Assessing Research Integrity in the Age of AI: A Longitudinal Analysis Using an AI Misuse Impact Index

Tichaona Chikore[1]
Farai Nyabadza[2]

AFFILIATIONS

[1&2]Department of Mathematics and Applied Mathematics, Faculty of Science, University of Johannesburg, Johannesburg, South Africa.

CORRESPONDENCE

Email: tichaona.chikore@gmail.com

REFERENCE

Chikore, T., & Nyabadza, F. (2026). Assessing research integrity in the Age of AI: A longitudinal analysis using an AI misuse impact index. In I. Kariyana & W. Sinkala (Eds.), *Artificial Intelligence and Postgraduate Supervision in Higher Education* (pp. 218–244). ERRCD Forum. https://doi.org/10.38140/obp4-2026-12

Abstract: The increasing adoption of artificial intelligence (AI) in academic research has reshaped scholarly practices while introducing complex ethical risks, particularly concerning research integrity and academic misconduct. This study proposes a comprehensive quantitative and empirical framework, adapted from the Cobb-Douglas production function, to model how the misuse of AI contributes to systemic quality degradation, using retractions as a proxy for integrity breaches. By leveraging longitudinal publication and retraction data from Retraction Watch and Scopus, we construct an AI misuse impact index to track the relationship between research output and integrity risks over time. Time series lag analysis reveals that retraction rates most strongly correlate with prior publication volumes at a one-year lag, indicating the rapid manifestation of AI-driven misconduct. To identify critical intervention points, we apply piecewise linear modelling to detect thresholds where retraction rates accelerate disproportionately relative to publication growth. A plagiarism tolerance threshold is established, beyond which research quality deteriorates unsustainably. Additionally, we introduce a probabilistic damage model, quantifying the risk of systemic integrity failure as AI adoption expands. Results highlight a pronounced post-2009 rise in AI-related integrity risks, with a sharp inflection in 2023 when misconduct indicators exceeded acceptable tolerance levels, signalling a system-wide ethical crisis. The study further proposes a dynamic, data-driven method for calibrating institutional plagiarism thresholds in alignment with evolving integrity risks and patterns of AI adoption. This model enables proactive monitoring and policy adjustments, linking integrity governance directly to empirical risk indicators. The findings underscore the urgent need for adaptive, transparent AI oversight frameworks within academia, ensuring that AI complements rather than undermines the ethical and intellectual foundations of research. Future research should extend this work by integrating discipline-specific AI use patterns and developing real-time academic integrity monitoring systems.

Keywords: Academic integrity, AI-Mediated supervision, AI policy, artificial intelligence, ethical considerations, impact index, plagiarism tolerance, responsible AI use.

1. Introduction

The increasing utilisation of artificial intelligence (AI) in postgraduate supervision is reshaping academic mentorship, presenting new opportunities to enhance research efficiency and streamline administrative tasks, particularly in guiding students through their complex research journeys (Ali, 2020; Altmäe et al., 2023). AI-driven tools are already transforming various aspects of academic work, such as automating literature reviews, managing administrative duties, and providing personalised, data-driven feedback to students (Kamalov et al., 2023; Ali et al., 2024). These innovations hold significant promise, but their integration into postgraduate supervision—an inherently personalised mentorship process—raises unique challenges. Postgraduate supervision extends beyond mere knowledge transfer; it fosters the development of critical research skills, ethical reasoning, and professional growth, elements that AI may struggle to replicate or support effectively (Zawacki-Richter et al., 2019; Mauti, 2025). Despite AI's efficiency, concerns are growing regarding its impact on traditional mentoring, which emphasises human interaction, ethical guidance, and individualised development, particularly in the context of postgraduate education (Nguyen & Vuong, 2024; Zhai et al., 2024).

AI tools offer clear benefits for supporting postgraduate research; however, concerns regarding their potential to undermine academic integrity, encourage over-reliance on technology, and depersonalise supervision remain underexplored (Crawford et al., 2024; Ahmad et al., 2023). The increasing integration of AI into research processes may diminish ethical reasoning and weaken the traditional mentorship bond. Additionally, the rise in article retractions since the COVID-19 pandemic, coinciding with the emergence of tools such as ChatGPT, raises pertinent questions about AI's role in academic misconduct (Nguyen & Vuong, 2024; Al-Jahwari & Yousif, 2025).

Despite the compelling qualitative evidence linking the rapid adoption of Generative AI to rising academic misconduct, the research community faces an urgent, quantifiable crisis that remains unaddressed. The most significant gap in the current literature is the critical lack of a clear, empirical, and quantitative framework to systematically measure how AI misuse translates into systemic degradation of research quality across the enterprise. Current studies acknowledge the ethical risks but fail to model the quantitative relationship between key research inputs (for example, human effort, institutional resources, and AI utilisation) and measurable outcomes of integrity failure (for example, retractions) over time (Acosta-Enriquez et al., 2025; Papagiannidis et al., 2025).

This deficiency leaves academic institutions effectively blind to the escalating risk of systemic integrity failure. Without a rigorous quantitative model, it is impossible to identify the critical inflection points, specifically the 'plagiarism tolerance thresholds', where the rate of integrity breaches begins to accelerate disproportionately, thereby exceeding institutional capacity for effective quality control. Consequently, existing governance policies remain largely reactive, failing to provide the predictive mechanisms necessary for timely, evidence-based intervention. This leads to the central, unmet need that this study addresses: the necessity for a framework to predict and quantitatively model the risk of integrity collapse in the AI-mediated research environment.

To overcome this methodological gap and provide the necessary predictive framework, we utilise a two-stage approach. This includes an analytical model adapted from the foundational

Cobb-Douglas production function (Cobb & Douglas, 1928), as applied in related research (Baulk, 2024), coupled with piecewise linear modelling to empirically detect these critical quantitative inflexion points. To explore temporal dynamics, we apply lagged correlation analysis to determine whether retractions follow publication spikes with a delay, thereby identifying optimal intervention timing. Using piecewise linear regression (Yang et al., 2016), we aim to detect thresholds where retractions accelerate disproportionately relative to output, thereby aiding in the definition of critical risk points. We also propose a plagiarism tolerance threshold to signal when misconduct exceeds acceptable limits and test this concept through simulations comparing capped and uncapped misuse scenarios. The model examines institutional enforcement patterns and adapts tolerance levels dynamically. Finally, we develop a probabilistic model to estimate the likelihood of integrity collapse due to cumulative AI misuse, integrating damage thresholds and failure probabilities to inform early institutional interventions.

This study aims to construct a quantitative integrity risk index for artificial intelligence (AI) misuse in research by developing an AI Misuse Impact Index that integrates publication and retraction data to quantify the relationship between the growth of AI-driven research output and integrity risks. The index will serve as an empirical tool for tracking vulnerabilities in research integrity as AI adoption expands. Using piecewise linear modelling and time series lag analysis, the study seeks to identify critical thresholds and tipping points where retraction rates accelerate disproportionately in relation to publication growth, signalling systemic misconduct. Furthermore, the research will establish empirically grounded plagiarism tolerance thresholds and determine critical inflection points that indicate the onset of widespread ethical degradation within academic systems. To enhance analytical precision, a probabilistic damage model will be introduced to quantify the risk of systemic research misconduct under varying scenarios of AI adoption, capturing the dynamic interplay between research output, AI misuse, and retraction patterns as a predictive tool for monitoring emerging academic integrity crises. Finally, the study proposes a dynamic, data-driven integrity governance framework that enables institutions to adapt plagiarism tolerance thresholds and integrity policies in real time, linking empirical indicators to institutional decision-making for proactive intervention and the sustained safeguarding of research integrity in AI-augmented academic environments.

We identify key points in postgraduate supervision where the use of AI should be regulated, utilising patterns of academic misconduct predicted by machine learning models. The study evaluates AI's impact on students' research skills, ethics, and development, proposing targeted interventions to mitigate risks. It offers a measurable framework for balancing AI efficiency with ethical oversight and supports policy recommendations such as ethical AI guidelines, adaptive plagiarism thresholds, and real-time integrity monitoring to assist universities in maintaining academic standards while adopting new technologies.

2. Materials and Methods

Our analysis follows a two-stage methodological design. This approach is fundamentally based on the Cobb-Douglas production function (Cobb & Douglas, 1928), which quantitatively models how the combined factors of AI, human effort, and institutional resources impact research quality within AI-mediated supervision. First, a general model defines the relationship between inputs and research output. Second, the model is extended to account for variations in AI adoption across individuals and institutions. This approach captures both direct and systemic effects of AI use, allowing for empirical validation and informing policy decisions.

2.1 The general case

We propose a model where research quality (Q) depends on study resources (R), combined effort (E), and AI use (A). Resources include infrastructure such as libraries and software, while effort reflects contributions from both students and supervisors. AI acts as a multiplier, enhancing how resources and effort affect research quality. The model assumes diminishing returns; beyond certain thresholds, increasing resources, effort, or AI yields smaller gains. AI tools improve efficiency and accuracy, while also reducing misconduct, thus helping to optimise research outcomes. To quantify these relationships, we introduce elasticities εR and εE, representing the proportional contributions of resources, effort, and AI to research quality, respectively. Therefore, we model research quality as a function of resources, effort, and the technological factor as follows:

$$Q = A^{\gamma} R^{\alpha} E^{\beta}, \tag{1}$$

such that $\gamma, \alpha, \beta \in [0,1]$, satisfying the condition that $\gamma + \alpha + \beta = 1$ always. We also assume that the initial values $A(0), R(0), E(0) \geq 0$. Through this model, we capture the interactions between these variables and can assess how changes in resources, effort, and technological enhancement influence the overall research quality. The condition that $\gamma + \alpha + \beta = 1$ enforces constant returns to scale, meaning that a proportional increase in all inputs results in a proportional increase in research quality. The model assumes that increased resources, effort, and AI use improve research output without excessive inflation. This helps evaluate how different inputs affect research quality and integrity, guiding supervision policies.

