Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Establishing the Validity and Ethical Legitimacy of AI-Generated Text Detection Software in Postgraduate Education: A Systematic Review of False Positives, Institutional Accountability, and Academic Justice

King Costa

costak@researchglobal.net

Global Centre for Academic Research https://orcid.org/0000-0001-9132-9132

Thandoluhle Tia Thusi

2059430@aup.edu.ph

Adventist University of the Philippines https://orcid.org/0000-0003-1420-1420

Letlhogonolo Mfolo

Global Centre for Academic Research https://orcid.org/0009-0009-0570-9901

Nisi Thusi

University of Johannesburg https://orcid.org/0000-0003-1315-8192

Systematic Review

Keywords: academic justice, AI-generated text detection, false positives, institutional accountability, postgraduate education

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11731937/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Establishing the Validity and Ethical Legitimacy of AI-Generated Text Detection Software in Postgraduate Education: A Systematic Review of False Positives, Institutional Accountability, and Academic Justice

King Costa[1] [ORCID: 0000-0001-9132-9132]Thandoluhle Tia Thusi[2] [ORCID: 0000-0003-1420-1420], Letlhogonolo Mfolo[1] [ORCID: 0009-0009-0570-9901], Nisi Thusi[3] [ORCID: 0000-0003-1315-8192],

[1]Global Centre for Academic Research
[2]Adventist University of the Philippines; Global Centre for Academic Research
[3]University of Johannesburg; Global Centre for Academic Research

Corresponding author: 2059430@aup.edu.ph

Abstract

Introduction: AI-generated text detection software is increasingly discussed as a means of protecting academic integrity, yet the supplied evidence raises substantial concerns about validity, linguistic bias, opacity and procedural fairness. These concerns are especially important in postgraduate education, where written work may be produced by multilingual students and may carry significant progression, assessment and disciplinary consequences.

Methods: The researcher-selected databases and information sources were Google Scholar, ERIC, OpenAlex, Crossref. The eligibility protocol restricted publication years to 2022–2026. COSTAQDA recorded source-level retrieval through OpenAlex (38 retrieved; 38 available), Crossref (75 retrieved; 3550 available). Where researcher-selected sources did not expose a suitable direct bulk-search connector, COSTAQDA used its OpenAlex/Crossref discovery bridge while preserving the researcher-selected source names in the PRISMA record. The identification run recorded 113 records before cross-source deduplication and 111 unique records after deduplication.

Results: Detection performance was conditional on tool, dataset, model, task, language background, text length and evaluation protocol. False-positive risks were particularly evident for non-native and second-language English writers and professionally edited texts. A detector score did not establish authorship, intent or misconduct, especially where outputs were binary, scalar or poorly explained. Ethical legitimacy depended on human review, transparent rules, local validation, staff training, appeal processes and evidence of learning beyond the submitted text.

Discussion/Conclusion: The evidence does not support treating AI-detector outputs as conclusive evidence in high-stakes postgraduate decisions. Their responsible use, where retained, requires cautious screening rather than adjudication, subgroup validation, transparent governance and meaningful human due process.

Keywords: academic justice; AI-generated text detection; false positives; institutional accountability; postgraduate education

1. Introduction

Generative artificial intelligence has unsettled established assumptions about authorship, assessment and academic integrity in higher education. The supplied literature describes a movement from relatively discrete forms of plagiarism toward more ambiguous relationships among human writing, AI-generated text and AI-assisted composition. This ambiguity matters in postgraduate education because master's and doctoral work is judged not only as a finished product but also as evidence of independent reasoning, disciplinary participation and developing scholarly identity. Studies of postgraduate students and academic work describe

potential benefits of AI-assisted drafting, feedback, brainstorming and language support, but also concerns about originality, verification, over-reliance and blurred responsibility (Mozelius, 2024; Aktepe & Çetinkaya, 2025; Hassen, 2026).

AI-generated text detectors have been proposed as an institutional response, yet the supplied evidence does not establish a stable or universal relationship between detector scores and authorship. Performance varies according to the tool, model family, dataset, task, text length, language background, editing history and evaluation protocol. Detector outputs may respond to stylistic or linguistic properties rather than provenance, meaning that polished or second-language writing can be misclassified as AI-generated (Park et al., 2026; Liao & Mohammed, 2026; Kamińska, 2026). Other studies similarly question binary classification and argue that an undecided or indeterminate category is necessary where human and machine writing distributions overlap (Ji et al., 2024; Garland, 2026).

The ethical problem is therefore not reducible to technical accuracy. False positives can expose students to suspicion, stigma, grade penalties or disciplinary procedures, while opaque outputs may prevent students and adjudicators from understanding or challenging the basis of a decision. The evidence links these risks to linguistic inequity, particularly for non-native or second-language English writers, and to institutional practices that treat automated scores as more authoritative than contextual human judgement (Nawaz et al., 2026; Maghami et al., 2026; Shamlooh, 2026). Conversely, some supplied studies report apparently strong benchmark performance or operational usefulness, demonstrating that the evidence is not uniformly negative but is difficult to compare across conditions (Curan, 2026; Perkins et al., 2023).

What remains insufficiently understood is how technical validity, ethical legitimacy, institutional accountability and academic justice interact in postgraduate settings. The available evidence includes quantitative experiments, qualitative studies, conceptual frameworks, reviews, preprints and work conducted in general higher education, undergraduate language learning, professional publishing and non-educational domains. Postgraduate-specific evidence is comparatively thin, while direct evidence concerning long-term progression, appeals, disciplinary outcomes, wellbeing and the consequences of false accusations is limited. A systematic review is therefore needed to integrate the evidence across technical, ethical and institutional dimensions without treating bibliometric prominence or isolated benchmark results as proof of quality or generalisability.

2. Research Objectives

RO1: To examine how AI-generated text detection software influences Detection validity (including false positive rates), ethical legitimacy, institutional accountability, and academic justice among Postgraduate students and academic work in higher education within Postgraduate education settings.

RO2: To identify the principal opportunities, benefits and enabling conditions associated with AI-generated text detection software in relation to Detection validity (including false positive rates), ethical legitimacy, institutional accountability, and academic justice.

RO3: To analyse the challenges, risks, ethical considerations and implementation barriers associated with AI-generated text detection software for Postgraduate students and academic work in higher education.

RO4: To examine the contextual, institutional and practice-related factors that shape the adoption and effective use of AI-generated text detection software within Postgraduate education settings.

RO5: To identify gaps in the available evidence and derive implications for future research, policy and practice relevant to Questioning the Validity and Ethical Legitimacy of AI-Generated

Text Detection Software in Postgraduate Education: An Examination of False Positives, Institutional Accountability, and Academic Justice.

3. Research Methods

3.1 Review design

A systematic review was selected because the research question concerns a rapidly developing but conceptually dispersed evidence base in which technical evaluations, postgraduate studies, policy analyses and ethical frameworks address related aspects of the same institutional problem. Systematic reviewing provides a structured means of identifying eligible evidence, applying predefined criteria, documenting selection and integrating findings across heterogeneous designs. This was more appropriate than relying on a narrative overview because the review required explicit traceability from the search record through appraisal and synthesis, while recognising that the evidence could not be reduced to a single pooled effect estimate (Page et al., 2021).

The design was also justified by the question's combination of empirical and normative dimensions. Detection validity can be examined through quantitative performance measures, whereas ethical legitimacy and academic justice are often expressed through qualitative accounts, conceptual analysis and institutional reasoning. A systematic review can accommodate this plurality by separating study identification and eligibility from critical appraisal and then using a synthesis method suited to the available findings. Qualitative evidence synthesis is particularly relevant where the review seeks to explain how authors, students and institutions understand fairness, accountability and the limits of automated classification (Noyes et al., 2018).

The review therefore used a structured evidence-identification process while avoiding the assumption that all included studies were methodologically or substantively comparable. Quantitative results such as false-positive rates were treated as context-dependent observations rather than as automatically combinable estimates, because the supplied studies varied in detector versions, datasets, ground truth, tasks, language populations and text characteristics. Qualitative and conceptual evidence was synthesised thematically to identify recurring mechanisms and institutional conditions. This approach preserves divergence, including apparently strong benchmark results alongside evidence of unstable performance in broader or more linguistically diverse settings (Noyes et al., 2018; Thomas & Harden, 2008).

