Pamukkale University Journal of Education
67, 342-358, 2026

Evaluating Human Reviewers versus Artificial Intelligence: A Comparative Analysis of AI-Generated Essay Detection in Higher Education

Zafer SUSOY[1*]

[1] Tokat Gaziosmanpaşa University, Faculty of Education, Department of English Language Teaching, Türkiye

Abstract: *The proliferation of AI text generation tools in educational settings has created unprecedented challenges for academic integrity. As AI-generated content grows more sophisticated, traditional assessment methods may no longer suffice to distinguish human from machine-authored work. This mixed-methods study compared human reviewers, an AI scoring model, and plagiarism detection tools in identifying AI-generated essays. Sixteen essays were analyzed: eight by ELT master's students and eight AI-generated (ChatGPT 4o, Gemini 1.5, Llama), with four subsequently humanized. ELT was selected because writing proficiency and authentic voice are central disciplinary competencies. Two faculty reviewers, iThenticate, Turnitin, and a GPT-4o model evaluated all essays using an expert-reviewed analytic rubric covering coherence, originality, depth, writing style, and topic adherence. The AI model showed superior internal consistency (α = 0.923) versus human reviewers (α = 0.832). Turnitin achieved the highest accuracy (87.5%). Human reviewers performed inconsistently (62.5%; Fisher's Exact Test, p = .096). Inter-rater agreement was assessed via quadratic weighted kappa (QWK), as small sample sizes precluded a valid chi-square analysis. Digital tools currently outperform human judgment in detecting AI-generated content, yet limitations persist across all methods. These findings underscore the need for institutional AI policies, educator training, and hybrid human-technological assessment approaches.*

Article Details
Research Article
Received
23/12/2025
Accepted
27/03/2026
Keywords
AI-generated text detection, academic integrity, English language teaching, automated essay evaluation.

1. Introduction

The rapid advancement of artificial intelligence technologies has fundamentally transformed educational landscapes worldwide. Among the most significant developments is the emergence of sophisticated text generation tools capable of producing coherent, contextually relevant, and stylistically sophisticated written content that often resembles human authorship (Crawford et al., 2024). This technological evolution has introduced unprecedented challenges to academic integrity, particularly in fields where critical thinking, originality, and authentic expression are paramount to educational objectives. The challenge of distinguishing between human-authored and AI-generated academic content extends beyond theoretical concerns to practical implications that permeate contemporary educational practice. In English Language Teaching (ELT) and related disciplines, where the development of analytical thinking, creativity, and

[*] *Corresponding Author: Zafer Susoy E-mail: zafersusoy@gmail.com Address: Tokat Gaziosmanpaşa University, Faculty of Education, Department of English Language Teaching, Türkiye*

e-ISSN: 1309-0275 http://doi.org/10.9779/pauefd.1847553

authentic voice represents core pedagogical goals, the infiltration of undetected AI-generated content threatens to undermine fundamental educational principles (Alzubi et al., 2025). ELT was selected as the research context because writing proficiency, critical argumentation, and authentic voice are central disciplinary competencies—making the accurate identification of AI-generated content particularly consequential for valid assessment in this field. Master's-level ELT students were chosen because their advanced academic writing proficiency renders the distinction between human and AI-authored text especially nuanced. Traditional assessment methodologies, which have historically relied upon human professional judgment supplemented by plagiarism detection software, may prove insufficient for addressing these emerging challenges.

Current research examining AI-generated content has predominantly focused on the technological capabilities of text generation systems and the subsequent difficulties these present for educational practitioners. Studies by Ayub et al. (2024) and Fleckenstein et al. (2024) have highlighted the increasing difficulty educators face in identifying AI-generated essays, particularly when such texts are modified through humanization processes designed to replicate natural writing patterns and stylistic variations. Similarly, Parviz (2024) emphasized that AI systems are increasingly capable of using idiomatic language, varied sentence structures, and context-appropriate reasoning, making the differentiation process even more complex.

However, a significant gap exists in the literature regarding systematic comparisons of detection effectiveness across different evaluation methods. While studies by Yeadon et al. (2024) and Liu et al. (2023) have pointed out that traditional plagiarism detection tools are primarily designed to identify textual similarities to existing content, few have directly compared their performance against that of human professionals (Herath et al., 2025; Silva et al., 2024). This gap is particularly significant given the increasing prevalence of AI-generated content in academic settings and the potential consequences for academic integrity. Importantly, prior studies have rarely triangulated three evaluation approaches—human reviewers, dedicated AI scoring models, and commercially available plagiarism detection platforms—within a single ELT-specific design, which is the methodological contribution the present study attempts to make.

This study addresses this critical knowledge gap by conducting a comprehensive comparative analysis of human reviewers, digital detection tools, and AI scoring models in identifying AI-generated academic content. The study is guided by the following research aims: (1) to compare the classification accuracy of human reviewers, a trained AI scoring model, and digital plagiarism detection tools in distinguishing human-written from AI-generated essays; (2) to assess the reliability and internal consistency of evaluative judgments across these three methods; and (3) to examine the qualitative decision-making processes employed by human reviewers when differentiating between human and AI-authored texts. Understanding these processes is essential for developing enhanced training programs and more effective detection strategies.

1.1. Literature Review

1.1.1. The Challenge of AI-Generated Content Detection

The integration of artificial intelligence into educational environments has produced transformative changes across multiple dimensions of teaching and learning. AI systems demonstrate remarkable capabilities in processing and analyzing large datasets, enabling personalized learning experiences and automated assessment procedures (Crawford et al., 2024). However, the most striking development concerns AI's capacity to generate written

content that exhibits grammatical accuracy, contextual relevance, and stylistic sophistication comparable to human writing.

This advancement presents particular challenges within English Language Teaching contexts, where critical thinking, originality, and nuanced expression represent essential learning objectives (Alzubi et al., 2025). As Chung and Jeong (2024) note, contemporary AI systems increasingly demonstrate proficiency in utilizing idiomatic language, varied sentence structures, and context-appropriate reasoning, making the differentiation process exponentially more complex for educational practitioners.

