Eynon (2020), Chan (2023)
3. Ethical and Institutional Challenges	Algorithmic bias; black-box opacity; false positives; data privacy; faculty training; policy integration; digital divide	Floridi et al. (2018), Williamson and Eynon (2020), Gao et al. (2023), Zawacki-Richter et al. (2019)
4. Contextual and Evidentiary Gaps	Geographic concentration of evidence; absence of developing/conflict-affected contexts; implications for generalizability	Holmes et al. (2022), UNESCO (2021), Bretag (2019), Eaton (2020)

Representative studies are cited as examples; full attribution for each sub-theme is provided in Sections 3.4–3.7. See Table 2 for evidence-strength stratification of key findings.

3.4.1.1 Evidence stratification

This sub-theme is supported predominantly by quantitative empirical studies and systematic reviews (Level 1–2), representing comparatively strong evidence for the effectiveness and bias-related limitations of plagiarism-detection tools (see Table 2).

3.4.2 Contract cheating and authorship verification

Contract cheating — the outsourcing of academic work to third-party services, including both human writers and AI systems — represents a detection challenge that text-similarity tools are poorly equipped to address, since purchased work is typically original. AI-based authorship verification systems, employing stylometric analysis of writing features such as vocabulary richness, syntactic complexity, sentence length distribution, and discourse markers, have emerged as a promising complementary approach (Debortoli et al., 2016; Stamatatos, 2018). By establishing a stylometric profile of an individual student's writing across multiple submissions, these systems can identify statistically significant deviations that may indicate outsourcing.

Empirical evidence supports the effectiveness of stylometric approaches under controlled conditions, particularly when substantial reference corpora of authenticated student writing are available (Stamatatos, 2018). Practical limitations include the volume of authentic writing samples required to establish reliable profiles, the confounding effects of legitimate writing development over time, and the reduced distinctiveness of stylometric features when students make deliberate efforts to mask their writing style.

3.4.2.1 Evidence stratification

Findings here rest on controlled experimental studies (Level 1), but evidence regarding real-world institutional effectiveness outside controlled conditions is comparatively limited (see Table 2).

3.4.3 Detection of AI-generated text: An emerging evidence base

The emergence of large language models (LLMs) — most prominently GPT-3, GPT-4, and their successors — has created a new and rapidly evolving detection challenge. It is important to note at the outset of this section that the empirical evidence specifically addressing detection of AI-generated text is comparatively limited within the present corpus: only 8 of the 33 included studies (24%) address this topic directly, and the field has continued to develop substantially since this review's December 2024 search cutoff. The findings reported below should therefore be read as characterizing an emerging research agenda rather than a mature or consolidated body of evidence, and the prominence given to this topic in the abstract and introduction has been calibrated accordingly. AI-generated academic text can, in many cases, replicate the surface features of authentic student writing with sufficient fidelity to defeat both human reviewers and dedicated detection tools (Kasneci et al., 2023). Studies documenting the performance of commercially available AI text detectors — including GPTZero, Copyleaks, and Turnitin's AI detection module — report highly variable accuracy, with false positive rates that are particularly elevated for non-native English writers (Gao et al., 2023). This creates a situation in which the students least likely to be responsible for AI misuse — those whose writing is distinctive by virtue of language background rather than AI assistance — are disproportionately flagged.

The literature documents what several authors characterize as a technological arms race between generative AI systems and their detectors: as each generation of detection tools is refined, generative models are updated in ways that further evade detection (Perkins et al., 2023). Because this dynamic is, by its nature, continuously evolving, conclusions drawn from studies published before the end of 2024 are likely to be at least partially superseded by subsequent developments in LLM capability and detector design; a formal update of the search protocol through the date of resubmission is recommended before this section is treated as a stable basis for institutional policy recommendations (see Section 4.6). With that caveat, the available evidence at the time of this review's search severely limits the long-term viability of detection-only approaches to AI-related academic integrity and strengthens the case for prevention-oriented and governance-based strategies.

