Asian Journal of Assessment in Teaching and Learning
*Vol 15, Issue 2, 2025 (21-42) /eISSN **2821-2916***

Is AI only to Blame? Assessing Teachers' Perceived Challenges in AI Detectability

Ahnaf Chowdhury Niloy[1*], Tazreen Huda[2]
[1]Department of Learning Sciences, Georgia State University, Georgia, United States of America
[2]Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh

*Corresponding author: aniloy1@gsu.edu

Received: 20 November 2025; Revised: 11 December 2025;
Accepted: 15 December 20xx; Published: 19 December 2025

To link to this article: https://doi.org/10.37134/ajatel.vol15.2.2.2025

Abstract

Existing research on teachers' ability to detect AI-generated texts has predominantly emphasized technical shortcomings, overlooking the behavioral and environmental factors that shape detection accuracy. As generative AI becomes embedded in education, understanding how institutional and personal contexts influence teachers' detection performance is crucial for ensuring academic integrity. This study aims to identify and analyze the key internal (behavioral) and external (institutional and contextual) factors affecting teachers' ability to distinguish AI-generated from human-written texts. It further seeks to examine how these factors interact across global regions to develop a more comprehensive framework for understanding detection challenges. An exploratory sequential mixed-method design was employed. The first phase involved 15 key informant interviews with educators from three continents to identify salient determinants of detection capability. Insights from this phase guided the development of a survey administered to 317 teachers across four continents. Data was analyzed using Structural Equation Modeling (SEM) to test interrelationships among identified factors. Findings revealed that rigid university policies significantly hinder teachers' detection ability, especially in Europe, both directly and indirectly through time limitations and content indistinguishability. By integrating behavioral and contextual dimensions, the study advances beyond technically centered perspectives and proposes a global framework for understanding AI detectability. The results have theoretical and practical implications for policymakers and AI developers. Limitations include reliance on perception-based data and lack of African representation, warranting broader, experimental validation in future research.

Keywords: *AI Text Detection, Teacher Perception, Academic Integrity, Mixed-Method Research, Educational Policy*

INTRODUCTION

Teachers stand at the forefront of education, serving as the primary agents through whom learning is delivered, interpreted, and sustained. Despite the evolution of influential educational philosophies like constructivism, flipped learning, and instructional scaffolding, none have replaced the central role of teachers in guiding learners, contextualizing knowledge, and shaping meaningful educational experiences (Özkan, 2022; Li, 2023). Research consistently shows that while educational philosophies and technologies have transformed classroom practices, they have not diminished the teacher's centrality (Paniagua & Istance, 2018; Harris & Jones, 2019; Maba et al., 2023). As the closest evaluators of student work, teachers also function as the first line of defense in protecting academic integrity, using their professional judgment, familiarity with students' abilities, and disciplinary expertise to assess the authenticity of learning evidence (Gottardello & Karabag, 2020). However, the rapid emergence of Generative AI has complicated this responsibility, introducing new forms of deception that challenge traditional cues used in evaluating student writing (Bozkurt, 2024; Kofinas et al., 2025).

Although the idea of Artificial Intelligence (AI) dates back to mid-20th-century work by scholars such as McCarthy and Turing (Cristianini, 2016; Niloy et al., 2024a), its impact on education has become urgent only with the emergence of modern Generative AI (GAI). Large Language Models (LLMs) including ChatGPT, Gemini, Copilot, Claude, Perplexity, and LLaMA, now produce fluent academic writing that closely resembles student work, posing immediate challenges for assessment and academic integrity. These models rely on neural networks and reinforcement learning rather than traditional rule-based algorithms (Duan et al., 2019; Guan et al., 2019; Kushwaha & Kar, 2021), enabling them to generate essays, reflections, and explanations that can be easily integrated into coursework. The release of ChatGPT 3.5 as a freely accessible tool in 2022 (OpenAI, 2022; Lambert & Stevens, 2023) dramatically shifted how students approach written tasks. Its rapid adoption across educational settings (Sier, 2022) and the subsequent introduction of more advanced versions, including GPT-5 and GPT-4o (Edwards, 2023; Rogers, 2024), have made GAI a constant presence in classroom assignments. Competing free tools such as Microsoft Copilot and Google Gemini further diversified students' options (Grant, 2023; Bocian, 2024; Field, 2024; Sadka, 2024). As a result, teachers must evaluate student work under growing uncertainty as a polished submission may raise questions about true authorship.

While there have been approximately 7.6 million studies published on AI and Education, according to Google Scholar database as of November 2024, only a limited studies have tried to focus on AI's impact on Teachers' assessment abilities (Fleckenstein et al., 2024; Lameras & Arnab, 2021; Simuţ et al., 2024; Singh & Ram, 2024; Swiecki et al., 2022). A majority of these studies have only tried to determine the factors through qualitative approaches (Celik et al., 2022; Chaka, 2024; De Wilde, 2024; Kumar & Mindzak, 2024; Weber-Wulff et al., 2023). The studies that have tried to claim that AI Chatbots might deceive teachers' detectability, are heavily dependent upon the data collected from students' perceptions (Ibrahim et al., 2023; Murray & Tersigni, 2024). While some studies have utilized experimental approaches to justify that AI Chatbots do deceive teachers (Chaka, 2023; Walters, 2023), but the studies could not assign specific weights to the causes and were forced to consider equal weights. Such uniform approach often challenges the findings' authenticity and reliability in the real world as it is less likely that all the causes could be equally influential. Moreover, the existing studies that followed structural equation modelling techniques have only tried to draw a link between the AI Chatbots characteristics and its impact on influencing a user's behavioral intentions (Niloy et al., 2024a), and to some extent, its potential to impact academic integrity (Niloy et al., 2024b). But why and how teachers' reviewing credibility is being affected is yet to be modelized statistically. Furthermore, the studies only hypothesize on the fact that certain abilities of AI Chatbots might be the cause of deception but do not quantify the assumption to determine the strength of the causes. As such, the suggested factors by existing authors are mostly scattered opinions lacking a statistical assessment of the viewpoints from the actual victims – the teachers themselves.

Interestingly, a mere 0.2% of studies on AI and Education have directly or indirectly discussed the factors that might affect detectability of AI Texts. As quantitative proof of the identified internal and external factors by prior authors remains unexplored, this study focuses on mitigating this gap by developing a statistical model to assess the valid constructs that could possibly affect the cognitive abilities of a teacher in terms of detecting Generative AI (GAI) produced texts. This study also measures the effect sizes of the causal factors to provide a more robust understanding of the factors' relationship to the teachers' undetectability, that current studies have not been able to determine. Furthermore, intermediatory relationships of the factors in the behavioral process of the teacher are also explored in this study to get a much broader understanding of how and why the teacher struggles to conduct an effective evaluation of a script and how the factors influence each other and create a multiplier effect, especially in the AI Chatbot era. Utilizing data collected from teachers around the globe and across continents, this study provides a robust understanding of the global phenomenon. Based on the gap that exists in current literature, this study formulates the following research questions for investigation into this domain:

RQ1: What factors do educators perceive as contributing to their difficulty in detecting AI-generated texts?
RQ2: Do these perceived struggles differ across geographical regions?

LITERATURE REVIEW

The recent rise of GAI tools has fundamentally altered written assessment by enabling students to generate high-quality text rapidly. While AI's development dates back decades (Cristianini, 2016), the educationally significant shift is the emergence of LLMs trained on massive datasets (Abdullah et al., 2022; Finnie-Ansley et al., 2022; Leah & Meroño-Peñuela, 2022). These tools can create content that aligns with academic conventions (Chan, 2023; Mathew, 2023; C. Zhou et al., 2023) and respond in natural language (Hughes, 2023; Cascella et al., 2023), making them powerful but also potentially deceptive within assessment contexts. Their widespread adoption by students has disrupted traditional assumptions about authorship and writing performance (Dwivedi et al., 2023; Hobert & von Wolff, 2019; Pérez-Marín, 2021; Wollny et al., 2021; AlAfnan et al., 2023; Coleman, 2023).

Digital deception, defined as the use of technology to mislead or obscure meaning (Hancock, 2007), has become increasingly relevant to academic integrity. Classic communication theories such as Media Richness Theory (Daft & Lengel, 1986) argued that richer channels facilitate deception; however, modern GAI tools now provide instantaneous, personalized, natural-language interactions that invert this assumption (Atlas, 2023; Fui-Hoon Nah et al., 2023; B. Williamson, 2024). Similarly, earlier research suggested that deceptive messages were longer or more linguistically distinctive (L. Zhou, Burgoon, Nunamaker et al., 2004; L. Zhou, Burgoon, Twitchell et al., 2004; L. Zhou & Zhang, 2004), but AI's ability to customize length, tone, and structure undermines the reliability of such cues. Interpersonal Deception Theory (Carlson et al., 2004) also suggests that trust between sender and receiver affects detection; in classrooms, this trust may cause instructors to over-interpret polished writing as student-generated.