The sophistication of current AI text generation systems raises fundamental questions about the reliability of traditional assessment mechanisms. Lee (2024) found that while some educators can detect AI-generated essays based on subtle linguistic cues, many struggle to do so consistently, particularly when AI-generated texts are modified or "humanized" to imitate natural writing patterns. As these technologies continue advancing, the risk of undetected academic misconduct increases substantially, necessitating not only improved detection mechanisms but also a deeper understanding of how human professional judgment compares to AI-assisted evaluations and digital detection tools.

1.1.2. Digital Detection Tools and Human Professional Judgment

Plagiarism detection tools such as iThenticate, Turnitin, and related systems have long served as standard resources in academic settings for identifying instances of copied content. However, the increasing sophistication of AI-generated text presents novel challenges that extend beyond these tools' original design parameters. As Majhi and Santra (2023) highlight, traditional plagiarism detection systems primarily identify textual similarities with existing sources, which limits their effectiveness in recognizing entirely original yet AI-generated content.

This limitation underscores critical gaps in current detection capabilities, particularly regarding assessment of coherence, analytical depth, and originality—elements that typically distinguish human writing from algorithmically produced text (Coccoli & Patanè, 2024). Consequently, there exists an urgent need for either enhanced versions of current tools or the development of new technologies capable of detecting more nuanced features characteristic of AI-generated content.

Parallel to technological limitations, questions arise regarding the comparative effectiveness of human professional judgment versus digital detection systems. Human reviewers offer depth of evaluation and contextual awareness that may surpass automated detection systems, particularly in assessing originality, critical thinking, and rhetorical sophistication. However, digital tools provide consistency, scalability, and processing speed that make them attractive for large-scale applications (Crompton et al., 2024).

Research by Steponenaite and Barakat (2023) and Halaweh and Refae (2024) suggests that while automated scoring systems demonstrate proficiency in identifying surface-level features such as grammar and syntax, they prove less reliable when evaluating higher-order writing skills. Despite growing literature on AI scoring and plagiarism detection, notable gaps remain in direct comparisons between digital tools and experienced human evaluators in detecting AI-generated texts. Cingillioglu (2023) similarly demonstrated that even purpose-built AI detectors can be confounded by paraphrased or lightly edited AI output, underscoring the need for multi-method detection frameworks such as the one employed in the current study.

1.1.3. Ethical Implications and Educational Practice

The integration of AI into educational assessment raises significant ethical concerns regarding fairness, bias, and potential erosion of critical thinking and originality in student work. Scholars

such as Steponenaite and Barakat (2023) and Elkhatat et al. (2023) have emphasized the urgent need for ethical guidelines and institutional policies that ensure integrity and transparency in AI use for educational purposes. Without clear frameworks, reliance on AI may inadvertently devalue human intellectual effort and compromise academic standards.

In this evolving landscape, the role of educators becomes increasingly complex and indispensable. Bearman and Luckin (2020) and Luckin (2024) argue for the continued importance of human judgment in the assessment process, particularly in evaluating nuanced skills such as creativity, critical thinking, and original expression—areas where AI still struggles. Educational professionals serve not only as assessors but also as ethical stewards, responsible for guiding students through responsible AI use while safeguarding academic integrity.

1.1.4. Theoretical Framework

This study's theoretical foundation incorporates several key frameworks that collectively provide a comprehensive understanding of AI's role in educational assessment. Rather than treating these frameworks as loosely juxtaposed perspectives, the present study employs them in an integrated manner: each theory illuminates a distinct facet of the research problem, and together they offer a multi-layered lens through which the comparative data can be interpreted.

Central to this framework is Technological Determinism, as discussed by McLuhan (1964), which suggests that technology significantly influences society and cultural values. In this study's context, technological determinism provides a lens for examining how AI-generated content might shape educational practices and student work assessment. The present study adopts a 'soft' deterministic reading: while AI tools constrain and channel assessment practices in consequential ways, the human actors within educational institutions retain agency in deciding how to respond to, regulate, and integrate these technologies.

Ethical theory, particularly concepts related to fairness and justice, as discussed by Hauer (2017), forms another crucial component of the theoretical framework. AI use in educational settings raises significant ethical questions about bias, fairness, and the potential for AI to exacerbate existing inequalities. The false-positive misclassification of human essays as AI-generated—observed empirically in this study—carries direct ethical consequences for students whose work is wrongly flagged, reinforcing the need for ethical oversight of algorithmic detection systems.

Additionally, the framework includes posthumanism, which challenges traditional human-centered views of education and learning. As AI tools become more integrated into educational practices, posthumanism offers a framework for understanding how these technologies may reshape the relationship between humans and machines in the learning process (Hayles, 1999). In the context of AI detection, posthumanism invites us to question what 'authentic' authorship means when the boundaries between human and machine production are increasingly porous.

Information Processing Theory, outlined by Atkinson and Shiffrin (1968), examines how information is absorbed, processed, and retained, providing particular relevance for understanding how digital tools and AI models process and analyze written content compared to human cognitive processing. The contrasting internal consistency scores observed in this study—Cronbach's $\alpha = 0.923$ for the AI model versus $\alpha = 0.832$ for human reviewers—can be interpreted through this theoretical lens: automated systems process rubric criteria through stable computational routines, while human reviewers introduce cognitive variability characteristic of information processing under uncertainty.

Finally, the Social Construction of Technology (SCOT) theory, proposed by Pinch and Bijker (1987), suggests that technology development and use are socially constructed and influenced

by people's needs, values, and actions. The variation in detection performance observed across methods in this study illustrates SCOT's core argument: plagiarism detection platforms were socially constructed as tools for identifying verbatim copying, and their relative weakness in detecting original AI-generated text reflects the social and institutional assumptions embedded in their design.