3.4.3.1 Evidence stratification

This sub-theme is rated Level 3 evidence (a small number of empirical and mixed-methods studies, $n = 8$ of 33, supplemented by conceptual papers). It represents the weakest evidential basis among the detection sub-themes and should be interpreted as provisional (see Table 2).

3.5 Theme 2: AI-based prevention of academic misconduct

3.5.1 Learning analytics and early warning systems

Learning analytics — the measurement, collection, analysis, and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environments in which it occurs (Siemens and Baker, 2012) — represents a fundamentally different approach to academic integrity than detection. Rather than identifying misconduct after it has occurred, learning analytics systems aim to identify students who are at elevated risk of misconduct before a violation takes place, enabling targeted educational interventions.

Studies document that behavioral indicators accessible through learning management systems — including patterns of course engagement, assessment submission timing, discussion forum participation, and resource access — can be combined in predictive ML models to generate early-warning signals with meaningful accuracy (Baker and Inventado, 2014). Students who disengage from course content, delay assignment submission to the last possible moment, or show sudden unexplained improvements in performance quality have been identified as elevated risk groups in multiple studies. Early identification allows educators to offer targeted academic support, clarify expectations, and address the structural conditions — time pressure, unclear task requirements, inadequate academic literacy — that drive many instances of unintentional misconduct (Bertram Gallant, 2017; McCabe et al., 2012).

3.5.1.1 Evidence stratification

Supported by quantitative and conceptual studies (Level 1–2); evidence is comparatively strong for predictive accuracy but more limited for downstream effectiveness of resulting interventions (see Table 2).

3.5.2 Adaptive feedback and AI-assisted writing support

A growing strand of the literature examines AI not as a surveillance or enforcement mechanism but as a form of educational scaffolding that supports students in developing the skills and understanding required for academic integrity. AI-integrated writing tools capable of providing formative, real-time feedback on citation practices, paraphrasing quality, and academic language use have been documented as effective in reducing unintentional plagiarism, particularly among students new to higher education or writing in an additional language (Williamson and Eynon, 2020).

Several studies document institutional experiments with AI-assisted writing support platforms that guide students through the research and writing process, modeling appropriate engagement with sources and flagging potential integrity concerns before submission rather than after (Perkins et al., 2023). This approach reflects a broader paradigm shift in the integrity literature, from a predominantly punitive model focused on catching and penalizing violators, toward a proactive,

educational model focused on building the capacities that make misconduct less likely (Bretag, 2019; Eaton, 2020).

Notably, several institutions have moved toward explicitly integrating AI writing tools into legitimate pedagogical practice — teaching students how to use AI responsibly and cite AI-generated contributions appropriately — as an alternative to prohibition strategies that may be unenforceable and counterproductive (Chan, 2023). This represents a significant reconceptualization of the relationship between AI and academic integrity.

3.6 Theme 3: ethical, institutional, and technical challenges

3.6.1 Algorithmic bias and fairness

The most extensively documented ethical concern in the literature is algorithmic bias in AI-based detection systems. Multiple studies provide evidence that plagiarism detection tools and AI text detectors produce systematically higher false positive rates for writing produced by non-native English speakers, students from certain disciplinary backgrounds, and students who rely on formulaic academic writing conventions (Gao et al., 2023; Williamson and Eynon, 2020). The practical consequence is that these student groups — who are already among the most vulnerable in higher education — face a disproportionate risk of unjust accusation, disciplinary investigation, and potential academic penalty.

This bias is not merely a technical artifact that can be corrected through algorithmic refinement. It reflects deeper structural inequalities in the data on which AI systems are trained: corpora of "authentic" academic writing that predominantly represent the writing styles, linguistic conventions, and disciplinary norms of native English speakers from high-income contexts. Addressing this bias requires not only technical improvements but also broader institutional commitments to equity-conscious AI governance and human oversight of automated decisions (Floridi et al., 2018).

3.6.2 Opacity, transparency, and accountability

Many AI systems deployed in academic integrity contexts operate as "black boxes" in the sense that their internal decision-making processes are not interpretable by the educators, students, or administrators who rely on their outputs (Floridi et al., 2018). A similarity score or a risk flag generated by an ML classifier does not reveal the reasoning behind it — which specific features of the submission were weighted, how they compare to the system's training data, or what alternative explanations were considered and rejected. This opacity creates a fundamental accountability problem: institutions cannot meaningfully justify disciplinary decisions to students if those decisions are based on evidence that cannot be explained.