To begin, we assess the impact of AI on research quality by comparing expected and actual research output within our proposed model. By analysing cases where AI leads to increased retraction rates or diminished originality, we can determine appropriate intervention strategies, such as adjusting AI usage policies or modifying academic integrity thresholds. Research quality follows the functional form given in (1), and if AI does not influence research quality, we define the expected research quality in AI's absence ($\gamma = 0$) as:

$$Q^{*} = R^{\alpha} E^{\beta}, \tag{2}$$

where $Q^{*} \leq Q$ represents the baseline research quality determined solely by resources and effort, and $\alpha + \beta = 1$. The difference between actual and expected research quality allows us to isolate the impact of AI:

$$\Delta Q = Q - Q^* = A^\gamma R^\alpha E^\beta - R^\alpha E^\beta, \tag{3}$$

Rearranging, we obtain:

$$\Delta Q = R^\alpha E^\beta (A^\gamma - 1). \tag{4}$$

The model is not mathematically tractable if $A = 1$, hence we consider the case when $A > 1$, then $\Delta Q > 0$, suggesting that AI enhances research output, potentially by improving access to information, automating writing assistance, or refining data analysis. Conversely, if $0 < A < 1$, then $\Delta Q < 0$, implying that AI negatively impacts research quality, possibly due to overreliance on automated systems, increased plagiarism, or decreased originality in academic work. In Figure 1, we demonstrate the nature of the trajectory of ΔQ when $A > 1$ revealing a monotonically increasing graph confirming that higher AI use tends to correlate with greater deviations in research quality relative to baseline expectations. This framework allows for the empirical measurement of AI's effect on research quality by estimating the deviation of actual output from expected output.

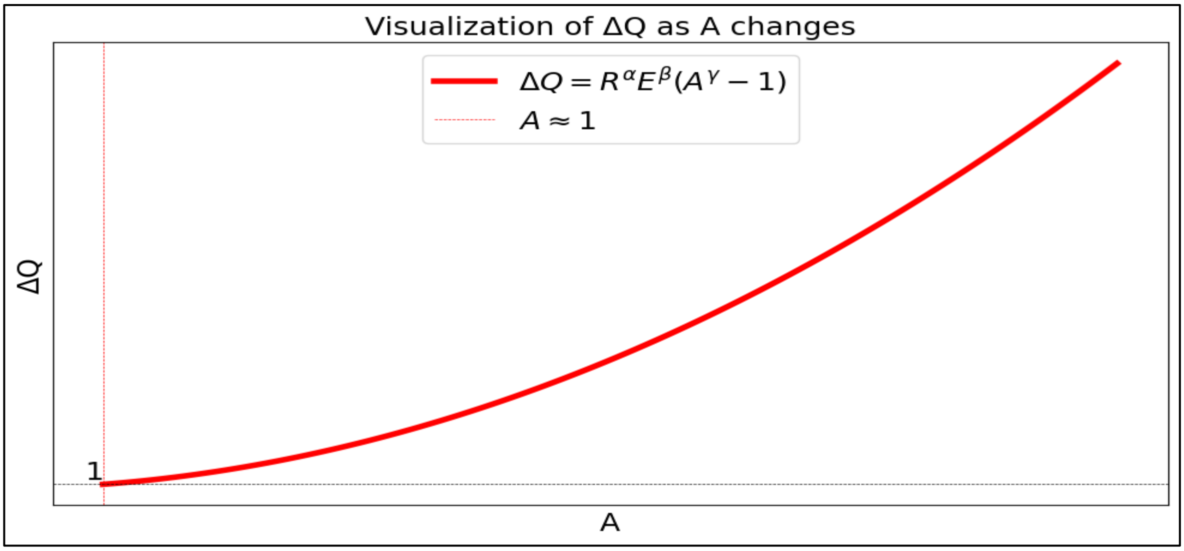

Figure 1: *Graph of ΔQ as A changes for a constant $\gamma > 0$: The symbolic graph indicates a positive correlation between A and ΔQ, suggesting that higher values of A lead to a greater increase in ΔQ.*

We derive the first-order conditions (FOCs) to determine the optimal allocation of AI, resources and effort for maximizing research quality. Given the research output function in Equation (1), we take the partial derivatives with respect to A, R and E to analyse how research quality responds to changes in these inputs. The partial derivative with respect to A is

$$\frac{\partial Q}{\partial A} = \gamma A^{\gamma-1} R^\alpha E^\beta. \tag{5}$$

Setting this equal to zero would imply an optimal AI level, but since AI usage is typically a decision rather than a resource that can be infinitely adjusted, we must incorporate its cost to make this equation meaningful. We assume that there is a cost associated with AI usage, denoted as $c(A)$ a cost per unit of AI. For simplicity, we assume a linear function $c(A) = kA$. Resources, which include expenses related to AI subscriptions, books, databases, computational tools, and AI-driven research assistants, have a cost r. Investing in more resources incurs higher costs, which must be balanced against their contribution to research quality. We also assume that the cost per unit of effort is w, representing the opportunity cost of time and energy expended by students and supervisors in conducting research. This can be interpreted as workload, time spent on revisions, or even financial costs such as research grants. The optimisation problem now becomes:

$$\max_{A,R,E}[Q - (rR + wE + c(A))]. \tag{6}$$

Taking the first-order condition of Equation (6) with respect to A from (1) and (5), we get that

$$c'(A) = k = \gamma A^{\gamma-1} R^{\alpha} E^{\beta}. \tag{7}$$

This equation states that the marginal benefit of AI on research quality should equal its marginal cost. If AI is costly or leads to unintended negative consequences (such as increased retractions or loss of originality), adjusting its usage becomes necessary. If the marginal benefit of AI exceeds k, (that is, $\gamma A^{\gamma-1} R^{\alpha} E^{\beta} > k$), increasing A raises net quality; if below k, decreasing A is optimal. If the optimal AI level is exceeded (that is, A is too high), it may indicate that AI is replacing genuine research effort rather than complementing it, necessitating intervention.

For resources, the first-order condition is:

$$\frac{\partial Q}{\partial R} = \alpha A^{\gamma} R^{\alpha-1} E^{\beta} . \tag{8}$$

This result implies that when α is high, resources such as library access, technological tools, and academic databases play a pivotal role in improving research quality. If R It is too low; investing in these resources becomes essential to maintain high research standards. Conversely, when α is low, the impact of additional resources is minimal, indicating that increasing R does not substantially enhance research quality. Instead, other factors such as researcher effort or the AI play a more dominant role.

For effort, the first-order condition is:

$$\frac{\partial Q}{\partial E} = \beta A^{\gamma} R^{\alpha} E^{\beta-1}. \tag{9}$$

This result suggests that if β is high, research quality is more dependent on human effort. A low level of effort, particularly due to an over-reliance on AI tools, may compromise academic rigour and originality. To determine the optimal balance between resources, effort, and AI usage, we introduce a cost function that includes the cost of resources (rR) and the cost of effort (wE). The objective is to maximise net research quality:

$$\max_{A,R,E}[Q - (rR + wE + c(A))]. \tag{10}$$

Taking the first-order conditions for this optimisation problem, we obtain that:

$r = \alpha A^{\gamma} R^{\alpha-1} E^{\beta}$ and $w = \beta A^{\gamma} R^{\alpha} E^{\beta-1}$.

These conditions establish equilibrium points where the marginal contribution of resources and effort to research quality equals their respective costs.

We introduce a utility framework that incorporates a temporal adjustment mechanism, recognising that the perceived value of research output evolves over time. As AI becomes more integrated into research practices, its marginal utility diminishes due to factors such as over-reliance, reduced originality, and ethical concerns. We introduce a discounting factor, $\delta \in [0,1]$, which adjusts the utility derived from each additional AI-enhanced research output, where $n(t)$ is the number of published articles on AI. The student's utility at the time t follows:

$$U(t) = U(Q_\infty) + \delta^{n(t)}[U(Q_0) - U(Q_\infty)] \tag{11}$$

where $U(Q_0)$ and $U(Q_\infty)$ represent the utility levels when AI use is minimal and at its maximum saturation, respectively and $Q_\infty = \max_{A,R,E}[Q - (rR + wE + c(A))]$ as given in Equation (10). This formulation captures the transition from initial AI adoption, where its benefits are pronounced, to a scenario where excessive reliance leads to diminishing returns. The discounting effect ensures that AI usage is dynamically regulated, preventing overuse that might compromise research integrity. By fitting this model to AI publication data, we obtained the following fitted parameters $\delta = 0.8123,\ U(Q_0) = 7.26 \times 10^{-16},\ U(Q_\infty) = 1$ and an R^2 value of 0.8735 indicating a solid fit to the observed trends.