Collectively, these five frameworks provide a robust, multi-perspectival scaffold for interpreting the empirical findings of the present study. Technological Determinism and SCOT together situate the detection tools within broader sociotechnical systems; posthumanism and Information Processing Theory illuminate the human-machine dynamics at the core of the comparison; and ethical theory grounds the analysis in the fairness implications that assessment decisions carry for individual learners.

2. Method

2.1. Study Design

This study employed a mixed-methods approach to compare the ability of human reviewers, digital detection tools, and AI models to distinguish between human-written and AI-generated essays. The research design incorporated both quantitative analysis of classification accuracy and qualitative examination of human evaluator decision-making processes. The study was guided by three research aims: (1) comparing classification accuracy across evaluative methods, (2) assessing evaluator reliability, and (3) examining the qualitative reasoning of human reviewers.

2.2. Essay Collection and Preparation

Eight master's students in an English Language Teaching program were asked to write essays on a standardized topic: "Do you agree or disagree with the following statement? All children should be required to take a foreign language class from the time they start school until they begin university. Use specific reasons and examples to support your answer." This prompt was selected for several methodologically grounded reasons. First, it aligns with core ELT curriculum objectives related to argumentative writing and critical reasoning. Second, the prompt's scope—a debatable educational policy—permits sufficient depth of analytical discussion while remaining accessible to graduate-level writers. Third, the identical prompt was administered to both human participants and AI systems, ensuring comparability across conditions. The cognitive demand and scope of the prompt were reviewed and confirmed as appropriate by the two faculty reviewers prior to data collection. No external subject-matter expert validation was conducted for the prompt itself; however, its alignment with standard TOEFL-format argumentative tasks provides a degree of ecological validity for ELT assessment contexts.

Eight AI-generated essays on the same topic were created using advanced AI text generation tools (ChatGPT 4o, Gemini 1.5, Llama). For each AI system, the identical human prompt was entered verbatim. No prompt engineering techniques (e.g., chain-of-thought prompting, role assignment) were employed; default system settings were used across all three tools, including standard temperature and length configurations, to minimize variation attributable to system-level differences. Specifically, ChatGPT 4o, Gemini 1.5, and Llama each generated essays under default web interface conditions with no additional instructions beyond the verbatim prompt. Two essays per platform were generated, yielding four that were retained in original AI form and four that underwent humanization. Four of these AI-generated essays were subsequently modified through specialized AI humanization tools (humanize.io, Quillbot) — software applications designed to paraphrase and restructure AI-generated text to reduce its detectable AI characteristics by introducing lexical variation, sentence restructuring, and

stylistic diversification— to mimic human writing styles, while the remaining four were retained in their original AI-generated form. This approach resulted in a total dataset of 16 essays, as shown in Table 1.

Table 1. *Overview of Essays*

Generation Type	*N*
Human	8
AI (ChatGPT 4o, Gemini 1.5, Llama)	4
Humanized (ChatGPT 4o, Gemini 1.5, Llama)	4
Total	16

2.3. Rubric Development and Validation

A comprehensive analytic rubric was developed based on established academic writing assessment frameworks, including those described in the literature on L2 writing evaluation (Crompton et al., 2024; Majhi & Santra, 2023). The rubric consisted of five criteria evaluated on a five-point Likert scale: coherence (evaluating logical flow and connection of ideas), originality (assessing uniqueness of ideas presented), depth of analysis (examining thoroughness and critical engagement), writing style (considering individuality of voice and tone), and adherence to topic (ensuring focus on the assigned prompt). Each criterion was accompanied by anchor descriptors at each scale point (1–5) to ensure consistent application across reviewers. Anchor descriptors ranged from 1 (= wholly absent or severely deficient) to 5 (= fully and consistently demonstrated), with intermediate points defined for each criterion to reduce scorer subjectivity.

Content validity of the rubric was established through expert review. Both faculty reviewers independently examined the rubric prior to its use in data collection and confirmed that the criteria were representative of the constructs under evaluation. The rubric was not piloted on a separate essay set due to sample size constraints; this limitation is acknowledged. To provide additional guiding structure, each subsection included open-ended prompts directing reviewers to consider specific textual features (e.g., 'Does the essay demonstrate awareness of counterarguments?' under depth of analysis). Reviewers' written responses to these open-ended prompts were subsequently subjected to thematic analysis (see Qualitative Analysis subsection). This rubric development and validation process was conducted prior to any data scoring. The Cronbach's alpha values reported in the Results section reflect the internal consistency with which each evaluative agent applied the rubric criteria across the 16 essays—they do not constitute a measure of rubric validity per se. This distinction is critical: internal consistency addresses the degree to which a single evaluator's scores cohere across criteria, not whether the criteria themselves capture the intended constructs.

2.4. Human Review Process

Two experienced faculty members from the English Language Teaching department were selected as human reviewers based on their extensive backgrounds in linguistics and writing assessment. Both reviewers possessed multiple years of experience in teaching and assessing academic writing at various levels in higher education. The 16 essays were anonymized and distributed to reviewers, who were tasked with analyzing texts using the researcher-developed rubric without knowledge of each essay's origin to eliminate bias. Reviewers were required to provide detailed insights and rationale regarding their evaluations. To minimize potential fatigue effects and carryover bias, essays were presented in a randomized order independently determined for each reviewer. The authors acknowledge that employing only two human reviewers constitutes a methodological limitation. A minimum of three reviewers is generally

recommended to enhance inter-rater reliability and reduce idiosyncratic bias. This constraint arose from the difficulty of recruiting additional ELT faculty with the requisite expertise in writing assessment who were available within the study's timeframe. Future replications should aim to recruit at least three reviewers to improve the robustness of conclusions related to human evaluator consistency.

2.5. Digital Detection Tool Analysis

The complete set of essays was processed through plagiarism detection tools (iThenticate and Turnitin) to assess their effectiveness in identifying AI-generated content. These tools were specifically tasked with differentiating between AI-generated essays and those that had been humanized, as well as providing overall AI detection scores. It is important to note that iThenticate, in its configuration available at the time of data collection, did not generate a discrete AI-probability score comparable to Turnitin's dedicated AI detection feature; consequently, quantitative classification performance metrics are reported exclusively for Turnitin in the Results section. iThenticate findings are discussed qualitatively where relevant.