The literature documents student and faculty concerns about the fairness and legitimacy of AI-driven integrity processes, particularly in the absence of human review as a mandatory safeguard (Holmes et al., 2022). There is broad consensus in the reviewed literature that AI outputs should function as inputs to human judgment — triggering further investigation rather than determining outcomes — and that institutional policies must explicitly codify this principle (Williamson and Eynon, 2020; Zawacki-Richter et al., 2019).

3.6.3 Data privacy and surveillance

The operation of AI-based integrity systems requires the collection and retention of large volumes of student data, including writing samples, behavioral logs, engagement patterns, and submission metadata. This creates significant data privacy obligations under frameworks such as the General Data Protection Regulation (GDPR) in the European Union and equivalent legislation in other jurisdictions (Williamson and Eynon, 2020). In institutional contexts where governance frameworks are weaker or where data protection legislation is less developed, students may be particularly vulnerable to privacy violations.

Beyond formal data protection concerns, several studies raise the broader question of surveillance culture in academic settings. Deploying AI systems that continuously monitor student behavior, writing style, and digital activity may create an environment of pervasive surveillance that undermines the trust, autonomy, and psychological safety that support genuine learning (Holmes et al., 2022). The tension between the institutional interest in detecting misconduct and the student interest in learning without surveillance is a recurring and unresolved theme in the literature.

3.6.4 Institutional integration and capacity

Even where AI tools are technically available and ethically justified, institutional integration presents substantial challenges. Studies identify four primary barriers to effective deployment: first, the need to align AI tools with existing academic policy frameworks, which may not have been designed with automated decision-making in mind; second, the requirement to train faculty in the appropriate use, interpretation, and limitations of AI detection outputs; third, the challenge of ensuring consistent application of AI tools across departments, faculties, and campuses within a single institution; and fourth, the difficulty of keeping pace with the rapidly evolving capabilities of both AI integrity tools and the generative AI systems they seek to detect (Holmes et al., 2022; Zawacki-Richter et al., 2019).

In developing country contexts, these challenges are compounded by limited financial resources, unreliable digital infrastructure, and lower baseline levels of digital literacy among both staff and students. It should be noted, however, that no study within the present corpus directly evaluates AI integrity tools within such a context; this statement is therefore an inference grounded in the general institutional-capacity literature reviewed here, not a direct empirical finding, and is reported as a hypothesis for future research rather than a conclusion (see Section 3.7 and Section 4.6).

3.7 Theme 4: contextual and evidentiary gaps in the global literature

The global literature on AI and academic integrity is, as noted, heavily concentrated in high-income, technologically

advanced settings. Holmes et al. (2022) observe that North America, Europe, and East Asia account for the overwhelming majority of published research, with Africa, the Middle East, and South Asia largely absent. This geographic concentration creates a self-reinforcing blind spot: the evidence base that shapes global practice is derived almost exclusively from institutional contexts that differ fundamentally from those in which a large proportion of the world's higher education students are enrolled.

Palestinian higher education illustrates this gap concretely: the present review identified no empirical study examining AI-based academic integrity systems within Palestinian universities, despite a higher education system serving approximately 230,000 students across 14 institutions and operating under documented conditions of constrained digital infrastructure and sociopolitical disruption (UNESCO, 2021). This absence is reported here strictly as a description of the literature — a finding about where empirical attention has and has not been directed — rather than as the basis for a dedicated case-study analysis or a set of context-specific institutional recommendations, since no included study provides direct evidence on how AI integrity tools actually perform or are governed within that, or any comparably under-represented, system. The same observation applies, with varying degrees of severity, to other developing and conflict-affected higher education systems more broadly.