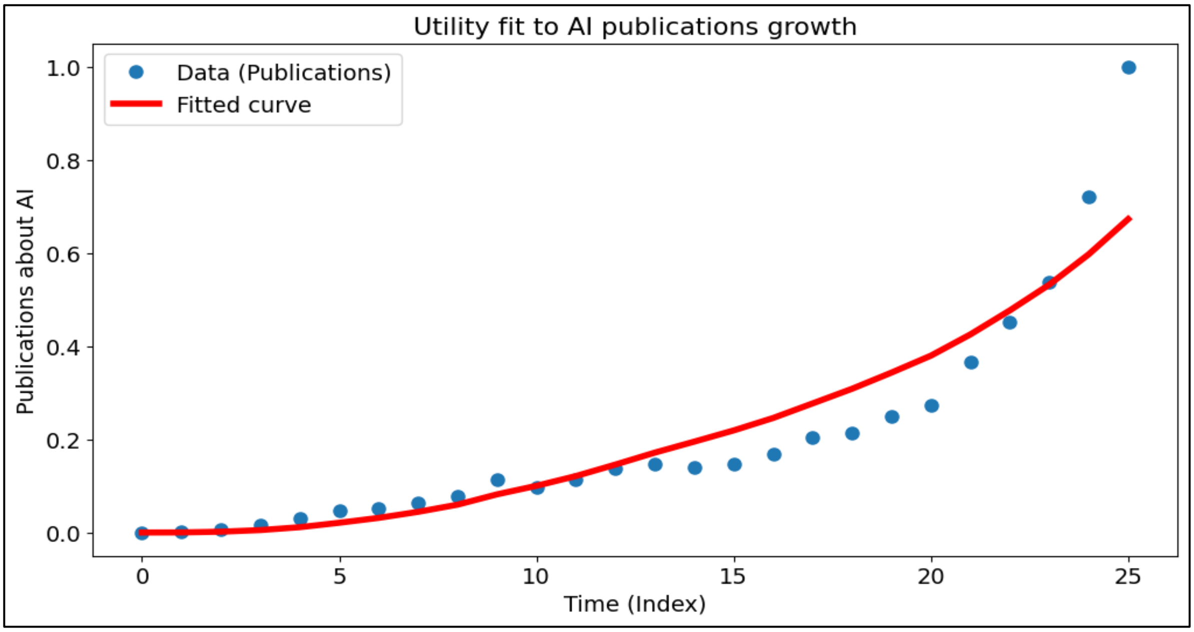

Figure 2: *Fitted utility curve modelling the evolving impact of AI-enhanced research over time, with estimated initial values for* $\delta, U(Q_0), U(Q_\infty)$ *being 1, 0, and 0.95, respectively*

The AI publications data is a proxy for AI awareness growth and spread annually. The results in Figure 2 suggest that while the initial perceived utility of AI-enhanced research was almost negligible, it increased rapidly with adoption and then gradually levelled off as the number of AI-related publications grew. The long-term utility stabilises at a positive value of 1, which means that even in a saturated research environment, AI maintains a lasting, meaningful contribution. However, its incremental benefits diminish over time, emphasising the importance of thoughtful and measured AI adoption in research to preserve both the value of AI and the integrity of the research field as it continues to evolve. For the rest of the study, without loss of generality, we work with the maximum utility 1, which is negligible as a factor and does not affect the model's outcomes.

2.2 The heterogeneous case

To model the heterogeneity of AI use within postgraduate research with maximum utility, we introduce an AI adoption distribution function that accounts for variations across researchers, disciplines, and institutions. This approach deviates from the assumption of uniform AI utilisation by acknowledging that AI integration into research workflows is influenced by factors such as familiarity with AI tools, institutional policies, ethical considerations, and field-specific norms. To understand individual student behaviour, we introduce distinct researcher types indexed by $i = 0,..,n$, where each researcher i uses AI at a distinct level A_i. The aggregate research quality across all researchers is then expressed as:

$$Q_{agg} = \sum_i A_i^{\gamma} R^{\alpha} E^{\beta} \ . \qquad (12)$$

In cases where AI heterogeneity arises due to institutional access, career incentives, or disciplinary norms, we define AI use for researchers i as $A_i = A_0 + \theta_i A_0$ where A_0 represents baseline AI access, and θ_i captures individual deviations due to external factors or preferences. If excessive AI use compromises research integrity, institutions may restrict its application in specific areas. Conversely, if underuse results from limited access or insufficient training, policies may prioritise AI literacy and equitable access to resources.

Building on the understanding that AI adoption varies across different research tasks, we model the effect of AI use on research quality by treating research output as a collection of individual research articles within a broader academic research space, denoted as B. The articles within the research space B are put into classes $b_1, b_2, \dots, b_j$, where each b_j corresponds to a specific aspect of research quality, such as novelty, proper citation practices, depth of analysis, and evaluative reasoning. For simplicity, we assume that each b_j is independent of the rest, hence, coupling or cross terms are not considered. To account for AI's influence, we assume a baseline set of AI usage levels $A_0 = \{z_1, z_2, \dots, z_k\}$ where each z_i represents the responsible, acceptable level of AI use that supports quality research in the corresponding component b_i without undermining academic standards. However, the actual AI use intensities, denoted as $\theta_i A_0 = \{s_1, s_2, \dots, s_k\}$ can deviate from these baselines, potentially introducing ethical risks to research quality. To quantify the deviation from responsible AI use, we define the difference between actual AI use and the baseline levels in each b_j as:

$$\{t_{11}, t_{12}, \dots, t_{kj}\} = \{s_{11}, s_{12}, \dots, s_{kj}\} - \{z_{11}, z_{12}, \dots, z_{kj}\}. \tag{13}$$

Each research component b_j exhibits two functions: $\Gamma_j(t_{kj})$, which captures the positive contribution of AI within an acceptable range of use, and $d_j(t_{kj})$, which quantifies the potential negative effects of AI overuse. These negative effects may include issues such as increased plagiarism, reliance on AI-generated content, reduced originality, or weakened engagement with research materials. To prevent AI use from exceeding acceptable thresholds, we introduce a damage limit φ_j for each research component b_j, enforcing the condition:

$$\sum_{t_{kj} \in T} d_j(t_{kj}) < \varphi_j \quad \text{for all } k. \tag{14}$$

where T represents the set of AI use deviations beyond the baseline. If the cumulative negative impact $D_j(t)$ for any component exceeds φ_j, research quality is considered compromised within that dimension.

This reasoning is illustrated in the model flow in Figure 3 that follows:

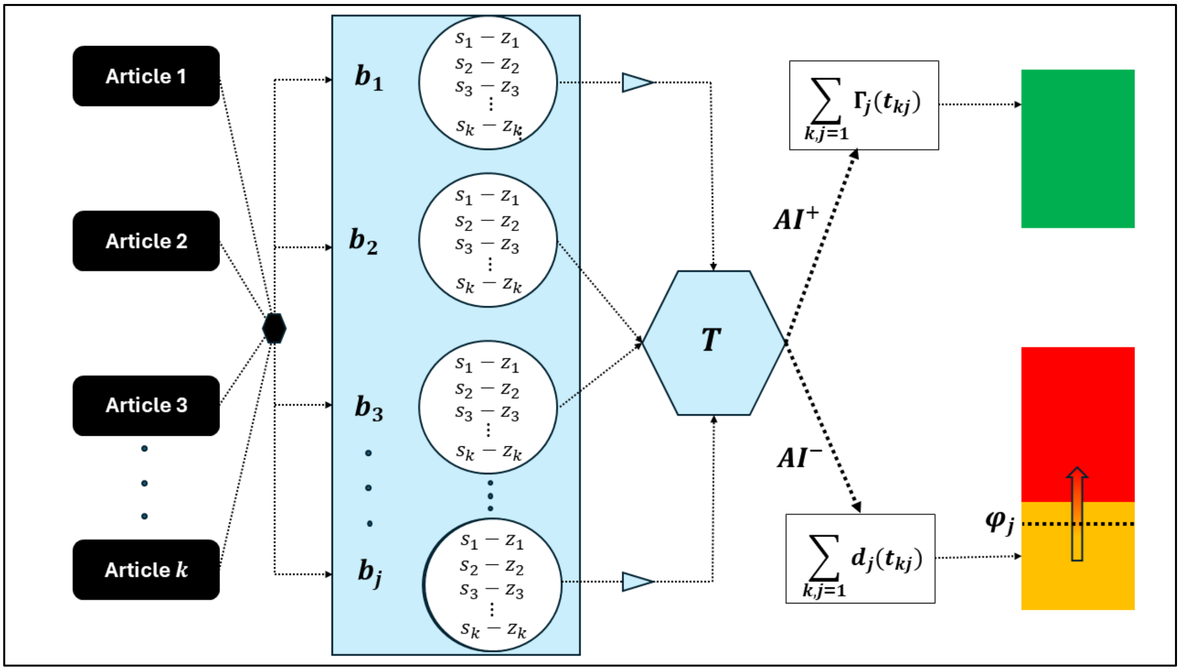

Figure 3: An illustration of how the model classifies and quantifies positive and negative gains by considering the research components b_j. The arrows for AI^+ and AI^- denote positive and negative benefits of AI use, respectively

By combining these ideas, we can express the total research quality as a sum of overall quality dimensions. Specifically, the overall research quality can be written as

$$Q_{agg} = \sum_k^K \left(\sum_{i=1}^N \left(\sum_i A_i^\gamma R^\alpha E^\beta \left[\Gamma_j(t_{kj}) - d_j(t_{kj})\right]\right)\right). \tag{15}$$

The inner summation over i aggregates the contributions from all researchers based on their individual AI use A_i (modulated $by\ A_i^\gamma$), research rigour R^α, and student effort E^β. The function $\Gamma_j(t_{kj})$ represents the beneficial contribution of AI to the k-th component when operating at its baseline level z_i, while $d_j(t_{kj})$ captures the detrimental impact arising from deviations t_{kj} beyond the baseline. The cumulative damage for each research component b_j is calculated based on the deviations t_{kj} and if the total negative impact $d_j(t_{kj})$ for any dimension exceeds its threshold φ_j, we consider that component to have experienced a compromise in research quality, allowing for the management of AI adoption in research.