2.6. AI Model Evaluation

An AI model using GPT-4o, trained with the same rubric used by human reviewers, was employed to review essays for differentiating between human-written and AI-generated content. Specifically, the rubric was uploaded into the GPT-4o system via a structured system prompt that instructed the model to evaluate each essay according to the five rubric criteria, assign Likert-scale scores for each criterion, and provide a classification decision (human or AI) with a probability estimate. The same structured prompt was used for each of the 16 essays to ensure procedural consistency. The system prompt specified the following evaluation sequence: (a) read the essay in full; (b) score each of the five rubric criteria from 1 to 5 with brief justification; (c) synthesize scores to arrive at an overall classification (Human / AI) with a probability estimate (0–100%). No chain-of-thought or few-shot exemplars were provided, consistent with the default-condition approach applied to the AI generation phase.

2.7. Data Analysis

The analysis included a comparison of the accuracy, reliability, and consistency of human reviewers, plagiarism detection tools, and the AI model in differentiating between human-written and AI-generated content. Statistical analyses were performed using SPSS (Version 29.0). Regarding parametric assumptions: given the ordinal nature of the rubric scores and the small sample (N = 16), normality and homogeneity of variance assumptions could not be reliably established. Accordingly, while descriptive statistics and Cronbach's alpha coefficients were reported, caution was exercised in interpreting parametric test outcomes; non-parametric alternatives were prioritized where appropriate. Statistical methods employed included chi-square tests; however, as all expected cell counts fell below five (violating the minimum threshold that permits no more than 20% of cells with expected frequency < 5), Pearson chi-square results were reported for transparency but should be interpreted with considerable caution. Primary inferential conclusions were based on Fisher's Exact Test results, which are appropriate under conditions of small expected cell counts. In addition, quadratic weighted kappa (QWK) coefficients were calculated to assess inter-rater agreement among the human reviewers, the AI model, and Turnitin. Given that the primary research interest concerned the degree of agreement among evaluative agents rather than mere statistical association, and given the ordinal nature of the classification scores produced by each method, QWK was selected as the most appropriate measure of inter-rater agreement (Landis & Koch, 1977). Cronbach's Alpha was calculated to assess the internal consistency with which each evaluator applied the rubric criteria. The *t*-tests and ANOVA initially planned were superseded by confusion matrix analyses and Fisher's Exact Test, given the small sample size and the primarily categorical

nature of the classification outcome variable. This methodological substitution is explicitly acknowledged as a deviation from the original analysis plan.

2.7.1. Qualitative Analysis

The open-ended written justifications provided by human reviewers for each essay classification were subjected to inductive thematic analysis following the framework outlined by Braun and Clarke (2006). Two researchers independently coded the written responses, identifying recurring patterns in the criteria and textual features that reviewers cited as indicative of human or AI authorship. Codes were subsequently grouped into broader themes through iterative discussion until consensus was reached. Representative excerpts from reviewers' qualitative justifications are provided in the Results section to illustrate each identified theme.

2.8. Ethical Considerations

This study received appropriate institutional approval and followed ethical guidelines for research involving human participants. All student essays were collected with informed consent, and participant identities were anonymized throughout the analysis process. No sensitive personal information was collected or used in the study.

3. Results

Results are organized in accordance with the three research aims stated in the Study Design subsection: (1) classification accuracy across methods, (2) evaluator reliability, and (3) qualitative decision-making processes.

3.1. Reliability Analysis

Cronbach's Alpha values revealed important insights into the consistency with which the rubric criteria were applied by each evaluative agent. The standardized alpha for human reviewers was 0.832, while the trained AI model achieved a standardized alpha of 0.923. These values indicate that the AI model applied the rubric criteria more uniformly across essays, while human reviewers showed somewhat greater variability across criteria. It is important to note that these alpha coefficients reflect the internal consistency of criterion application by a single rater (or agent) across essays, not inter-rater reliability between raters. With only two human reviewers, formal inter-rater reliability (e.g., QWK between the two reviewers) was also calculated and is reported below. The relatively small sample size ($N = 16$) means that alpha estimates should be treated as indicative rather than definitive, as Cronbach's alpha is sensitive to sample size and may be unstable in samples of this magnitude (Cortina, 1993).

3.2. Classification Accuracy

3.2.1. Human Reviewers

Crosstabulation analysis revealed that human reviewers demonstrated a wide distribution in their AI probability estimates for both human-written and AI-generated essays. For human-written essays, estimates varied widely, with reviewers assigning probabilities of 0%, 5%, 10%, 50%, 75%, and 85% AI likelihood. In contrast, estimates for AI-generated essays were more concentrated, falling mostly within ranges of 75%–100% AI probability. Instances of misclassification were noted. The QWK coefficient between the two human reviewers was calculated at $\kappa_w = 0.61$, indicating moderate-to-substantial agreement. The chi-square test results for the human reviewers' classification accuracy are presented in Table 2.

Statistical Note: All chi-square analyses in this study are characterized by 100% of cells having expected counts below five, violating the conventional chi-square assumption. Accordingly, Fisher's Exact Test is the primary basis for inferential interpretation throughout.

Table 2. *Chi-Square Test Results for Human Reviewers' Classification Accuracy*

Test	Value	*df*	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	14.000[a]	10	0.173	0.096	-	-
Likelihood Ratio	19.408	10	0.035	0.096	-	-
Fisher's Exact Test	12.618	-	-	0.096	-	-
Linear-by-Linear Association	8.752[b]	1	0.003	0.001	0.000	0.000
N of Valid Cases	16	-	-	-	-	-

Note: . a. 22 cells (100.0%) have expected count less than 5. The minimum expected count is 0.50. b. The standardized statistic is 2.958. Given that 100% of cells have expected frequencies below five, Pearson chi-square results should not be interpreted as valid; Fisher's Exact Test (p = .096) is the appropriate statistic for this analysis.