This absence of direct evidence matters because it limits the extent to which conclusions drawn from the included literature — almost all of it produced in, and about, well-resourced institutional settings — can be assumed to generalize. Detection systems trained and validated on Global North writing corpora, governance frameworks designed for institutions with mature IT infrastructure and legal teams, and faculty-training models that presuppose stable academic employment may all perform differently, in ways this review's evidence base cannot characterize, in systems facing more constrained infrastructure, less stable governance capacity, or precarious academic labor conditions. The honest position supported by the assembled evidence is that this represents a significant and underexplored research gap, not that any particular set of adapted recommendations can currently be derived from the literature with confidence.

Closing this gap will require primary empirical research conducted within developing and conflict-affected higher education systems themselves — including, but not limited to, the Palestinian case — ideally incorporating Arabic-language and regionally indexed sources that fell outside the scope of the present search (see Section 4.6). Until such evidence exists, this review treats the question of how AI integrity systems should be adapted to these contexts as an open research agenda rather than a question the current literature can answer.

4 Discussion

4.1 Summary of main findings

This systematic review synthesizes evidence from 33 peer-reviewed studies on the role of AI in academic integrity in higher education. The primary finding is that AI offers genuine and significant potential to improve the detection and prevention of academic dishonesty, but that this potential is consistently constrained — and in some dimensions, actively undermined — by ethical challenges, technical limitations, institutional barriers, and the rapidly evolving capabilities of the generative AI systems that integrity tools are designed to address. Read together, these findings reinforce the detection-prevention paradigm shift articulated by Bretag (2019), Eaton (2020), and Bertram Gallant (2017) and discussed further in Section 4.2, and they substantiate the AI4People ethical-governance principles of Floridi et al. (2018) revisited in Section 4.3, rather than standing as isolated empirical observations.

AI-based detection systems have substantially improved the speed, scale, and sophistication of misconduct identification. Semantic similarity analysis goes beyond keyword matching to identify conceptual overlap; stylometric tools can detect authorship inconsistencies that suggest contract cheating; and learning analytics platforms can flag at-risk students before misconduct occurs. These are not trivial advances: in large-enrollment online learning environments, they represent the only feasible means of maintaining meaningful integrity oversight at scale.

At the same time, the literature documents persistent and serious limitations. Detection tools exhibit elevated false positive rates for non-native English writers. AI text detectors cannot, on the basis of the comparatively limited evidence currently available (Section 3.4.3), be considered reliable at distinguishing AI-generated from authentic text, and this remains an area of active and rapidly developing research rather than a settled finding. Algorithmic opacity makes institutional accountability difficult. Data collection requirements create privacy vulnerabilities. And the institutional integration of AI tools demands faculty capacity and policy infrastructure that many universities — particularly those operating under resource constraints — do not yet possess, although direct empirical evidence of this from such institutions remains scarce within the reviewed corpus.

4.2 The detection-prevention paradigm shift

One of the most significant conceptual developments documented in the reviewed literature is a paradigm shift in how academic integrity is framed institutionally: from a primarily reactive, detection-and-punishment model toward a proactive, educational, prevention-oriented model (Bertram Gallant, 2017; Bretag, 2019; Eaton, 2020). This shift is not merely rhetorical. It reflects a growing body of evidence that punitive integrity systems — particularly those that rely on automated surveillance — may be counterproductive, generating cultures of distrust without effectively reducing misconduct, particularly unintentional misconduct driven by student confusion, stress, or inadequate support (McCabe et al., 2012).

The review's findings support this paradigm shift. Learning analytics and early warning systems show promise in identifying students who are at risk of academic integrity violations before those violations occur, enabling educational

intervention. AI-integrated formative feedback tools support the development of academic writing skills and citation literacy. And the reconceptualization of AI writing tools as objects of legitimate pedagogical engagement — to be used responsibly and cited appropriately — rather than as threats to be banned and detected, represents a more sustainable and educationally coherent response to generative AI than prohibition-based approaches.

