International Journal for Educational Integrity

RESEARCH Open Access

How strong is the evidence in generative AI-related academic misconduct allegations? A mixed-methods analysis

Albert Munoz[1*], Mercedez Hinchcliff[1], Cameron Langfield[1] and Ann Rogerson[1]

*Correspondence:
Albert Munoz
amunoz@uow.edu.au
[1]University of Wollongong, Wollongong, Australia

Abstract

Since the public release of ChatGPT in late 2022, higher education institutions have experienced an increase in academic misconduct allegations related to suspected generative AI (GenAI) misuse. A necessary part of these processes is the inclusion of evidence to support and strengthen allegations. To date, no systematic framework exists for understanding trends and patterns in GenAI student academic misconduct and classifying the types and probative value of evidence presented in these allegations. This study addresses that gap through a mixed-methods analysis of 1,162 GenAI-related misconduct case records spanning January 2023 to December 2025 at one regional Australian university. Analysis confirmed an increase in case volumes over the study period and produced an empirically derived 15-code evidence taxonomy, which was applied across the 1,855 evidence items against three probative quality credentials drawn from legal evidence scholarship: relevance, credibility, and inferential force. Results point to patterns across time that differ across evidence types, including an increasing use of fabricated references and student admissions of AI use, and a decrease in allegations being raised without support. These findings highlight a critical gap in current misconduct policies: institutions lack explicit criteria for determining what constitutes reliable evidence in GenAI misconduct cases. To address this, the study offers three contributions: first, an empirically derived taxonomy that categorises the types of evidence used in GenAI misconduct proceedings. Second, a structured framework for assessing the quality and reliability of that evidence, designed for direct application in institutional decision-making. Third, the first empirical analysis at scale of how evidence is currently gathered and evaluated in GenAI misconduct cases.

Keywords Academic misconduct, Generative AI, Evidence quality, Probative value, Evidence taxonomy

Background

Higher education scholars have raised concerns about misuse of generative AI (GenAI) in student assessments since the public release of ChatGPT in late 2022 (Newton 2025). In parallel, institutions have reported a substantial increase in academic misconduct

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

allegations related to suspected GenAI misuse (Goodier 2025; reporting UK sector-wide figures compiled via freedom-of-information requests). While some institutions have recently shifted to accommodate GenAI use, the scale and trajectory of these allegations raise further questions about the quality and probative value of the evidence upon which they rest. Procedural frameworks for misconduct exist widely, yet we are aware of no published institutional standards specifying what constitutes sufficient or probative evidence in GenAI misuse cases.

University misconduct panels routinely make consequential evidence-weighing decisions in the absence of a systematic analytical method for evaluating evidence. Despite ongoing debate about how much and what type of evidence is required to impose sanctions (Eaton 2025), these panels often rely on professional judgement about the allegation and the evidence supporting it (Bassett et al. 2026). These proceedings nonetheless operate under the civil standard of proof (i.e. the balance of probabilities) as directed by government accreditation bodies (for example, TEQSA, 2025). This standard requires that, on the available evidence, it is more likely than not that the alleged misconduct occurred, with the burden of proof resting with the institution and the academic raising the allegation rather than the student. Where that burden is not discharged, the allegation should not succeed regardless of whether misconduct occurred. Panels rarely include staff with formal legal training (Eaton 2025), and when combined with the comparatively slow development of institutional guidance in this area (Kumar et al. 2024; Michel-Villarreal et al. 2023), it raises concerns that the quality of evidence influencing disciplinary outcomes may vary considerably.

The academic integrity literature has invested considerably in developing policy and procedural frameworks that inform how institutions design and conduct misconduct investigations (e.g., Bretag et al. 2011), and regulatory standards require that such procedures be fair, documented, and consistently applied (Department of Education, 2021; TEQSA, 2024). Ellis et al. (2020) proposed evidentiary stages and thresholds to assist in managing academic misconduct investigations – (1) suspicion at the point of marking; (2) turning suspicion into evidence and (3) findings of fact and upholding allegations. They called for further work into ways reliable and relevant evidence can be identified so that outcomes are not determined on "guesswork, pre-conceptions or assumptions" (Ellis et al. 2020, p. 147). At the point of writing, the authors are not aware of a systematic framework for evaluating the quality and probative value of academic misconduct evidence. However, legal scholarship provides some clarity about the ways experts judge the probative value of evidence, in terms of how strongly it favours or disfavours some claim about the case at hand (Anderson et al. 2005). Specifically, legal scholarship advocates for the evaluation of evidence across three dimensions: relevance (whether it matters to the question at issue), credibility (whether it can be trusted), and inferential force (how strongly it supports the allegation). Collectively, the three dimensions provide an analytical framework, scalable to non-legal disciplines and extendable to other practical decision-making contexts (Twining 2006). At the point of writing, the authors are not aware of a systematic framework for evaluating the quality and probative value of academic misconduct evidence. This gap is particularly consequential in the context of GenAI-related allegations, where the evidentiary challenges are different from those encountered in other forms of misconduct. This study addresses that gap through analysis of 1,162 GenAI-related academic misconduct cases investigated at a regional

Australian university between 2023 and 2025. The data captures the first three years of institutional response to GenAI-related misconduct at scale, providing an empirical basis for examining how evidence practices have developed in the absence of a guiding framework. From the cases, a taxonomy of evidence types and their credential profiles across cases was developed and frequencies were measured through the three-year period. This study is guided by three research questions:

RQ1: What types of evidence are submitted, and how do they vary in probative value?
RQ2: What are the temporal trends in caseload and evidence practices over 2023–2025?
RQ3: To what extent are evidence characteristics associated with case outcomes?

Evidentiary challenges in GenAI misconduct allegations

Large language models (LLMs) such as ChatGPT are built on transformer architecture, in which text is generated through a probabilistic process, selecting the most statistically likely next token, repeating this process until the output is complete (Vaswani et al. 2017). Because the model generates novel word sequences rather than retrieving or reproducing existing text, the output does not carry sequences of words that tools like Turnitin® are designed to detect. Empirical testing of dedicated AI detection tools confirmed their inaccuracy and lack of reliability (Elkhatat et al. 2023; Weber-Wulff et al. 2023), diverging from vendor claims of accuracy that persist despite this evidence; independent evaluations report no material improvement in detector reliability between 2023 and 2025 (Hadra et al. 2026). The challenge is compounded as model capability advances, producing content that more closely approximate human writing patterns (Masih et al. 2025). Human detection offers no reliable alternative. Controlled experiments demonstrate that both novice and experienced teachers are unable to identify AI-generated text among student submissions, and that both groups are overconfident in their judgements (Fleckenstein et al. 2024). Furthermore, making claims of accurate detection of AI generated content risks misclassifying legitimate student work (Hadra et al. 2026), and may never have sufficient evidentiary weight to match the necessary standard of proof for academic misconduct (Bassett et al. 2026). One possible alternative lies in comparing submitted work against authenticated samples of a student's prior writing, where stylistic and linguistic inconsistencies may serve as indicators of authorship concern (Perkins et al. 2024). However, this approach is contingent on the prior collection of verified writing samples, which introduces its own challenges (Perkins et al. 2020) and is further complicated by the natural development of a student's writing over time, which can make legitimate improvement difficult to distinguish from anomalies attributable to misuse.

The same probabilistic nature of LLMs that undermines textual detection also produces an identifiable evidential signal in the form of fabricated references. Because LLMs generate references through next-word prediction rather than retrieval from a verified database, they can produce citations that are statistically plausible. These are often correctly formatted, with author names, journal titles, and volume numbers that follow recognisable patterns, but are not real scholarly works (Walters and Wilder 2023). Unlike textual analysis, references included in a student submission can be verified to

either exist or not by checking against academic databases (Rogerson 2017). However, fabricated references may not always be conclusive of GenAI-related misconduct: a student may fabricate independently, and authorised or incidental GenAI use (for example, using a tool to edit or format text) can inadvertently introduce a fabricated reference (Bassett et al. 2026). This is part of why reference-based evidence attracts moderate rather than strong inferential force in the credential analysis. Furthermore, reference hallucination can only be detected in submissions that contain references; assessments involving short-answer responses, reflective writing, or examinations produce no such signal, regardless of whether AI was used.