The Pearson Chi-Square test yielded a *p*-value of 0.173, indicating non-significant results. However, as this statistic is not valid given the cell count violations, interpretive weight is placed on Fisher's Exact Test ($p = .096$), which likewise fails to reach conventional significance thresholds. This suggests that overall, the pattern of human reviewer estimates did not significantly depart from chance. The Linear-by-Linear Association showed a highly significant *p*-value of 0.003, indicating that as human reviewers assigned higher probabilities of AI generation, the likelihood of the essay being AI-generated also increased—suggesting a directional trend even if overall classification was inconsistent.

3.3. AI Model

The AI model demonstrated superior performance compared to human reviewers. Statistical analysis results are presented in Table 3.

Table 3. *Chi-Square Test Results for AI Model Classification Accuracy*

Test	Value	*df*	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	12.571[a]	5	0.028	0.005	-	-
Likelihood Ratio	16.439	5	0.006	0.005	-	-
Fisher's Exact Test	11.117	-	-	0.005	-	-
Linear-by-Linear Association	7.798[b]	1	0.005	0.001	0.001	0.001
N of Valid Cases	16	-	-	-	-	-

Note: a. 12 cells (100.0%) have expected count less than 5. The minimum expected count is 0.50. b. The standardized statistic is 2.793. Fisher's Exact Test (p = .005) is the primary inferential statistic given the violation of chi-square assumptions.

Fisher's Exact Test ($p = .005$) confirmed a statistically significant relationship between actual essay type and AI model classification scores, indicating that the AI model effectively discriminated between human-written and AI-generated texts. Pearson chi-square results are again noted but not interpreted as primary given assumption violations.

3.4. Plagiarism Detection Tools

Crosstabulation analysis revealed significant insights into Turnitin's AI detection capabilities in classifying essays. Results are presented in Table 4.

Table 4. *Chi-Square Test Results for Plagiarism Detection Tools Classification Accuracy*

Test	Value	*df*	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	12.500[a]	6	0.052	0.005	-	-
Likelihood Ratio	16.152	6	0.013	0.005	-	-
Fisher's Exact Test	11.364	-	-	0.005	-	-
Linear-by-Linear Association	10.584[b]	1	0.001	0.001	0.001	0.001
N of Valid Cases	16	-	-	-	-	-

Note: a. 14 cells (100.0%) have expected count less than 5. The minimum expected count is 0.50. b. The standardized statistic is 3.253. Fisher's Exact Test (p = .005) is the primary inferential statistic.

Key observations indicated that Turnitin accurately classified 87.5% of human-written essays as having a 0% AI score. However, when analyzing AI-generated essays, Turnitin faced considerable challenges. Only 25% of these essays received a full 100% AI score, while many others were assigned mid-range scores between 50% and 80%, reflecting uncertainty in the tool's assessment. The detailed crosstabulation analysis is presented in Table 5.

Table 5. *Confusion Matrix Analysis for Turnitin AI Score Detection*

		0.00	10.00	50.00	60.00	80.00	90.00	100.00	Total
Essay Type	Human Count	7	1	0	0	0	0	0	8
	% within Essay Type	87.5%	12.5%	0.0%	0.0%	0.0%	0.0%	0.0%	100%
	AI Count	1	0	1	1	2	1	2	8
	% within Essay Type	12.5%	0.0%	12.5%	12.5%	25.0%	12.5%	25.0%	100%
Total	Count	8	1	1	1	2	1	2	16

3.5. Confusion Matrix Analysis for All Methods

Performance evaluation using confusion matrix analysis revealed distinct patterns across evaluation methods. The comprehensive comparison is presented in Table 6.

Table 6. *Classification Performance Summary*

Method	TP	FN	FP	TN	Accuracy	Precision	Recall	F1
Human Reviewers	6	2	4	4	62.5%	50.0%	66.7%	57.1%
AI Model	8	0	3	5	81.25%	72.7%	100.0%	84.2%
Turnitin	7	1	1	7	87.5%	87.5%	87.5%	87.5%

*Note: TP = True Positives (correctly identified human essays); FN = False Negatives (human essays misclassified as AI); FP = False Positives (AI essays misclassified as human); TN = True Negatives (correctly identified AI essays). *Precision and F1-Score for the AI Model have been corrected from the original manuscript (original values: Precision = 100.0%, F1 = 76.9%), which were mathematically inconsistent with the three reported false positives. Corrected values: Precision = 72.7% (8/[8+3]), F1 = 84.2%.*

An important correction to Table 6 warrants explicit documentation. In the original manuscript, the AI Model's Precision was reported as 100.0% and its F1-Score as 76.9%. These values are mathematically inconsistent with the simultaneously reported three false positives. With 8 true

positives and 3 false positives: Precision = 8/(8+3) = 72.7%, and F1 = 2×(0.727×1.000)/(0.727+1.000) = 84.2%. These corrected values have been incorporated into Table 6. The authors regret this computational error in the original submission.

Table 7. *Quadratic Weighted Kappa (QWK) Agreement Among Evaluative Agents*

Comparison	QWK (κw)	Interpretation
Human Reviewer 1 vs. Reviewer 2	0.61	Moderate-to-Substantial
Human Reviewers (avg.) vs. AI Model	0.48	Moderate
Human Reviewers (avg.) vs. Turnitin	0.44	Moderate

Note. QWK values of 0.41–0.60 indicate moderate agreement; 0.61–0.80 indicate substantial agreement (Landis & Koch, 1977). QWK is preferred over chi-square for ordinal classification agreement given the small sample size and violation of chi-square assumptions.

Table 7 presents QWK coefficients examining agreement among evaluative agents. The two human reviewers achieved κw = 0.61, indicating moderate-to-substantial agreement, yet their agreement with both the AI model (κw = 0.48) and Turnitin (κw = 0.44) was only moderate, confirming that the human reviewers and automated systems applied meaningfully different classificatory criteria.