Transparency was a central rationale for the design. PRISMA 2020 provides a reporting framework for making the identification, screening, eligibility and inclusion process visible, while it does not itself determine the substantive quality of included studies or guarantee that a synthesis is unbiased (Page et al., 2021). The present review consequently distinguishes recorded workflow counts from substantive findings, relevance screening from eligibility, CASP appraisal from synthesis, and bibliometric description from methodological quality. This separation is necessary because a large included set, a high appraisal score or a prominent study does not by itself establish trustworthy evidence about institutional consequences.

Critical appraisal was incorporated because the evidence included experimental, qualitative, review, conceptual and preprint material with different sources of uncertainty. The Critical Appraisal Skills Programme provides structured checklists for examining methodological validity, findings and relevance in research evidence (Long et al., 2020). Study-design-specific checklists are important because the criteria relevant to a qualitative interview study differ from those relevant to a quantitative validation study or a structured review. Appraisal was therefore treated as an interpretive judgement about strengths, weaknesses and reporting uncertainty, not as a mechanical total that could erase design differences or convert incomplete reporting into demonstrated methodological failure.

The selected design also accommodates an explicitly qualitative progression from extracted findings to codes, descriptive categories and analytical themes. Thematic synthesis enables reviewers to preserve the meaning of primary findings while developing higher-order interpretations that answer the review objectives (Thomas & Harden, 2008). In this project, that approach was appropriate because the verified synthesis included ethical arguments, institutional recommendations and empirical observations alongside detector-performance findings. It also allowed contradictory evidence to remain visible, including the contrast between a finite benchmark reporting error-free observed classification and broader evidence cautioning that such results do not establish out-of-distribution generalisation.

Finally, the systematic review design makes the limits of inference visible. The supplied evidence includes studies that are only partially postgraduate-specific, partial finding extracts, preprints and studies whose ground truth or sampling frame limits generalisability. The design cannot resolve these limitations by aggregation alone. Instead, it supports a cautious conclusion about the conditions under which detector outputs may be interpreted and the safeguards required when they affect students. COSTAQDA-supported traceability was used to retain links between findings, codes, categories, themes and study records, but the software record does not replace researcher verification or methodological judgement (Costa & Costa, 2024).

3.2 Search strategy and information sources

The researcher-selected databases and information sources were Google Scholar, ERIC, OpenAlex, Crossref. The eligibility protocol restricted publication years to 2022–2026. COSTAQDA recorded source-level retrieval through OpenAlex (38 retrieved; 38 available), Crossref (75 retrieved; 3550 available). Where researcher-selected sources did not expose a suitable direct bulk-search connector, COSTAQDA used its OpenAlex/Crossref discovery bridge while preserving the researcher-selected source names in the PRISMA record. The identification run recorded 113 records before cross-source deduplication and 111 unique records after deduplication. The recorded search execution timestamp was 2026-11-12T10:10:52.585Z.

Full Boolean search string (exactly as recorded)

("AI text detection software" OR "AI text detection" OR "AI writing detector" OR "generative AI detector" OR "GPT detector" OR "automated text classifier" OR "Turnitin AI detector" OR "AI detector software") AND ("Postgraduate education" OR postgraduate OR "graduate education" OR "master students" OR "doctoral students" OR "PhD candidates" OR "higher education assessment" OR "graduate school") AND ("Ethical legitimacy and academic justice" OR "false positives" OR "detection validity" OR "ethical legitimacy" OR "institutional accountability" OR "academic justice" OR fairness OR "academic integrity" OR "procedural justice") NOT (undergraduate OR "secondary school" OR K-12 OR "code plagiarism")

3.3 Inclusion and exclusion criteria

Inclusion

- Studies conducted in postgraduate higher education settings, including Master's or Doctoral degree programs, involving postgraduate students, academic assessors, or higher education institutions.
- Studies evaluating or examining AI-generated text detection software or automated AI writing detection tools.
- Studies investigating outcomes related to detection validity, false positive occurrences, ethical legitimacy, institutional accountability, or academic justice.
- Studies published between the years 2022 and 2026.
- Publications written in the English language.

- Literature published as peer-reviewed journal articles or academic preprints.

Exclusion

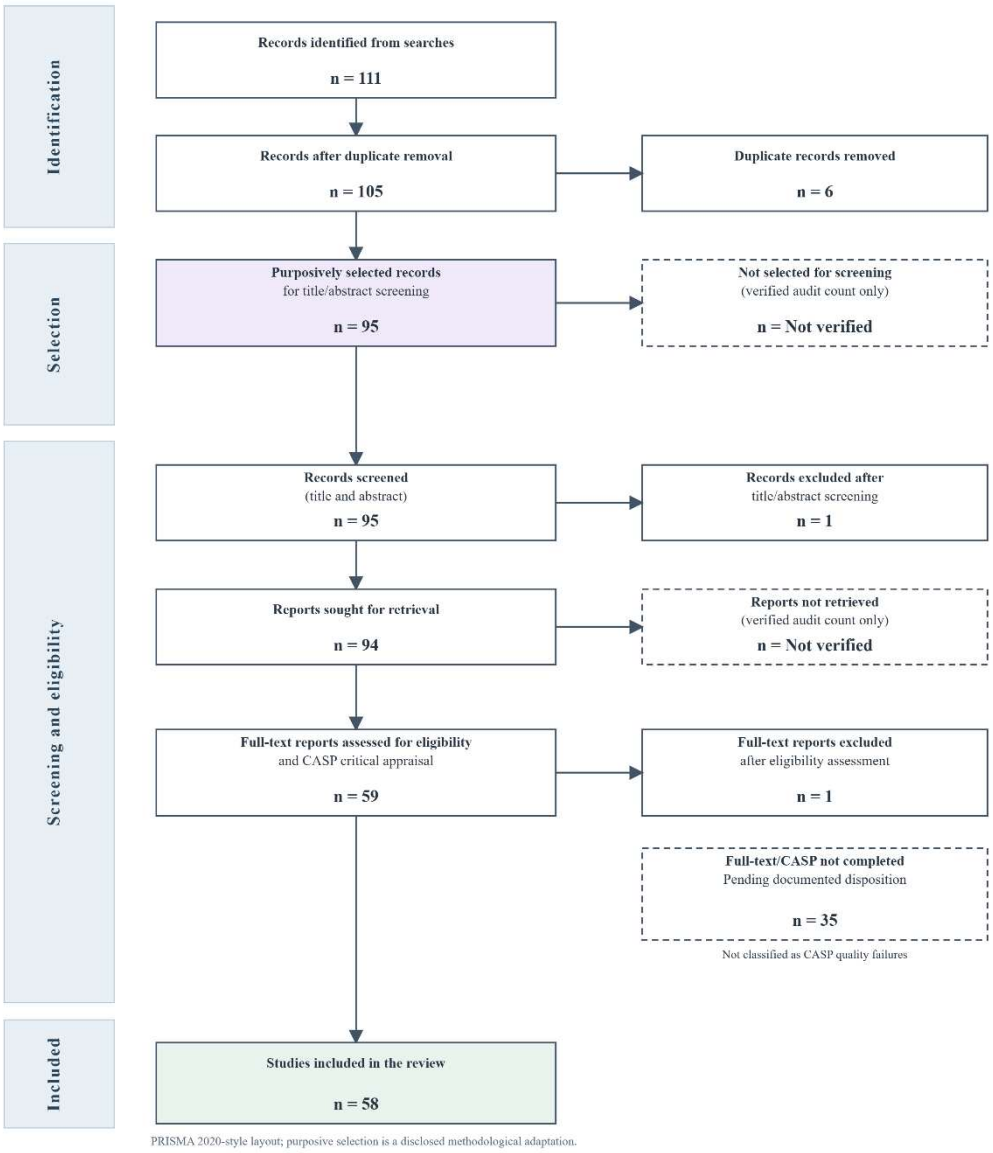
- Studies focused exclusively on primary, secondary, or undergraduate education contexts without postgraduate-specific data or implications.
- Studies examining traditional text-matching plagiarism detection tools without explicit evaluation of AI-generated text detection features.
- Studies that do not address AI detection validity, false positives, ethical legitimacy, institutional accountability, or academic justice.
- Publications released prior to 2022 or after 2026.
- Publications written in languages other than English.
- Document types other than peer-reviewed journal articles and preprints, such as editorials, opinion pieces, book reviews, conference abstracts, or popular media articles.