Empirical studies paint a concerning picture. Gao et al. (2023) found that human reviewers incorrectly classified 32% of AI-generated abstracts as human-authored and mislabeled 14% of genuine human abstracts as AI-generated. Other evaluations show that teachers remain vulnerable to deceptive affordances of AI regardless of training or experience (Farazouli et al., 2024; Chaka, 2023; Walters, 2023). Meanwhile, AI detection tools, such as: GPTZero, Copyleaks, Writer, Crossplag, and Turnitin, exhibit inconsistent accuracy and nontrivial rates of false positives and negatives (Elkhatat et al., 2023; Andrews, 2023; Aw, 2024; Ivanov, 2023; Barton, 2024). This inconsistency creates risks for both failing to detect misconduct and wrongly penalizing genuine student work.

Technology alone, however, cannot explain teachers' challenges. Research indicates that workload, insufficient technological training, and institutional ambiguity contribute to educators' difficulty in evaluating AI-assisted writing (Niloy et al., 2024a; Basu, 2023). Additional scholarship notes that AI's realistic content generation and adaptive personalization amplify deception (Schmitt & Flechais, 2024; S. M. Williamson & Prybutok, 2024; Black, 2024; Natale, 2023). Yet these studies tend to describe the problem rather than model its underlying causes. Despite isolated contributions, literature lacks a comprehensive examination of how technological, situational, and behavioral factors jointly shape educators' perceived difficulty in AI detectability. This gap is especially pronounced at global scale, as cultural and pedagogical norms may influence both writing expectations and instructors' interpretations of student work (Biener & Waeber, 2024). By focusing explicitly on classroom assessment and modeling the factors that contribute to teachers' perceived struggle, the present study addresses a critical need in the evolving landscape of AI-mediated learning.

METHODOLOGY

1. Research Design

This study undertakes a mixed-method approach and an exploratory sequential research design to investigate the causes behind teachers' inability in detecting AI-generated texts. For the first phase, exploratory Key Informant Interviews (KIIs) with educators and learning sciences researchers identify key factors contributing to detection difficulties. Insights from these interviews lay the foundation for identifying key factors affecting the detection. Later, a Likert scale survey targeting a broader sample of teachers, is conducted to quantify the relevance of each identified factor. A single cross-sectional survey design is used to facilitate data collection, followed by Structural Equation Modelling (SEM) to analyze the relationships among factors. This two-phase approach ensures both depth and rigor, combining qualitative exploration with quantitative validation, which is the fundamental approach of exploratory sequential research designs.

2. Sampling and Data Collection

For the qualitative identification of factors, the study follows a purposive sampling technique. The study conducts thorough Key Informant Interviews (KIIs) with 15 respondents to identify potential factors of digital deception in text-based communication. Saunders & Townsend (2016) opined that, for qualitative interviews a sample size between 15-60 is adequate; thus, the sample size of this study is well-suited for extracting results. 3 Open Ended questions are asked to the respondents, and the interviews are transcribed in a Microsoft Word document as meeting minutes for analysis. The interviews are conducted via virtual calls, in-person meetings, and Microsoft Forms, whichever is applicable and feasible for the respondent. The respondents are coded as Rx, where x represents a number, from Respondent R1 to Respondent R12, where Respondent R4, R8, R9, R10, R11, and R14 form the Asia Group, Respondent R5, R6, R7, R12 and R13 form the Europe Group, and Respondent R1, R2, R3, and R15 form the Australia Group. A detailed breakdown of the respondents is provided in Table 1.

Table 1 Key informant interview (KII) respondent details and coding

Respondent Group	**Respondent Code**	**Designation & Affiliation**
Asia Group	R4	Professor, Islamic University of Technology, Bangladesh
	R8	Professor, Indian Institute of Technology (IIT) - Delhi, India
	R9	Professor, Nanyang Technological University, Singapore
	R10	Assistant Professor, Tsinghua University, China
	R11	Professor, Kyoto University, Japan
	R14	Assistant Professor, University of Malaya, Malaysia
Europe Group	R5	Staff Scientist, University of Eastern Finland, Finland
	R6	Professor, University of Gothenburg, Sweden
	R7	Professor, Durham University, UK
	R12	Assistant Professor, University of Oslo, Norway
	R13	Assistant Professor, University of Nottingham, UK
Australia Group	R1	Assistant Professor, University of Wollongong, Australia
	R2	Assistant Professor, The University of Adelaide, Australia
	R3	Professor, Monash University, Australia
	R15	Assistant Professor, University of Queensland, Australia

After the factors have been initially identified, following a stratified sampling technique, a survey is conducted consisting of 28 closed ended Likert scale questions for the 3 identified factors. The questionnaire also includes 6 questions for the demographic analysis. Here, the continents serve as the stratums. Administrative members and faculty members of the universities in the respective continents are contacted, and the purpose is communicated beforehand regarding sharing and filling up the questionnaire by the faculty members only. For ensuring that the sample is valid, respondents who did not list them as a "Teacher" were omitted from the analysis.

The purpose of this survey is to collect data for the quantitative confirmation of the identified factors. As the targeted survey population is the university faculties, the study analyzes a valid sample of 317 samples out of 394 respondents across 4 continents – Asia, Europe, Africa, and Australia. For conducting statistical analysis, it is recommended that the sample be minimum 200. As the valid sample used in this study is beyond 200, the study opines the sample to be adequate. The study involves samples from several countries across multiple continents in both qualitative (3 continents) and quantitative (4 continents) phases to get a broader understanding of the global challenges faced by educators.

3. Tools and Instruments

The study utilizes Microsoft Teams, Google Meet, and Zoom for conducting virtual KIIs. Microsoft Forms is used for both Qualitative and Quantitative data collection. Microsoft Word is used for preparing the transcribed Meeting Minutes. KIIs that have been conducted in a language other than English has been transcribed and reviewed by 2 independent Faculty members of English Language and Linguistics to ensure that language translation is appropriate. Atlas.ti version 9 is used for the narratives analysis for the identification of possible factors. Microsoft Excel 365, IBM SPSS 27, and IBM AMOS 18 are used for conducting quantitative analysis including Exploratory Factor Analysis, Confirmatory Factor Analysis, and Path Analysis.

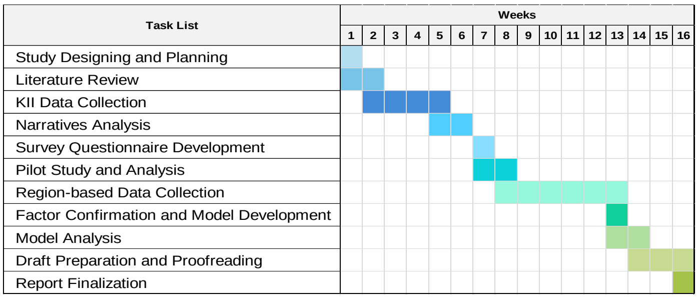
4. Procedure

The 15 KII respondents provided their responses to the 3 open-ended questions during the live meeting, taken in either physical or virtual form. Similarly, the respondents who opt for convenience were provided with a Microsoft Form containing the same questions where the respondents can answer in a written format. The responses were transcribed to Microsoft word documents and were analyzed based on keywords to identify key factors. The 15 respondents were clustered to get a deeper understanding regarding the regional differences in opinions. The 3 clusters were based on geographic location of the respondents – 6 from Asia, 4 from Australia, and 5 from the European region. As such, the clusters were named as – Asia Group (AG), Europe Group (EG), and Australia Group (AuG). The narratives of the responses on the 3 open ended questions were carefully analyzed to derive the 6 common codes which later formed the 3 identified core factors of this study. The factors were further validated quantitatively through a 5-point Likert Scale survey, following the Confirmatory Factor Analysis. As the factors are identified and confirmed, the conceptual model was developed and the relationship amongst the factors were analyzed through a Path Model analysis, ideally followed in Structural Equation Modelling techniques, using IBM AMOS. Total effect sizes were also calculated to understand the strength of the relationship amongst the factors.

5. Timeline

The study is conducted, including the submission of the final report, within 16 weeks. A detailed breakdown of the tasks is shown in

Figure ***1*** using a Gantt Chart.