A different category of evidence comes from direct observation of student behaviour during the production of the work. For example, online proctoring captures data such as screen activity, webcam recordings, and keystroke patterns. Such data documents how the student engaged with the task, rather than attempting to infer misconduct retrospectively from the completed output (Nicola-Richmond et al. 2024). Behaviours such as minimal keyboard activity, extended periods of inactivity followed by the sudden insertion of substantial text, or navigation to external applications can be judged as consistent with misconduct behaviour. The inference is strongest when the recorded behaviour is evaluated against the declared conditions of the assessment: conduct documented as departing from the permitted use of tools or resources provides a direct, contemporaneous basis for inferring misconduct, rather than an inference drawn retrospectively from the finished work. For example, proctored online examinations generate continuous behavioural records, while take-home assignments and essays submitted without supervision generate no such system-recorded behavioural evidence, although other process evidence (for example, drafts, supervision meetings, or presentations) may be available where such practices are in place (Dawson et al. 2024).

A final category of evidence is generated through the investigation process. Principles of natural justice require that students be informed of the allegation and provided with an opportunity to respond before a determination is made (Ellis et al. 2020). These obligations are both codified in Australian regulatory standards (Department of Education, 2021; TEQSA, 2025) and essential to good academic integrity policy (Bretag et al. 2011). This response opportunity typically takes the form of an investigative meeting or panel interview. Whatever the student communicates in that context becomes part of the evidentiary record available to the decision-maker. For example, a student may provide an admission, a denial, a partial explanation, or an account of mitigating circumstances.

Overall, evidence types in academic misconducts differ substantially in their origin, verifiability, availability, and most importantly in their probative value. Yet no taxonomy of evidence exists in GenAI-related misconduct proceedings, or empirical work in weighing them to guide panels in imposing fair and equitable outcomes. In the absence of such a framework, the evaluation of evidence remains with unstructured professional judgement (Bassett et al. 2026). This is not a trivial concern, especially when the overall narrative of a case remains constant and leads decision-makers to respond to the overall story more than to changes in the probative weight of individual items. Such scenarios raise the possibility that misconduct investigators are influenced more by narrative coherence than by the evidentiary strength of what is before them (Glöckner and Engel 2013). This coherence-based reasoning risk is not new, but GenAI cases sharpen it, because a plausible overall narrative of misuse can readily outrun the discrete, verifiable

evidence actually available. It is here that we borrow analytical frameworks from legal scholarship for evaluating the probative value of evidence (Anderson et al. 2005; Twining 2006). This study draws on this scholarship to respond to the research questions, applying the evidence credentials framework to empirical data from GenAI-related misconduct cases to evaluate the evidential basis upon which these consequential decisions are made.

Methods

Study design

This study employs a mixed-methods design integrating qualitative framework analysis with quantitative statistical analysis applied to an academic misconduct case record corpus. The design is motivated by the nature of the research questions: describing temporal patterns and associations with case outcomes required quantitative analysis, while deriving and applying an evidence taxonomy and assessing probative quality required qualitative interpretive work. Framework analysis (Gale et al. 2013) was selected as the primary qualitative method to produce a structured matrix output of an evidence taxonomy against the evidence credentials across the corpus of cases. The resulting framework takes a balanced approach that combines theory-driven and data-driven reasoning. It uses established evidence criteria as a guiding structure for analysis (Anderson et al. 2005), while also allowing the classification of evidence to emerge directly from the data. This pragmatist approach is particularly well suited to applied, policy-focused institutional research (Bailey 1994; Doty and Glick 1994). Document analysis (Bowen 2009) provided the analytical approach for case-record documents.

The unit of analysis is the individual evidence item within each student's academic misconduct case record. Individual case files frequently contained multiple discrete evidence items, including, for example, similarity reports, examiner commentary, internal inconsistencies in writing style, or fabricated or non-verifiable references cited by the student. Each evidence item was coded independently according to evidence type and three probative quality dimensions. Cases were also analysed holistically, based on the combination of evidence types presented and their aggregated evidentiary credentials.

This study does not adjudicate whether individual students engaged in the misuse of generative artificial intelligence. Instead, it evaluates the quality of the evidence documented in institutional decision-making, namely its relevance, credibility, and inferential force. The civil standard of proof thus serves as an interpretive benchmark that motivates the rating logic rather than as a standard the study itself applies. The study examines patterns in evidentiary practice across cases from 2023 to 2025.

Setting and data source

The study was conducted at one regional Australian university. The institution operates under the Australian Higher Education Standards Framework (HESA) and is registered with the Tertiary Education Quality and Standards Agency (TEQSA). Academic misconduct allegations at the institution are managed through an inhouse case management system introduced in 2017 that records all stages of the investigation process from initial allegation through to outcome, in structured case records. The allegations are managed through a staged investigation pipeline governed by a procedure related to the institutional academic integrity policy. At the initial stage, the teaching staff review the initial

evidence and interview the student. At this stage, the staff member may dismiss the allegation, make a preliminary finding of poor academic practice (indicating the alleged breach was not extensive, blatant and did not result in the student gaining a significant academic advantage), or uphold the allegation and refer it to an Academic Integrity Officer (AIO). From here, the AIO conducts an independent review of the evidence, may obtain further evidence, and may dismiss the allegation, find poor academic practice, apply a low-level outcome, or escalate the matter to the Faculty Investigation Committee (FIC). The FIC is composed of experienced AIOs, an executive of the faculty as chair, with other members of the FIC having no prior involvement in the case. The FIC conducts a formal hearing, gives an opportunity for the student to discuss their case and ask questions (who may or may not have a support person), and determines the outcome. From the evidence gathered and the hearing proceedings, the FIC may dismiss the allegation, find poor academic practice, or uphold the allegation and apply a disciplinary outcome. At each stage, the applicable standard of proof is based on the balance of probabilities and the burden of proof rests with the institution rather than the student. Across the corpus, the median interval from allegation to finalisation was 12 days (interquartile range 6–20), so students were typically asked to account for their work close in time to submission.

The source dataset used in this analysis comprised all academic misconduct case records lodged in the institutional system between January 2023 and December 2025, totalling 5,127 records. Each case record contains structured fields capturing allegation details, assessment characteristics, investigation stage determinations, outcome classifications, and free-text fields recording investigator observations, interview summaries, and outcome reasoning.

The corpus is not restricted to coursework; it includes cases arising from proctored online examinations, and the taxonomy accordingly contains examination and proctoring evidence categories. All data were processed on secure local institutional infrastructure; no case data was shared with external services or cloud-based systems.

During the study period, the institution's formal position on detection tooling evolved. The assessment policy in force through 2024 contained no provisions on artificial intelligence or detection software; from 1 January 2025, a revised policy permitted authenticity checks through institutionally approved software while expressly prohibiting the upload of student work to third-party tools, including AI detection software. Student- and staff-facing guidance on acceptable GenAI use expanded over the same period commencing with changes to the academic integrity policy in January 2023. AI detection-tool outputs, comprising text-similarity reports (e.g., Turnitin) and standalone AI-detection tools, appear predominantly in a subset of case records from 2023 to 2024, predating the institution's 2025 policy provisions on AI detection software. Consistent with the sector-wide pattern of early detector adoption ahead of settled institutional guidance (Moorhouse et al. 2023; Perkins and Roe 2024), these outputs were treated in the present analysis as low-credibility evidence.