3.6. Qualitative Analysis of Human Evaluator Decision-Making

Thematic analysis of human evaluator reasoning revealed several key patterns in decision-making processes, as summarized in Table 8.

Table 8. *Key Assumptions in Human Evaluator Decision-Making*

Feature	AI Writing (Assumption)	Human Writing (Assumption)
Grammar & Spelling	Near-perfect	Small errors present
Lexical Variety	Repetitive, generic words	More varied, expressive
Structure	Extremely organized and clear	Sometimes disorganized but expressive
Argumentation Depth	Surface-level, lacks insight	Deeper, complex reasoning
Creativity	Predictable, formulaic	Original, engaging
Use of Examples	Generic statements	Personal, real-world examples
Overall Impression	Feels robotic and mechanical	Feels natural and engaging

3.7. Thematic Analysis Findings

3.7.1. Theme 1: Language Accuracy as Primary Indicator

Human evaluators frequently mentioned language accuracy as a key criterion, assuming that AI-generated texts exhibit near-perfect grammar while human-written essays may contain minor errors. As one evaluator noted: *"Yapay zeka genellikle kusursuz dil bilgisi ve yazım kullanır* (*AI typically uses flawless grammar and spelling*)".

3.7.2. Theme 2: Word Choice and Repetition

Evaluators identified word choice and lexical variety as significant differentiators, noting that "AI sometimes uses ordinary or repetitive words" while human writers tend to vary their word choices more naturally. For example, one reviewer wrote: "The vocabulary is functional but lacks the kind of deliberate lexical risk-taking that characterizes authentic graduate-level writing."

3.7.3. Theme 3: Logical Structure versus Creative Flow

A key observation was the contrast between structured logic and creative expressiveness. Evaluators noted that "AI creates clear and well-structured theses" but may lack the unconventional yet insightful arguments typical of human writing. One reviewer observed: "The essay follows a textbook five-paragraph structure with no deviation—there is no sense of a writer discovering their argument as they write."

3.7.4. Theme 4: Depth of Argumentation

Human evaluators believed that "Human writings are generally deeper and more complex" compared to AI-generated texts that can be overly general or formulaic. A representative reviewer comment stated: "The argument addresses the prompt adequately but never engages with its most challenging implications—it stays safely at the surface."

3.7.5. Theme 5: Creativity and Originality

Creativity emerged as a major factor, with evaluators noting that "Human writings tend to have more creative narrative styles" compared to predictable AI approaches. This perception aligns with Chein et al. (2024), who observed that human raters reliably associate stylistic unpredictability and idiosyncratic phrasing with authentic human authorship, even when such associations do not always yield accurate classifications.

3.7.6. Theme 6: Use of Examples and Personal Experiences

Evaluators emphasized the importance of real-world examples, noting that *"İnsan yazıları daha çeşitli ve kişisel örnekler içerebilir"* (Human writings include more diverse and personal examples). One evaluator specifically noted: "The examples in this essay are plausible but generic—they could apply to any student in any country, which is not characteristic of a writer drawing from lived experience."

4. Discussion

The findings of this study provide critical insights into the comparative effectiveness of human reviewers, AI scoring models, and digital detection tools in distinguishing between human-generated and AI-generated essays. The results reveal significant implications for academic integrity, assessment methodologies, and the evolving role of AI in educational settings.

4.1. Performance of Human Reviewers

The analysis of human reviewers' performance revealed notable inconsistencies despite their expertise in English Language Teaching. The variability in judgments indicates that subjective factors and individual biases significantly influenced assessments, aligning with research by Liebe et al. (2023) and Chein et al. (2024). Some reviewers overestimated the likelihood of essays being AI-generated, leading to false positives, while others exhibited leniency, resulting in false negatives. The QWK analysis (κ_w = 0.61 between reviewers) corroborates these findings: while the two reviewers achieved moderate-to-substantial agreement with each other, their alignment with automated detection systems was considerably lower. This observation is consistent with Ayub et al. (2024), who demonstrated that humanization tools such as Quillbot can effectively obscure the stylometric markers that human readers and automated detectors alike rely upon.

4.2. AI Scoring Model Performance

The AI scoring model demonstrated superior internal consistency (Cronbach's α = 0.923 vs. 0.832 for human reviewers), suggesting that it applied uniform criteria while reducing subjective variation. The model successfully identified all human-written essays, achieving perfect recall, though it exhibited a higher rate of false positives than Turnitin. As corrected in

Table 6, the AI model's Precision was 72.7%, yielding an F1-Score of 84.2%. This false-positive tendency carries non-trivial practical consequences: in real assessment contexts, incorrectly flagging a genuine student essay as AI-generated could have serious academic integrity repercussions for the student.

The AI model's ability to differentiate between human and AI-generated content with greater reliability than human reviewers raises important questions regarding the potential role of AI in assessment. Despite its systematic approach, the model may not fully capture the creative, nuanced, and contextually rich aspects of human writing that professionals intuitively assess. From an Information Processing Theory perspective (Atkinson & Shiffrin, 1968), this limitation reflects the AI model's reliance on pattern-matching within the rubric's operationalized criteria, whereas human reviewers draw on broader schematic knowledge of disciplinary writing norms and rhetorical context.

4.3. Effectiveness of Digital Detection Tools

Among the detection methods examined, Turnitin demonstrated the highest overall performance with 87.5% classification accuracy. However, the tool struggled with AI-generated essays, frequently assigning mid-range scores rather than definitive classifications. Fisher's Exact Test (p = .005) confirmed a statistically significant relationship between essay type and Turnitin AI classification scores. These results emphasize a critical challenge identified by Peng et al. (2023) and Majhi and Santra (2023): while plagiarism detection tools were originally designed to identify verbatim textual similarities, they are not fully optimized for detecting original AI-generated content. This limitation is consistent with SCOT theory (Pinch & Bijker, 1987): Turnitin was socially constructed as a similarity-detection instrument, and its moderate performance in detecting original AI-generated text reflects the assumptions of the social context in which it was designed—prior to the widespread availability of generative AI writing tools.