The plagiarism tolerance P is directly related to the acceptable threshold for negative effects arising from AI use. In this context, we define damage tolerance φ_j as the upper limit for damage within each research component b_j, and plagiarism tolerance P as the threshold that ensures research quality remains uncompromised. The relationship between damage tolerance and plagiarism tolerance can be expressed as follows:

$$d_j(t_{kj}) \leq \varphi_j \leq P \tag{16}$$

This means that the cumulative damage resulting from AI overuse should not exceed the plagiarism tolerance P, which serves as the limit that ensures the integrity of research is maintained. When introducing plagiarism tolerance P into the overall research quality Q_{agg}, we must account for the effect of AI overuse on the research quality. The cumulative damage for each research component b_j influences the research quality, as seen in Equation (15). The role of plagiarism tolerance P is to prevent the cumulative damage from exceeding the acceptable limits. This ensures that $d_j(t_{kj})$ does not exceed the threshold at which research quality would be compromised. If $d_j(t_{kj})$ exceeds P, strict intervention is required to mitigate the overuse of AI and prevent further damage to research quality(for example, the faculty can reject a thesis submission). If $d_j(t_{kj}) \leq \varphi_j \leq P$, the research quality remains unaffected by AI overuse, and the total quality Q_{agg} is not compromised.

To account for plagiarism tolerance in adjusting the overall research quality Q_{agg}, a correction factor is introduced into the formula for total quality. This reflects the impact of the acceptable damage due to AI overuse:

$$Q_{agg} = \sum_k^K \left(\sum_{i=1}^N \left(\sum_i A_i^\gamma R^\alpha E^\beta \left[\Gamma_j(t_{kj}) - \min(d_j(t_{kj}), P)\right]\right)\right). \tag{17}$$

The minimum function ensures that if the damage $d_j(t_{kj})$ exceeds φ_j, the damage is capped at P to prevent further degradation of research quality. If the damage is less than or equal to φ_j, it remains as is, contributing to the overall quality without requiring intervention. The inclusion of plagiarism tolerance P in the formula for Q_{agg} ensures that AI overuse, particularly regarding plagiarism or reduced originality, does not undermine research quality beyond an acceptable threshold. This threshold is determined by PIntervention occurs when damage exceeds this value, adjusting the total quality to account for the negative impacts of AI. Therefore, P sets the upper limit for the damage $d_j(t_{kj})$ that is acceptable to maintain research quality, and the total research quality Q_{agg} is dynamically adjusted to prevent AI overuse from undermining important aspects such as originality, rigor, and academic integrity.

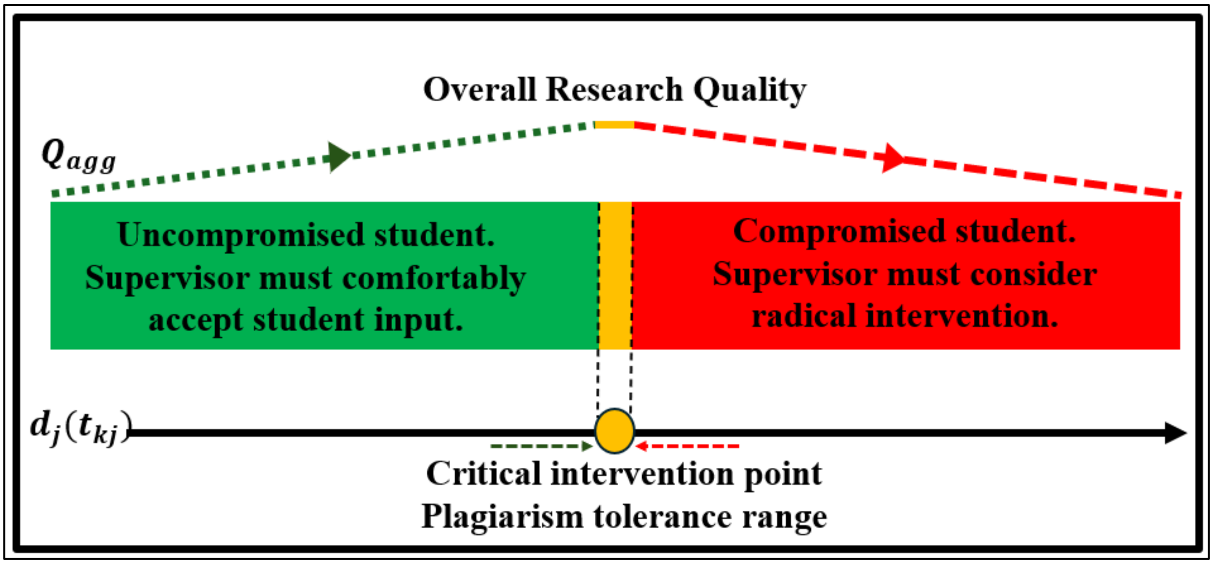
Overall Research Quality
Q_{agg}
Uncompromised student. Supervisor must comfortably accept student input.
Compromised student. Supervisor must consider radical intervention.
$d_j(t_{kj})$
Critical intervention point
Plagiarism tolerance range

Figure 4: A schematic illustration of the effect of AI use and abuse. Initially students are presumed to be responsible and at the end, the damage calls for radical intervention

Figure 4 presents two scenarios: in the case of an uncompromised student ($d_j(t_{kj}) < \varphi_j$), the supervisor can comfortably accept the student's input, which contributes positively to the aggregate research quality (Q_{agg}). Conversely, if the student is compromised ($d_j(t_{kj}) > P$), the supervisor may need to consider more drastic measures, as indicated by the red zone, which poses a threat to research quality. The intervention point serves as a critical boundary (when $\varphi_j < d_j(t_{kj}) < P$), prompting the supervisor to decide whether intervention is necessary based on the student's performance within the plagiarism tolerance range. This visual representation reinforces the delicate balance between student autonomy and the need for oversight in ensuring high-quality research output.

2.3 Modelling with data

This study relies on two core datasets, as shown in Figure 5: the number of peer-reviewed published articles and the number of retracted articles. Each dataset is selected based on its alignment with the central aim of analysing the relationship between research capacity, research output, and research integrity. The temporal scope of all datasets spans from 2000 to 2024, providing a 24-year window that enables the capture of long-term structural and systemic trends in global scientific production.

The publication and retraction datasets are derived from the Scopus database. Scopus is selected due to its broad disciplinary coverage, standardised indexing practices, and global reach, making it a suitable source for analysing trends in scholarly output across time and regions. Scopus includes metadata on authorship, affiliations, and document types, which supports disaggregation and further analysis. Its reliability and comprehensiveness make it an appropriate proxy for global research activity and integrity.

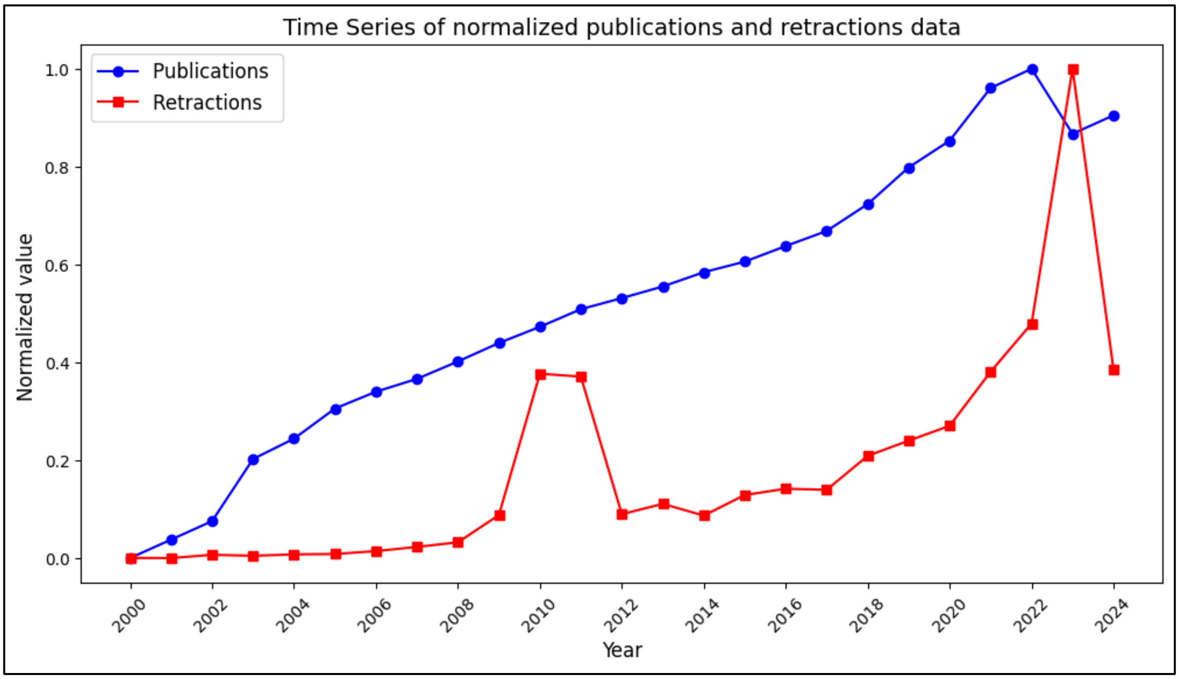

Figure 5: *The plot illustrates the evolving relationship among the three variables. A notable rise in retractions is observed after 2011–2012, following earlier increases in publications, suggesting a possible quality–quantity trade-off*

Retractions are formal notices that a publication has been withdrawn due to error, misconduct, or other violations of publication ethics. While retractions do not capture all dimensions of research integrity, they are one of the few standardised and publicly traceable signals of compromised scientific quality. Including this metric allows the study to empirically assess how the growth of research output may affect the prevalence of integrity failures. The global scope of retraction data is intentional, as research misconduct and correction practices are not limited to a specific country or system, and cross-national trends in retractions are evidence of systemic pressures affecting scientific rigour.