GenAI-related cases were identified from the full corpus through a multi-pass classification procedure. The first pass applied a categorical filter on misconduct type fields, capturing cases with GenAI misuse as a misconduct category label (introduced mid-2023). The second pass applied keyword search across the two primary free-text fields (Summary of Allegation and Summary of Evidence). These two fields were specifically

selected as they are mandatory fields populated by the alleging staff member and therefore most likely to contain early-stage evidence signals of GenAI misuse potential and identified GenAI related cases logged before the specific misuse category was introduced. The keyword search in these two fields involved using terms organised into three signal groups: (*i*) tools producing original content, (*ii*) tools modifying existing text, and (*iii*) terms signalling fabricated citation evidence. The third and final pass applied rule-based classification to resolve ambiguous cases (e.g., those triggering keyword signals without categorical confirmation) through a deterministic five-rule decision sequence. The rule sequence prioritised categorical override, applied reference-proximity checks for fabricated citation signals, excluded transformative tool signals in exam misconduct contexts, and applied co-occurrence logic for generic artificial intelligence references. Cases unresolved by the rule sequence were routed to manual review (12 in total) to be evaluated by a member of the research team. The full classification procedure produced a GenAI case set of 1,162 cases, representing 22.7% of the total institutional misconduct corpus across the study period. The full keyword lists for the three signal groups, the verbatim five-rule decision sequence, the corpus-to-dataset flow reporting exclusions at each pass, and a discussion of false-positive and false-negative risks are provided in Additional file 1.

Key terms used throughout this paper are defined in Table 1, following the institution's academic integrity policy and academic misconduct procedure. A case is included as GenAI-related when the multi-pass classification procedure identifies suspected generative-AI misuse from the misconduct-category label and/or free-text signal terms; inclusion denotes topical relevance to the study, not a finding of misconduct.

Framework analysis implementation

Framework analysis was conducted in five stages, following Gale et al. (2013): familiarisation, initial coding, development of the working analytical framework, application of the framework, and charting. Familiarisation involved immersion in a cross-section of case records drawn from across the three-year period and representing the range of investigation stages, assessment types, and outcomes present in the corpus. Prior to engaging in the research, the author team had extensive experience in raising, investigating, and adjudicating academic misconduct cases. As such, much of the familiarisation work required reviewing a purposively selected sample of cases in detail, attending

Table 1 Definitions

Term	Definition (as used in this study; wording follows the institution's academic integrity policy and misconduct procedure)
Allegation	A report that a student may have engaged in academic misconduct, made to the relevant Subject Coordinator or the Office of Academic Integrity.
Case	The record of a single allegation in the institutional case-management system, tracked through the staged investigation.
Upheld misconduct	A finding, on the balance of probabilities, that academic misconduct occurred, academic misconduct being conduct that breaches the institution's academic integrity requirements as defined in its academic integrity policy.
Poor academic practice	A determination that the conduct represents poor academic practice rather than academic misconduct, applied only where the breach is not extensive or blatant and does not result in the student gaining a significant academic advantage; it attracts an educational rather than a disciplinary response.
Dismissal	An outcome in which the decision-maker found insufficient evidence of academic misconduct, and poor academic practice had not occurred, so the allegation was not upheld.

to how evidence was described, what types of observational claims investigators made, and where the boundaries between allegation narrative and discrete evidentiary content were unclear. This stage generated the initial analytical framing that informed the development of the analytical framework and identified recurring ambiguities. Specifically, familiarisation allowed the formal distinction between an investigator's characterisation of a submission and a discrete, independently assessable evidence item. Development of the working analytical framework required the production of an evidence taxonomy, which was later used in full coding. The taxonomy was developed through a hybrid deductive-inductive classification procedure informed by Bailey (1994). The deductive component was provided by the evidence credentials framework (Anderson et al. 2005), which constrained the classification space to distinct evidence types whose probative properties could be meaningfully differentiated across the three credential dimensions. The inductive component involved empirical derivation of specific categories from the corpus of data, while building on patterns identified during familiarisation.

The evidence taxonomy was developed iteratively (as recommended by Nickerson et al. 2013). The meta-characteristic was defined as the evidentiary function of each item in assessing the quality of evidence bearing on GenAI-related misconduct, with reference to the civil standard of proof. Three formalisation criteria were applied at each iteration: internal coherence, requiring that items within a category share a common evidentiary function and would receive comparable credential ratings under similar case circumstances; distinctiveness, requiring that items cannot be equally placed in an adjacent category without loss of meaning; and recurrence, requiring that categories capture systematic features of practice rather than isolated anomalies. The iteration process produced category merges, splits, and removals across successive versions. Objective ending conditions, (following Nickerson et al. 2013) required that every evidence item could be assigned to at least one category, that no new category was required in the final iteration, and that each of the eight substantive categories contained multiple instances, the single-instance Third-Party Source Verification code (3.3) being the documented exception. Subjective ending conditions assessed conciseness, robustness, comprehensiveness, extendibility, and explanatory power (Kundisch et al. 2022). Both sets of conditions were satisfied, and the finalised taxonomy contained 15 codes organised into eight substantive categories.

Alongside the taxonomy, the second component of the framework analysis was the assessment of evidence credentials as three independent rating dimensions applied to each discrete evidence item. Each dimension of evidence credentials was evaluated using an ordinal scale (High/Moderate/Low for relevance and credibility; Strong/Moderate/Weak for inferential force) to capture meaningful gradations in evidentiary quality rather than binary present/absent judgements. Relevance was assessed as whether the item logically bears on the question of GenAI misuse, credibility was assessed as the trustworthiness of the evidence source accounting for documented limitations such as AI detector false positive rates, the hearsay character of peer statements, or the self-serving nature of student denials, and inferential force assessed as what the item establishes when taken alone without relying on other items in the case. A rating rationale note was recorded for each dimension where the assignment was not self-evident from the evidence type.

Application of the framework to the full corpus of 1,162 cases was conducted through a two-stage hybrid coding procedure. The first stage used rule-based pattern matching

to identify evidence codes that could be reliably identified from keyword signals and field provenance. For example, rules were applied to detect fabricated reference evidence when accompanied by statements about how database searches returned no result, or a student admission in an interview was noted in an interview summary. The second stage used a locally hosted LLM to apply coding (following guidance by Dunivin 2025) that required reading evidence descriptions in context, which was a type of judgement that keyword matching alone could not reliably produce, and to assign credential ratings for every coded evidence item. The model was deployed in a closed, local environment, meaning that case data were processed entirely on institutional infrastructure and were not transmitted to external services. The LLM (Qwen 2.5:14b) was configured to produce identical outputs for identical inputs, ensuring that the same case text would always receive the same code regardless of when it was processed. A set of automated post-processing rules were applied after each model response to enforce coding decisions that required uniform application across the full corpus. The model configuration and the complete set of automated post-processing rules are provided in Additional file 1 (Section D); the full coding prompt (approximately 3,800 words) is available from the corresponding author upon request.

Use of LLMs for coding qualitative data often requires close inspection of implementation choices, as even subtle changes in a configuration can alter interpretations (Carlson and Burbano 2026). Initial LLM coding results required a calibration step in which human coding output was used as the target for iterative pipeline adjustment until LLM coding converged with human practice (Burgess et al. 2026; Than et al. 2025). The calibration included independent coding of a purposive sample of twelve cases by four co-investigators using the evidence taxonomy and credential rating framework. Agreement among the four human coders was assessed using inter-coder reliability (ICR) measures (O'Connor and Joffe 2020). Results indicated a majority agreement rate of 94.6%, establishing the baseline level of agreement achievable by trained human coders. ICR measures confirmed the baseline level of agreement; Fleiss' Kappa ($\kappa = 0.619$) and mean pairwise Cohen's Kappa (0.625) both indicated substantial agreement (Landis and Koch 1977), supported by mean pairwise percent agreement of 91.7% and mean pairwise Jaccard similarity of 0.630 as supplementary consistency indicators.

