International Journal of Research in English Education

Online ISSN: 2538-3027

Homepage: https://ijreeonline.com

Pitfalls of a Second Language Writer and AI Detection: Evaluating Plagiarism Checking Tools in Iranian EFL Students' Academic Writings

Mohammad Amin Dehghan Baghi[1], **Seyedeh Elham Elhambakhsh**[2]✉, and **Ali Mohammad Fazilatfar**[3]

1. Department of English Language and Literature, Faculty of Foreign Language, Yazd University, Yazd, Iran. Email: aminnima@gmail.com
2. Corresponding author, Department of English Language and Literature, Faculty of Foreign Language, Yazd University, Yazd, Iran. Email: s.e.elhambakhsh@yazd.ac.ir
3. Department of English Language and Literature, Faculty of Foreign Language, Yazd University, Yazd, Iran. Email: afazilatfar@yazd.ac.ir

Article Info

Article type:
Research Article

Article history:

Received:
June 12, 2026

Received in revised form:
September 20, 2026

Accepted:
September 29, 2026

Available online:
September 30, 2026

***Keywords*:**
AI detection tools,
Iranian EFL students,
false negatives,
false positives,
plagiarism

ABSTRACT

obective: This study evaluated AI detection tools' efficacy in assessing EFL students' writing, focusing on distinguishing AI-generated from human text without misclassifying L2 work.

Methods: A two-phase quantitative design was used. Phase 1 involved 30 BA students from Yazd University producing one non-AI and one ChatGPT-assisted essay on identical topics. Phase 2 compared two MA thesis corpora: 30 from the pre-AI era (2010–2015) and 30 from the post-AI era (2020–2025). All texts were analyzed using Turnitin and Grammarly Premium.

Results: Results, analyzed via Wilcoxon Signed-Rank Tests and one-sample t-tests, revealed significant performance divergence. Turnitin demonstrated near-perfect accuracy in identifying AI-assisted essays, whereas Grammarly showed a statistically significant false negative rate, failing to detect some AI content. Both tools accurately analyzed short, human-written essays. However, the thesis analysis revealed increased AI-detection scores in the post-AI corpus, contrasting with near-zero scores in the pre-AI corpus.

Conclusion: This finding suggests either actual, undisclosed AI use or algorithmic bias misclassifying Iranian EFL linguistic patterns as AI-generated. The study concludes that while Turnitin is more robust than Grammarly, uncritical tool application risks systematic misjudgment against L2 writers. False positives and negatives threaten academic equity. Thus, AI-detection outputs should trigger pedagogical dialogue and human-led investigation rather than serving as sole misconduct evidence.

Keywords: AI detection tools, Iranian EFL students, false negatives, false positives, plagiarism

Cite this article: Dehghan Baghi, MA., Elhambakhsh, E., and Fazilatfar, AM. (2026). Pitfalls of a Second Language Writer and AI Detection: Evaluating Plagiarism Checking Tools in Iranian EFL Students' Academic Writings. ***International Journal of Research in English Education***, *11*(3), 139-155 .

© The author(s) retain the copyright.
Publisher: Science Academy Publications.

[Downloaded from mail.ijreeonline.com on 2026-10-30]

1. Introduction

The rapid integration of Generative Artificial Intelligence (GenAI), particularly large language models (LLMs) like ChatGPT, into the educational landscape represents a paradigm shift with profound implications for academic writing and assessment. These tools offer unprecedented support to language learners, facilitating idea generation, drafting, and editing (Mizumoto et al., 2024; Yuan et al., 2024). However, this technological advancement is a double-edged sword, introducing complex challenges to principles of academic integrity. In response, educational institutions are increasingly deploying AI detection software as technological gatekeepers to preserve authenticity and originality.

1.1 Statement of the Problem

A critical, yet underexplored, issue in this dynamic is the diagnostic validity and contextual equity of automated detection software when applied to non-native English writing. A growing consensus demonstrates that algorithmic classifiers frequently struggle to evaluate L2 academic discourse accurately (Chaka, 2024a; Weber-Wulff et al., 2023). The unique linguistic and rhetorical profiles of L2 writers—characterized by formulaic phrasing, predictable phraseology, and reduced lexical variation—deviate from native corpora, frequently triggering false-positive accusations of AI generation (Liang et al., 2023). This technological vulnerability poses a severe threat to educational fairness by disproportionately penalizing authentic non-native scholarship.