Task List	Weeks															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Study Designing and Planning	■															
Literature Review	■	■														
KII Data Collection		■	■	■	■											
Narratives Analysis					■	■										
Survey Questionnaire Development							■									
Pilot Study and Analysis							■	■								
Region-based Data Collection								■	■	■	■	■	■			
Factor Confirmation and Model Development													■			
Model Analysis													■	■		
Draft Preparation and Proofreading														■	■	■
Report Finalization																■

Figure 1 Gantt chart of timeline

6. Ethical Consideration

a. Informed Consent

All the participants in the study provided informed consent in written format before participating in the study and the purpose of the study was not communicated prior to the participation in order to ensure that any bias does not influence the response. After participation, all the participants were informed of the purpose of the study for transparency.

7. Analysis & Findings

a. Factor Identification

The first phase of the analysis involved identifying key factors through a systematic narrative analysis. Six recurring codes emerged from the responses of 15 participants across three groups, which were then consolidated into three overarching themes. Findings showed that respondents from both the Asia group and the Australia group consistently viewed the human-like and customized nature of AI chatbot outputs as the primary reason for their difficulty in distinguishing human-written text from AI-generated text. The Australia group further emphasized limited access to reliable AI detection tools and the perceived inefficiency of existing detection platforms. They also highlighted insufficient training in AI detectability and in the use of detection systems. Representative statements supporting the development of these codes and themes are presented in Table 2.

Table 2 Some reference narratives of defined codes

Code	Respondents	Key Narratives
Access to reliable detection tool	R4, R5, R7, R10, R11	"Many teachers do not have access to advanced tools or the time to verify the originality of every submission. These challenges are further exacerbated by institutional constraints that limit alternative assessment methods." "AI tools are capable of producing highly tailored responses that fit specific assignment criteria, leaving educators with little resource to verify originality." "Institutions must support teachers by providing reliable detection tools and reducing the emphasis on traditional grading." "Assignments completed with AI tools often meet or exceed the expected standards, making it nearly impossible to differentiate them from authentic submissions without advanced detection technologies." "Teachers are rarely provided with adequate training in using detection tools"

continued

Cognitive Miserliness of Teacher	R2, R8	“Many teachers, including myself, lack the training or resources to effectively use AI-detection tools.” “This problem is my limited experience with detection tools and the time constraints of grading large volumes of assignments.”
Customized Response	R1, R7, R10, R15	“The main challenge I face is distinguishing between authentic student work and AI-generated content. With AI producing human-like responses, traditional evaluation methods fall short. When students prompt chatbots to generate assignment responses tailored to their specific topics, the resulting work often appears indistinguishable from genuine effort. This puts teachers in a difficult position, as we lack reliable means to assess the originality of such submissions.” “When students use AI tools to produce unique, creative content, it becomes nearly impossible to discern their actual level of understanding or effort, especially in subjective assessments.” “This tailored generation of responses poses a direct challenge to educators.” ““When students use chatbots to produce work that appears intellectually robust.”
Detection Tool Inefficiency	R3, R4, R6, R14	“AI-detection tools are not always accurate and manually assessing every submission is impractical given the lack of time.” “Students can easily use chatbots to produce work that closely aligns with their own linguistic style, making it indistinguishable from authentic writing.” “A significant challenge lies in the limitations of AI-detection tools. These systems are far from perfect and often flag legitimate content as AI-generated or miss actual AI-written material.” “Without robust tools to verify originality, educators are at a significant disadvantage.”
Human-Like Response	R2, R11, R13	“The ability to produce polished and context-specific responses poses a significant threat. Students can exploit these capabilities to generate work that aligns closely with their academic requirements, making it nearly impossible for teachers to distinguish between genuine effort and AI-assisted work.” “This mimicry becomes a threat when students use chatbots to produce work that appears authentic.”
Lack of Training	R2, R5, R8, R10, R11	“This problem is compounded by procrastination—both on the part of students submitting last-minute work and teachers delaying their evaluations.” “Evaluating assignments thoroughly requires more effort when AI tools are involved. Procrastination on my part, as well as institutional pressures to grade quickly, often result in missed instances of AI-generated content.” “This problem is my limited experience with detection tools and the time constraints of grading large volumes of assignments.” “Time is a significant factor; with multiple responsibilities, it’s hard to thoroughly review every submission. Additionally, existing university

		policies do not accommodate flexible assessment methods that could mitigate these challenges." "AI tools are becoming more sophisticated, and my familiarity with detection tools is limited."

EG placed greater emphasis on teachers' cognitive miserliness, suggesting that instructors seeking convenience may fail to engage deeply with student work, either due to negligence or limited effort. Although this perspective was acknowledged, only a small number of respondents from the AuG and AG supported this view. Similar to the AG, the AuG primarily attributed detection challenges to the human-like quality and personalized nature of AI-generated text. However, the AuG regarded teachers' cognitive miserliness and the inefficiency of detection tools as the least influential factors in their inability to differentiate between human and AI-authored scripts. Across all groups, six recurrent codes emerged, which were organized into three overarching themes: rigid university policy, lack of time, and content inseparability. The Sankey diagram (Figure 2) illustrates how the six codes map onto these three factors, suggesting potential interrelationships among them. Whether these factors operate as distinct constructs is examined more rigorously in the subsequent quantitative analysis.

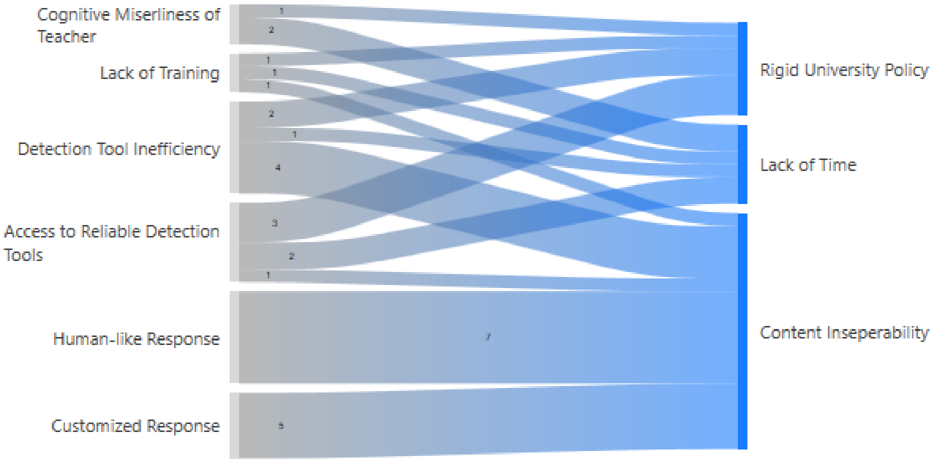


Figure 2 Sankey diagram of codes and factors

Further exploration of the narratives revealed that although 3 of the factors are considered as core factors, the narratives analysis shows that *Content Inseparability* is regarded as the most significant and impactful factor that is affecting the detectability of teachers (See, *Figure 3*). Regardless of the groups, all 3 groups commonly argued that *Content Inseparability* is the most daunting factor (See, Figure 3), giving it more emphasis compared to the other mentioned causes. However, all three groups also identified a *lack of time* and *rigid university policies* as two other major factors affecting their detectability, although the relative emphasis placed on these factors by the groups, as evident from their wording or statements, arguably varied.

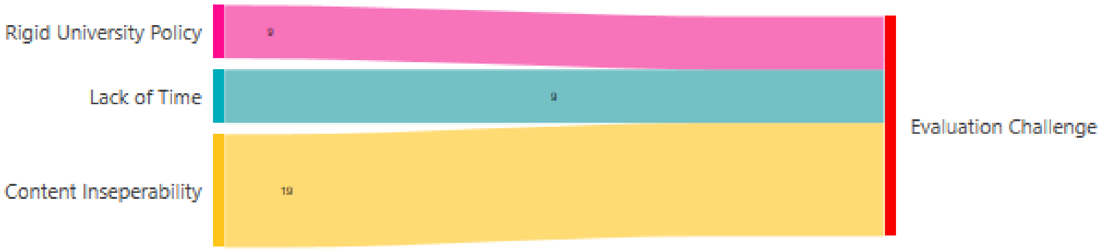


Figure 3 Sankey diagram of identified causes and the effect of evaluation challenge

b. Factor Confirmation

Although the factors have been identified through narratives, and the Sankey Diagrams give a visualized perspective over the relative weights of the factors, the analysis requires a quantitative justification for a more logical and robust understanding of the factors to confirm them as a valid factor and determine their relationships. Hence, the quantitative phase of the study is conducted.

c. Demographic Profile Analysis

In quantitative research, demographic analysis involves systematic examination of participant characteristics to describe the sample, detect patterns, and evaluate the appropriateness and representativeness of the study population. Although not the sole method of validating a sample, it is essential for assessing whether the respondents reflect the population of interest. In this study, demographic analysis indicated that across six continents, the largest proportion of valid respondents were Senior Lecturers or Lecturers. Because the analysis focused exclusively on teachers, 80.5 percent of total responses (n = 317) were included. Additionally, 29.9 percent of teachers held the rank of Assistant Professor or higher. Table 8 presents the continental distribution of the valid sample.