The LLM pipeline was then calibrated, using first-run LLM coding data (generated from a locally hosted run using anonymised data) and human coded data to calibrate LLM coding with human coding data. LLM calibration occurred through iteration of the coding prompt and acceptance criteria set that its inclusion should not degrade ICR measures established by the four human coders. Following ten iterative refinement cycles, the pipeline's coding had converged with human coding on the calibration sample. Because this sample was also the tuning target, convergence here establishes calibration rather than generalisation; generalisation to unseen records is assessed separately through a held-out validation exercise which we detail below. The convergence was confirmed by an increase in Fleiss' Kappa from 0.619 to 0.646, mean Cohen's Kappa from 0.625 to 0.651, and the majority agreement rate from 94.6% to 100.0%. Cases where divergence between LLM and human coders persisted were reviewed by humans during the charting stage. As the final stage of the framework analysis, charting aggregated the item-level coding output to produce a case-level matrix in which each case is represented as a row, with columns capturing the evidence types present, probative quality

ratings on each credential dimension, the maximum inferential force rating across all evidence items in the case, case outcome, assessment method, assessment weighting, and temporal period. This matrix structure enabled direct integration with the quantitative case-level dataset, with evidence-quality variables entered as predictors in the outcome association analyses described in the following section.

Held out validation

Upon conclusion of the calibration exercise, the LLM coder was further evaluated using a held-out sample of cases not used during the calibration. Such practice provides a practical safeguard against overfitting a coding procedure to the cases used to develop it (Than et al. 2025; Carlson and Burbano 2026). The held-out validation exercise comprised fifty-four cases, comprising 93 discrete evidence items, drawn from the case corpus using a fixed random seed, stratified by year (2023–2025) and by evidence category, with deliberate top-up sampling of low-prevalence codes so that rare evidence types would be represented. The sample was coded independently by two members of the research team using the finalised taxonomy and credential-rating framework. Both coders were blind to the LLM coder results. Our acceptance criterion required human–LLM agreement to fall within 0.10 of human–human agreement, applied to overall and for the most frequent codes, with rare codes reported descriptively rather than as pass or fail given their small cell counts.

Additional file 1 (Section B, Tables B.1–B.4) reports the results: per-code and per-dimension agreement, human–human versus human–LLM comparisons, and an error analysis of divergent codings. Overall human–human agreement was AC1 = 0.896 and human–LLM agreement AC1 = 0.918; AC1 is treated as the primary coefficient because low code prevalence deflates κ. The non-inferiority criterion was met ($\Delta = -0.02$, 95% CI – 0.04 to – 0.00).

Results

Across the 1,162 GenAI-related misconduct cases, 1,855 discrete evidence items were coded against the 15-code taxonomy (see Table 2). Fifty-eight cases (5.0%) yielded no identifiable evidence item and received only the Assertion Only code, indicating allegations progressed without any supporting evidentiary content. In a small number of these cases, multiple instances of assertion were recorded, but none was supported by any other type of evidence.

Evidence types and probative value profiles

The taxonomy distribution reveals a high frequency of interview admissions and reference-based evidence, with co-occurrence of fabricated references and interview admissions observed in 10.8% of cases. Textual and detection-based evidence types were present but less frequent. The concentration of evidence in interview admissions is notable given that this evidence type is generated through the investigation process itself rather than submitted at the point of allegation. Such a finding indicates that in a substantial proportion of cases, the strongest evidence available emerged only after the institution had already determined the allegation was sufficient to proceed.

The probative quality profiles reveal a pattern of systematic misalignment between the evidence types commonly used at allegation and those carrying the strongest inferential

force. Detection tool outputs, comprising text-similarity reports (e.g., Turnitin) and standalone AI-detection tools (e.g., GPTZero), were present but produced uniformly weak inferential force profiles, consistent with their documented limitations (e.g., Weber-Wulff et al. 2023). Textual evidence types, including intrinsic AI-typical patterns and comparative style anomalies, similarly clustered at weak to moderate inferential force despite their frequent appearance in allegation narratives, reflecting that these evidence types are frequent in misconduct proceedings but not exclusive to GenAI misuse.

By contrast, the evidence types carrying the strongest inferential force profiles differ from detection-based approaches in a critical respect: they do not depend on probabilistic classification of text characteristics (e.g., interview and written admissions, observed prohibited exam behaviour, and fabricated references). While admissions directly address the question of whether GenAI was used, observed exam behaviour, captured during remote-proctored online examinations in which access to prohibited tools was recorded in the system logs, is documented by objective system records, and fabricated references are independently verifiable against external databases. Despite near-universal high relevance and credibility ratings of fabricated references, this category received predominantly moderate rather than strong inferential force ratings. This pattern is a product of the calibrated rating logic that distinguishes between a reliable indicator of GenAI involvement and evidence sufficient on its own to discharge the balance of probabilities standard.

Notably, evidence related to interview inconsistencies exhibits near-uniform probative quality ratings (Moderate relevance, High credibility, Moderate inferential force) across 98.5% of items, reflecting the narrow epistemic range of this evidence type. Every instance describes and documents a student's inability to explain their own work during a formal interview, which invariably constitutes a credible institutional record (High) that is consistent with GenAI use but equally consistent with alternative explanations such as poor comprehension or test anxiety (Moderate relevance and inferential force). This uniformity is structurally paralleled by other epistemically constrained codes, such as Turnitin/Similarity Reports (100% Weak inferential force) and Interview Admissions (97.2% Strong). In contrast, codes with operational definitions that describe diverse evidence scenarios (e.g., Misattributed/Mismatched Sources and Matching/Near Matching Responses) tended to produce greater rating variation.

Temporal trends in caseload and evidence practices

GenAI-related allegations grew across the period of the study, rising from 11.9% of all misconduct cases in 2023 to 20.0% in 2024 and 36.9% in 2025, against a background of comparatively stable total misconduct volume (see Fig. 1). A chi-square test confirmed that this proportional growth across the years is not attributable to random variation ($\chi^2(2) = 320.93$, $p < .001$), with Cramér's V reported as the effect size measure given its ability to standardise the chi-square statistic against both sample size and table dimensions ($V = 0.250$, approaching a moderate effect; Cohen, 2013). The trajectory suggests that GenAI-related allegations are increasingly displacing or co-occurring with other allegation types rather than adding to the overall caseload.

The structure of evidence submitted within those cases shifted considerably across the same period (see Fig. 2), and in ways that carry interpretive significance beyond the caseload growth itself. The most pronounced change is the rise of fabricated reference