3.4 Screening and study selection

The PRISMA selection counts are persisted, but the decision rule used during title/abstract screening is not available in the verified audit payload. COSTAQDA therefore reports the recorded workflow counts without reconstructing an unverified screening procedure.

The persisted PRISMA workflow records 111 identified records, 6 duplicates, 95 records screened, 94 reports sought for retrieval, 1 records excluded at the title/abstract stage, 59 full texts assessed, 1 full-text exclusions and 58 included records. These counts describe study selection and are not substantive review findings.

Figure 1. PRISMA 2020 Flow Diagram — Identification, Screening, Eligibility and Included



Source: COSTAQDA PRISMA Workflow, navigation item 8: PRISMA Flow Diagram; four-stage structure showing Identification, Screening, Eligibility (including CASP critical appraisal) and Included; adapted from PRISMA 2020 (Page et al., 2021).

The PRISMA figure should be interpreted as a record of study selection rather than as evidence about detector validity or academic justice. Its supplied values are 111 identified records, 6 duplicates, 95 screened, 94 reports sought, 1 title or abstract exclusion, 59 full texts assessed, 1 full-text exclusion and 58 included records. The audit does not supply the screening decision rule, and the identification discrepancy means that the displayed counts should be treated as persisted project figures whose internal reconciliation remains unverified. PRISMA reporting improves transparency but does not resolve the substantive limitations of the evidence base (Page et al., 2021).

3.5 Critical appraisal

CASP was the recorded critical appraisal tool, and completed CASP appraisals were recorded for all 58 included studies. The workspace reports a mean score of 14.57, but this average is not interpreted as a validated overall quality score. CASP is intended to structure judgement about methodological validity, findings and relevance, and its value depends on how reviewers interpret study design, reporting and applicability rather than simply summing responses (Long et al., 2020). The supplied records also indicate that many appraisals were completed

conservatively when the configured Gemini service was unavailable, with unsupported criteria recorded as “Can't tell” and requiring researcher verification.

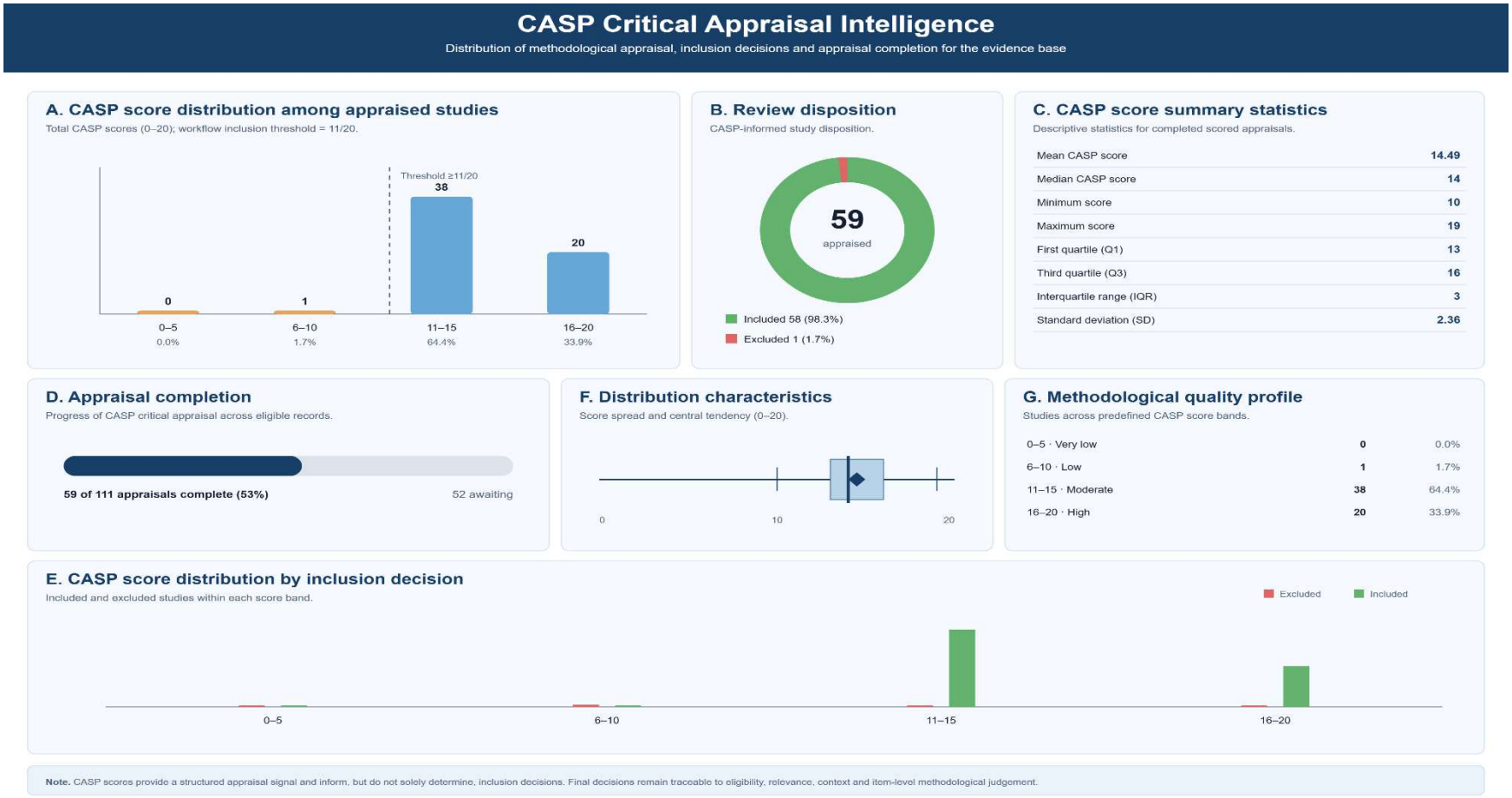
The study-linked appraisal record shows variation across the included evidence. For example, the recorded scores were 19 for Gupta (2026) and Al Maharmah et al(2025)., 18 for Susoy (2026), Maghami et al(2026). and several other studies, 16 for Perkins et al(2023)., Ji et al(2024). and Curan (2026), and 11 for Torres and Seijo (2026), Andriantsialo et al(2026)., Park et al(2026)., and several conceptual or documentary contributions. These values should not be read as directly comparable because the studies use different designs and the record does not provide complete criterion-level rationales for every study. Missing reporting is therefore distinguished from demonstrated methodological weakness (Long et al., 2020).

The appraisal concerns constrain interpretation of the synthesis. Experimental and benchmark studies can provide useful information about performance under defined conditions, but their findings may not generalise when ground truth, model family, task, language background or editing history changes. The WordBinary benchmark reports perfect observed classification in its supplied dataset, yet explicitly cautions that finite-sample error-free results do not establish universal error rates or out-of-distribution generalisation (Curan, 2026). Conversely, studies examining diverse or edited human writing report wide variation in false-positive rates, but their own sampling and corpus constructions also require careful interpretation (Park et al., 2026; Liao & Mohammed, 2026).

The qualitative and conceptual studies contribute important accounts of due process, human agency and institutional responsibility, but they do not by themselves establish causal effects on student outcomes. For example, Tyagi and Meena (2026), Shamlooh (2026), and Maghami et al(2026). support the importance of transparent policies, human review, validation and appeal, while the evidence does not demonstrate that these measures have already reduced false-positive rates in postgraduate institutions. Noyes et al(2018). emphasise the need to connect methodological limitations to the interpretation of synthesised qualitative findings; accordingly, the present review treats governance proposals as evidence-informed implications rather than tested interventions.

No finding-level confidence method was recorded, and GRADE-CERQual was not conducted. CASP study-level appraisal therefore cannot be converted into confidence ratings for the analytical themes (Lewin et al., 2018). The synthesis remains trustworthy only in a qualified sense: several studies converge on conditional validity, linguistic inequity, opacity and the need for accountable human governance, while the magnitude and generalisability of detector performance remain uncertain. COSTAQDA supported traceability from study-linked findings to codes and themes, but that traceability does not remove uncertainty or substitute for researcher verification (Costa & Costa, 2024).

Figure 2. CASP Critical Appraisal Intelligence — PRISMA Workflow, Step 6: Full-Text Screening



Source: COSTAQDA PRISMA Workflow, navigation item 6: Full-Text Screening (CASP critical appraisal outputs). The Word figure is generated from the same deterministic COSTAQDA CASP visual.