Table 3 Demographic profile of survey respondents

		Frequency	**Percentage**
Age	Below 18	0	0.0
	18-27	31	7.9
	28-37	187	47.5
	38-47	13	3.3
	48-57	108	27.4
	Above 57	55	14.0
Sex	Male	226	57.4
	Female	150	38.1
	Intersex	18	4.6
Profession	Teacher	317	80.5
	Teaching Assistant	9	2.3
	Other	68	17.3
Current Designation	Professor/Professor Emeritus/Distinguished Professor	39	9.9
	Associate Professor/Assistant Professor	79	20.0
	Senior Lecturer/Lecturer	190	48.2
	Adjunct Faculty/Visiting Faculty	9	2.3
	Teaching Assistant	9	2.3
	Other	68	17.3
Active Teaching Experience	0-2 Years	96	24.4
	2-5 Years	103	26.1
	5-10 Years	74	18.8
	10+ Years	121	30.7

d. Exploratory Factor Analysis

The Exploratory Factor Analysis (EFA) was conducted using Principal Component Analysis (PCA) for extraction and Varimax rotation to clarify the underlying factor structure. Applying the eigenvalue greater than one criterion, four factors emerged, explaining 64.406 percent of the total variance. The decision to use Varimax rotation, an orthogonal technique, was intended to maximize the variance of squared loadings within each factor, thereby reducing cross-loadings and enhancing the interpretability of the factor solution. Sampling adequacy was confirmed through the Kaiser-Meyer-Olkin (KMO) measure, which yielded a value of 0.827, indicating strong suitability of the data for factor analysis. In addition, Bartlett's test of sphericity was significant ($p < .001$), demonstrating that the correlation matrix contains sufficient shared variance among items to justify proceeding with factor extraction (see Table 4). The resulting factor solution was stable and theoretically coherent. No items required removal during the EFA, and the Varimax-rotated structure produced a clean, interpretable grouping of items that aligns well with the initial conceptual framework. These findings provide a solid empirical foundation for subsequent validation steps and further interpretation in later stages of the analysis.

Table 4 EFA results

	F1	F2	F3	F4
RUP1	.846			
RUP2	.774			
RUP3	.867			
RUP4	.808			
RUP5	.886			
RUP6	.873			
RUP7	.813			
CI1		.669		
CI2		.718		
CI3		.662		
CI4		.787		
CI5		.782		
CI6		.781		
LT1			.756	
LT2			.871	
LT3			.753	
LT4			.854	
LT5			.826	
LT6			.906	
LT7			.634	
PEC1				.761
PEC2				.794
PEC3				.804
PEC4				.844
PEC5				.810
PEC6				.783
PEC7				.784
PEC8				.794

Note: Rotation Sums of Squared Loadings = 64.406, KMO measure of sampling adequacy = 0.827, Bartlett's test of Sphericity = 0.000

e. Confirmatory Factor Analysis

Confirmatory Factor Analysis (CFA) is a statistical technique used to validate the hypothesized relationships between observed variables and their underlying latent constructs. Unlike Exploratory Factor Analysis (EFA), which identifies potential factor structures, CFA is hypothesis-driven and assesses how well the proposed model fits the observed data. CFA is a crucial analysis to confirm factors whereas EFA is effective for identifying factors.

In this study, CFA was conducted to confirm the four-factor structure identified in the EFA. To improve the model's fit, 12 items were removed, while ensuring that all factors retained at least three items, maintaining their theoretical integrity. The model exhibited good fit across key indices, including comparative fit (Cmin/df = 2.99 < 3; CFI = 0.936 > 0.9) and absolute fit (SRMR = 0.0488 < 0.1; RMSEA = 0.079 < 0.085; GFI = 0.903 > 0.85). All the remaining items' factor loading exceeded 0.65, with most items surpassing a loading value of 0.7 (See, Figure 4).

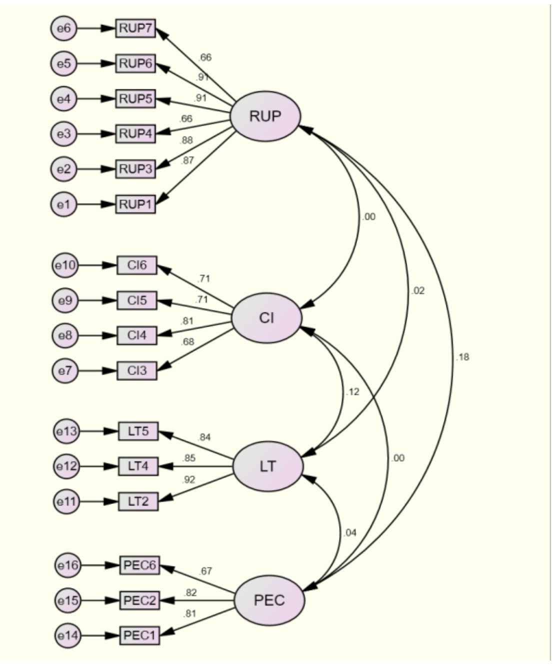


Figure 4 CFA model with results
Note: Cmin/df = 2.99 (P = 0.000), CFI = 0.936, GFI = 0.903, SRMR = 0.0488, RMSEA = 0.079

f. Reliability and Validity

The factor loadings obtained from the Confirmatory Factor Analysis (CFA) were further evaluated for reliability and validity. Composite Reliability (CR), a measure of internal consistency that assesses the shared variance among observed variables within a construct, was calculated for each factor. Unlike Cronbach's Alpha, CR accounts for the varying factor loadings of individual items, providing a more precise reliability estimate. All factors demonstrated adequate reliability, with CR values exceeding the minimum threshold of 0.7, confirming strong internal consistency. Convergent validity was also established, as all constructs achieved Average Variance Extracted (AVE) values above 0.5. This indicates that the items effectively represent their respective latent constructs and that the constructs capture more variance from their items than from measurement error (See, Table 5).

Table 5 Reliability and convergent validity

Factor	**Item**	**Factor Loadings**	**CR**	**AVE**
Rigid University Policies (RUP)	RUP1	0.875	0.926	0.681
	RUP3	0.880		
	RUP4	0.661		
	RUP5	0.915		
	RUP6	0.911		
	RUP7	0.665		
Content Inseparability (CI)	CI3	0.683	0.818	0.529
	CI4	0.808		
	CI5	0.705		
	CI6	0.708		
Lack of Time (LT)	LT2	0.920	0.904	0.759
	LT4	0.852		
	LT5	0.839		
Perceived Evaluation Challenge (PEC)	PEC1	0.812	0.815	0.597
	PEC2	0.823		
	PEC6	0.673		

Discriminant validity was assessed using two widely accepted methods: the Fornell-Larcker Criterion and the Heterotrait-Monotrait (HTMT) ratio test. The Fornell-Larcker Criterion evaluates discriminant validity by comparing the square root of the Average Variance Extracted (AVE) for each construct with its correlations with other constructs. The results confirmed that the square root of the AVE for each factor exceeded its corresponding inter-construct correlation values, indicating adequate discriminant validity (See, Table 6).

Table 6 Discriminant validity using the Fornell-Larcker criterion

	RUP	**CI**	**LT**	**PEC**
RUP	**0.825**			
CI	0.002	**0.727**		
LT	0.029	0.094	**0.871**	
PEC	0.178	0.006	0.020	**0.773**

Note: Diagonal values represent the squared root of AVE and corresponding off-diagonal values are correlation coefficients.

The HTMT ratio test further supported these results, revealing that the HTMT values for all paired constructs were below the acceptable threshold of 0.9 as presented in Table 7. This indicates that the constructs are sufficiently distinct from one another, providing additional evidence of discriminant validity. These results confirm the distinctiveness of the constructs, ensuring the robustness and validity of the measurement model.