Table 2 Evidence taxonomy, counts, and probative value ratings

Category	Operational Definition	Example Indicator Language	Items (*n*, %)	Cases (*n*, %)	Relevance	Credibility	Inf. Force
Non-Probative Codes							
Assertion Only	The allegation is recorded but no identifiable evidence item is described in any available case field. Mutually exclusive with all other codes; excluded from probative quality rating.	*"The student is alleged to have submitted work generated by an AI tool."*	63 (3.4%)	58 (5.0%)	N/A	N/A	N/A
Category 1: Student Work / Textual Evidence							
Style/Level Anomalies (Comparative)	Anomalies in writing style or quality assessed relative to the student's expected level (year of study, course level) or prior submitted work. Requires explicit comparison to a student-specific reference point.	*"The writing is substantially more sophisticated than work submitted earlier in the session and is inconsistent with the expected standard of a first-year undergraduate."*	70 (3.8%)	65 (5.6%)	M:100.0%	H:10.0% M:88.6% L:1.4%	M:51.4% W:48.6%
AI-Typical Content Patterns (Intrinsic)	Intrinsic features of the text recognised as characteristic of GenAI output, assessed independently of any knowledge of the specific student's history or level.	*"The submission contains generic transitional phrases, unnaturally uniform paragraph structure, and unsolicited definitions of basic terms not required by the task."*	302 (16.3%)	239 (20.6%)	H:1.4% M:98.6%	H:1.8% M:91.5% L:6.7%	M:29.4% W:70.6%
Category 2: Detection Tool Outputs							
Turnitin / Similarity Reports	Evidence derived from Turnitin or equivalent text-matching tools, including similarity percentages, flags, or report findings. Codes the use of the report as evidence, not policy thresholds.	*"The Turnitin report returned a similarity score of 1%, which is unusually low for a research-based assessment of this type."*	73 (3.9%)	71 (6.1%)	L:86.3% M:13.7%	M:94.5% L:5.5%	W:100.0%
AI-Detection Tool Outputs	Evidence from tools specifically designed to detect AI-generated content (e.g., GPTZero, ZeroGPT), including numerical probability scores or classifier outputs.	*"GPTZero returned an AI probability score of 94% across the submitted work."*	69 (3.7%)	67 (5.8%)	H:82.1% M:17.9%	L:100.0%	W:92.5% M:7.5%
Category 3: Reference and Source Evidence							
Non-Existent / Fabricated References	Evidence that cited sources cannot be located, do not exist, or are fabricated, regardless of the mechanism of production. The fact of non-existence is the evidence item; AI mechanism is contextual, not definitional.	*"Three of the five references in the submission cannot be located in any library database, institutional repository, or via Google Scholar."*	414 (22.3%)	356 (30.6%)	H:99.2% M:0.8%	H:91.6% M:8.4%	S:2.0% M:98.0%
Misattributed / Mismatched Sources	Sources that are real but cited for propositions they do not support, described with inaccurate details, or referenced in a manner unrelated to the task.	*"The article cited in support of the central argument does not discuss this topic and makes no such claim."*	196 (10.6%)	196 (16.9%)	H:20.4% M:79.6%	H:58.0% M:42.0%	S:19.3% M:27.1% W:53.6%
Third-Party Source Verification	Independent confirmation from a named third party (not the alleging investigator) that cited sources are fabricated or misused. Requires identification of a specific verifier.	*"The subject librarian confirmed that none of the three cited sources exist in any accessible database or journal index."*	1 (0.1%)	1 (0.1%)	H:100.0%	H:100.0%	M:100.0%

Table 2 (continued)

Category	Operational Definition	Example Indicator Language	Items (*n*, %)	Cases (*n*, %)	Relevance	Credibility	Inf. Force
Category 4: Documentary Admissions							
Written Student Admissions	Any written statement from the student acknowledging GenAI misuse. Applies to written channels only (email, LMS submission, written response to allegation). Oral admissions are coded 5.1.	*"The student acknowledged in their written response that they used ChatGPT to draft the initial sections of the assignment."*	10 (0.5%)	10 (0.9%)	H:100.0%	H:100.0%	S:100.0%
Category 5: Interview Evidence							
Interview Admissions (Oral)	Verbal admission during an SC, AIO, or FIC/SCC interview, recorded in interview summaries or outcome comments. Code only when the narrative explicitly states oral acknowledgement.	*"During the viva voce interview, the student admitted to using ChatGPT to produce the first draft of the submission."*	410 (22.1%)	399 (34.3%)	H:98.6% M:1.1% L:0.2%	H:99.3% M:0.7%	S:97.2% M:1.4% W:0.2%
Interview Inconsistencies	Evidence from interview that the student cannot demonstrate knowledge of, or provide a coherent account of, the submitted work. Negative capability indicator, not an admission. Requires explicit inability signal.	*"The student was unable to explain the analytical framework described in their submission or identify any of the sources cited therein."*	67 (3.6%)	63 (5.4%)	M:100.0%	H:100.0%	M:100.0%
Category 6: Examination and Proctoring Evidence							
Proctoring Records (Carrier Evidence)	The system-generated record itself — Proctorio logs, screen captures, tab-change events, camera flags. Documents the existence of the record as an evidence carrier; the behaviour it documents is coded separately under 6.2.	*"Proctorio screen recordings and browser navigation logs were generated during the examination period and are attached to the case file."*	40 (2.2%)	37 (3.2%)	H:88.9% M:8.3% L:2.8%	H:100.0%	S:86.1% M:11.1% W:2.8%
Observed Prohibited Exam Behaviour	The actual prohibited behaviour described as occurring during the examination, as documented by the proctoring record. Code only when the behaviour itself is described, not merely inferred from a log entry.	*"The Proctorio recording shows the student accessing the ChatGPT interface at 14:32 during the examination window."*	57 (3.1%)	51 (4.4%)	H:84.2% M:14.0% L:1.8%	H:77.2% M:22.8%	S:68.4% M:29.8% W:1.8%
Category 7: Collusion Evidence							
Matching / Near-Matching Responses	Evidence that submissions from two or more students are identical or near identical in structure or content at a level inconsistent with independent work.	*"Comparative review identified three submissions from the same tutorial group that were identical in argument structure and phrasing across multiple paragraphs."*	55 (3.0%)	51 (4.4%)	H:18.2% M:74.5% L:7.3%	H:38.2% M:61.8%	S:7.3% M:41.8% W:50.9%
Category 8: System Records							
System Metadata / File Properties	Document metadata, LMS timestamps, author fields, or file properties constituting objective system-generated evidence bearing directly on authorship or submission integrity. Unlike 8.1 (procedural), this code carries substantive probative value.	*"Document metadata identifies a different author name from the enrolled student, with a file creation timestamp post-dating the submission deadline."*	28 (1.5%)	26 (2.2%)	H:21.4% M:71.4% L:7.1%	H:3.6% M:89.3% L:7.1%	M:67.9% W:32.1%

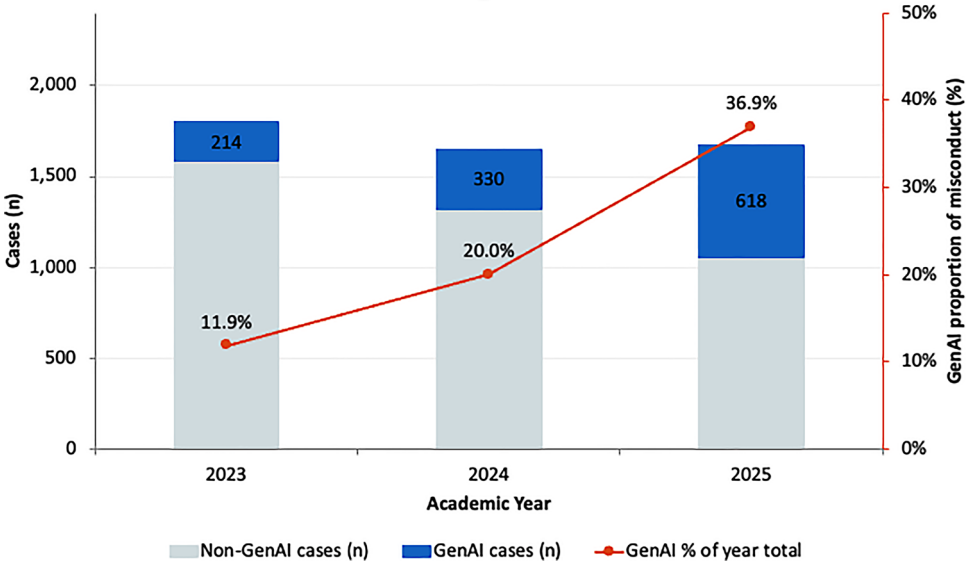


Fig. 1 GenAI misconduct cases per year at the Authors' Institution

Evidence Category	2023	2024	2025
Interview Admissions	17.7%	22.7%	23.2%
Fabricated References	10.4%	14.7%	30.0%
AI-Typical Content Patterns	19.5%	17.8%	14.5%
Misattributed Sources	13.4%	9.4%	10.2%
Similarity Reports	4.0%	5.5%	3.1%
AI Detection Tool Outputs	5.8%	8.8%	0.5%
Style/Level Anomalies	5.8%	2.5%	3.7%
Interview Inconsistencies	4.9%	4.1%	3.0%
Assertion Only	7.6%	2.5%	2.5%
Prohibited Exam Behaviour	1.8%	5.5%	2.3%
Matching Responses	3.0%	2.5%	3.1%
Proctoring Records	3.4%	2.5%	1.6%
System Metadata	1.5%	1.4%	1.6%
Written Student Admissions	1.2%	0.0%	0.6%
Third-Party Source Verification	0.0%	0.0%	0.1%

Fig. 2 Code frequency per Year. Values are the percentage of coded evidence items in each year

evidence, which grew from 10.4% of all coded items in 2023 to 30.0% in 2025. In addition, interview admissions grew steadily in absolute terms across the period, suggesting that the interview was increasingly used to gather evidence in the form of admissions. Furthermore, increases in fabricated reference evidence are also seen in the co-occurrence with interview admissions, which grew from 3.1% to 15%. By 2025, approximately

1 in 7 cases contained this co-occurrence indicating that students presented with fabricated references evidence also admitted to GenAI use during the interview. This trajectory may reflect staff raising allegations becoming more attuned to fabricated references as a signal of GenAI misuse and utilising this evidence to present the student with an opportunity to admit GenAI misuse.