We normalise all datasets using min-max scaling ([0,1]) to control for systemic factors like population growth and expansion in higher education, ensuring trends reflect real shifts in research behaviour, not structural changes. This allows proportional comparisons across time, revealing whether retraction trends persist when adjusted for rising publication volumes. Normalisation is essential to accurately assess the link between research output and integrity. The datasets feed into a utility function that models how increased activity may trade off with quality, incorporating time lags to capture delayed effects. This approach ensures methodological rigour and interpretive clarity.

3. Presentation of Results

We provide a comprehensive analysis of research integrity risks, including the estimation of retraction-based risks and the identification of key thresholds. Lag analysis highlights critical intervention points, while the quantification of plagiarism tolerance underscores the impact of AI misuse on academic quality. Additionally, we explore the dynamics of institutional enforcement and model the time-to-failure of research quality under sustained AI-related misconduct.

3.1 Estimating retraction-based integrity risks

To indirectly estimate the impact of AI misuse on research integrity, we define an AI misuse impact index, denoted as $AI_x(t)$. This index is based on the assumption that retractions are primarily due to some unethical conduct on the part of the authors, and it is calculated as the ratio of the normalized number of retractions to the normalized number of publications in a given year, with a small positive constant ϵ (set to 0.001) added to the denominator to avoid division by zero in years with negligible publication output. The formula is expressed as:

$$AI_x(t) = \frac{\text{Number of retractions}}{\text{Number of Publications} + \epsilon} \tag{18}$$

A rising value of $AI_x(t)$ indicates an increasing integrity risk relative to research output, potentially reflecting the growing overuse or misuse of AI tools in academic research. Analysing the data, we observe that the $AI_x(t)$index remains negligible in the early 2000s, with values such as 0.0878 in 2002 and 0.0230 in 2003. However, a sharp increase appears around 2010, where the index jumps to 0.7973, signalling a substantial rise in integrity risks. This trend continues, with the index reaching 0.3963 in 2021 and 0.4792 in 2022. Notably, it peaks dramatically at 0.7975 in 2023, suggesting a critical point where AI misuse may have exceeded acceptable thresholds, likely corresponding to the plagiarism tolerance P in the model.

This pattern in Figure 6 aligns closely with the framework's assumptions that before 2010, AI's influence on academic norms appears minimal, with low integrity risks. Between 2010 and 2020, rising AI adoption coincided with steadily increasing retraction rates relative to publications, indicating growing concerns. From 2020 to 2023, the proxy index spikes rapidly, suggesting significant overuse or unethical reliance on AI tools in the model where integrity risks surpass tolerable limits. The finding that $AI_x(t)$ exceeds 1.0 in 2023 strongly suggests that the cumulative damage from AI misuse likely breached both the damage limit ϕ_j and plagiarism tolerance P, necessitating strict intervention to preserve research integrity. This pattern aligns closely with the framework's assumptions that before 2010, AI's influence on academic norms appears minimal, with low integrity risks. Between 2010 and 2020, rising AI adoption coincides with steadily increasing retraction rates relative to publications, indicating growing concerns. From 2020 to 2023, the proxy index spikes rapidly, suggesting significant overuse or unethical reliance on AI tools in the model where integrity risks surpass tolerable limits.

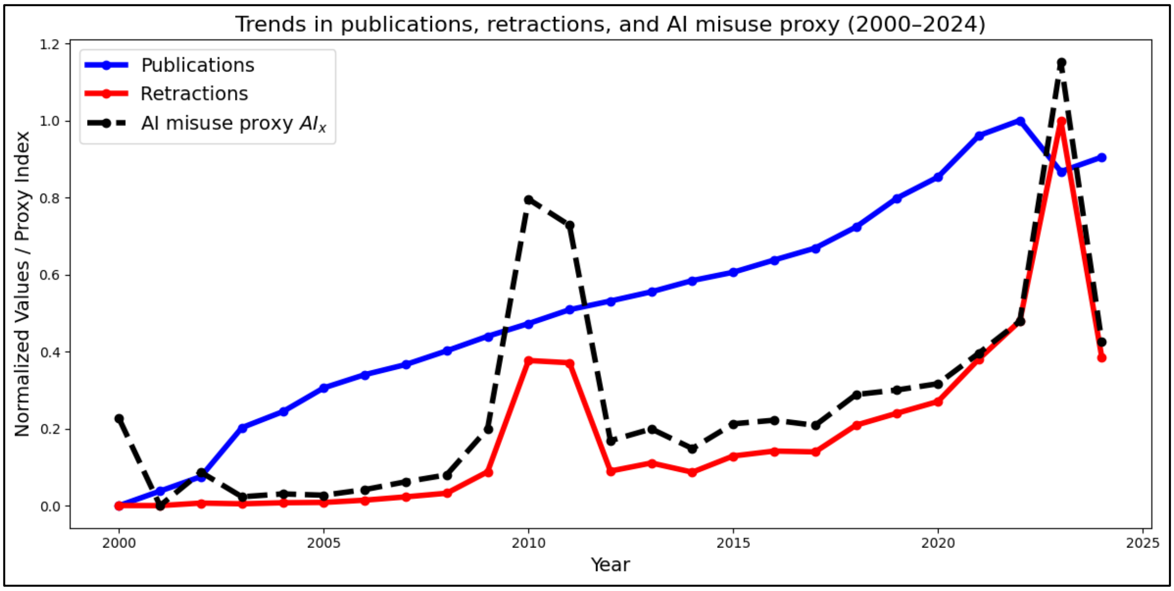

Figure 6: *Normalized trends in publications, retractions, and AI misuse proxy (2000–2024). Notice the sharp AI proxy spikes post-2010 and peaking in 2023, indicating rising integrity risks alongside AI adoption*

3.2 Lag analysis of retraction rates and thresholds estimation

Using time series correlation and lag analysis, we explore whether spikes in retraction rates follow increases in publication output. The results show that retraction rates are most strongly correlated with submissions one year later (Pearson correlation coefficient of 0.8576 at lag 1), indicating that unethical AI use, such as plagiarism or manipulation, manifests within a year. As the lag increases, the correlation weakens (0.6207 at lag 2 and 0.4501 at lag 3), supporting the hypothesis that AI-driven misconduct has a relatively short-term impact on retraction decisions.

Given this observation, which is illustrated in Figure 7, it is essential to define thresholds for intervention based on the most relevant time window and, more specifically, the 1-year lag period in which correlations are strongest. We use piecewise linear models rather than nonlinear alternatives like exponential or quadratic functions to identify explicit threshold points where the relationship between normalised publications and retractions changes behaviour. Piecewise linear models estimate these breakpoints directly, making them well-suited for detecting shifts associated with integrity risk thresholds in our theoretical framework. Nonlinear models can capture general trends but do not provide interpretable points of change. The data exhibits distinct phases of growth that align naturally with segmented linear behaviour, supporting the use of this model form.

We apply two models: a 2-segment and a 3-segment piecewise linear fit. The 2-segment model, with an R-squared value of 0.5446, identifies breakpoints at 0.0 and 0.6585 publications, capturing general retraction acceleration. The 3-segment model, with a higher R-squared value of 0.7517, identifies breakpoints at 0.0, 0.8525, 0.8615, and 1.0, offering a more detailed analysis. We use the 3-segment model, as it provides a more accurate reflection of retraction rate shifts, identifying the P-threshold where retraction rates significantly increase, and interventions may be needed to protect research integrity. In Figure 7, the spline reveals whether changes in retraction rates occur abruptly at the breakpoints or follow a continuous, gradual trend, helping assess if the segmented thresholds reflect real inflexion points or oversimplify a nonlinear relationship. These inflections represent key moments when the system's response to publication volume changes, suggesting areas where the retraction rate might either increase or slow down based on trends in AI misuse.

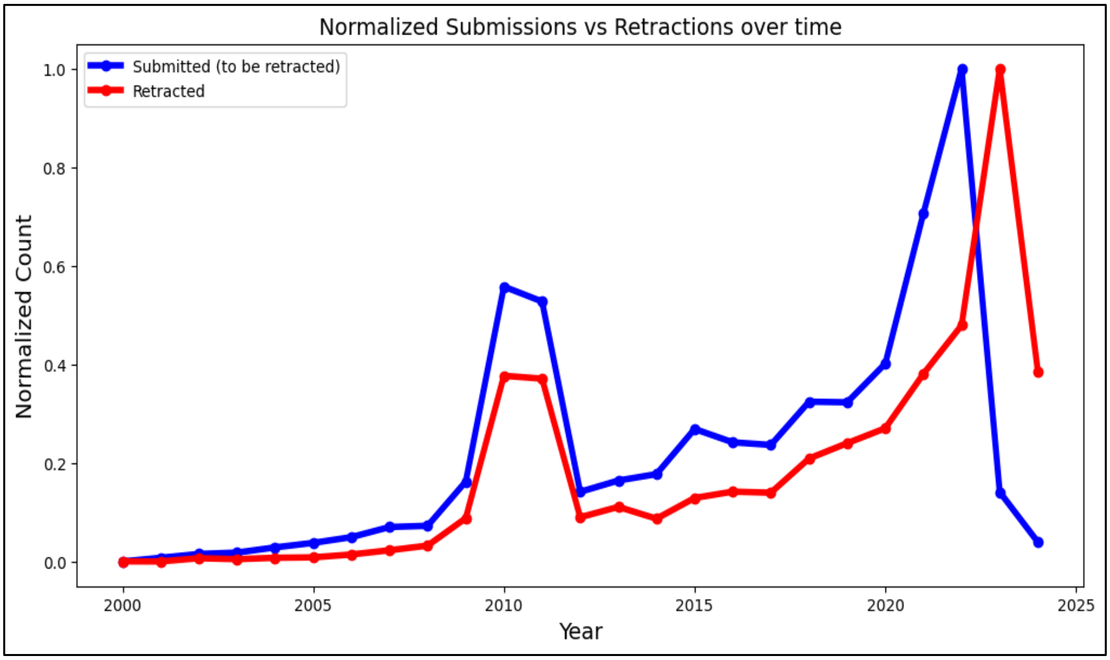

Figure 7: *A comparison of the trends in the number of submissions and retractions from 2000 to 2024*

We conduct a piecewise linear analysis to identify thresholds in the relationship between normalised publications and retractions, aiming to pinpoint when retraction rates accelerate.