Table 7 Discriminant validity using HTMT ratio

	RUP	**CI**	**LT**	**PEC**
RUP	■			
CI	.011	■		
LT	.030	.114	■	
PEC	.205	.008	.023	■

Jointly, EFA and CFA validated that a model involving *perceived evaluation challenges* (PEC) and the 3 factors identified during the qualitative phase – *Rigid University Policies* (RUP), *Content Inseparability* (CI), and *Lack of Time* (LT), can be developed.

g. Path Analysis: The Conceptual Model

A conceptual model serves as a theoretical framework that illustrates how the key constructs in a study are expected to relate to one another. It clarifies the underlying theory, guides the development of measurement instruments, and provides the foundation for empirical testing through techniques such as Structural Equation Modeling (SEM).

In this study, the conceptual model positions RUP as the exogenous construct and PEC as the primary endogenous construct (see Figure 5). Alongside RUP, both CI and LT are hypothesized to exert direct effects on PEC. The model further proposes that LT and CI operate as mediating variables. Insights drawn from narrative analysis and discussions with key informant interview (KII) respondents suggest additional pathways: a potential relationship between RUP and LT, and another between LT and CI. Together, these linkages create a sequential mediating pathway from RUP to PEC, with LT and CI acting as mediators. To examine the perceived challenges faced by teachers in the context of AI integration, the study articulates seven hypotheses derived from this conceptual structure.

H1: Rigid University Policy (RUP) significantly influences Perceived Evaluation Challenge (PEC)
H2: Content Inseparability (CI) significantly influences Perceived Evaluation Challenge (PEC)
H3: Lack of Time (LT) significantly influences Perceived Evaluation Challenge (PEC)
H4: Lack of Time (LT) significantly influences Content Inseparability (CI)
H5: Rigid University Policy (RUP) significantly influences Lack of Time (LT)

H6: Rigid University Policy (RUP) significantly influences Perceived Evaluation Challenge (PEC), when mediated through Lack of Time (LT)
H7: Rigid University Policy (RUP) significantly influences Perceived Evaluation Challenge (PEC), when mediated through Lack of Time (LT) and Content Inseparability (CI)

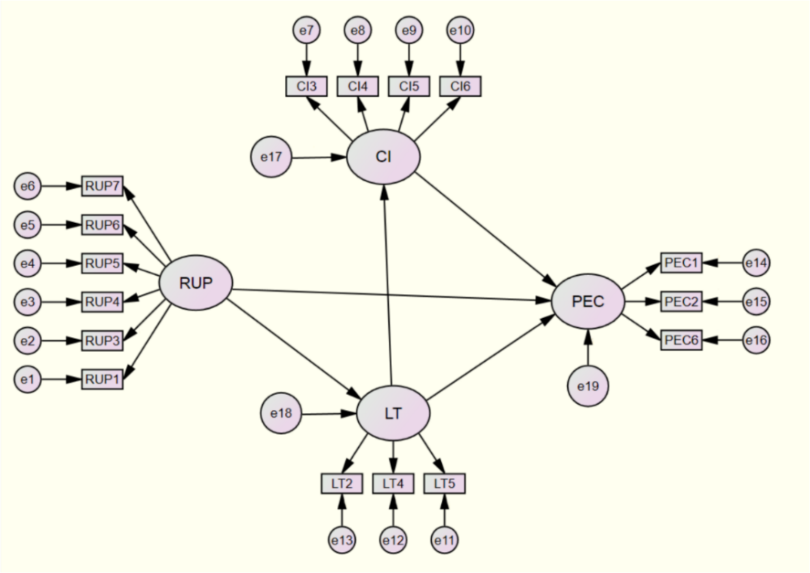


Figure 5 Conceptual Model

h. Model Fit and Path Effect Analysis

A valid sample of 317 respondents was collected from six continents to provide a global perspective. The largest proportion of respondents came from Asia (53.62%), while the smallest proportion was from Africa (8.51%). Europe and Australia accounted for 19.55% and 18.29%, respectively. Due to less than 10% of the total sample was represented by Africa region, the African sample was not considered for continent-specific analysis due to its limited size. However, the African sample was included in the global analysis phase (See, Table 8). This study refers "Global" as the sample size of 317 respondents that includes the sample of all the valid respondents across the 4 continents.

Table 8 Region-wise valid sample

Region	Continent	Frequency	Percentage
Asia	Asia	170	53.62
Australia and Oceania	Australia and Oceania	62	19.55
Europe	Europe	58	18.29
Africa	Africa	27	8.51
Total		**317**	**100**

The conceptual model was evaluated across multiple datasets representing three continents, with Africa excluded from the initial regional analyses. Model fit was then assessed for the full global dataset, which incorporated all regional responses and consisted of 317 participants. The results showed that the model achieved strong fit for the Asia, Australia, and global samples, supported by both absolute and relative fit indices. In contrast, the European sample demonstrated only partial fit, with GFI and RMSEA values falling slightly below recommended thresholds. To ensure the robustness of the estimates, each regional dataset, as well as the global sample, was bootstrapped to 2000 samples at a 95 percent confidence interval (see Table 9).

Table 9 Region-wise model fit results

Region	Cmin/df	CFI	GFI	RMSEA	SRMR	Decision
Asia	1.844	.923	.851	.084	.069	Good Fit
Australia	1.053	.953	.771	.005	.002	Good Fit
Europe	2.054	.847	.772	.136	.089	Partial Fit
Global*	2.960	.936	.903	.079	.049	Good Fit

Note: The sample is bootstrapped to 2000 at 95% bias-corrected confidence interval. * = The global analysis includes the samples of the Africa region.

Direct effect hypotheses were tested at the 5 percent significance level. For the Australia and Oceania region, all hypotheses (H1–H5) were rejected, as none of the paths produced p-values below 0.05. A similar pattern emerged in the Asia region, where hypotheses H1–H5 were also rejected; however, the LT–CI relationship (H4) approached significance and was accepted at the 10 percent level. In the Europe sample, hypotheses H2–H5 were rejected, while H1 was supported, indicating a significant direct relationship between RUP and PEC. The Europe region also demonstrated a slightly higher explained variance ($R^2 = 0.173$) compared to Asia ($R^2 = 0.149$). The standardized beta for the RUP–PEC path was notably stronger in Europe ($\beta = 0.402$), although the remaining hypotheses were not supported at the 5 percent level. Nonetheless, H4 and H5 reached significance at the 10 percent threshold. At the global level, H1 was accepted, confirming a significant direct effect from RUP to PEC ($\beta = 0.18$). All other hypotheses (H2–H5) were rejected at the 5 percent level, yet the LT–CI relationship (H4) was again accepted at the 10 percent level. These results are summarized in Table 10.

Table 10 Region-wise direct effect and hypothesis results

Region	Direct Path	Direct Effect	P Value	R^2	Decision
Asia	RUP =>PEC	.130	.214	.029	Reject H1
	CI=>PEC	.037	.743		Reject H2
	LT=>PEC	.079	.474		Reject H3
	LT=>CI	.191**	.075**	.037	Reject H4
	RUP=>LT	.143	.153	.021	Reject H5
Australia	RUP =>PEC	.040	.788	.084	Reject H1
	CI=>PEC	.268	.124		Reject H2
	LT=>PEC	.083	.584		Reject H3
	LT=>CI	.113	.477	.013	Reject H4
	RUP=>LT	-.130	.367	.017	Reject H5
Europe	RUP =>PEC	.402	.008	.173	**Accept H1**
	CI=>PEC	-.087	.577		Reject H2
	LT=>PEC	.004	.977		Reject H3
	LT=>CI	.283**	.055**	.080	Reject H4
	RUP=>LT	-.230**	.090**	.053	Reject H5
Global*	RUP =>PEC	.180	.004	.034	**Accept H1**
	CI=>PEC	-.003	.966		Reject H2
	LT=>PEC	.031	.625		Reject H3
	LT=>CI	.119**	.067**	.014	Reject H4
	RUP=>LT	.023	.703	.001	Reject H5

Note: The sample is bootstrapped to 2000 at 95% bias-corrected confidence interval. * = The global analysis includes the samples of the Africa region. ** = Accepted at 10% level of significance.

Indirect effects were tested through hypotheses H6 and H7 using 2,000-sample bootstrapping for all regional and global models. An indirect path was considered significant if the p-value was below 0.05 and the confidence interval excluded zero. Across all analyses, none of the indirect effects reached significance at the 5 percent level, as zero fell within all confidence intervals and p-values exceeded

0.05. The only exception was in the Europe sample, where the indirect effect from RUP to PEC through LT was significant at the 10 percent level. Full results are presented in Table 11.