The near disappearance of AI detection tool outputs as submitted evidence is noteworthy in the opposite direction. Having grown from 5.8% of items in 2023 to 8.8% in 2024, AI-detection tool outputs fell to 0.5% in 2025 ($n = 5$ in 2025). Given that AI detection tools carried uniformly weak probative quality profiles in the credential analysis, this withdrawal is consistent with the growing institutional recognition that such outputs cannot discharge the balance of probabilities standard, a position increasingly reflected in the literature (e.g., Elkhatat et al. 2023; Hadra et al. 2026; McKie 2026; Van Vlasselaer et al. 2026).

Finally, the proportion of evidence items recorded as Assertion Only declined from 7.6% of items in 2023 to approximately 2.5% by 2025, indicating that investigators became more consistent in documenting discrete evidentiary content over time rather than progressing allegations based on the allegation narrative alone.

Evidence characteristics and institutional outcomes: scope of inference

The analysis also examined whether evidence type, quantity, and probative quality are associated with outcome dimensions, such as escalation depth, outcome (misconduct vs. poor academic practice vs. dismissal), and dismissal probability. The results demonstrate that while statistically significant associations exist at the bivariate level, the explanatory power of evidence characteristics across all model specifications is modest, and the interpretive constraints are substantial. Full model output, including coefficients, 95% confidence intervals, model comparison, and diagnostics, is reported at the case level (one observation per case) in Additional file 1 (Section C, Tables C.1–C.8). For example, the relationship between evidence credentials and escalation depth is more consistent with the policy architecture of the institutional misconduct process rather than evidence type and probative value. Although evidence quality may influence whether the Subject Coordinator determines that misconduct has occurred, it is that finding, not the evidence credentials, that triggers escalation under institutional policy. In terms of outcome severity, the distinction between a finding of misconduct and a finding of poor academic practice is also governed by factors largely external to the evidence record. While first time misconducts are more likely to result in an assignment of poor academic practice and require an educational intervention (as discussed in Munoz et al. 2025), repeat offenders or those found to have acted deliberately are more likely to receive a misconduct finding rather than be determined by evidence type and credentials. The evidence types present in a case may be correlated with these factors (for example, cases with interview admissions may involve more cooperative students who could be consequently assessed more leniently), but this correlation does not establish that evidence quality determines outcome severity.

Further investigations into linkages between evidence characteristics and dismissal did not return strong inference, presumably because the confounds described above constrain these to descriptive associations rather than causal claims. For example, cases with stronger inferential force and greater numbers of high-credibility items were less

likely to be dismissed (e.g., number of evidence items at the AIO stage: OR 0.31, 95% CI 0.06–0.89; Additional file 1, Table C.7), but the model could not establish that these evidence characteristics causally determined those outcomes. For instance, dismissals at the FIC may occur because students provide compelling accounts that clear them from wrongdoing, because procedural deficiencies are identified, or because welfare or equity considerations are exercised at the panel's discretion. The evidence records capture what was documented in the case file, not what was argued, tested, or weighed during the hearing. The dismissal findings are accordingly treated as descriptive associations: they indicate that cases with certain evidence profiles are more likely to end in dismissal, but they do not establish that evidence quality was the reason.

The absence of a strong predictive relationship is a finding about the institutional system rather than a limitation of the analysis. Outcomes of student academic misconduct proceedings are influenced by the properties of the institutional structure that is processing these matters, as evidence quality is not incorporated into the relevant procedures. Cases proceed, escalate, and are resolved based on policy triggers, investigation procedures, outcome guidelines, and panel discretion, with evidentiary assessment occurring informally and without structured criteria. The probative value framework developed in this study provides a more structured way to assess evidence quality beyond what currently exists in institutional practice. Its contribution is not as a predictor of outcomes within the current system, but as a diagnostic tool that reveals the gap between the evidentiary basis of misconduct cases and the formal decision architecture through which those cases are processed. This is especially valuable as GenAI technologies become more embedded in the student experience in higher education.

Discussion

The growth in GenAI-related misconduct allegations across the study period is consistent with broader trends in the literature and reflects the continued expansion of generative AI use in higher education rather than any development specific to this institution. In addition to the narrative around the volume of cases, a further finding exists in the changing composition of the evidence underpinning them. To our knowledge, the 15-code evidence taxonomy presented here is the first empirically derived classification of evidence in GenAI misconduct proceedings to incorporate systematic probative quality assessment of cases in higher education settings. Derived directly from institutional case records rather than academic discourse, it provides a grounded account of what evidence appears in proceedings and what that evidence is worth under the balance of probabilities standard. The central finding points to alignment between evidence type and its respective credential profile, with AI detection outputs consistently attracting weak ratings and admission-based evidence consistently attracting strong ones. These results shed light on the evidentiary landscape of GenAI misconduct and emphasise that different types of evidence make different contributions in establishing whether misconduct occurred.

A further observation is the temporal dynamism of probative value. The inferential force of any GenAI-specific signal is contingent on the state of LLM capabilities at the time a submission was made, and those capabilities have shifted considerably across the study period. The same evidence type therefore carries different weight depending on when it was identified. This reflects a deeper structural feature of GenAI misconduct

proceedings: investigative practice is reactive, seeking to rely on the weaknesses of LLM outputs to detect and support allegations of misconduct in a setting where those weaknesses are progressively diminishing. The implication is that any evidence framework applied in this space will require regular recalibration to remain fit for purpose, and the evidentiary landscape is expected to continue shifting beyond the study window as tool capability and authorised-use norms evolve. A further implication is that reliance on any single evidentiary signal will continue to carry inherent risk, however reliable it appears at a given moment.

For example, the decline in assertion-only cases and the near-disappearance of AI detection tool outputs as submitted evidence is an encouraging sign. Such a finding suggests that educators have become more realistic about what such tools can indicate. On the other hand, the rise of fabricated reference evidence deserves more careful treatment. At 30.0% of coded items in 2025, up from 10.4% in 2023, it has become the single largest structural shift in investigative practice across the period. This reflects a real, independently verifiable signal of GenAI involvement where it was previously only a limited indicator of contract cheating (Ellis et al. 2022) or use of file sharing sites or paraphrasing tool use (Rogerson 2017). The concern is not with the current quality of this evidence type but with its trajectory and longevity of use. LLMs are improving at generating plausible, retrievable citations at a documented rate, and a signal whose probative force rests on the unreliability of LLM citation generation will weaken as that generation becomes more reliable. If institutional practice continues to rely on reference verification as the core underpinning to allegations, it risks becoming dependent on an evidentiary signal that is being progressively eroded by the same technological development driving the caseload itself.