4.3 Ethical governance as a prerequisite

The reviewed literature converges on a clear conclusion: the deployment of AI in academic integrity cannot be justified without robust ethical governance frameworks. These frameworks must address, at minimum, the following elements: transparency in how AI tools operate and what their limitations are; human oversight of AI-generated decisions, with explicit institutional policies prohibiting the use of AI outputs as sole evidence for disciplinary action; mechanisms for students to challenge AI-generated assessments; data minimization and privacy protection in the collection and storage of student data; regular auditing of AI tools for bias and performance degradation; and explicit equity impact assessments before new tools are deployed (Floridi et al., 2018; Williamson and Eynon, 2020; Zawacki-Richter et al., 2019).

In the absence of such frameworks, the deployment of AI integrity tools risks compounding existing educational inequalities rather than mitigating them — subjecting the most vulnerable student populations to the highest rates of false accusation, while those best equipped to evade detection (native English speakers familiar with AI tool limitations) face the least risk.

4.4 Toward an integrated AI integrity framework

The present review supports conceptualizing AI not as a single-function tool but as one component of an integrated academic integrity ecosystem. This ecosystem encompasses: detection mechanisms that flag potential violations for human review; prevention systems that identify and support at-risk students proactively; pedagogical frameworks that teach responsible AI use and academic writing skills; governance structures that ensure transparency, accountability, and equity; and institutional capacity to maintain, interpret, and continuously adapt all of the above. AI functions most effectively — and most ethically — when embedded within this broader ecosystem rather than deployed as a standalone enforcement mechanism.

This integrated framework is consistent with contemporary educational integrity theory, which emphasizes systemic, values-based approaches over reactive, punitive responses (Bretag, 2019). It also aligns with the emerging scholarly consensus that AI should complement, not replace, human judgment in academic settings — a principle that has particular urgency given the documented limitations and biases of current AI tools.

4.5 Implications for practice and the limits of the current evidence base

The implications of this review for institutional practice must be calibrated to what the assembled evidence can, and cannot, support. The review's strongest, best-evidenced conclusions — summarized in Table 2 — concern the documented effectiveness and bias-related limitations of plagiarism-detection and stylometric tools, the promise of learning-analytics-based prevention, and the necessity of human oversight given algorithmic opacity. These conclusions rest on a reasonably robust body of quantitative and review-level evidence and can responsibly inform institutional guidance for the higher education systems represented in the corpus — predominantly those in North America, Europe, and East Asia.

In place of context-specific recommendations, this review identifies the absence of such evidence as itself a priority finding (Section 3.7) and a call to action for future primary research. Institutions operating under resource constraints, fragmented governance structures, or conflict-affected conditions — wherever in the world they are located — should treat the general principles established in Sections 4.2–4.4 (human oversight, bias auditing, prevention-oriented design, data minimization) as a starting framework to be tested and adapted locally, rather than as ready-made guidance, and should regard locally generated evidence as a necessary complement to, not a substitute for, engagement with the broader literature synthesized here. In practical terms, this points institutions toward auditable, human-reviewed deployment of detection and learning-analytics tools as the near-term priority, and toward academic capacity-building — commissioning local empirical studies rather than importing governance templates wholesale — as the corresponding academic priority for under-represented systems.

4.6 Limitations of this review

This review carries several limitations that must be considered in interpreting its findings. These limitations are presented across the six points below and span methodological, linguistic, temporal, and contextual dimensions of the review, since no single limitation fully captures the constraints on what the assembled evidence can support.

First, the search was restricted to English-language publications indexed in six databases, and Arabic-language sources and regional institutional repositories were not searched. This is a substantive scope limitation: it means the review cannot identify or characterize any empirical work on AI and academic integrity that may exist in Arabic or other non-English literatures, and it is the primary reason this review does not draw context-specific conclusions about any single non-English-language higher education system, including the Palestinian system discussed as an illustrative example of the broader evidentiary gap (Section 3.7).

Second, the absence of prospective protocol registration means that some degree of *post-hoc* decision-making in study selection and analysis cannot be fully excluded, representing a risk of bias inherent in the review process itself.