Table 11 Region-wise indirect effect results

Region	Indirect Path	Indirect Effect	C.I.	P	Decision
Asia	RUP =>LT=>PEC	0.011	-0.034 – 0.088	0.544	Reject H6
	RUP=>LT=>CI=>PEC	0.001	-0.011 – 0.092	0.295	Reject H7
Australia	RUP =>LT=>PEC	-0.104	-0.140 – 0.022	0.340	Reject H6
	RUP=>LT=>CI=>PEC	-0.004	-0.161 – 0.016	0.286	Reject H7
Europe	RUP =>LT=>PEC**	-0.000	-0.228 – 0.002	0.060	Reject H6
	RUP=>LT=>CI=>PEC	0.006	-0.075 – 0.089	0.805	Reject H7
Global*	RUP =>LT=>PEC	0.003	-0.008 – 0.024	0.460	Reject H6
	RUP=>LT=>CI=>PEC	0.001	-0.004 – 0.015	0.536	Reject H7

Note: The sample is bootstrapped to 2000 at 95% bias-corrected confidence interval. * = The global analysis includes the samples of the Africa region. ** = Significant at 10% level of significance.

In path analysis, the total effect reflects the overall influence of an exogenous construct on an endogenous construct through all direct and indirect pathways. It is calculated by summing the direct effect and all indirect effects operating through mediating variables. In this model, RUP is an exogenous construct and PEC is the endogenous construct. Thus, the total effect of RUP on PEC includes three components: the direct path (RUP → PEC), the indirect path through LT (RUP → LT → PEC), and the sequential indirect path through LT and CI (RUP → LT → CI → PEC).

Results indicate that the Europe region shows a statistically significant total effect, as the p-value is below 0.05 and the confidence interval excludes zero. The global analysis also demonstrates a statistically significant total effect. However, the Asia and Australia regions do not meet these criteria, indicating nonsignificant total effects.

Albers (2017) notes that effect size can be more informative than statistical significance alone. Following Cohen’s (2013) guidelines, effects below 0.02 are considered small, those above 0.35 are large, and intermediate values represent moderate effects. Asia and Australia show small total effects, whereas Europe shows a large effect (0.407), consistent with its statistical significance. The global effect size is moderate (0.181) and statistically significant (p = 0.01). These findings highlight meaningful regional differences and reinforce the value of considering both significance and effect size when interpreting total effects (see Table 12).

Table 12 Region-wise total effect results

Region	Total Effect Path	Total Effect	C.I.	P	Effect Category
Asia	RUP => PEC	0.142	-0.071 – 0.357	0.166	Small
Australia		0.025	-0.271 – 0.342	0.820	Small
Europe		0.407	0.085 – 0.684	0.023	Large
Global*		0.181	0.050 – 0.313	0.010	Moderate

Note: The sample is bootstrapped to 2000 at 95% bias-corrected confidence interval. * = The global analysis includes the samples of the Africa region.

DISCUSSION

This study offers a new lens for understanding teachers' struggles with detecting AI-generated content by demonstrating that these challenges arise not solely from the deceptive power of generative AI, but from the intersection of institutional structures, time constraints, and the increasingly blended nature of human–AI writing. Prior research has emphasized AI deception as a technological issue, noting that the natural language fluency of AI chatbots enables them to mimic human writing in ways that mislead educators (Grazioli & Jarvenpaa, 2003a, 2003b; Hancock, 2007; Niloy et al., 2024a; Niloy et al., 2024b). The findings of this study extend that discussion by revealing that deception becomes more consequential in environments shaped by rigid university policies. The SEM analysis demonstrated that among all modeled factors, Rigid University Policies (RUP) emerged as the only consistent and significant predictor of teachers' Perceived Evaluation Challenge (PEC), both globally and within certain regions. This indicates that institutional structures, rather than AI alone, play a central role in shaping the difficulties educators encounter during assessment.

Translating these findings into classroom realities shows why detection problems persist. Strict marking deadlines, fixed assessment formats, and large class sizes, all typical features of rigid institutional frameworks, restrict the time available for careful evaluation. When the SEM results showed that RUP increases Lack of Time (LT), the finding reflected what many educators experience in practice: compressed grading windows force teachers to read superficially, preventing them from noting subtle inconsistencies between a student's usual writing and a highly polished AI-generated submission. For example, lecturers who must grade dozens of essays within 48 to 72 hours cannot compare writing across assignments or examine abrupt shifts in tone, argumentation, or conceptual depth, which are critical indicators of AI involvement. The statistical relationship between LT and Content Inseparability (CI) further illustrates how time shortages intensify the difficulty of parsing mixed-authorship texts. Students increasingly blend AI-generated passages with original writing, and when teachers are pressed for time, this blended content becomes virtually inseparable. In other words, CI does not function independently; it becomes problematic when exacerbated by workload pressures stemming from rigid policies.

These results refine existing scholarship by illuminating how institutional environments magnify the cognitive challenges teachers face. Earlier work emphasized technological deception, but this study shows that teachers' struggles arise from a systemic chain of influences: rigid policies constrain time; limited time reduces evaluative depth; reduced depth makes blended AI–human writing harder to detect. This layered mechanism explains why detection difficulties persist even when teachers have strong disciplinary knowledge or prior exposure to AI tools. It also helps explain the regional differences observed: in Europe and the global sample, RUP had a strong total effect on PEC, suggesting that detection difficulties are embedded in broader policy cultures, not merely in individual teaching practices.

The findings echo long-standing concerns regarding the pressures that institutional rigidity places on both teachers and learners (Cuban, 1984; Duah & McGivern, 2024). They also support arguments that effective evaluation in emerging technological contexts requires both flexibility and training (McConnell & Fry, 1972). In the era of generative AI, these needs become even more urgent. Teachers cannot be expected to safeguard academic integrity if structural conditions deny them the time and autonomy required for meaningful assessment. Rather than attributing detection failure solely to AI's sophistication or teachers' limitations, this study positions the issue within a broader institutional ecosystem that must be re-examined if academic integrity is to be preserved.

IMPLICATIONS

The findings of this study have important theoretical and practical implications for researchers, educators, institutions, and policymakers. By demonstrating that RUP exerts the strongest influence on PEC, with LT and CI functioning as key intermediaries, this study contributes a more comprehensive understanding of the psychological, technological, and institutional factors shaping teachers' ability to detect AI-generated content. Existing literature often examines these constructs in isolation; the multidimensional model developed here reveals how they operate as an interconnected system. This contributes to theory by shifting the focus from AI's deceptive capacity to the broader institutional contexts that either mitigate or exacerbate its impact. The global scope of the sample further extends knowledge by showing how cultural and policy environments influence teacher perceptions, underscoring the need for region-sensitive approaches to academic integrity in AI-mediated learning spaces.

The practical implications of these findings highlight the importance of equipping lecturers with both the structural support and the professional skills required for effective detection. Lecturer training must evolve beyond general discussions of AI ethics and instead focus on concrete evaluative practices, such as building baseline writing profiles for students, analyzing stylistic inconsistencies that emerge in hybrid AI–human texts, and understanding the limitations and error patterns of AI detection tools. Such training becomes essential in time-constrained environments, allowing lecturers to make informed judgments even when in-depth review is not possible.

Institutional policy reform emerges as a critical implication of this research. Because RUP significantly influences PEC, universities must reconsider policies that impose unrealistic grading timelines, prescribe uniform assignment formats, or restrict assessment flexibility. Extending marking windows, reducing assessment loads in writing-intensive courses, and providing dedicated academic integrity support units can substantially reduce LT and improve evaluative accuracy. Policies must shift from compliance-driven approaches toward structures that support pedagogical judgment, recognizing that academic integrity work is cognitively demanding and cannot be compressed without compromising quality.

A further implication concerns the design of AI-resilient assessments. To counteract CI, assignments should incorporate processes that reveal the evolution of student thinking, such as draft iterations, annotated reflections, oral defenses, or personalized components tied to local contexts or class discussions. These approaches reintroduce elements of authenticity that are difficult for AI to replicate and allow teachers to verify authorship more confidently. The creation of such assessments requires institutional endorsement, as lecturers cannot redesign tasks without the flexibility and authority granted by policy.

Finally, the findings carry broader implications for developers and policymakers. The persistence of CI and PEC despite existing detection tools suggests that technical solutions alone cannot safeguard academic integrity. Developers may explore watermarking or traceable patterns in AI-generated outputs, but these innovations must be complemented by institutional structures that support human evaluation. Policymakers should treat academic integrity as a shared responsibility across technological, institutional, and pedagogical domains.