The expectation implicit in the probative value framework is that cases with stronger evidence should produce more consistent determinations. This implies stronger evidence will lead to fewer dismissals and more robust findings of whether academic misconduct occurred. The analysis found that the relationship between evidence quality and factors such as escalations, misconduct outcomes, and determinations of misconduct severity is not formally structured in this way. However, the absence of a formally structured relationship between evidence quality and outcomes is not a limitation of the analysis. It is a description of how the system currently operates. There is no requirement for investigators to assess the probative quality of evidence before progressing an allegation, no minimum evidentiary threshold at any stage of the pipeline, and no policy mechanism that distinguishes between evidence types based on what can be established. A process that relies on individual expertise for its evidentiary rigour will produce outcomes that vary with that expertise, with allegations raised by staff with less expertise increasing the risk of false positives. This variation is consequential for students, who may be required to resort to appeal mechanisms if they face disciplinary proceedings whose evidentiary basis is weak, and for institutions, whose findings are correspondingly difficult to defend as consistent and fair.

Practical implications

The evidence credentials framework (Anderson et al. 2005) provides a structured basis for assessing whether the evidence assembled in support of an allegation can discharge the balance of probabilities standard before the case is formally progressed. The

question inferential force asks refers to how much support an item, taken on its own, lends to the proposition of GenAI misuse, which is the same question the civil standard of proof poses at the point of determination and is applied here to gauge evidentiary weight rather than to reach a determination. Applying that question at the point of allegation, using the taxonomy to classify evidence and the credential framework to assess its weight, would bring investigative practice into closer alignment with the evidentiary demands of the process it initiates.

At the level of institutional policy, individual application of the credentials framework is necessary but insufficient while evidentiary standards remain uncodified. Inconsistency in evidentiary values used in misconduct proceedings is a structural feature of the current system that individual good judgement cannot reliably overcome at scale. Codifying evidentiary standards in institutional misconduct procedures would make the evidentiary basis of proceedings more transparent, more consistent, and more defensible. This is not simply mechanically determining outcomes from evidence scores; instead, it is ensuring consequential decisions affecting students are made on a basis that is explicit, auditable, and sufficient to meet the standard of proof the process is required to apply.

Limitations and future research

Several limitations apply to this study. The analysis is based on retrospective case file documentation, and the evidence items coded represent what was formally documented in institutional records rather than the full evidentiary basis on which decisions were made. Institutional policy related to inappropriate use of GenAI evolved over the study period as more about the implications of misuse became apparent. Assessors may have relied on contextual knowledge, informal conversations, or other forms of tacit reasoning that did not enter the written record, and codes with lower observed frequencies may be subject to greater documentation underreporting bias.

The calibration sample of twelve cases is small; to address the concern that tuning and evaluation not rest on the same data, generalisation was assessed on a separate held-out sample of 54 cases coded blind to the pipeline. Agreement between the pipeline and human coders on that sample was statistically indistinguishable from agreement between the two human coders. A residual limitation is that several evidence codes remain rare even in the held-out sample, so per-code estimates for those codes are imprecise and are reported with that caveat rather than as pass-or-fail outcomes.

The analysis is not disaggregated by student cohort (e.g., undergraduate/postgraduate, domestic/international) or by discipline; the de-identified dataset does not carry these fields and reporting at that level fell outside the study's ethics approval, so generalisability across cohorts and disciplines was not assessed. The salience of some stylistic signals (e.g., AI-typical writing patterns) may also differ across disciplines.

Finally, the study is based on data from a single Australian university over a three-year period. Institutional policy, disciplinary composition, student population, and assessor training contexts vary across institutions, and the temporal patterns identified here may not generalise beyond this setting. Multi-institution replication, particularly across different regulatory and policy environments, is a priority for future research. Future research should extend this framework to multi-institution samples, incorporate assessment context as a covariate, and develop prospective study designs capable of capturing the assessor reasoning that retrospective case file analysis cannot access.

Conclusion

This study presents the first empirically grounded taxonomy of evidence in GenAI-related academic misconduct proceedings, comprising 15 evidence codes organised into eight substantive categories with probative value profiles across three credential dimensions. Applied to 1,162 cases and 1,855 evidence items from a single Australian university over 2023–2025, the taxonomy revealed that the type of evidence largely determines its credibility and inferential force profile. However, these credential profiles cannot be reliably linked to case outcomes. Furthermore, AI detection tool outputs, the most publicly prominent evidentiary instrument in institutional responses to GenAI misuse, consistently attract the weakest credential ratings. The temporal analysis demonstrated that investigative practice is developing, but in ways that introduce new vulnerabilities as they close existing ones. The decline of assertion-only cases and the retreat from AI detection tool evidence are genuine improvements in evidentiary judgement. The concurrent rise of fabricated reference evidence reflects growing investigator sophistication and an emerging dependency: as LLMs become more capable of generating plausible and retrievable citations, the discriminating power of reference verification as a GenAI-specific signal will diminish. Institutional practice is currently converging on an evidence type whose probative value is being progressively eroded by the same technological development that is driving the caseload growth it is being called upon to address. The taxonomy and credential framework developed in this study offer a structured vocabulary to assess evidence quality. The taxonomy lays important groundwork for the codification of evidentiary standards into institutional policy, ensuring consistent application across cases and open to be scrutinised for fairness. The taxonomy and credential framework's contribution is not as a predictor of outcomes within the current system, but as an instrument for building a more evidence-informed misconduct process as institutions seek to protect academic integrity in an age of GenAI.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s40979-026-00235-9.

Supplementary Material 1

Author contributions

A.M. wrote the main manuscript, conducted the analysis, and reviewed the literature. M.H. made conceptual contributions to the work and made extensive contributions to the manuscript clarity, coherence and cohesion. C.L. helped scrutinise and confirm the taxonomy codes, contributed to the testing regime for the local LLM and confirmed the validity of the scripts used to conduct the analisys. A.R. contributed to the development of the institutional policies, data repository and case management system used to collect the data. All authors reviewed and contributed to the development and finalisation of the manuscript.

Data availability

In accordance with the research ethics approval conditions imposed by the authors' institution, data used in this analysis cannot be shared with entities outside the institution. Ethics approval to access academic misconduct records was provided by the authors institution (HREC H2025-0735).

Declaration

Ethics approval and consent to participate

Approval for use, analysis, waiver of participants consent, and access to the datasets was granted by the University of Wollongong's Human Research Ethics Committee (H2025-0735) and the research was conducted in line with institutional and national guidelines. The datasets analysed during the current study are not publicly available due to the sensitive nature of academic misconduct case records and the conditions of the ethics approval, which restrict data sharing to protect student privacy. The computational scripts used for data processing, analysis, and evidence coding are available from the corresponding author on request. The researchers are employees of the institution whose misconduct data was analysed. This positionality enabled access to the data and domain knowledge but introduces potential bias,

especially during the calibration exercise. This bias was mitigated by the anonymisation of data, except during the calibration exercises where the sample of cases did not involve any of the researchers as decision makers or investigators.

Competing interests
Prof. Ann Rogerson is an existing editorial board member of IJEI. No other interests exist.

Received: 10 May 2026 / Accepted: 17 August 2026
Published online: 02 September 2026