This crisis is particularly acute in higher education contexts such as Iran, where postgraduate scholars face immense institutional pressure to publish in indexed venues while navigating Persian-English cross-linguistic transfer and conventional disciplinary norms (Ahmadi, 2014; Saeedi & Baleghizadeh, 2026). While international benchmarks have raised alarms regarding algorithmic bias, empirical scrutiny within the Iranian EFL context remains scarce. In the absence of calibrated local thresholds and institutional guidelines (Hosseini & Ebadi, 2025), Iranian graduate students are vulnerable to automated gatekeeping tools that conflate developmental phraseological dependence with illicit AI generation (Asadi et al., 2025). Therefore, an empirical evaluation of how these detection mechanisms process localized EFL corpora is an urgent imperative.

1.2 Objectives of the Study and Research Questions

This study directly addresses this gap by investigating the diagnostic performance of two widely deployed tools, Turnitin and Grammarly Premium, in evaluating academic texts produced by Iranian EFL postgraduate students. Specifically, the investigation pursues the following research questions:

1. How do Turnitin and Grammarly compare in their accuracy (including false-positive and false-negative rates) in detecting AI-generated essays by Iranian EFL students?

[Downloaded from mail.ijreeonline.com on 2026-10-30]

2. How do Turnitin and Grammarly compare in their accuracy (including false-positive and false-negative rates) in identifying human-written essays by Iranian EFL students?
3. What differences exist in the AI-detection scores generated by Turnitin and Grammarly for Iranian MA theses written in a pre-AI era (2010–2015) compared to those written in a post-AI era (2020–2025)?

By problematizing the uncritical institutional reliance on automated classifiers, this study provides actionable empirical evidence to inform pedagogical policies, mitigate automated assessment bias, and safeguard academic integrity in the generative AI era.

2. Literature Review

The proliferation of sophisticated Large Language Models (LLMs) like ChatGPT has altered the academic landscape, presenting both unprecedented opportunities for learning and formidable challenges to traditional notions of authorship and integrity. This review synthesizes the relevant literature across three key areas: the established challenges of second language (L2) writing, the evolving definition of plagiarism in the age of AI, and the current state of AI detection technologies, ultimately identifying a critical gap in the context of Iranian EFL learners.

2.1 The Unique Challenges of Second Language Writing

Academic writing in a second language is a complex endeavor, fraught with linguistic, cognitive, and sociocultural hurdles. L2 writers often grapple with limited vocabulary, grammatical inaccuracies, and difficulties with sentence structure, which can affect the overall fluency and complexity of their text (Hamdani & Abid, 2023). Beyond sentence-level issues, they face significant challenges in organizing ideas logically, maintaining coherence, and adhering to the specific rhetorical conventions of English academic discourse (Hyland, 2007; Swales & Feak, 2004). These linguistic and rhetorical features collectively form a distinct "fingerprint" that differentiates L2 writing from that of native speakers. This fingerprint is not indicative of a lack of effort or originality but is a natural characteristic of the language acquisition process. As this review will argue, this very fingerprint may become a liability in the face of automated AI detection.

2.2 Plagiarism and the New Frontier of AI-Generated Content

The definition of plagiarism has expanded beyond the verbatim copying or improper paraphrasing of human-authored text (Kouam, 2024). The advent of generative AI has introduced a new category: AI-generated content presented as one’s own original work (Amirzhanov et al., 2025). This evolution complicates academic integrity, as AI can produce text that is not a direct match to existing sources, thereby evading traditional plagiarism checkers that rely on textual overlap.

Research in EFL contexts, particularly in Iran, highlights that students often resort to textual borrowing due to a complex interplay of factors, including low linguistic proficiency, pressure

[Downloaded from mail.ijreeonline.com on 2026-10-30]

to publish, unfamiliarity with Western academic conventions, and sometimes, cultural attitudes toward textual ownership (Ahmadi, 2014; Pennycook, 1996; Rezanejad & Rezaei, 2013). This existing propensity, combined with the easy accessibility of AI tools, creates a perfect storm where the use and potential misuse of AI writing assistance become a significant concern for educators.