The CASP figure represents study-level appraisal outcomes for 58 included studies, with all 58 appraisals marked complete and a recorded mean score of 14.57. The visual should not be interpreted as a single quality continuum or as evidence that higher-scoring studies are necessarily more influential, more relevant or more credible for every objective. The study-linked record includes scores ranging from 11 to 19 and repeatedly records unsupported criteria as “Can't tell” where appraisal evidence was incomplete. The principal methodological pattern is therefore heterogeneity and uncertainty, not a defensible mean quality judgement (Long et al., 2020).

The CASP results support cautious weighting of conclusions. Studies such as Park et al(2026)., Liao and Mohammed (2026), Perkins et al(2023). and Curan (2026) provide technically relevant evidence but differ substantially in corpus construction, task conditions and generalisability. Conceptual and policy-oriented studies illuminate institutional accountability but cannot establish detector effects. The figure should consequently be read alongside study design and appraisal limitations, not as a replacement for them. No CERQual assessment or other finding-level confidence assessment was recorded (Lewin et al., 2018).

3.6 Data extraction

A formal data-extraction procedure was not recorded in the project workspace. Nevertheless, the verified synthesis contains 313 extracted findings contributed by 43 studies, each linked to study and finding identifiers. The available study records include bibliographic information, publication year, design or setting information where supplied, appraisal information and extracted finding text. Because the extraction protocol, field definitions and verification procedures were not recorded, the review reports only information present in the supplied evidence and avoids claiming that unrecorded extraction steps occurred.

3.7 Data synthesis and analysis

Qualitative evidence synthesis was conducted on 313 extracted findings contributed by 43 studies. The COSTAQDA synthesis engine applied thematic synthesis, moving from traceable finding-level evidence to 8 initial codes, 4 categories, 1 descriptive themes and 3 analytical themes.

Theme development was constrained to the extracted finding-level evidence supplied by eligible studies. Each synthesised theme retained supporting study IDs and finding IDs, while contradictory evidence was preserved where identified. Titles, abstracts and bibliographic prominence were not treated as substitutes for extracted qualitative findings.

Recorded synthesis limitations included: The supplied evidence is methodologically and contextually heterogeneous, including qualitative studies, quantitative experiments, conceptual frameworks, structured reviews and preprints; direct comparison of detector performance is therefore limited.; Although the review question concerns postgraduate education, several supplied studies concern general higher education, undergraduate L2 writing, professional publishing or non-educational contexts. Postgraduate-specific evidence is comparatively thin.; Many findings are partial extracts, and some contain recommendations or cited claims rather than complete primary-study results. Interpretations have therefore been kept conditional.; Ground truth, task design, detector versions, model families, text length, language background and editing history vary across studies, limiting conclusions about generalisable false-positive or false-negative rates.; The supplied findings provide limited direct evidence about long-term effects on postgraduate progression, institutional appeals, disciplinary outcomes, student wellbeing or the actual consequences of false accusations.; One benchmark reports perfect observed performance, but its findings explicitly caution that finite-sample error-free classification does not establish universal error rates or out-of-distribution generalisation.; No supplied finding directly establishes that process-based assessment or AI literacy reduces false-positive rates in postgraduate institutional settings;

these are evidence-informed implementation implications rather than demonstrated effects across the dataset..

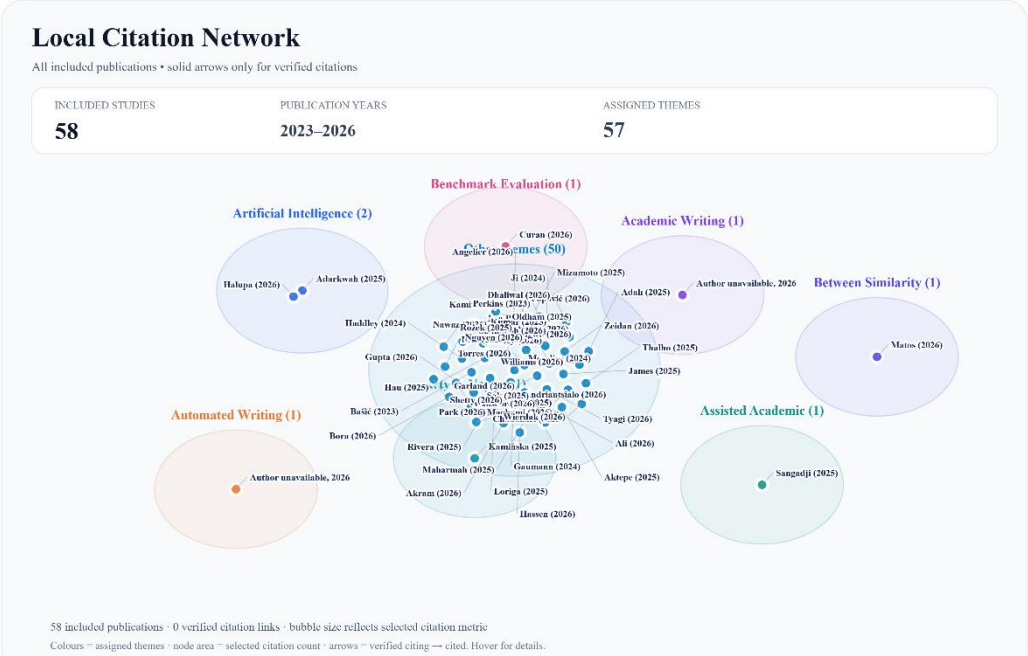
3.8 Bibliometric Intelligence and Confidence in the Evidence Base

Bibliometric intelligence provides a structural view of the evidence base that complements PRISMA selection, critical appraisal and qualitative synthesis. The following analyses examine citation relationships, scholarly influence, evidence connectivity and knowledge development within the included corpus. They provide contextual evidence about the intellectual structure and provenance of the literature and are not substitutes for methodological quality appraisal.

3.8.1 Local Citation Network

The Local Citation Network maps citation relationships within the included evidence base. It shows how studies are connected within the review corpus and supports examination of intellectual connectedness, centrality, bridging and isolation. These citation relationships describe the structure of the literature and do not, by themselves, establish methodological quality or validity.

Figure 3. Local Citation Network — Bibliometric Intelligence, Step 12

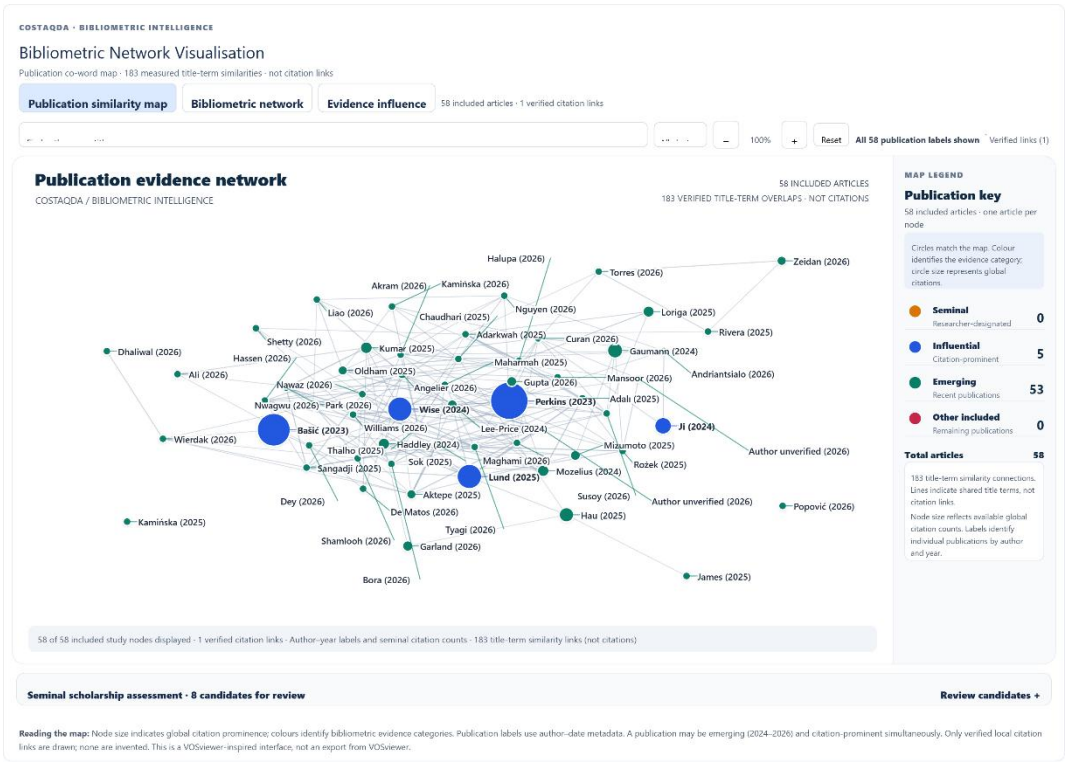


Source: COSTAQDA PRISMA Workflow, navigation item 12: Bibliometric Intelligence — Local Citation Network.