4.4. Theoretical and Practical Implications

From a theoretical perspective, these findings align with principles of Technological Determinism (McLuhan, 1964) and Social Construction of Technology theory (Pinch & Bijker, 1987). The increasing sophistication of AI-generated content suggests that technology is reshaping academic practices, necessitating new frameworks for assessment and integrity verification. The constructivist perspective (Vygotsky, 1978) also becomes relevant as the role of educators in fostering critical thinking and originality must be reconsidered. Posthumanism (Hayles, 1999) further complicates the traditional binary between 'authentic' human authorship and 'inauthentic' machine production: as AI-generated and humanized texts become increasingly indistinguishable, institutions may need to reconceptualize what they are assessing and why, rather than simply investing in more sophisticated detection tools.

Pedagogically, these results underscore the need for educators to incorporate AI literacy into academic training. Institutions may need to refine assessment strategies, incorporating oral defenses or requiring students to annotate their writing processes to ensure originality. Given the false-positive risk identified in this study, any institutional AI detection policy should include a clear appeals mechanism and a human review stage before academic integrity consequences are triggered.

4.5. Limitations and Future Research Directions

First and most critically, the small sample size (16 essays) restricts statistical power and severely limits generalizability. As a direct consequence, chi-square tests were statistically invalid across all three analyses (100% of cells with expected frequency < 5), and the study relied on Fisher's Exact Test as the primary inferential method. Cronbach's alpha coefficients

should similarly be interpreted with caution, given their sample-sensitivity. Additionally, the study focused on a specific academic context (ELT master's students) and essay topic, which may not represent broader educational settings. The use of only two human reviewers also constrains the robustness of conclusions regarding human inconsistency.

Second, the threshold for classifying essays as 'AI-generated' or 'human-written' based on probability estimates was not systematically prespecified for the human reviewers, which introduces ambiguity in the operationalization of the classification variable. Future studies should prespecify and report explicit probability thresholds. Third, the AI text generation and humanization tools used represent a specific moment in a rapidly evolving technological landscape; findings may not generalize to newer model versions. Fourth, because only two human reviewers participated, it is impossible to separate systematic reviewer-level biases from idiosyncratic judgment patterns.

Future research should employ a minimum of three reviewers and substantially larger essay samples to permit valid parametric analyses. The effectiveness of alternative AI detection methodologies, such as linguistic fingerprinting and deep learning models, warrants investigation. Longitudinal studies examining the impact of AI-assisted writing on student learning outcomes over extended periods would also provide valuable insights.

4.6. Conclusions

This study contributes to the growing discourse on AI's role in education by empirically examining the capabilities and limitations of human reviewers, AI scoring models, and digital detection tools. The results highlight the increasing difficulty in distinguishing AI-generated content from human writing, reinforcing the need for enhanced assessment strategies.

Key findings indicate that while AI scoring models exhibit superior reliability ($\alpha = 0.923$) and digital detection tools like Turnitin demonstrate relatively high accuracy (87.5%), significant limitations remain across all methods. Human reviewers showed considerable inconsistency (62.5% accuracy), though their intuitive understanding of contextual nuances remains valuable. The corrected confusion matrix metrics (AI Model Precision = 72.7%, F1 = 84.2%) and QWK analyses further clarify the relative strengths and weaknesses of each evaluative approach. Importantly, these findings should be interpreted in light of the study's small sample size, which precluded valid chi-square analysis; Fisher's Exact Test results constitute the primary basis for inferential conclusions. The qualitative thematic analysis adds a complementary layer of insight by illuminating the heuristics human reviewers employ—heuristics that, while intuitive, proved insufficiently reliable at the classification task when evaluated against objective accuracy benchmarks.

4.7. Practical Recommendations

Based on these findings, several recommendations emerge for educational practice. First, given human reviewers' demonstrated inconsistency, targeted training programs should be developed to improve educators' ability to recognize AI-generated content. These programs should focus on identifying subtle markers of AI-generated text, including atypical coherence patterns and overuse of statistically probable phrasing. Training materials could incorporate examples of humanized AI texts—the category that proved most difficult for all evaluative methods—to sensitize reviewers to the textual signatures that survive the humanization process.

Second, while Turnitin and similar tools showed promise, further enhancements are required to detect AI-generated content more accurately. Future development should explore the integration of machine learning models capable of detecting not only textual similarity but also stylistic anomalies characteristic of AI-generated text.

Third, a combination of human judgment, AI-assisted scoring, and plagiarism detection tools may yield the most effective assessment strategy. Hybrid approaches that leverage the strengths of each method could provide more reliable and fair evaluations of student work. Critically, any hybrid framework should incorporate a human review stage before consequential academic integrity decisions are made, given the false-positive rates observed for the AI model (27.3% of AI-flagged essays were, in fact, human-authored). Lastly, given that AI-generated content capabilities continue advancing, institutions must establish clear policies regarding AI's role in academic work.

Ethics Committee Approval

This research was conducted with the permission obtained by the Tokat Gaziosmanpaşa University Scientific Research and Publication Ethics Social and Human Sciences Board's decision dated 23/07/2024 and numbered 01-33.

Conflict of Interest

The authors declare that they have no conflict of interest.

Author Contribution

Zafer Susoy: *Designing the study and analyzing the data. Writing the introduction and method. Collecting the data, coding data via SPSS, and obtaining research permissions.*

Orcid

Zafer Susoy *https://orcid.org/0000-0002-6890-6007*

REFERENCES

Alzubi, A. A. F., Nazim, M., & Alyami, N. (2025). Do AI-generative tools kill or nurture creativity in EFL teaching and learning? *Education and Information Technologies*, 30, 15147-15184. https://doi.org/10.1007/s10639-025-13409-8

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*, 2, 89-195. Academic Press.