2.3 AI Detection Tools: Promise, Performance, and Proven Bias

In response, a market of AI detection services has rapidly emerged. These tools, such as Turnitin's AI writing detector and Grammarly's AI checker, utilize machine learning and natural language processing (NLP) algorithms to analyze textual features like syntactic patterns, lexical diversity, and semantic coherence to distinguish between human and AI-authored text (Chaka, 2024b; Weber-Wulff et al., 2023).

Comparative studies generally indicate that commercial, paid tools (e.g., Turnitin, Copyleaks, iThenticate) outperform free alternatives in terms of accuracy (Chaka, 2024a ; Rahman et al., 2024, 2025). For instance, Malik and Amjad (2025) found Turnitin maintained high accuracy even against paraphrased AI text. However, a critical and consistent flaw has been identified across numerous studies: a significant bias against non-native English writers. Research by Mizumoto et al. (2024) and Chaka (2024a) has demonstrated that AI detectors exhibit alarmingly high false positive rates when analyzing texts written by EFL learners, with one study reporting rates as high as 61.3%. The detectors appear to misinterpret the linguistic simplifications, rhetorical structures, and formulaic phrasing common in L2 writing as hallmarks of AI generation (Weber-Wulff et al., 2023). This bias raises profound ethical questions about equity and fairness in automated assessment.

2.4 The Shifting Paradigm: GenAI and Plagiarism Detection in EFL Academic Writing (2021–2026)

In recent years, the rapid proliferation of Large Language Models (LLMs) such as ChatGPT has fundamentally challenged traditional notions of textual borrowing, transforming academic integrity from simple cut-and-paste detection into complex algorithmic attribution (Weber-Wulff et al., 2023; Yuan et al., 2024). Conventional plagiarism detection software (PDS)—such as Turnitin and iThenticate—traditionally relied on string-matching and n-gram overlap algorithms to identify uncredited text reuse (Rahman et al., 2024, 2025). However, as Malik and Amjad (2025) and Chaka (2024a) argue, these conventional tools struggle when encountering machine-generated paraphrasing and synthetic linguistic signatures, creating severe challenges for academic gatekeepers.

In EFL contexts, the distinction between unintentional plagiarism, formulaic phraseological dependence, and illicit AI assistance has become increasingly contested. Mizumoto et al. (2024) demonstrated that non-native writers frequently produce repetitive syntactic structures and predictable lexical combinations that machine classifiers frequently misinterpret as AI-generated text, leading to disproportionate rates of false-positive accusations. Concurrently,

[Downloaded from mail.ijreeonline.com on 2026-10-30]

empirical evaluations by Chaka (2024b) and Amirzhanov et al. (2025) demonstrate that AI detection instruments display marked discrepancies in sensitivity and specificity when applied across disciplinary genres. Consequently, researchers caution against relying unquestioningly on automated similarity percentages, advocating instead for transparent institutional policies and ethical writing pedagogy that integrate critical digital literacy alongside algorithmic accountability (Hamdani & Abid, 2023; Kouam, 2024).

2.5 The Impetus for Present Study

While the global discourse on AI detection is expanding, a glaring gap exists in the literature regarding its application in specific, culturally situated EFL contexts. Previous studies have largely been conducted in broad, often Western, settings, leaving the unique linguistic and rhetorical profile of populations like Iranian EFL students underexplored. The specific syntactic structures, lexical choices, and L1 (Persian) transfer effects that characterize Iranian academic writing may interact with algorithmic detection logic in ways that are not yet understood.

Therefore, this study seeks to address this critical gap by evaluating the performance and potential biases of Turnitin and Grammarly specifically within the Iranian EFL academic context. It moves beyond broad comparisons to ask a more urgent question: In the high-stakes environment of academic evaluation, can these tools reliably distinguish between AI-assisted and human-written text without systematically penalizing the authentic, effortful, and linguistically distinct work of Iranian L2 writers? The findings will provide much-needed empirical evidence to inform pedagogical practices, institutional policies, and the development of more equitable assessment technologies.

3. Materials and Methods

3.1 Research Design

This study employed a two-phase, quantitative research design to evaluate the performance of AI detection tools on Iranian EFL students' academic writing. The first phase utilized a within-subjects experimental design. Thirty undergraduate students each produced two essays on an identical topic: one composed independently without AI assistance and one composed with the assistance of ChatGPT. This design controlled for inter-participant variability in writing proficiency and style, allowing for a direct comparison of tool outputs under different conditions.