Taken together, these implications underline a central conclusion: improving teachers' ability to detect AI-generated content requires coordinated action at multiple levels. Training alone is insufficient; institutional support must align with the cognitive realities of assessment, and assignment design must adapt to the changing landscape of student writing. By illuminating the layered mechanisms through which RUP, LT, and CI influence PEC, this study provides a roadmap for strengthening academic integrity in an era where generative AI will continue to evolve.

CONCLUSION

This study offers a clearer, system-level explanation for why teachers struggle to evaluate student work in the Gen-AI era. Using a mixed-method design, it proposes a conceptual model showing that perceived evaluation challenges stem not only from the blending of human and AI writing but from the combined influence of both controllable and uncontrollable contextual factors.

The study is limited by its reliance on self-reported perceptions, which future research should validate through experimental or observational methods. Further work could also examine more diverse contexts and integrate additional institutional, pedagogical, and technological variables to capture the complexity of educators' evaluative behaviors.

The significance of this study lies in moving the field beyond a narrow focus on AI tools themselves. By demonstrating that teachers' difficulties emerge from a broader social and organizational system, it underscores that addressing only the technological half of the problem will inevitably lead to solving this daunting issue partially.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENTS

The authors have no acknowledgement to provide to any particular individual or institution, other than to the participants of this study.

AVAILABILITY OF DATA

Data will be made available on request.

FUNDING

This study was not supported by any grants from funding bodies in the public, private, or not-for-profit sectors. The authors declare that no financial support was received for the research, authorship, and publication of this article.

DISCLOSURE OF AI USE

During the preparation of this work, the authors used Text Generative AI tools to assist with language refinement and structural editing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the integrity and accuracy of the presented text. No generative AI tools were used for generating or writing any original piece of the work (either completely or partially), data analysis, interpretation, or the drawing of conclusions.

REFERENCES

Abdullah, M., Madain, A., & Jararweh, Y. (2022). ChatGPT: Fundamentals, Applications and Social Impacts. 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), 1–8. https://api.semanticscholar.org/CorpusID:257537378

AlAfnan, M. A., Samira Dishari, Marina Jovic, & Koba Lomidze. (2023). ChatGPT as an Educational Tool: Opportunities, Challenges, and Recommendations for Communication, Business Writing, and Composition Courses. Journal of Artificial Intelligence and Technology, 3(2 SE-Research Articles), 60–68. https://doi.org/10.37965/jait.2023.0184

Andrews, E. (2023). Comparison of different programs that claim to detect AI-generated text. Colorado State University. https://tilt.colostate.edu/comparing-ai-detection-tools-one-instructors-experience/

Atlas, S. (2023). ChatGPT for Higher Education and Professional Development: A Guide to Conventional AI. DigitalCommons@URI. https://digitalcommons.uri.edu/cba_facpubs/548/?utm_source=digitalcommons.uri.edu%2Fcba_facpubs%2F548&utm_medium=PDF&utm_campaign=PDFCoverPages

Aw, B. (2024). 12 Best AI Detectors in 2024: From Over 180 Tests. Brandan Aw. https://brendanaw.com/best-ai-detector

Barton, R. (2024). Turnitin adding AI writing detection, but instructors should use it with caution. Purdue University. https://www.purdue.edu/online/turnitin-adding-ai-writing-detection-but-instructors-should-use-it-with-caution/

Basu, B. (2023). ChatGPT and its impact on education sector. Daily Sun. https://www.daily-sun.com/printversion/details/673514/ChatGPT-and-its-impact-on-education-sector

Biener, C., & Waeber, A. (2024). Would I lie to you? How interaction with chatbots induces dishonesty. Journal of Behavioral and Experimental Economics, 102279. https://doi.org/https://doi.org/10.1016/j.socec.2024.102279

Black, J. (2024). Can AI Lie? Chabot Technologies, the Subject, and the Importance of Lying. Social Science Computer Review, 08944393241282602.

Bocian, Z. (2024). Key Chatbot Statistics You Should Follow in 2024. Chatbot. https://www.chatbot.com/blog/chatbot-statistics/

Bozkurt, A. (2024). GenAI et al.: Cocreation, Authorship, Ownership, Academic Ethics and Integrity in a Time of Generative AI. *Open Praxis*. https://doi.org/10.55982/openpraxis.16.1.654.

Carlson, J. R., George, J. F., Burgoon, J. K., Adkins, M., & White, C. H. (2004). Deception in Computer-Mediated Communication. Group Decision and Negotiation, 13(1), 5–28. https://doi.org/10.1023/B:GRUP.0000011942.31158.d8

Cascella, M., Montomoli, J., Bellini, V., & Bignami, E. (2023). Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. Journal of Medical Systems, 47(1), 33. https://doi.org/10.1007/s10916-023-01925-4

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. TechTrends, 66(4), 616–630.

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. Journal of Applied Learning and Teaching, 6(2).

Chaka, C. (2024). Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. Journal of Applied Learning and Teaching, 7(2).

Chan, A. (2023). GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. AI and Ethics, 3(1), 53–64. https://doi.org/10.1007/s43681-022-00148-6

Coleman, T. (2023). 2023: the year of the AI boom. The Week. https://theweek.com/tech/2023-ai-boom

Cristianini, N. (2016). Intelligence Reinvented. New Scientist, 232(3097), 37–41. https://doi.org/10.1016/S0262-4079(16)31992-3

Cuban, L. (1984). Policy and Research Dilemmas in the Teaching of Reasoning: Unplanned Designs. Review of Educational Research, 54(4), 655–681. https://doi.org/10.3102/00346543054004655

Daft, R., & Lengel, R. (1986). Organizational Information Requirements, Media Richness and Structural Design. Management Science, 32, 554–571. https://doi.org/10.1287/mnsc.32.5.554

De Wilde, V. (2024). Can novice teachers detect AI-generated texts in EFL writing? ELT Journal, 78(4), 414–422. https://doi.org/10.1093/elt/ccae031

Duah, J. E., & McGivern, P. (2024). How generative artificial intelligence has blurred notions of authorial identity and academic norms in higher education, necessitating clear university usage policies. The International Journal of Information and Learning Technology, 41(2), 180–193. https://doi.org/10.1108/IJILT-11-2023-0213

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. International Journal of Information Management, 48, 63–71. https://doi.org/10.1016/J.IJINFOMGT.2019.01.021

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642. https://doi.org/10.1016/J.IJINFOMGT.2023.102642

Edwards, B. (2023). OpenAI's GPT-4 exhibits "human-level performance" on professional benchmarks. Arstechnica. https://arstechnica.com/information-technology/2023/03/openai-announces-gpt-4-its-next-generation-ai-language-model/

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. International Journal for Educational Integrity, 19(1), 17. https://doi.org/10.1007/s40979-023-00140-5

Farazouli, A., Cerratto-Pargman, T., Bolander-Laksov, K., & McGrath, C. (2024). Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. Assessment & Evaluation in Higher Education, 49(3), 363–375. https://doi.org/10.1080/02602938.2023.2241676

Field, H. (2024). OpenAI launches new AI model GPT-4o and desktop version of ChatGPT. CNBC. https://www.cnbc.com/2024/05/13/openai-launches-new-ai-model-and-desktop-version-of-chatgpt.html

Finnie-Ansley, J., Denny, P., Becker, B. A., Luxton-Reilly, A., & Prather, J. (2022). The Robots Are Coming: Exploring the Implications of OpenAI Codex on Introductory Programming. Proceedings of the 24th Australasian Computing Education Conference, 10–19. https://doi.org/10.1145/3511861.3511863

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. Computers and Education: Artificial Intelligence, 6, 100209.

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. Journal of Information Technology Case and Application Research, 25(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. Npj Digital Medicine, 6(1), 75. https://doi.org/10.1038/s41746-023-00819-6

Gottardello, D., & Karabag, S. (2020). Ideal and actual roles of university professors in academic integrity management: a comparative study. *Studies in Higher Education*, 47, 526 - 544. https://doi.org/10.1080/03075079.2020.1767051.

Grant, N. (2023). Google Builds on Tech's Latest Craze With Its Own A.I. Products. The New York Times. https://www.nytimes.com/2023/05/10/technology/google-ai-products.html

Grazioli, S., & Jarvenpaa, S. (2003a). Consumer and Business Deception on the Internet: Content Analysis of Documentary Evidence. International Journal of Electronic Commerce, 7, 93–118.