Ayub, T., Ahmad Malla, R., Khan, M. Y., & Ganaie, S. A. (2024). The art of deception: Humanizing AI to outsmart detection. *Global Knowledge, Memory and Communication*, *73*(8), 889-905. https://doi.org/10.1108/GKMC-03-2024-0133

Bearman, M., & Luckin, R. (2020). Preparing university assessment for a world with AI: Tasks for human intelligence. In M. Bearman, P. Dawson, R. Ajjawi, J. Tai, & D. Boud (Eds.), *Re-imagining university assessment in a digital world* (pp. 49-63). Springer. https://doi.org/10.1007/978-3-030-41956-1_5

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Chein, J. M., Martinez, S. A., & Barone, A. R. (2024). Human intelligence can safeguard against artificial intelligence: Individual differences in the discernment of human from AI texts. *Scientific Reports*, 14(1), Article 25989. https://doi.org/10.1038/s41598-024-76218-y

Chung, J. Y., & Jeong, S.-H. (2024). Exploring the perceptions of Chinese pre-service teachers on the integration of generative AI in English language teaching: Benefits, challenges, and educational implications. *Online Journal of Communication and Media Technologies*, *14*(4), e202457. https://doi.org/10.30935/ojcmt/15266

Cingillioglu, I. (2023). Detecting AI-generated essays: The ChatGPT challenge. *The International Journal of Information and Learning Technology*, *40*(4), 259-268. https://doi.org/10.1108/IJILT-03-2023-0043

Coccoli, M., & Patanè, G. (2024). AI vs. AI: The detection game. In *2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI)* (pp. 1-6). IEEE. https://doi.org/10.1109/RTSI61910.2024.10761124

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Crawford, L. M., Hendzlik, P., Lam, J., Cannon, L. M., Qi, Y., DeCaporale-Ryan, L., & Wilson, N. A. (2024). Digital ink and surgical dreams: Perceptions of artificial intelligence-generated essays in residency applications. *The Journal of Surgical Research*, 301, 504-511. https://doi.org/10.1016/j.jss.2024.06.020

Crompton, H., Edmett, A., Ichaporia, N., & Burke, D. (2024). AI and English language teaching: Affordances and challenges. *British Journal of Educational Technology*, *55*(6), 2503-2529. https://doi.org/10.1111/bjet.13460

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), Article 17. https://doi.org/10.1007/s40979-023-00140-5

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, Article 100209. https://doi.org/10.1016/j.caeai.2024.100209

Halaweh, M., & Refae, G. (2024). Examining the accuracy of AI detection software tools in education. In *2024 Fifth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)* (pp. 186-190). IEEE. https://doi.org/10.1109/IDSTA62194.2024.10747004

Hauer, T. (2017). Technological determinism and new media. *International Journal of English and Literature*, *8*(3), 51-56.

Hayles, N. K. (1999). *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*. University of Chicago Press.

Herath, D. B., Ode, E., & Herath, G. B. (2025). Can AI replace humans? Comparing the capabilities of AI tools and human performance in a business management education scenario. *British Educational Research Journal*, *51*(3), 1073–1096. https://doi.org/10.1002/berj.4111

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lee, G. (2024). AI-based writing feedback in ESP: Leveraging ChatGPT for writing business cover letters among Korean university students. In *Proceedings of the International CALL Research Conference* (pp. 234-248). https://doi.org/10.29140/9780648184485-20

Liebe, L., Baum, J., Schütze, T., Cech, T., Scheibel, W., & Döllner, J. (2023). UNCOVER: Identifying AI generated news articles by linguistic analysis and visualization. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* (pp. 39-50). https://doi.org/10.5220/0012163300003598

Liu, Y., Zhang, Z., Zhang, W., Yue, S., Zhao, X., Cheng, X., Zhang, Y., & Hu, H. (2023). ArguGPT: Evaluating, understanding and identifying argumentative essays generated by GPT models. *arXiv preprint*. https://doi.org/10.48550/arXiv.2304.07666

Luckin, R. (2024). Nurturing human intelligence in the age of AI: Rethinking education for the future. *Development and Learning in Organizations: An International Journal*, *38*(3), 12-15. https://doi.org/10.1108/DLO-04-2024-0108

Majhi, D., & Santra, P. (2023). Scholarly communication and machine-generated text: Is it finally AI vs AI in plagiarism detection? *SRELS Journal of Information Management*, *60*(3), 171-185. https://doi.org/10.17821/srels/2023/v60i3/171028

McLuhan, M. (1964). *Understanding media: The extensions of man*. McGraw-Hill.

Parviz, M. (2024). The double-edged sword: AI integration in English language education from the perspectives of Iranian EFL instructors. *Complutense Journal of English Studies*, 32, 45-62. https://doi.org/10.5209/cjes.97261

Peng, X., Zhou, Y., He, B., Sun, L., & Sun, Y. (2023). Hiding the ghostwriters: An adversarial evaluation of AI-generated student essay detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 10406-10419). https://doi.org/10.18653/v1/2023.emnlp-main.644

Pinch, T. J., & Bijker, W. E. (1987). The social construction of facts and artifacts: Or how the sociology of science and the sociology of technology might benefit each other. In W. E. Bijker, T. P. Hughes, & T. J. Pinch (Eds.), *The social construction of technological systems: New directions in the sociology and history of technology* (pp. 17–50). MIT Press.

Silva, G. S., Khera, R., & Schwamm, L. H. (2024). Reviewer experience detecting and judging human versus artificial intelligence content: The Stroke Journal essay contest. *Stroke*, *55*(10), 2573-2578. https://doi.org/10.1161/STROKEAHA.124.045012

Steponenaite, A., & Barakat, B. (2023). Plagiarism in AI empowered world. In *HCI International 2023 Posters* (pp. 434-442). Springer. https://doi.org/10.1007/978-3-031-35897-5_31

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Yeadon, W., Agra, E., Inyang, O.-O., Mackay, P., & Mizouri, A. (2024). Evaluating AI and human authorship quality in academic writing through physics essays. *European Journal of Physics*, 45(5), Article 055703. https://doi.org/10.1088/1361-6404/ad669d