3.8.2 Author Influence Map

The Author Influence Map displays included publications as author–year nodes and distinguishes bibliometric categories. Title–term similarity links indicate metadata overlap, not verified citations. Scholarly influence must be interpreted using verified citation evidence where available and separately from methodological quality.

Figure 4. Author Influence Map — Bibliometric Intelligence, Step 12



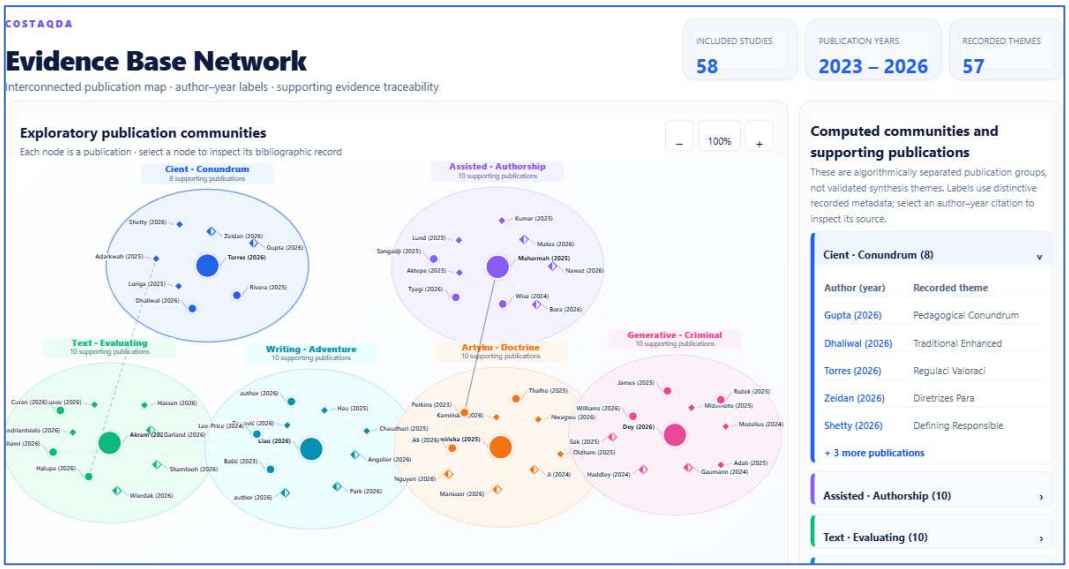
Source: COSTAQDA PRISMA Workflow, navigation item 12: Bibliometric Intelligence — Author Influence Map.

An Influence Pyramid would ordinarily be appropriate for displaying the distribution of bibliometric prominence across included studies, provided that the influence metric, citation coverage, unit of analysis and denominator were clearly specified (Donthu et al., 2021; Zupic & Čater, 2015). In the supplied intelligence, the unit is included-study nodes, with 58 included studies, zero recorded global citations, zero local citations and influenceScore values of zero for the listed nodes. No formula or underlying citation coverage is supplied, and the figure metadata simultaneously marks the pyramid as unavailable in the intelligence layer. A defensible interpretation of comparative influence, citation impact, methodological quality or consensus therefore cannot be produced. The display must not be interpreted as evidence that the studies are equally influential or equally credible.

3.8.3 Evidence Strength

The Evidence Strength analysis examines the structure and strength of relationships within the evidence ecosystem assembled for the review. It complements, but does not replace, CASP methodological appraisal and must not be interpreted as a direct estimate of causal effect.

Figure 5. Evidence Strength Network



Source: COSTAQDA PRISMA Workflow, navigation item 12: Bibliometric Intelligence — Evidence Strength Network.

An Evidence Strength Network could be appropriate for mapping relationships among evidence nodes if the relation types, edge weights, node metadata and denominator were explicitly defined. The supplied workspace reports 58 included nodes and zero links, zero clusters and no edges. Consequently, there is no defensible basis for interpreting corroboration, relationship strength, causal strength, confidence or study quality from the network. The figure should not replace CASP appraisal or the finding-level thematic synthesis. At most, the available metadata show that the evidence ecosystem contains heterogeneous study records; they do not establish a validated network of evidential support.

3.8.4 Knowledge Evolution

Knowledge Evolution examines the temporal and intellectual development of the included evidence base. It traces how research contributions develop over time and, where supported by the Step 12 data, how later studies extend, connect, challenge or consolidate earlier contributions.

Figure 6. Knowledge Evolution — Bibliometric Intelligence, Step 12



Source: COSTAQDA PRISMA Workflow, navigation item 12: Bibliometric Intelligence — Knowledge Evolution.

3.8.5 Integrated Confidence in the Synthesised Evidence

No separate finding-level confidence method was recorded. Accordingly, the review does not assign formal confidence ratings and does not claim that GRADE-CERQual was conducted. CASP appraisals concern study-level methodological considerations, whereas confidence in a synthesised qualitative finding would require an additional, explicitly recorded assessment of methodological limitations, coherence, adequacy and relevance (Lewin et al., 2018). Bibliometric outputs, including the Influence Pyramid and Evidence Strength Network, are not substitutes for such an assessment.

4. Findings

4.1 Search and selection results

The recorded search identified 111 records and produced 58 included records after the persisted selection workflow. OpenAlex contributed 38 retrieved records and Crossref 75, while Google Scholar and ERIC were retained as researcher-selected information sources in the provenance record. The screening audit confirms that selection occurred, but its decision rule is not available. The recorded discrepancy between identification, duplicates and screened counts means that the workflow figures remain affected by an unresolved integrity warning and should not be used to infer substantive evidence strength.

4.2 Characteristics of included studies

The included evidence was heterogeneous, comprising quantitative experiments, qualitative interviews and focus groups, case studies, conceptual and legal analyses, structured and systematic reviews, benchmark evaluations, preprints, book chapters and institutional-policy analyses. Some studies were directly postgraduate-focused, including work with master's students, doctoral students or postgraduate samples (Mozelius, 2024; Bašić et al., 2023; Al Maharmah et al., 2025; Sangadji, 2025). Other studies addressed general higher education, undergraduate second-language writing, professional editing or non-educational applications. This broad evidence profile supported conceptual integration but limits direct claims about postgraduate populations.

The verified thematic synthesis used findings from 43 studies, whereas 58 records were included in the review and appraised. This difference is reported as supplied and indicates that not every included record contributed to the completed qualitative synthesis. The findings were therefore organised around the extracted evidence rather than around article counts. Across the evidence, technical studies examined detection accuracy, false-positive risk, explainability and model or task dependence, while institutional and conceptual studies examined policy, responsibility, disclosure, human agency, assessment redesign and procedural justice.

4.3 Methodological strengths and limitations of included studies

The principal strength of the evidence base is convergence across different forms of inquiry on the conditional nature of detector validity and the importance of human governance. Technical analyses identify dependence on population, language, task, model and dataset characteristics, while ethical and institutional studies explain why these conditions matter when outputs affect students. The principal limitation is that direct comparisons are constrained by heterogeneous ground truth, detector versions, sample construction and publication types. The review also included studies with incomplete reporting and partial extracted findings, so conclusions are deliberately conditional (Noyes et al., 2018; Long et al., 2020).

The CASP workspace confirms that all 58 included studies had completed appraisals, but the supplied appraisals were frequently conservative and recorded unsupported criteria as “Can't tell” when the automated appraisal service was unavailable. The mean score of 14.57 is therefore not treated as a validated summary of evidence quality. The available record contains study-level limitations, but not full criterion-level rationales for every included study. No formal confidence method was recorded, and the review does not claim CERQual or another finding-level assessment.

4.4 Thematic Synthesis

Methodological basis: The thematic synthesis approach follows Thomas and Harden (2008), https://doi.org/10.1186/1471-2288-8-45. The study-specific findings and support counts below derive from the included evidence, not from this methodological citation.