The second phase involved a comparative corpus analysis of longer academic texts. Two independent corpora of MA theses were compiled: one from a pre-AI era (2010-2015, n=30) and one from a post-AI era (2020-2025, n=30). This comparative approach aimed to identify potential false positives by analyzing texts from a period where AI use was not feasible. The primary data for both phases consisted of the quantitative AI-detection probability scores (0-100%) generated by Turnitin and Grammarly.

3.2 Participants

[Downloaded from mail.ijreeonline.com on 2026-10-30]

For Phase 1, participants consisted of 30 Iranian BA-level EFL students from Yazd University, selected through convenience sampling. Participation was voluntary and informed consent was obtained. To control for linguistic proficiency as a confounding variable, participants were required to demonstrate an intermediate to advanced level of English, as determined by the Oxford Quick Placement Test (OQPT). Based on their scores, participants were categorized into intermediate (n=19, CEFR B1-B2) and advanced (n=11, CEFR C1-C2) groups. The sample included 8 males and 22 females, aged 18-22.

For Phase 2, the participants were the authors of the 60 selected MA theses from Yazd university. As these were archival documents, no direct demographic data was collected; however, the theses are representative of the typical population of TEFL MA students at the university during the specified time periods.

3.3 Instruments and Materials

3.3.1 Writing Prompts and Texts

Phase 1: A single prompt ("The causes and effects of divorce in Iran") was used to generate 60 essays (30 non-AI-assisted and 30 AI-assisted), each between 350-500 words.

Phase 2: The corpora consisted of 60 complete MA theses (excluding references and appendices), each between 25,000-50,000 words.

1. **AI Writing Tool:** The AI-assisted essays in Phase 1 were generated using the GPT-4 version of ChatGPT (OpenAI, 2023). Participants were instructed to use the tool for assistance with outlining, drafting, and writing, with the interaction (e.g., starting a new chat) standardized for consistency.
2. **AI Detection Tools:** The premium versions of two industry-standard platforms were used:

Turnitin: The Turnitin similarity checker with its integrated AI writing detection feature (Turnitin, LLC, 2023) was used to generate an AI Writing Score (percentage likelihood of AI generation).

Grammarly: The Grammarly Premium plagiarism detector (Grammarly, Inc., 2023) was used, which provides a percentage-based score indicating potential AI-generated content.

3. **Proficiency Test:** The Oxford Quick Placement Test (Version 1) was administered to ensure a homogeneous sample in Phase 1 and to categorize participants by proficiency level.

3.4 Data Collection Procedure

The procedure for Phase 1 began with the administration of the OQPT. Following the test, each participant wrote their first essay independently without any AI support. Subsequently, they produced a second essay on the same topic using ChatGPT for assistance, following

[Downloaded from mail.ijreeonline.com on 2026-10-30]

standardized instructions. All 60 essays were then submitted to both Turnitin and Grammarly for analysis.

For Phase 2, the 60 theses (30 pre-AI, 30 post-AI) were digitally gathered from the Yazd university's library. All five chapters of each thesis (excluding references and appendices) were submitted to both Turnitin and Grammarly, generating an AI-detection score for each document.

3.5 Data Analysis

The primary data for analysis were the numerical AI-detection scores (0-100%) from both tools. Prior to formal analysis, the data were screened for normality using skewness and kurtosis statistics. As the data for most conditions were highly skewed, non-parametric tests were employed.

To answer research questions comparing the performance of Turnitin and Grammarly on the same set of texts (e.g., AI-assisted essays, theses from a specific era), Wilcoxon Signed-Rank Tests were conducted. Effect sizes (r) were calculated to determine the magnitude of any observed differences.

To evaluate the false positive and false negative rates of the tools, a series of one-sample t-tests were conducted. The mean AI-detection score for each set of texts was compared against a known true value: 0% for confirmed human-written texts (non-AI-assisted essays, pre-AI theses) and 100% for confirmed AI-assisted texts (AI-assisted essays). A statistically significant deviation from these values would indicate a systematic error. All analyses were performed using IBM SPSS Statistics, Version 27.