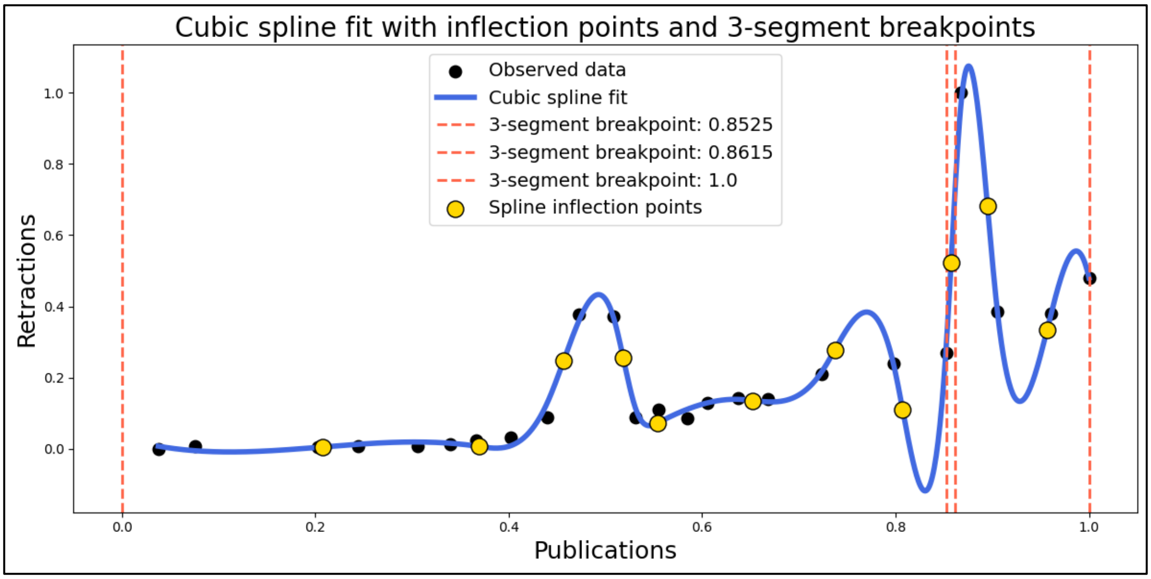

Figure 8: *Comparison of a cubic spline fit to normalised publication and retraction data, overlaid with vertical lines representing the 3-segment model breakpoints*

3.3 AI-induced ethical risks and plagiarism tolerance thresholds

Our model accounts for varying AI adoption across disciplines, institutions, and researchers by introducing individual AI use levels (Equation 12). Using longitudinal data, we calculate a marginal damage ratio, which links retractions to increased scholarly output. We also determine the plagiarism tolerance value by calculating an annual AI misuse impact index, reflecting misconduct relative to academic productivity.

We determine the plagiarism tolerance threshold P by examining the AI misuse impact index, $AI_x(t)$and identifying the first year in which it exceeds a selected inflection point value derived from the data. This inflection point represents a moment where the relationship between retractions and publication volume begins to shift noticeably, signalling growing systemic risk. The corresponding $AI_x(t)$value in that year is then set as P. If no year surpasses the selected inflection point, the highest observed $AI_x(t)$value is used instead. This empirically derived tolerance acts as a cap on AI misuse's damage in subsequent simulations, limiting its degrading effect on overall research quality. Rather than being a theoretical parameter, P is a data-driven indicator of when misconduct begins to accelerate disproportionately relative to academic output, marking a threshold where research integrity is at risk.

To reflect this in the heterogeneous model, we extend the concept of quality to include specific components of research such as originality, citation practices, and rigorous analysis (each represented by subspaces b_j). Each of these components has a baseline AI use z_k, a beneficial range $\Gamma_j(t_{kj})$ and a damage function $d_j(t_{kj})$, as given in Equations (13). The total quality is adjusted by the minimum of the damage and the threshold P, as shown in Equation (15), a formula that operationalises the plagiarism tolerance P as a regulatory boundary. If AI overuse in any dimension b_j causes damage $d_j(t_{kj})$ to exceed P, the system triggers intervention to cap further degradation. When damage is within tolerance (see Equation (16)), no correction is needed. This creates a natural "stop" mechanism on the decline of academic quality, ensuring that research output remains credible even in an AI-augmented environment.

To do this, we calculate the marginal damage from academic retractions relative to the growth in and articles to estimate a plagiarism tolerance threshold P. We then simulate overall academic quality Q_{agg} under two scenarios: one where damage is unchecked, and another where it is capped at P, to illustrate how enforcing a threshold preserves systemic research integrity. We assign equal weights $\gamma = \alpha = \beta = \frac{1}{3}$ to reflect the assumption that , publications, and effort equally add to academic quality in the absence of detailed empirical data on their relative impacts. This approach ensures a balanced contribution while maintaining neutrality in the absence of bias toward any one factor. We set $E = 1$ to normalise effort, assuming it remains constant across years, which is reasonable since academic supervisors with tenure do not have a high turnover rate.

This simplification isolates the effects of publications, leaving focus specifically on how damage (via retractions) impacts quality, with and without the plagiarism tolerance threshold. To get Figure 8, we first compute year-over-year changes in retractions, and published articles, then divide the retraction change by the article change to get marginal damage ratio. Using these ratios, we identify the threshold P where damage exceeds a critical value and simulate academic quality Q_{agg} both with and without capping damage at P, showing the impact on system stability through a comparative plot. In the unconstrained scenario, academic quality Q_{agg} erodes rapidly as retractions accelerate past sustainable levels. However, when we enforce the plagiarism tolerance P, quality is preserved, and the system stabilises, demonstrating that this threshold has not only mathematical elegance but also real-world utility. A negative aggregate research quality score indicates that AI is being overused to the point where its associated damage (such as plagiarism or loss of originality) outweighs its benefits. This suggests that either the plagiarism tolerance P is too low or the damage function $d_j(t_{kj})$ escalates too rapidly, signalling systemic quality degradation and the need for immediate intervention.

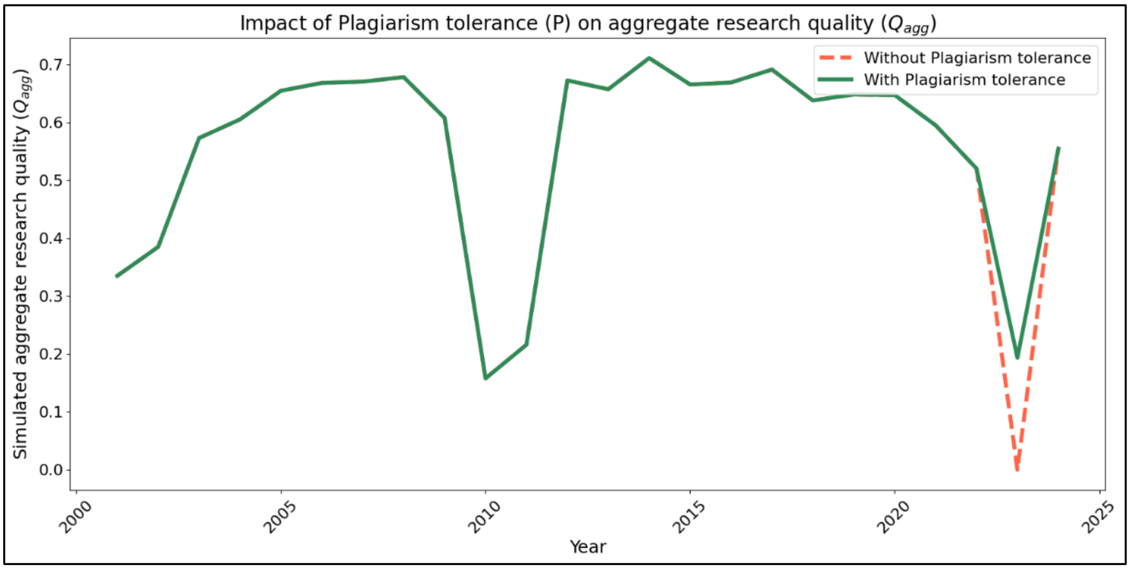

Figure 8: *Simulated academic quality (Q_{agg}) with and without plagiarism tolerance*

The gap between the two curves (green and red) in Figure 8 reflects the potential benefit of enforcing a tolerance threshold in controlling the erosion of research integrity

3.4 Quantifying institutional plagiarism enforcement

To further contextualise P, we map it to familiar academic integrity tools, such as Turnitin which is simply the acceptable Similarity index at an institution S. We begin by considering the institutional plagiarism tolerance as a function of both the institution's stated standards and its actual performance in curbing misconduct. The ratio $\frac{AI_x(t)}{S}$ compares the actual impact of AI misuse $AI_x(t)$, which reflects the institution's acceptable plagiarism score, to the declared tolerance threshold (S), which is measured by the ratio of retractions to publications. This ratio helps assess how well an institution's policies align with enforcement. A ratio greater than 1 suggests effective oversight or low misconduct. A ratio near 1 indicates misconduct approaching the declared threshold, requiring stronger oversight. A ratio less than 1 shows the retraction rate exceeds the declared tolerance, signalling inadequate enforcement. We introduced earlier the multiplier P, which represents the system-wide empirically observed plagiarism tolerance threshold. This threshold is based on the broader trends in AI misuse, where research quality begins to degrade when misconduct rises above a critical level. By multiplying the ratio $\frac{AI_x(t)}{S}$ by P, we scale the institution's tolerance dynamically, aligning it with the system-wide risk of misconduct. This adjustment ensures that institutional thresholds are sensitive to both the institution's own performance and the broader academic environment. Thus, we arrive at the final formula for the adjusted institutional plagiarism tolerance (the ideal one, S_{ideal}):

$$S_{ideal} = P \times \frac{AI_x(t)}{S} \tag{19}$$

This formula provides a dynamic and data-driven approach to adjusting an institution's plagiarism tolerance based on both its own conduct and broader systemic trends in academic misconduct.