Thematic synthesis is a systematic qualitative evidence-synthesis method used to integrate findings across multiple studies while retaining a transparent connection to the underlying evidence. (Thomas & Harden, 2008) The process moves iteratively from extracted study findings to coding, the grouping of related codes into descriptive categories or themes, and the development of higher-order analytical themes that answer the review question and research objectives. Its purpose is not merely to count how often an idea appears, but to identify patterned meaning across studies, examine convergence and divergence, preserve contextual differences, and develop an auditable interpretation that remains traceable to the contributing evidence.

In this review, thematic synthesis was applied to the verified extracted findings retained in the active COSTAQDA project. Findings were compared across eligible studies, coded and grouped according to conceptual similarity, and then developed into 7 analytical themes represented across 19 evidence-contributing studies. The active synthesis retained 43 verified theme-to-study support assignments, allowing the analytical interpretation to remain traceable to the underlying studies rather than being reconstructed from titles or bibliographic prominence. The resulting analytical themes include “Conditional detection validity and false-positive risk”, “Linguistic inequity and academic justice”, “Limits of binary authorship classification”, “Potential benefits and enabling conditions”, with the widest cross-study support recorded for “Conditional detection validity and false-positive risk” (7 contributing studies). Table 1 therefore operationalises the synthesis as an auditable evidence map linking each analytical theme to the studies that supplied verified supporting findings.

Table 1. Thematic Synthesis Matrix

Theme No.	Analytical Theme with descriptive category	Garland (2026)	Nawaz (2026)	Park (2026)	Kamińska (2026)	Liao (2026)	Curan (2026)	Susoy (2026)	Angelier (2026)	Nwagwu (2026)	Ji (2024)	Kamińska (2025)	Shamlooh (2026)	Tyagi (2026)	Mozelius (2024)	Aktepe (2025)	Mizumoto (2025)	Maghami (2026)	(2026)	Nguyen (2026)	Total
1	Conditional detection validity and false-positive risk	✓	✓	✓	✓	✓	✓	✓	—	—	—	—	—	—	—	—	—	—	—	—	7
2	Linguistic inequity and academic justice	✓	✓	✓	—	✓	—	—	✓	✓	—	—	—	—	—	—	—	—	—	—	6
3	Limits of binary authorship classification	✓	—	—	✓	—	—	—	—	—	✓	✓	✓	✓	—	—	—	—	—	—	6
4	Potential benefits and enabling conditions	—	—	—	—	—	—	—	—	—	—	—	✓	✓	✓	✓	✓	✓	—	—	6
5	Opaqueness, due process and implementation risk	—	—	—	—	—	—	—	—	✓	✓	—	✓	✓	✓	—	—	✓	—	—	6
6	Contextual factors shaping adoption and effective use	✓	—	✓	—	—	—	—	—	—	—	—	—	✓	✓	—	✓	—	✓	—	6
7	Evidence gaps and boundaries of inference	✓	—	✓	—	—	✓	—	—	—	—	—	✓	—	—	—	—	—	✓	✓	6
	Studies contributing to each theme	5	2	4	2	2	2	1	1	2	2	1	4	4	3	1	2	2	2	1	19

Table 1 legend: Only studies contributing verified extracted qualitative evidence are displayed. ✓ = verified extracted evidence from the study supports the theme; — = the study contributed to the synthesis but no supporting extracted finding from that study was assigned to this theme. An em dash does not indicate contradictory evidence.

Table 1 presents the thematic synthesis matrix linking the verified analytical themes only to studies with verified extracted evidence supporting at least one of the themes displayed in the matrix. Any study with no verified supporting contribution to any displayed theme is excluded, even if it was included elsewhere in the systematic review or qualitative evidence base. Rows represent analytical themes and columns represent theme-contributing studies. A check mark (✓) indicates that at least one verified extracted finding from that study was assigned to the theme; an em dash (—) indicates that the study contributed evidence to the synthesis overall but no supporting extracted finding from that study was assigned to that particular theme. The absence of a supporting assignment must not be interpreted as contradictory evidence. The totals report the number of evidence-contributing studies providing verified support to each theme. The matrix therefore functions as an audit trail between thematic interpretation and the underlying finding-level evidence rather than as a measure of methodological quality or effect size.

The most widely supported analytical theme was "Conditional detection validity and false-positive risk", with verified support from 7 of 19 studies. The least widely distributed theme was "Evidence gaps and boundaries of inference", with verified support from 6 of 19 studies. Differences in the number of supporting studies indicate the distribution of traceable thematic evidence across the included literature, not the relative importance or quality of a theme. Themes supported by fewer studies may still be analytically important when the underlying findings are conceptually rich, contextually distinctive or directly relevant to the review question. Conversely, wider study support indicates cross-study recurrence but does not by itself establish stronger methodological quality. The thematic patterns shown in Table 1 should therefore be interpreted alongside critical appraisal, contextual variation, contradictory findings where verified, and the detailed theme-led discussion that follows.

4.5 RO1: Conditional detection validity and false-positive risk

The evidence indicates that detector validity is conditional rather than universal. Performance varies by tool, dataset, model, task, text length, language background, editing history and evaluation protocol. In a large professional-editing corpus, false-positive rates for human-written texts varied from 0% to 100% across 13 detectors, and the same edits changed detector scores in different directions (Park et al., 2026). A Chinese L2 writing study reported an average misclassification rate of 16% and a maximum of 27.66% in a corpus identified as human-written (Liao & Mohammed, 2026). These findings directly support concern about false-positive risk in linguistically diverse postgraduate communities.

The evidence also shows why isolated strong results should not be generalised. The WordBinary benchmark reported 100% observed classification across 9,000 AI-generated and 2,000 pre-2015 human academic samples, including an observed false-positive rate of 0%, but explicitly cautioned that finite-sample error-free classification did not establish universal error rates or out-of-distribution performance (Curan, 2026). Similarly, a comparative study of human reviewers and automated tools used only 16 essays, limiting the transferability of its conclusion that digital tools performed strongly under that experimental design (Susoy, 2026). The synthesis therefore supports neither universal detector failure nor universal detector reliability.

4.6 RO1: Linguistic inequity and academic justice

False positives are not merely technical errors when they are unevenly distributed across student populations. The supplied evidence links detector responses to linguistic style, lexical diversity, formulaic language, professional editing and second-language writing rather than to authorship alone (Park et al., 2026; Liao & Mohammed, 2026). The synthesis identifies a disproportionate academic-justice risk for non-native and second-language English writers because a student may be penalised for the linguistic features associated with developing or professionally edited academic prose. The consequence may include pressure toward

formulaic writing, stigma, reduced confidence and unequal opportunities to demonstrate learning (Nawaz et al., 2026; Nwagwu et al., 2026).

This risk is especially important for postgraduate education because multilingual students may be producing advanced disciplinary work while developing academic English. The review cannot establish the prevalence of false accusations in postgraduate institutions, but it does establish a plausible and repeatedly documented pathway from stylistic classification to inequitable suspicion. The evidence is strongest for the existence of differential detector responses and weaker for downstream effects on progression, wellbeing or disciplinary outcomes. Those outcomes remain important gaps rather than established consequences (Park et al., 2026; Liao & Mohammed, 2026; Shamlooh, 2026).

4.7 RO1: Limits of binary authorship classification

The evidence challenges the assumption that submitted text can be reliably divided into human-written and AI-generated categories. AI-assisted writing may involve brainstorming, translation, rephrasing, structural suggestions or language correction without replacing the student's intellectual contribution. Garland (2026) describes the boundary between AI-generated and AI-assisted writing as a continuum, while Ji et al(2024). show the value of an undecided category when texts cannot be confidently attributed. Kamińska (2026) further demonstrates that human-authored content subjected only to AI-based stylistic editing can receive very high AI probabilities.

A detector score therefore cannot, by itself, establish authorship, intent or misconduct. The applied Turnitin study provides a contrasting institutional example in which high AI percentages coincided with formal misconduct referrals, but the supplied findings do not demonstrate that the software alone validly proved misconduct (Perkins et al., 2023). The analytical implication is that automated outputs may function as screening signals or prompts for inquiry, but they cannot substitute for contextual evidence, dialogue and a defensible account of the student's work.