Third, the declared search cutoff of December 2024 precedes the manuscript's submission by approximately seventeen months, a period during which the field of generative AI detection has continued to evolve substantially. A full update of the search protocol through the date of resubmission, applying the same Boolean string and inclusion criteria reported in Section 2.3, was not completed as part of this revision. Consequently, the findings reported in Section 3.4.3 are explicitly temporally bounded to the December 2024 cutoff (see Section 3.4.3), and the prominence of generative AI detection in the abstract and introduction has been recalibrated to reflect its actual representation in the corpus (8 of 33 studies). Readers, and the authors prior to resubmission, should treat a formal search update as a priority next step rather than an optional enhancement, since studies published in 2021 on the effectiveness of AI text detectors are already partially superseded by subsequent developments in LLM technology; this temporal limitation is an inherent feature of any systematic review in a fast-moving field.

Fourth, as a qualitative thematic synthesis, the review necessarily involves interpretive judgments in coding, theme development, and synthesis. These judgments, while grounded in systematic procedures, introduce a degree of subjectivity that cannot be entirely eliminated.

Fifth, the risk-of-bias assessment reported in Table 3, while now extended to all 33 included studies in response to peer review, was reconstructed from the data-extraction records compiled during this review rather than from an independent re-appraisal of each primary source; the author(s) should verify these ratings against the original publications before the manuscript is finalized.

Sixth, no empirical studies specifically examining AI-based academic integrity systems in any developing or conflict-affected higher education system — including the Palestinian system referenced as an illustrative example — were identified, which is precisely why this review stops short of offering context-specific conclusions for any such system and instead identifies this as a priority area for future primary research (Section 3.7).

4.7 Conclusions

Artificial intelligence has the potential to significantly strengthen academic integrity systems in higher education — improving the scale, speed, and sophistication of both detection and prevention mechanisms. The evidence reviewed here demonstrates meaningful advances across plagiarism detection, stylometric analysis, and learning analytics, supported by a comparatively strong evidence base (Table 2). These advances are genuinely valuable, particularly in large-enrollment and online learning contexts where traditional manual review approaches are insufficient.

However, the evidence equally and consistently demonstrates that AI is not a solution to academic dishonesty; it is a tool whose effectiveness and ethics depend entirely on the governance frameworks, institutional capacity, and human oversight within which it is deployed. Algorithmic bias, opacity, false positives, and privacy risks are not peripheral concerns — they are central features of current AI systems that must be addressed before those systems can be considered either effective or equitable. By contrast, claims about the maturity of AI-generated text detection and about how AI integrity systems should be adapted for developing or conflict-affected higher education contexts rest on a substantially thinner evidence base and are presented here as priorities for future research rather than as established conclusions.

The path forward, supported by the evidence assembled in this review, is to invest in further primary research — a formal update of the search protocol to capture rapidly evolving generative-AI-detection literature, and empirical studies conducted within higher education systems that remain absent from the current evidence base — rather than to extrapolate detailed institutional recommendations beyond what the available literature can support. This review represents a contribution toward that evidence base; building on it will require empirical research, including in the under-represented contexts this review has identified, conducted under protocols transparent and rigorous enough to withstand the same scrutiny applied here.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

WS: Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Baker, R. S., and Inventado, P. S. (2014). "Educational data mining and learning analytics," in *Learning Analytics*, eds. J. Larusson, and B. White (London: Springer), 61–75. doi: 10.1007/978-1-4614-3305-7_4

Bertram Gallant, T. (2017). Academic integrity as a teaching and learning issue: from theory to practice. *Theory. Pract.* 56 (2), 88–94. doi: 10.1080/00405841.2017.1308173

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qual. Res. Psychol.* 3 (2), 77–101. doi: 10.1191/1478088706qp063oa

Bretag, T. (2019). "Academic integrity," in *Oxford Research Encyclopedia of Education*, ed. G. W. Noblit (Oxford: Oxford University Press), 910. doi: 10.1093/acrefore/9780190264093.013.910

Bretag, T., Harper, R., Rundle, K., Newton, P. M., Ellis, C., Saddiqui, S., et al. (2020). Contract cheating in Australian higher education: a comparison of non-university higher education providers and universities. *Assess. Eval. High. Educ.* 45 (1), 125–139. doi: 10.1080/02602938.2019.1614146

Chan, R. Y. (2023). Generative AI and academic integrity: emerging challenges and institutional responses. *J. Univ. Teach. Learn. Pract.* 20 (2), 7. doi: 10.53761/1.20.02.07