The ratio between the actual retraction rate and the institution's acceptable plagiarism tolerance $\frac{AI_x(t)}{S}$, serves as a diagnostic indicator of policy enforcement and research integrity management. This metric captures the gap between declared academic standards and real-world accountability. For example, if an institution claims an acceptable plagiarism threshold of 20% while the actual retraction rate is 0.1% (equivalent to 1 retraction per 1,000 publications), the resulting ratio would be 0.005. Such a value suggests that the institution's retraction rate is only 0.5% of its stated institution's plagiarism score ($S = 20\%$), suggesting either weak enforcement, low actual misconduct, or underreporting. When the ratio exceeds 1, it signals that misconduct is rising beyond what the institution claims to tolerate, indicating a critical gap between policy and practice that demands immediate attention.

3.5 A probabilistic integrity decay framework

Our model extends to incorporate temporal dynamics that capture how research quality degrades progressively due to sustained AI misuse. For each research quality dimension $b_1, b_2, \dots, b_j$, deviations from the ideal AI baseline z_{kj}(quantified as $t_{kj} = s_{kj} - z_{kj}$) introduce incremental damage. These deviations, as defined in Equation (19), manifest as erosion in originality, increased plagiarism, or weakened analytical depth. The cumulative damage to the dimension b_j over t years,

$$D_k(t) = \sum_{t_j \in T} d_j\left(t_{kj}\right) < \varphi_j, \tag{20}$$

triggers a need for intervention in that dimension if it surpasses the threshold φ_k. At the system level, total accumulated damage

$$C_t = \sum_{k=1}^{K} D_k(t) \tag{21}$$

is measured against a plagiarism tolerance threshold P. System failure occurs when $C_t > P$, with the time to intervention marking the earliest point at which corrective actions (consider as examples policy reforms or audits) become necessary. This framework dynamically links AI misuse to quality decay, enabling real-time monitoring of integrity breaches. For damage accumulation, we assume each article's contribution

$$D_i - D_{i-1} = S_i \psi(D_{i-1}), \tag{23}$$

where S_i is the i-th article's vulnerability and $\psi(D_{i-1})$ models ethics decay from prior damage. With $\psi(D) = \omega D$ (assume linear decay for simplicity), solving $\sum_{i=1}^{k} S_i$ yields:

$$\ln\left(\frac{D_k}{D_0}\right) \sim \mathcal{N}(\mu_A, \sigma_A^2), \tag{24}$$

revealing log-normally distributed damage. This shows minor, initially imperceptible deviations compound exponentially, culminating into abrupt ethics failure once $C_t > P$. Consequently, the time to intervention becomes a probabilistic metric, emphasising that integrity breaches often emerge nonlinearly, necessitating proactive monitoring long before thresholds are visibly exceeded.

The risk of failure (when $C_t > P$) can now be probabilistically modelled using the cumulative damage distribution:

$$\text{Prob}(C_t > P) = 1 - \Phi\left(\frac{\ln\left(\frac{P}{C_0}\right) - \mu_A}{\sigma_A}\right). \qquad (25)$$

This equation allows us to quantify the probability of exceeding the threshold based on the AI adoption behaviour. In this expression, $\Phi(\cdot)$ represents the cumulative distribution function (CDF) of the standard normal distribution. The term C_0 stands for the initial level of damage, which we can normalize to 1 for simplicity. The parameters μ_A and σ_A capture the mean and dispersion of AI usage and misuse across researchers, as specified in the AI adoption model. This formulation allows us to probabilistically estimate the risk of breaching the system-wide tolerance threshold based on observed trends in AI misuse and damage accumulation. Hence, we modify Equation (19) to include the probability of exceeding the critical plagiarism tolerance in Equation (25) and get:

$$S_{ideal} = P \times \text{Prob}(C_t > P) \times \frac{AI_x(t)}{S}. \qquad (26)$$

This adjustment makes S_{ideal} responsive to the probabilistic risk of exceeding the critical plagiarism threshold, scaling tolerance based on both AI adoption and integrity risk. It ensures dynamic regulation, tightening controls as $\text{Prob}(C_t > P)$ rises.

4. Discussion of Findings

Integrating AI in postgraduate supervision boosts research efficiency but introduces costs, supervision time, and academic integrity risks. The optimisation framework (Equation 6) balances AI use, resources, and human effort to maximise quality while minimising costs. It guides decisions on when to use AI or rely on mentorship, ensuring that AI enhances quality without compromising ethics. This framework supports responsible AI integration, safeguarding both academic standards and mentorship.

The plagiarism tolerance threshold $P = 0.797$, with the threshold year being 2009-10, serves as a data-driven intervention point, marking where retractions begin to rise disproportionately relative to scholarly output. This value, empirically derived from the relationship between academic retractions and output, highlights a critical inflection point where misconduct threatens systemic integrity. Rising retractions, especially when reinforced by lagged effects, suggest the need to lower P to prevent broader ethical failures. A predictive model using lagged retraction data can forecast when this threshold be exceeded, signalling the need for corrective action. By integrating this threshold into governance practices, institutions can proactively manage academic risk, maintaining research credibility in the face of evolving AI-driven challenges.

The model's accommodation of heterogeneous AI adoption across disciplines (for the same institution, different students behaviour) represents another important advancement. By incorporating individual-specific AI use levels (as expressed in Equation (12)), the framework recognizes that AI's effects are context-dependent. This allows for more tailored policy interventions rather than relying on uniform, oversimplified assumptions. Complementing this is the dynamic damage capping mechanism, wherein P is enforced as a regulatory boundary as in Equation (16), functioning as a circuit breaker that prevents unchecked degradation of research quality. This feature, illustrated in Figure 3, demonstrates how capping damage at the empirically derived threshold stabilises aggregate research quality (Q_{agg}), offering institutions a pragmatic safeguard.

The marginal damage ratio quantifies the cost of AI misuse per unit of academic productivity, linking retractions to publication growth. This framework incorporates probabilistic risk assessment, modelling integrity breaches as log-normally distributed events via Equations (24) and (25). Due to delays in detection and retraction processes, current retraction rates may underestimate misconduct, particularly in fields adopting AI. This emphasises the importance of the plagiarism tolerance threshold P as an early warning indicator of rising systemic risks before retractions accumulate.

In this integrated model, the log-normal distribution emerges as the natural fit for capturing the accumulated damage due to the heterogeneity in AI adoption, where the system's overall quality degradation reflects the skewed distribution of AI reliance. This way, both individual and collective damage from AI use are modelled, with interventions triggered once cumulative damage C_t surpasses a threshold P, marking system failure. This log-normal behaviour of cumulative damage implies that while most research systems may remain below the plagiarism tolerance threshold P for an extended period, a smaller subset (due to compounding misuse and prior degradation) can experience a rapid, disproportionate escalation in damage. As a result, the system-wide risk of crossing the threshold P is not gradual but can manifest suddenly and unexpectedly. This underscores the importance of early monitoring and intervention, since once the cumulative damage C_t enters the long tail of the log-normal distribution, the probability of surpassing P increases non-linearly, making recovery more difficult and requiring more severe corrective action.

Despite these strengths, the model also presents critical limitations and challenges. One issue lies in the assumption of constant academic effort ($E = 1$), which normalises researcher input and ignores changing trends such as productivity inflation, burnout, or shifts in time allocation driven by AI tools. A more realistic approach would model effort as a function of AI adoption, capturing possible declines in human scholarly contribution. Another limitation is the uniform weighting of quality components ($\gamma = \alpha = \beta = \frac{1}{3}$) in the simulations, which, while practical, may not empirically reflect the relative importance of publications, citations, and effort across different fields. Field-specific weighting schemes, informed by expert surveys or citation network analyses, could address this imbalance.

The model's reliance on retraction data as a proxy for misconduct also poses challenges, as retractions are noisy indicators and many cases remain unreported, particularly in institutions with weak oversight. To improve robustness, alternative metrics such as post-publication peer review flags and preprint controversies can be incorporated in future work. Furthermore, the assumption of linear ethics decay ($\psi(D) = \omega D$) likely underestimates the risk of cascading failures, where trust in a journal or institution collapses suddenly. Nonlinear decay functions, such as sigmoidal or piecewise models, could better capture these tipping points. Finally, the conceptual subspaces representing originality, citation rigour, and other dimensions (b_j) lack operational definitions. These could be grounded in existing bibliometric indices, such as the OpenCitations Index of Crossref open DOI-to-DOI citations (COCI) for citation rigour or textual reuse detection for originality, to ensure empirical clarity.