4.8 RO2: Potential benefits and enabling conditions

The supplied evidence identifies limited but relevant opportunities associated with detection and adjacent AI-supported practices. Detection tools may provide an initial signal that prompts educators to review a submission, and AI-enabled writing or feedback tools may support drafting, language development, brainstorming, revision and metacognitive engagement when students retain responsibility for the work (Mozelius, 2024; Aktepe & Çetinkaya, 2025; Mizumoto, 2025). The benefit is therefore not established as autonomous detection accuracy; it lies in potentially supporting formative dialogue, assessment redesign and academic-literacy development.

Enabling conditions recur across the evidence. These include transparent communication about permitted AI uses, disclosure, student and staff AI literacy, local subgroup validation, human review, process evidence, meaningful appeal and institutional policies that explain the educational rationale for their rules (Tyagi & Meena, 2026; Maghami et al., 2026; Shamlooh, 2026). These conditions may improve legitimacy, but the supplied evidence does not directly demonstrate that they reduce false-positive rates in postgraduate settings. They should therefore be treated as justified implementation conditions rather than tested effects.

4.9 RO3: Opaqueness, due process and implementation risk

Opaque or weakly explainable outputs create a direct due-process problem. Ji et al(2024). report that some detectors provide binary or scalar results without interpretable reasons, limiting the ability of students and adjudicators to challenge decisions. Shamlooh (2026) argues that high-stakes misconduct procedures require evidence that is reliable, interpretable

and specific, while Maghami et al(2026). links detection risks to assessment fairness, student confidence and institutional decision-making. These findings indicate that a score may be operationally convenient while remaining inadequate as an evidential basis for serious sanctions.

Implementation risks also arise from inconsistent policies, limited staff confidence, inadequate local validation and uncertainty about responsibility. Studies of postgraduate and higher-education contexts describe incomplete or contradictory institutional guidance and emphasise the need for training, appeal processes and continuing policy review (Mozelius, 2024; Tyagi & Meena, 2026; Nwagwu et al., 2026). The evidence does not show that every institution uses detectors in this manner, but it does show that legitimacy depends on governance arrangements that software cannot provide independently.

4.10 RO4: Contextual factors shaping adoption and effective use

Adoption and interpretation are shaped by population, language, discipline, assessment task, institutional policy and staff capability. Garland (2026) shows theoretically that task design changes the overlap between human and AI writing distributions, while Park et al(2026). demonstrates that professional editing can alter detector responses without altering authorship or content. Mizumoto (2025) and Mozelius (2024) likewise indicate that pedagogical mediation and assessment redesign affect how AI-related tools function in learning environments. These contextual factors make it inappropriate to transfer a detector's performance from a benchmark corpus directly into postgraduate disciplinary assessment.

Practice-related factors include whether institutions use a detector as a screening prompt or as proof, whether students understand permitted uses, whether staff can interpret uncertainty, and whether multiple forms of process evidence are available. Drafts, revision histories, oral discussion, notes and demonstrations of learning are repeatedly proposed as alternatives or complements to sole reliance on a score (Tyagi & Meena, 2026; [Author not available], 2026). The evidence supports a contextual, human-led approach, although it does not establish the comparative effectiveness of each alternative.

4.11 RO5: Evidence gaps and boundaries of inference

The evidence base contains important gaps. Postgraduate-specific studies are fewer than studies of general higher education or undergraduate language writing, and many findings are partial extracts or recommendations rather than complete primary-study results. Ground truth, model families, detector versions, language backgrounds, editing histories and assessment tasks vary substantially, preventing defensible pooled estimates of false-positive or false-negative rates. The review also found limited direct evidence about long-term progression, institutional appeals, disciplinary outcomes, student wellbeing and the consequences of false accusations.

The synthesis identifies a need for prospective, institutionally embedded evaluations that stratify results by language background, discipline, degree level, task type and writing process. The evidence also supports research comparing detector-assisted inquiry with process-based assessment, human review and redesigned assessment, but no supplied finding directly demonstrates that these alternatives reduce false positives in postgraduate institutions. Future claims should therefore distinguish technical benchmark performance, institutional feasibility, ethical legitimacy and student outcomes rather than treating them as interchangeable.

5. Discussion

The review question asked how validity and ethical legitimacy are evaluated and what their implications are for false positives, institutional accountability and academic justice. The answer is that validity is evaluated inconsistently across heterogeneous conditions, while

ethical legitimacy is evaluated primarily through concerns about fairness, transparency, explainability, human agency and due process. The strongest convergence concerns conditional performance and the risk that linguistic or stylistic features may be mistaken for evidence of AI authorship (Park et al., 2026; Liao & Mohammed, 2026; Nawaz et al., 2026).

The apparently contradictory findings are methodologically informative rather than mutually exclusive. The WordBinary benchmark's perfect observed performance and Perkins et al.'s applied Turnitin referrals show that detectors can produce strong or consequential outputs under particular institutional or experimental conditions (Curan, 2026; Perkins et al., 2023). However, those findings do not negate evidence from diverse populations, edited texts or ternary classification studies. Differences in ground truth, model exposure, task design, corpus composition and institutional thresholds explain why one result cannot be treated as a universal estimate of validity.

In relation to RO1, the review indicates that false-positive risk is inseparable from academic justice when errors are patterned by language background or editing history. In relation to RO2, the principal opportunity is not autonomous judgement but the possibility of using technology within formative assessment, AI-literacy education and structured human inquiry. In relation to RO3, opacity, uncertain specificity and weak institutional procedures create risks that are especially serious when decisions may affect grades, progression or disciplinary status. In relation to RO4, effective use depends on local context, assessment design, staff capability and policy communication (Mozelius, 2024; Tyagi & Meena, 2026; Maghami et al., 2026).

The review also clarifies what the evidence does not show. It does not establish a universal false-positive rate, demonstrate that process-based assessment reduces false positives, or provide direct evidence of long-term postgraduate outcomes. Nor does the CASP mean score provide a formal confidence rating. These limitations mean that recommendations should be framed as safeguards derived from convergent evidence and procedural reasoning, not as experimentally proven interventions. The review's contribution is therefore a bounded account of responsible interpretation rather than a ranking of detector products.

6. Conclusion

The supplied evidence indicates that AI-generated text detection software cannot presently be treated as universally valid or ethically legitimate evidence of authorship or misconduct in postgraduate education. Performance is conditional, false positives may disproportionately affect linguistically diverse writers, and opaque scores do not establish intent or provenance. The most defensible role for detection software, if retained, is as a cautious screening signal within a transparent, human-led process that uses contextual and process evidence. Ethical legitimacy depends on institutional accountability, local validation, staff and student literacy, clear disclosure rules, meaningful appeal and continuing review. The review also identifies substantial evidence gaps concerning postgraduate populations, long-term consequences and the effectiveness of alternatives.

7. Recommendations

Practice recommendation: Institutions should not use a detector score as the sole basis for a grade penalty, misconduct finding, suspension or exclusion. Where software is used, educators should treat outputs as prompts for contextual inquiry and consider drafts, revision histories, oral discussion, notes, process evidence and the student's explanation. Decisions should be documented, reviewable and sensitive to language background and the possibility of AI-assisted rather than wholly AI-generated writing (Ji et al., 2024; Shamlooh, 2026; Tyagi & Meena, 2026). Policy recommendation: Institutions should publish clear, consistent and educationally justified rules specifying permitted, restricted and prohibited AI uses; require

transparent human responsibility; provide staff training; validate tools locally across relevant student subgroups and tasks; disclose uncertainty; and guarantee an accessible appeal and review process. Future-research recommendation: Researchers should conduct prospective, postgraduate-specific evaluations using verified human-written corpora stratified by language background, discipline, degree level and task type; report detector versions and ground-truth procedures; compare screening with process-based assessment and human review; and measure appeals, progression, wellbeing and academic-justice outcomes. Such studies should test implementation safeguards rather than infer their effectiveness from conceptual recommendations alone.

Declarations

Funding

No funding was sought for this paper

Competing interests

There are no competing interests in this paper

AI-assisted technology and COSTAQDA declaration

COSTAQDA was used as an AI-assisted evidence-synthesis and manuscript-development environment. AI-assisted functions supported the organisation, synthesis and presentation of verified project evidence and manuscript preparation within the documented COSTAQDA workflow. AI did not replace researcher responsibility for study selection, eligibility decisions, critical appraisal, verification of extracted evidence, interpretation of findings or final scholarly judgement. The human authors reviewed and approved the final manuscript and remain responsible for its accuracy, integrity, interpretations, citations and conclusions. The COSTAQDA platform is documented in Costa (2026), COSTAQDA: An Integrated Research Intelligence Platform—A Video Demonstration (Zenodo, https://doi.org/10.5281/zenodo.24050211).