4. Results

The performance of Turnitin and Grammarly was evaluated across two phases, yielding quantitative data on their accuracy, false negative rates, and false positive rates. Due to the non-normal distribution of the data, as indicated by high skewness and kurtosis values, non-parametric tests were primarily employed. The results are structured around the key comparisons from the research questions.

4.1 Performance on AI-Assisted and Non-AI-Assisted Essays

4.1.1 AI-Assisted Essays (False Negative Analysis)

A total of 30 essays written with ChatGPT assistance were analyzed. As shown in Table 1, Turnitin's detection scores were heavily clustered at the high end of the scale (M=99.33%, SD=2.17, Mdn=100%), with a range of 90% to 100%. In contrast, Grammarly's scores showed greater variability (M=90.00%, SD=13.13, Mdn=100%), with a range of 65% to 100%.

[Downloaded from mail.ijreeonline.com on 2026-10-30]

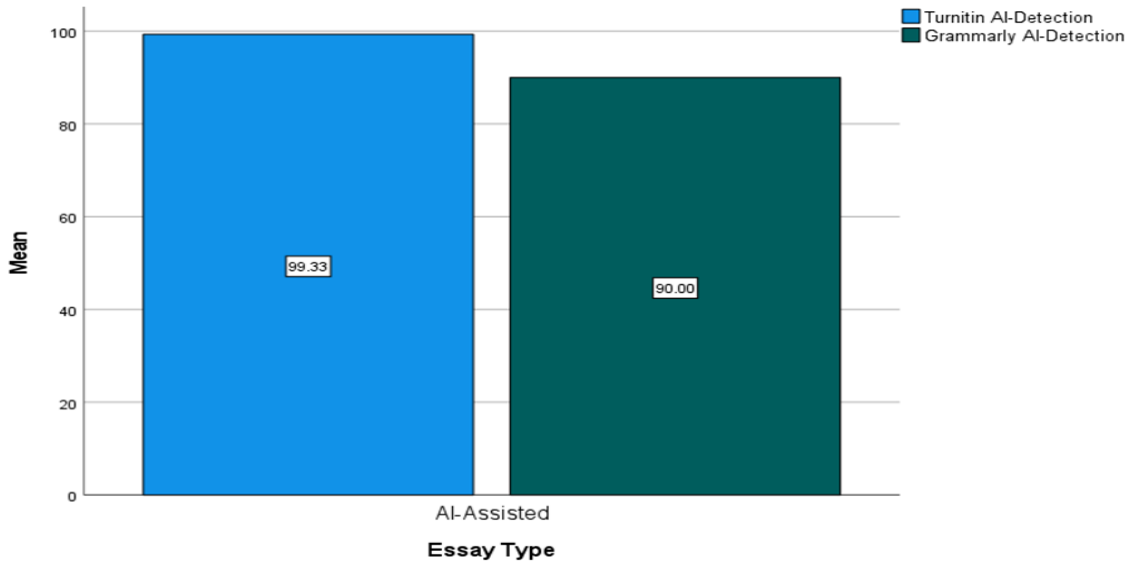

Figure 1. AI detection means for turnitin and grammarly on ai-assisted essays

Table 1. Descriptive statistics for AI-assisted essays

Tool	N	Mean%	Std. Deviation	Median	Minimum	Maximum	Skewness
Turnitin	30	99.33	2.17	100	90	100	-3.50
Grammarly	30	90	13.13	100	65	100	-0.89

A Wilcoxon Signed-Rank Test indicated that the median detection score for Turnitin was statistically significantly higher than for Grammarly, $Z = -3.18$, $p = .001$, with a large effect size ($r = .41$). To assess the false negative rate, one-sample t-tests were conducted against the test value of 100% (representing perfect detection of fully AI-written text). The results confirmed that Turnitin's mean score was not significantly different from 100%, $t(29) = -1.68$, $p = .103$. However, Grammarly's mean score was significantly lower than 100%, $t(29) = -4.17$, $p < .001$, indicating a statistically significant and substantial false negative rate for Grammarly.

4.2 Non-AI-Assisted Essays (False Positive Analysis)

Analysis of the 30 human-written essays revealed that both tools performed perfectly in this condition. As presented in Table 2, the AI-detection scores for both Turnitin and Grammarly were 0% for all 30 essays. Consequently, no statistical comparison was necessary, and the false positive rate for short, human-written essays was zero for both tools.