References

Anderson T, Schum D, Twining W (2005) Analysis of evidence. Cambridge University Press
Bailey KD (1994) Typologies and taxonomies: An introduction to classification techniques. Sage 12
Bassett MA, Bradshaw W, Bornsztejn H, Hogg A, Murdoch K, Pearce B, Webber C (2026) Heads we win, tails you lose: AI detectors in education. J High Educ Policy Manag 1–16. https://doi.org/10.1080/1360080X.2026.2622146
Bowen GA (2009) Document Analysis as a Qualitative Research Method. Qualitative Res J 9(2):27–40. https://doi.org/10.3316/QRJ0902027
Bretag T, Mahmud S, Wallace M, Walker R, James C, Green M, East J, McGowan U, Patridge L (2011) Core elements of exemplary academic integrity policy in Australian higher education. Int J Educ Integr 7(2)
Burgess R, Waters K, Spray E, Prieto-Rodriguez E (2026) Using large language models to complement humans for the coding of social media interactions between science teachers. Discover Educ 5(1):81. https://doi.org/10.1007/s44217-025-00868-x
Carlson NA, Burbano V (2026) The use of LLMs to annotate data in management research: Foundational guidelines and warnings. Strateg Manag J 47(3):699–725. https://doi.org/10.1002/smj.70023
Cohen J (2013) Statistical power analysis for the behavioral sciences. Academic, rev. ed
Dawson P, Nicola-Richmond K, Partridge H (2024) Beyond open book versus closed book: a taxonomy of restrictions in online examinations. Assess Evaluation High Educ 49(2):262–274. https://doi.org/10.1080/02602938.2023.2209298
Department of Education, S. a. E (2021) *Higher Education Standards Framework (Threshold Standards)* (F2021L00488). Federal Register of Legislation, Canberra. https://www.legislation.gov.au/F2021L00488 Australian Government Retrieved from
Doty DH, Glick WH (1994) Typologies As a Unique Form Of Theory Building: Toward Improved Understanding and Modeling. Acad Manage Rev 19(2):230–251. https://doi.org/10.5465/amr.1994.9410210748
Dunivin ZO (2025) Scaling hermeneutics: a guide to qualitative coding with LLMs for reflexive content analysis. EPJ Data Sci 14(1):28. https://doi.org/10.1140/epjds/s13688-025-00548-8
Eaton SE (2025) Global Trends in Education: Artificial Intelligence, Postplagiarism, and Future-focused Learning for 2025 and Beyond – 2024–2025 Werklund Distinguished Research Lecture. Int J Educational Integr 21(1):12. https://doi.org/10.1007/s40979-025-00187-6
Elkhatat AM, Elsaid K, Almeer S (2023) Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. Int J Educational Integr 19(1):17. https://doi.org/10.1007/s40979-023-00140-5
Ellis C, van Haeringen K, House D (2020) Technology, policy and research: establishing evidentiary standards for managing contract cheating cases. In: T. Bretag (Ed.) A Research Agenda for Academic Integrity (pp. 138–151). Edward Elgar Publishing. https://doi.org/10.4337/9781789903775.00017
Ellis C, Rogerson AM, House D, Murdoch K (2022) (Im)possible to prove: formalising academic judgement evidence in contract cheating cases using bibliographic forensics. Contract cheating in higher education: global perspectives on theory, practice, and policy. Springer; p. 185–198
Fleckenstein J, Meyer J, Jansen T, Keller SD, Köller O, Möller J (2024) Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. Computers Education: Artif Intell 6:100209. https://doi.org/10.1016/j.caeai.2024.100209
Gale NK, Heath G, Cameron E, Rashid S, Redwood S (2013) Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 13(1):117
Glöckner A, Engel C (2013) Can We Trust Intuitive Jurors? Standards of Proof and the Probative Value of Evidence in Coherence-Based Reasoning. J Empir Legal Stud 10(2):230–252. https://doi.org/10.1111/jels.12009
Goodier M (2025) Revealed: thousands of UK university students caught cheating using AI. The Guardian. Retrieved March 31 2026, from: https://www.theguardian.com/education/2025/jun/15/thousands-of-uk-university-students-caught-cheating-using-ai-artificial-intelligence-survey
Hadra M, Cambridge K, Mesbah M (2026) Evaluating the accuracy and reliability of AI content detectors in academic contexts. Int J Educational Integr 22(1):4. https://doi.org/10.1007/s40979-026-00213-1
Kumar S, Rao P, Singhania S, Verma S, Kheterpal M (2024) Will artificial intelligence drive the advancements in higher education? A tri-phased exploration. Technol Forecast Soc Chang 201:123258
Kundisch D, Muntermann J, Oberländer AM, Rau D, Röglinger M, Schoormann T, Szopinski D (2022) An Update for Taxonomy Designers. Bus Inform Syst Eng 64(4):421–439. https://doi.org/10.1007/s12599-021-00723-x
Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174
Masih A, Afzal B, Firdoos S, Mahmood J, Ali A, Abdulnabi MS, Balungu DM (2025) Classifying human vs. AI text with machine learning and explainable transformer models. Sci Rep 15(1):43310. https://doi.org/10.1038/s41598-025-27377-z
McKie A (2026) Universities are relying on AI-detection software to catch cheating. How well do the programs work? Nature 655(8122):535–537
Michel-Villarreal R, Vilalta-Perdomo E, Salinas-Navarro DE, Thierry-Aguilera R, Gerardou FS (2023) Challenges and opportunities of generative AI for higher education as explained by ChatGPT. Educ Sci 13(9):856
Moorhouse BL, Yeo MA, Wan Y (2023) Generative AI tools and assessment: Guidelines of the world's top-ranking universities. Computers Educ Open 5:100151. https://doi.org/10.1016/j.caeo.2023.100151
Munoz A, Hinchcliff M, Mackay J, Rogerson A (2025) Examining student reflections on academic misconduct: insights for academic integrity in higher education. Stud High Educ 1–15. https://doi.org/10.1080/03075079.2025.2591419

Newton PM (2025) How vulnerable are UK universities to cheating with new GenAI tools? A pragmatic risk assessment. Assess Evaluation High Educ 50(8):1332–1343. https://doi.org/10.1080/02602938.2025.2511794
Nickerson RC, Varshney U, Muntermann J (2013) A method for taxonomy development and its application in information systems. Eur J Inform Syst 22(3):336–359. https://doi.org/10.1057/ejis.2012.26
Nicola-Richmond K, Dawson P, Partridge H (2024) Online proctored exams: rhetoric vs reality. High Educ Res Dev 43(2):392–405. https://doi.org/10.1080/07294360.2023.2234310
O'Connor C, Joffe H (2020) Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. Int J Qualitative Methods 19:1609406919899220. https://doi.org/10.1177/1609406919899220
Perkins M, Roe J (2024) Decoding Academic Integrity Policies: A Corpus Linguistics Investigation of AI and Other Technological Threats. High Educ Policy 37(3):633–653. https://doi.org/10.1057/s41307-023-00323-2
Perkins M, Gezgin UB, Roe J (2020) Reducing plagiarism through academic misconduct education. Int J Educational Integr 16(1):3. https://doi.org/10.1007/s40979-020-00052-8
Perkins M, Roe J, Postma D, McGaughran J, Hickerson D (2024) 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. https://doi.org/10.1007/s10805-023-09492-6
Rogerson AM (2017) Detecting contract cheating in essay and report submissions: process, patterns, clues and conversations. Int J Educational Integr 13(1):10. https://doi.org/10.1007/s40979-017-0021-6
TEQSA (2024) Guidance note: academic and research integrity (Version 2.0). Australian Government. Retrieved from https://www.teqsa.gov.au/guides-resources/resources/guidance-notes/guidance-note-academic-and-research-integrity
TEQSA (2025) Student academic misconduct – outcomes and penalties. Australian Government. Retrieved from https://www.teqsa.gov.au/students/student-academic-misconduct-resources/outcomes-and-penalties
Than N, Fan L, Law T, Nelson LK, McCall L (2025) Updating The Future of Coding: Qualitative Coding with Generative Large Language Models. Sociol Methods Res 54(3):849–888. https://doi.org/10.1177/00491241251339188
Twining W (2006) Evidence as a multi-disciplinary subject. In: Twining W, editor. Rethinking Evidence: Exploratory Essays. 2nd ed. Cambridge University Press; p. 436–456. https://doi.org/10.1017/CBO9780511617249.016
Van Vlasselaer M, Van Droogenbroeck F, Spruyt B (2026) Who wrote this? Evaluating the reliability of AI detection tools in higher education. Int J Educational Integr 22(1):16. https://doi.org/10.1007/s40979-026-00226-w
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
Walters WH, Wilder EI (2023) Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep 13(1):14045. https://doi.org/10.1038/s41598-023-41032-5
Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltýnek T, Guerrero-Dib J, Popoola O, Šigut P, Waddington L (2023) Testing of detection tools for AI-generated text. Int J Educational Integr 19(1):26. https://doi.org/10.1007/s40979-023-00146-z

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.