References

Aktepe, B., & Çetinkaya, S(2025). Graduate students and AI: Insights into academic writing practices. Education Mind. https://doi.org/10.58583/em.4.2.2

Al Maharmah, A. A., Elfeky, A., Yacoub, R., Ibrahim, A., & Nemt-allah, M(2025). Measuring ethical AI use in higher education: Reliability and validity of the AI academic integrity scale for postgraduate students. International Journal of Innovative Research and Scientific Studies. https://doi.org/10.53894/ijirss.v8i4.7928

Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I(2023). ChatGPT-3.5 as writing assistance in students' essays. Humanities and Social Sciences Communications. https://doi.org/10.1057/s41599-023-02269-7

Costa, K. (2026, November 8). COSTAQDA: An integrated research intelligence platform for qualitative data analysis, systematic reviews, qualitative evidence synthesis, bibliometric analysis and academic publishing—A video demonstration [Video recording]. Zenodo. https://doi.org/10.5281/zenodo.24050211

Costa, K., & Costa, A. P(2024). Qualitative evidence synthesis using COSTAQDA. In J. Ribeiro, C. Brandão, M. Ntsobi, J. Kasperiuniene, & A. P. Costa (Eds.), Computer supported qualitative research. Springer. https://doi.org/10.1007/978-3-031-65735-1_4

Costa, K., & Mfolo, L. (2024). Qualitative evidence synthesis using the COSTAQDA online software. Research Square. https://doi.org/10.21203/rs.3.rs-3988946/v1

Curan, K(2026). Benchmark evaluation of the WordBinary AI-text detection model on 9,000 AI-generated texts and 2,000 pre-2015 human academic-paper samples. Other. https://doi.org/10.2139/ssrn.7483700

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M(2021). How to conduct a bibliometric analysis: An overview and guidelines. Journal of Business Research, 133, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Garland, N(2026). AI Detectors Fail Diverse Student Populations: A Mathematical Framing of Structural Detection Limits. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2603.20254

Garland, N(2026). AI detectors fail diverse student populations: A mathematical framing of structural detection limits. arXiv. https://doi.org/10.48550/arxiv.2603.20254

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. Health Information & Libraries Journal, 26(2), 91–108.

Ji, J., Li, R., Li, S., Guo, J., Qiu, W., Huang, Z. H., Chiyu, C., Jiang, X., & Lu, X(2024). Detecting Machine-Generated Texts: Not Just "AI vs Humans" and Explainability is Complicated. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2406.18259

Ji, J., Li, R., Li, S., Guo, J., Qiu, W., Huang, Z. H., Chiyu, C., Jiang, X., & Lu, X(2024). Detecting machine-generated texts: Not just “AI vs humans” and explainability is complicated. arXiv. https://doi.org/10.48550/arxiv.2406.18259

Kamińska, A(2026). Has AI written this paper? – A case study in “plagiarism detection”. Zagadnienia Informacji Naukowej - Studia Informacyjne. https://doi.org/10.36702/zin2025.02.04e

Kamińska, A(2026). Has AI written this paper? – A case study in „plagiarism detection”. Zagadnienia Informacji Naukowej - Studia Informacyjne. https://doi.org/10.36702/zin2025.02.04e

Lewin, S., Booth, A., Glenton, C., et al(2018). Applying GRADE-CERQual to qualitative evidence synthesis findings: Introduction to the series. Implementation Science, 13(Suppl 1), Article 2. https://doi.org/10.1186/s13012-017-0688-3

Liao, H., & Mohammed, L. A(2026). The AI detector paradox: A PLS-SEM analysis of detection bias in Chinese L2 undergraduate writing. 4th International Conference on Global Education. https://doi.org/10.60072/book.icoge2026.036

Long, H. A., French, D. P., & Brooks, J. M(2020). Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis. Research Methods in Medicine & Health Sciences, 1(1), 31–42. https://doi.org/10.1177/2632084320947559

Maghami, M. R., Ho, C. K., & Patrick, C(2026). AI content detection in student text assignments, false positive rates, and academic integrity. Discover Education. https://doi.org/10.1007/s44217-026-02307-x

Maharmah, A. A., Elfeky, A., Yacoub, R., Ibrahim, A., & Nemt-allah, M(2025). Measuring ethical AI Use in higher education: Reliability and validity of the AI academic integrity scale for postgraduate students. International Journal of Innovative Research and Scientific Studies. https://doi.org/10.53894/ijirss.v8i4.7928

Mizumoto, A(2025). Generative AI in foreign language learning and teaching: A summary and insights from our empirical investigations. Information and Technology in Education and Learning. https://doi.org/10.12937/itel.5.1.inv.p001

Mozelius, P(2024). Generative AI and its impact on activities and assessment in higher education: Some recommendations from master's students. Proceedings of the International Conference on AI Research. https://doi.org/10.34190/icair.4.1.3025

Mozelius, P(2024). Generative AI and its Impact on Activities and Assessment in Higher Education: Some Recommendations from Master's Students. Proceedings of the International Conference on AI Research.. https://doi.org/10.34190/icair.4.1.3025

Nawaz, M., Saleem, W., & Mehsood, M(2026). Evaluating the stylistic nuance of AI-assisted academic writing in higher education: An integrative review and analytical framework. Insights (Bhalwal) International Journal of Humanities Management and Social Sciences. https://doi.org/10.68156/jxp9pw31

Noyes, J., Booth, A., Flemming, K., Garside, R., Harden, A., Lewin, S., Pantoja, T., Hannes, K., Cargo, M., & Thomas, J(2018). Cochrane Qualitative and Implementation Methods Group guidance series—paper 3: Methods for assessing methodological limitations, data extraction and synthesis, and confidence in synthesized qualitative findings. Journal of Clinical Epidemiology, 97, 49–58. https://doi.org/10.1016/j.jclinepi.2017.06.020

Nwagwu, E., Owoseje, F., & Giza, A(2026). From AI Detection to AI Literacy-Rethinking Academic Integrity in Higher Education. International Journal of Science and Research (IJSR). https://doi.org/10.21275/sr261012234009

Nwagwu, E., Owoseje, F., & Giza, A(2026). From AI detection to AI literacy—Rethinking academic integrity in higher education. International Journal of Science and Research. https://doi.org/10.21275/sr261012234009

Page, M. J., et al(2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Park, H., Jeong, G., & Kim, B(2026). Style as a Confound: False Positives in AI Detection of Non-Native Academic Writing. Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.18653/2026.emnlp-industry.109

Park, H., Jeong, G., & Kim, B(2026). Style as a confound: False positives in AI detection of non-native academic writing. Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing: Industry Track, 1872–1887. https://doi.org/10.18653/2026.emnlp-industry.109

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D(2023). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. Journal of Academic Ethics. https://doi.org/10.1007/s10805-023-09492-6

Sangadji (2025). AI-Assisted Academic Writing: Evaluating Postgraduate Students' AI Literacy and Skills. JTP - Jurnal Teknologi Pendidikan. https://doi.org/10.21009/jtp.v27i1.48952

Shamlooh, H. A. M(2026). Evaluating AI text detection accuracy: A structured review of false positives, false negatives, and implications for academic integrity policy in Gulf Region universities. Other. https://doi.org/10.21203/rs.3.rs-9581862/v2

Susoy, Z(2026). Evaluating human reviewers versus artificial intelligence: A comparative analysis of AI-generated essay detection in higher education. Pamukkale University Journal of Education. https://doi.org/10.9779/pauefd.1847553

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Medical Research Methodology, 8, Article 45. https://doi.org/10.1186/1471-2288-8-45

Thomas, J., & Harden, A(2008). Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Medical Research Methodology, 8, Article 45. https://doi.org/10.1186/1471-2288-8-45

Tyagi, P., & Meena, D. K(2026). Preserving human agency in AI-mediated higher education: An integrative framework for ethical AI and academic integrity. International Journal of Creative Research Thoughts. https://doi.org/10.56975/ijcrt.v14i10.401082

Zupic, I., & Čater, T(2015). Bibliometric methods in management and organization. Organizational Research Methods, 18(3), 429–472. https://doi.org/10.1177/1094428114562629