The framework offers valuable practical implications and directions for refinement. One application is institutional benchmarking, where the ratio $\frac{AI_x(t)}{S}$ could serve as a public accountability metric, revealing gaps between declared policy tolerance and actual retraction rates. Institutions with disproportionately high ratios might undergo audits to align policy with practice. Early warning systems could also be developed by integrating the model with institutional data pipelines, such as plagiarism detection software and internal review reports, to flag departments where the probability of crossing the plagiarism tolerance threshold (P) exceeds safe levels. A dashboard visualising trends in academic damage metrics ($D_k(t)$) would help prioritise interventions.

5. Conclusions and Recommendations

This study explored the ethical implications and academic integrity risks associated with AI-mediated postgraduate supervision. Through a combination of theoretical modelling and empirical data analysis, several notable patterns emerged. While AI has the potential to greatly enhance research efficiency, accessibility, and analytical depth, the findings show that excessive or unguided use introduces significant risks. These include reduced originality in student work, heightened retraction vulnerability, and the erosion of meaningful mentorship. The study further demonstrates that the relationship between AI integration and research quality is non-linear, governed by diminishing marginal returns, reinforcing the need for deliberate and balanced adoption.

5.1 Institutional recommendations

To safeguard the integrity of postgraduate supervision in the era of AI, institutions must adopt a coherent and ethically grounded approach. This begins with the development of robust ethical AI frameworks that provide discipline-sensitive guidelines for appropriate AI use in academic work. Such frameworks should be reviewed regularly to reflect evolving technological and disciplinary realities (Papagiannidis, Mikalef, & Conboy, 2025). Alongside this, institutions should implement dynamic academic integrity systems capable of monitoring plagiarism tolerance thresholds and tracking key indicators of misconduct (including retraction rates) so that early intervention becomes both possible and consistent.

The study also highlights the importance of structured AI literacy training for both students and supervisors. This training should emphasise ethical AI use, citation conventions, and responsible academic writing practices to prevent overreliance on generative tools. Equally essential is the preservation of human-centred supervision. AI should complement, rather than replace, the interpersonal guidance, ethical mentorship, and scholarly dialogue that form the basis of quality postgraduate supervision (Koeszegi, 2024). Ensuring equitable access to AI tools and related training is another critical dimension; without such support, disparities in research quality may widen between students with differing levels of technological access.

5.2 Social Implications

The integration of AI into postgraduate supervision also carries important social implications. Equitable access to AI resources is vital to prevent the emergence of new academic divides that may disadvantage students from resource-constrained backgrounds. Moreover, the study underscores that human engagement remains central to fostering academic identity, confidence, and intellectual growth—elements that AI systems cannot replicate. Institutions should therefore prioritise supervisory models that balance technological assistance with sustained interpersonal mentorship, ensuring that social cohesion and inclusivity remain integral to the postgraduate experience. Furthermore, these findings challenge institutions to go beyond simple policy prohibitions and invest in pedagogical shifts that foster critical digital literacy among postgraduate students, moving them from mere functional use to sophisticated ethical reasoning. Ultimately, safeguarding the integrity of the supervisory process is paramount for maintaining public trust in the credentials awarded and ensuring the quality of the next generation of academic and industry research leaders.

5.3 Practical implications

From a practical perspective, institutions must adopt systematic mechanisms to monitor AI's long-term impact on research quality. This includes tracking how AI influences critical thinking, originality, and methodological rigour, and adjusting institutional policies as needed (Wits Centre for Learning, Teaching, and Development, 2024). Continuous evaluation enables institutions to detect emerging risks, refine training programmes, and ensure that AI remains a tool that enhances rather than undermines scholarly standards. A data-driven monitoring approach is, therefore, indispensable for maintaining quality assurance across postgraduate programmes. Specifically, the 'plagiarism tolerance threshold' identified through our econometric modelling offers a critical quantitative metric for policymakers, clearly signalling when reactive measures are no longer adequate and systemic change is immediately required. This necessitates the proactive development of transparent, adaptable institutional guidelines and the implementation of robust technological infrastructure capable of identifying and mitigating integrity risks before they escalate into widespread academic misconduct.

In conclusion, while AI offers powerful opportunities to enhance postgraduate research supervision, its integration must be accompanied by vigilant oversight, equitable support structures, and an unwavering commitment to academic integrity. By adopting ethical frameworks, strengthening mentorship, and instituting robust monitoring mechanisms, institutions can ensure that AI contributes positively to postgraduate education without compromising the values that underpin scholarly excellence.

6. Declarations

Funding: This research did not receive any external funding

Conflicts of Interest: The authors declare no conflicts of interest.

Use of Artificial Intelligence: The current work was supported with the assistance of artificial intelligence technologies (Grammarly) to assist with refining language for clarity, as confirmed by the authors.

References

Acosta-Enriquez, B. G., Arbulu Ballesteros, M., Vilcapoma Pérez, C. R., Huamaní Jordan, O., Martin Vergara, J.A., Martel Acosta, R., Arbulu Perez Vargas, C. G., & Arbulú Castillo, J. C. (2025). AI in academia: How do social influence, self-efficacy, and integrity influence researchers' use of AI models? *Social Sciences & Humanities Open*, 11, https://doi.org/10.1016/j.ssaho.2025.101274

Ahmad, S. F., Han, H., Alam, M. M., Rehmat, M. K., Arraño-Muñoz, M., & Ariza-Montes, A. (2023). Impact of artificial intelligence on human loss in decision making, laziness and safety in education. *Humanities and Social Sciences Communications, 10*, 311. https://doi.org/10.1057/s41599-023-01787-8

Ali, O., Murray, P. A., Momin, M., Dwivedi, Y. K., & Malik, T. (2024). The effects of artificial intelligence applications in educational settings: Challenges and strategies. *Technological Forecasting and Social Change*, 199, 123076. https://doi.org/10.1016/j.techfore.2023.123076

Al-Jahwari, M., & Yousif, M. J. (2025). The impact of AI tools on education: ChatGPT in focus. *Artificial Intelligence & Robotics Development Journal, 4*(4), 314–336. https://doi.org/10.52098/airdj.20244430

Ali, Z. (2020). Artificial intelligence (AI): A review of its uses in language teaching and learning. *IOP Conference Series: Materials Science and Engineering, 769*(1), 12043. https://doi.org/10.1088/1757-899X/769/1/012043

Altmäe, S., Sola-Leyva, A., & Salumets, A. (2023). Artificial intelligence in scientific writing: A friend or a foe? *Reproductive BioMedicine Online, 47*(1), 3-9. https://doi.org/10.1016/j.rbmo.2023.04.009

Balk, B. M. (2024). Why is the Cobb-Douglas production function so popular? *Evolutionary and Institutional Economics Review, 21*(1), 1–20. https://doi.org/10.1007/s40844-024-00279-x

Bearman, M., Ryan, J., & Ajjawi, R. (2023). Discourses of artificial intelligence in higher education: A critical literature review. *Higher Education, 86*(2), 369–385. https://doi.org/10.1007/s10734-022-00937-2

Cobb, C. W., & Douglas, P. H. (1928). A theory of production. *The American Economic Review, 18*(1), 139–165.

Crawford, J., Allen, K. A., Pani, B., & Cowling, M. (2024). When artificial intelligence substitutes humans in higher education: the cost of loneliness, student success, and retention. *Studies in Higher Education*, *49*(5), 883–897. https://doi.org/10.1080/03075079.2024.2326956

Kamalov, F., Santandreu Calonge, D., & Gurrib, I. (2023). New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution. *Sustainability*, 15(16), 12451. https://doi.org/10.3390/su151612451

Koeszegi, S. T. (2024). AI @ work: Human empowerment or disempowerment?. In H. Werthner et al. (Eds.), *Introduction to digital humanism* (pp. 175–196). Springer. https://doi.org/10.1007/978-3-031-45304-5_12

Malik, A. R., Pratiwi, Y., Andajani, K., Numertayasa, I. W., Suharti, S., Darwis, A., & Marzuki. (2023). Exploring artificial intelligence in academic essay: Higher education student's perspective. *International Journal of Educational Research Open, 5*, 100296. https://doi.org/10.1016/j.ijedro.2023.100296

Mauti, J. M. (2025). Ethical implications of artificial intelligence in university education. *East African Journal of Education Studies, 8*(1), 159-167. https://doi.org/10.37284/eajes.8.1.2583

Nguyen, M.-H., & Vuong, Q.-H. (2024). Artificial intelligence and retracted science. *AI & Society*. https://doi.org/10.1007/s00146-024-02090-z

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems, 34*(2), 101885. https://doi.org/10.1016/j.jsis.2024.101885

Wits Centre for Learning, Teaching, and Development. (2024). *Guidelines for GAI use in learning, teaching, and research*. University of the Witwatersrand. https://www.wits.ac.za/media/wits-university/learning-and-teaching/cltd/documents/Wits-CLTD-Guidelines-for-GAI-use-in-Learning-Teaching-Research-Dec2024.pdf

Yang, L., Liu, S., Tsoka, S., & Papageorgiou, L. G. (2016). Mathematical programming for piecewise linear regression analysis. *Expert Systems with Applications, 44*, 156–167. https://doi.org/10.1016/j.eswa.2015.08.034

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 1-27. https://doi.org/10.1186/s41239-019-0171-0

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments, 11*, 28. https://doi.org/10.1186/s40561-024-00316-7

Disclaimer: The views, perspectives, information, and data contained within all publications are exclusively those of the respective author(s) and contributor(s) and do not represent or reflect the positions of ERRCD Forum and/or its editor(s). ERRCD Forum and its editor(s) expressly disclaim responsibility for any damages to persons or property arising from any ideas, methods, instructions, or products referenced in the content.