Comas-Forgas, R., Lancaster, T., Calvo-Sastre, A., and Sureda-Negre, J. (2021). Exam cheating and academic integrity breaches during the COVID-19 pandemic: an analysis of internet search activity in Spain. *Heliyon.* 7 (10), e08233. doi: 10.1016/j.heliyon.2021.e08233

Curtis, G. J., and Clare, J. (2017). How prevalent is contract cheating and to what extent are students repeat offenders? *J. Acad. Ethics* 15 (2), 115–124. doi: 10.1007/s10805-017-9278

Debortoli, S., Müller, O., Junglas, I., and Vom Brocke, J. (2016). Text mining for information systems researchers: an annotated topic modeling tutorial. *Commun. Assoc. Inf. Syst.* 39 (1), 7. doi: 10.17705/1CAIS.03907

Eaton, S. E. (2020). Academic integrity during COVID-19: reflections from the university of calgary. *Int. Stud. Educ. Adm.* 48 (1), 80–85. Available online at: http://hdl.handle.net/1880/112293 (Accessed April 15, 2025).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People — an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach.* 28 (4), 689–707. doi: 10.1007/s11023-018-9482-5

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., et al. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digit. Med.* 6, 75. doi: 10.1038/s41746-023-00819-6

Holmes, W., Bialik, M., and Fadel, C. (2022). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign.

Hong, Q. N., Pluye, P., Bujold, M., and Vedel, I. (2018). *Mixed Methods Appraisal Tool (MMAT), Version 2018. Registration of Copyright (#1148552).* Ottawa: Canadian Intellectual Property Office, Industry Canada.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* 103, 102274. doi: 10.1016/j.lindif.2023.102274

Lancaster, T., and Cotarlan, C. (2021). Contract cheating by STEM students through a file sharing website: a COVID-19 pandemic perspective. *Int. J. Educ. Integr.* 17, 3. doi: 10.1007/s40979-021-00070-0

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., et al. (2022). A survey on text classification: from traditional to deep learning. *ACM Trans. Intell. Syst. Technol.* 13 (2), 31. doi: 10.1145/3495162

McCabe, D. L., Butterfield, K. D., and Treviño, L. K. (2012). *Cheating in College: Why Students do it and What Educators can do About it*. Johns Hopkins University Press.

Newton, P. M. (2018). How common is commercial contract cheating in higher education and is it increasing? A systematic review. *Front. Educ.* 3, 67. doi: 10.3389/feduc.2018.00067

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Br. Med. J.* 372, n71. doi: 10.1136/bmj.n71

Perkins, M., Roe, J., Postma, D., McGaughran, J., and Hickerson, D. (2023). Detection of GPT-4 generated text in higher education: combining academic judgement and software to identify generative AI tool misuse. *J. Acad. Ethics* 22 (1), 89–113. doi: 10.1007/s10805-023-09492-6

Rogerson, A. M., and McCarthy, G. (2017). Using internet based paraphrasing tools: original work, patchwriting or facilitated plagiarism? *Int. J. Educ. Integr.* 13, 2. doi: 10.1007/s40979-016-0013-y

Siemens, G., and Baker, R. S. (2012). "Learning analytics and educational data mining: towards communication and collaboration," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)*, 252–254. doi: 10.1145/2330601.2330661

Stamatatos, E. (2018). Masking topic-related information to enhance authorship attribution. *J. Assoc. Inf. Sci. Technol.* 69 (3), 461–473. doi: 10.1002/asi.23968

Tranfield, D., Denyer, D., and Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *Br. J. Manag.* 14 (3), 207–222. doi: 10.1111/1467-8551.00375

UNESCO (2021). *Education in Emergencies: Palestine — overview*. UNESCO Publishing.

Williamson, B., and Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learn. Media. Technol.* 45 (3), 223–235. doi: 10.1080/17439884.2020.1798995

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education — where are the educators? *Int. J. Educ. Technol. High. Educ.* 16, 39. doi: 10.1186/s41239-019-0171-0