Grazioli, S., & Jarvenpaa, S. (2003b). Deceived: Under target Online. Commun. ACM, 46, 196–205. https://doi.org/10.1145/953460.953500

Guan, C., Wang, X., Zhang, Q., Chen, R., He, D., & Xie, X. (2019). Towards a Deep and Unified Understanding of Deep Neural Models in NLP. 36th International Conference on Machine Learning, 2454–2463. https://proceedings.mlr.press/v97/guan19a.html

Hancock, J. T. (2007). Digital deception. Oxford Handbook of Internet Psychology, 61(5), 289–301.

Harris, A., & Jones, M. (2019). Teacher leadership and educational change. *School Leadership & Management*, 39, 123 - 126. https://doi.org/10.1080/13632434.2019.1574964.

Hobert, S., & von Wolff, R. M. (2019). Say Hello to Your New Automated Tutor - A Structured Literature Review on Pedagogical Conversational Agents. Wirtschaftsinformatik. https://api.semanticscholar.org/CorpusID:201114924

Hughes, A. (2023). ChatGPT: Everything you need to know about OpenAI's GPT-4 upgrade. BBC Science Focus. https://www.sciencefocus.com/future-technology/gpt-3/

Ibrahim, H., Liu, F., Asim, R., Battu, B., Benabderrahmane, S., Alhafni, B., Adnan, W., Alhanai, T., AlShebli, B., & Baghdadi, R. (2023). Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. Scientific Reports, 13(1), 12187.

Ivanov, V. (2023). Which Is The Best AI Content Detector? [13 Tools Tested]. Trickmenot.Ai. https://trickmenot.ai/which-is-the-best-ai-content-detector/

Kofinas, A., Tsay, C., & Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13585.

Kumar, R., & Mindzak, M. (2024). Who wrote this? Detecting artificial intelligence–generated text from human-written text. Canadian Perspectives on Academic Integrity, 7(1).

Kushwaha, A. K., & Kar, A. K. (2021). MarkBot – A Language Model-Driven Chatbot for Interactive Marketing in Post-Modern World. Information Systems Frontiers. https://doi.org/10.1007/s10796-021-10184-y

Lambert, J., & Stevens, M. (2023). ChatGPT and Generative AI Technology: A Mixed Bag of Concerns and New Opportunities. Computers in the Schools, 1–25. https://doi.org/10.1080/07380569.2023.2256710

Lameras, P., & Arnab, S. (2021). Power to the teachers: an exploratory review on artificial intelligence in education. Information, 13(1), 14.

Leah, H., & Meroño-Peñuela, A. (2022). The Hermeneutics of Computer-Generated Texts. Configurations, 30(2), 115–139. https://doi.org/10.1353/con.2022.0008

Li, H. (2023). Effects of a ChatGPT-based flipped learning guiding approach on learners' courseware project performances and perceptions. *Australasian Journal of Educational Technology*. https://doi.org/10.14742/ajet.8923.

Maba, W., Mantra, I., & Widiastuti, I. (2023). TEACHERS OF 21ST CENTURY: TEACHERS' ROLES, STRATEGIES INNOVATION AND CHALLENGES. *International Journal of Social Science*. https://doi.org/10.53625/ijss.v2i6.5473.

Mathew, A. (2023). Is artificial intelligence a world changer? A case study of OpenAI's Chat GPT.

McConnell, T. R., & Fry, M. A. (1972). Flexibility or rigidity: university attitudes towards the James Report. Higher Education Review, 4(3), 13.

Murray, N., & Tersigni, E. (2024). Can Instructors Detect Ai-Generated Papers? Postsecondary Writing Instructor Knowledge and Perceptions of Ai. Journal of Applied Learning & Teaching, 7(2), 1–13. https://doi.org/10.37074/jalt.2024.7.2.12

Natale, S. (2023). AI, human-machine communication and deception. The SAGE Handbook of Human-Machine Communication, 401–408.

Niloy, A. C., Bari, M. A., Sultana, J., Chowdhury, R., Raisa, F. M., Islam, A., Mahmud, S., Jahan, I., Sarkar, M., Akter, S., Nishat, N., Afroz, M., Sen, A., Islam, T., Tareq, M. H., & Hossen, M. A. (2024). Why do students use ChatGPT? Answering through a triangulation approach. *Computers and Education: Artificial Intelligence*, *6*(6), 100208. https://doi.org/10.1016/j.caeai.2024.100208

Niloy, A. C., Hafiz, R., Hossain, B. M., Gulmeher, F., Sultana, N., Islam, K. F., Bushra, F., Islam, S., Hoque, S. I., Rahman, M., & Kabir, S. (2024). AI chatbots: A disguised enemy for academic integrity? *International Journal of Educational Research Open*, *7*, 100396. https://doi.org/10.1016/j.ijedro.2024.100396

OpenAI. (2022). Introducing ChatGPT. Openai.Com. https://openai.com/index/chatgpt/

Özkan, E. (2022). Scaffolding as Teachers' Guidance Role in the Context of Constructivist Learning Approach. Journal of Educational Issues. https://doi.org/10.5296/jei.v8i1.19690.

Paniagua A. & Istance, D. (2018). Teachers as Designers of Learning Environments: The Importance of Innovative Pedagogies. . https://doi.org/10.1787/9789264085374-en.

Pérez-Marín, D. (2021). A Review of the Practical Applications of Pedagogic Conversational Agents to Be Used in School and University Classrooms. Digital. https://api.semanticscholar.org/CorpusID:234079366

Rogers, R. (2024). With OpenAI's Release of GPT-4o, Is ChatGPT Plus Still Worth It? Wired. https://www.wired.com/story/with-gpt-4o-is-chatgpt-plus-still-worth-it/

Sadka, A. (2024). What to expect from the next generation of chatbots: OpenAI's GPT-5 and Meta's Llama-3. The Conversation. https://theconversation.com/what-to-expect-from-the-next-generation-of-chatbots-openais-gpt-5-and-metas-llama-3-228217

Saunders, M. N. K., & Townsend, K. (2016). Reporting and justifying the number of interview participants in organization and workplace research. British Journal of Management, 27(4), 836–852.

Schmitt, M., & Flechais, I. (2024). Digital Deception: Generative artificial intelligence in social engineering and phishing. Artificial Intelligence Review, 57(12), 1–23.

Sier, J. (2022). Search engine AI ChatGPT takes the internet by storm, bad poetry and all. Financial Review. https://www.afr.com/technology/chatgpt-takes-the-internet-by-storm-bad-poetry-and-all-20221207-p5c4hv

Simuţ, R., Simuţ, C., Badulescu, D., & Badulescu, A. (2024). ARTIFICIAL INTELLIGENCE AND THE MODELLING OF TEACHERS'COMPETENCIES. Amfiteatru Economic, 26(65), 181–200.

Singh, V., & Ram, S. (2024). Impact of Artificial Intelligence on Teacher Education. Shodh Sari-An Internafional Mulfidisciplinary Journal.

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. Computers and Education: Artificial Intelligence, 3, 100075.

Walters, W. H. (2023). The Effectiveness of Software Designed to Detect AI-Generated Writing: A Comparison of 16 AI Text Detectors. Open Information Science, 7(1). https://doi.org/doi:10.1515/opis-2022-0158

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. International Journal for Educational Integrity, 19(1), 26.

Williamson, B. (2024). The Social life of AI in Education. International Journal of Artificial Intelligence in Education, 34(1), 97–104. https://doi.org/10.1007/s40593-023-00342-5

Williamson, S. M., & Prybutok, V. (2024). The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation. Information, 15(6), 299.

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are We There Yet? - A Systematic Literature Review on Chatbots in Education. Frontiers in Artificial Intelligence, 4, 654924. https://doi.org/10.3389/frai.2021.654924

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., & He, L. (2023). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. ArXiv Preprint ArXiv:2302.09419.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating Linguistics-Based Cues for Detecting Deception in Text-Based Asynchronous Computer-Mediated Communications. Group Decision and Negotiation, 13(1), 81–106. https://doi.org/10.1023/B:GRUP.0000011944.62889.6f

Zhou, L., Burgoon, J., Twitchell, D., Qin, T., & Jr, J. (2004). A Comparison of Classification Methods for Predicting Deception in Computer-Mediated Communication. J. of Management Information Systems, 20, 139–165. https://doi.org/10.1080/07421222.2004.11045779

Zhou, L., & Zhang, D. (2004). Can online behavior unveil deceivers? - an exploratory investigation of deception in instant messaging. 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of The, 9 pp. https://doi.org/10.1109/HICSS.2004.1265079