[Downloaded from mail.ijreeonline.com on 2026-10-30]

Table 2. AI-detection results on non-AI-assisted essays

Tool	N	Mean%	Std. Deviation	Median	Minimum	Maximum
Turnitin	30	0	0	0	0	0
Grammarly	30	0	0	0	0	0

4.3 Performance on MA Theses: Pre-AI vs. Post-AI Era

4.3.1 Pre-AI Era Theses (2010-2015) - Baseline False Positives

The analysis of theses written before the widespread adoption of generative AI provides a crucial baseline. The data were highly skewed, with most theses receiving a score of 0% from both tools. As shown in Table 3, the mean scores were very low (Turnitin: M=0.53%, Grammarly: M=0.43%), driven by a small number of outliers.

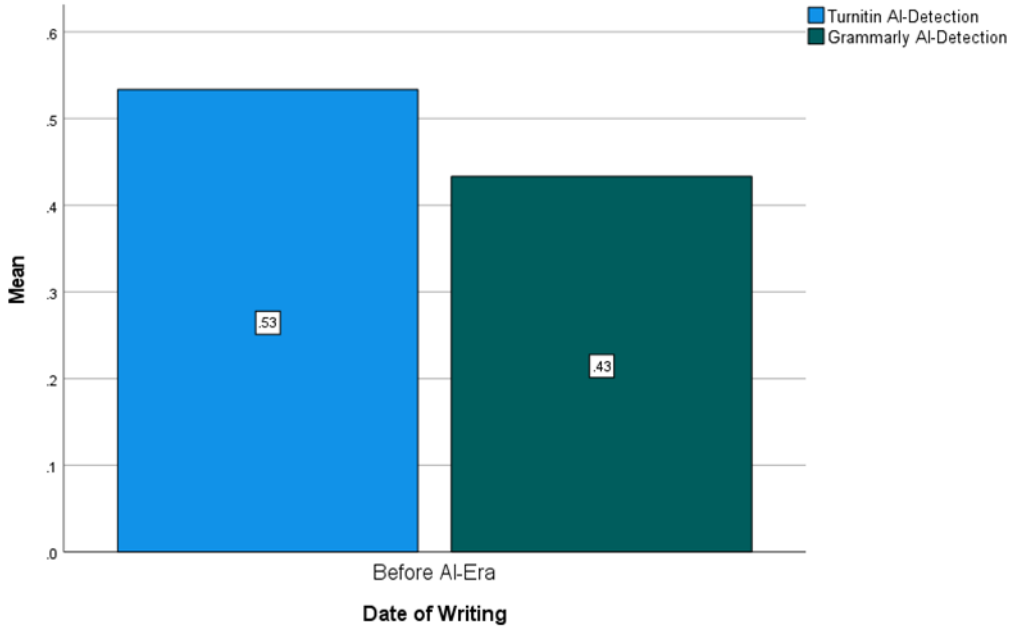

Figure 2. AI detection means for turnitin and grammarly on theses (2010-2015)

[Downloaded from mail.ijreeonline.com on 2026-10-30]

Table 3. Descriptive statistics for theses (2010-2015)

Tool	N	Mean%	Std. Deviation	Median	Range
Turnitin	30	0.53	2.24	0	0-12
Grammarly	30	0.43	1.68	0	0-9

A Wilcoxon Signed-Rank Test found no statistically significant difference between the detection scores of Turnitin and Grammarly for this corpus, Z = -0.82, p = .414. One-sample t-tests against a test value of 0% showed that the mean scores for both Turnitin, t(29) = 1.30, p = .202, and Grammarly, t(29) = 1.42, p= .167, were not significantly different from zero. This indicates that both tools exhibited a very low and statistically insignificant false positive rate on confirmed human-written theses from this era.

4.3.2 Post-AI Era Theses (2020-2025) - The Central Finding

A marked contrast was observed in the theses from the contemporary, post-AI era. While many theses still scored 0%, the overall distribution of scores shifted significantly upward. Descriptive statistics (Table 4) show that the mean detection scores increased substantially for both tools compared to the pre-AI corpus. Notably, the maximum scores were remarkably high, with one thesis being flagged as 62% AI-written by Turnitin and 55% by Grammarly.

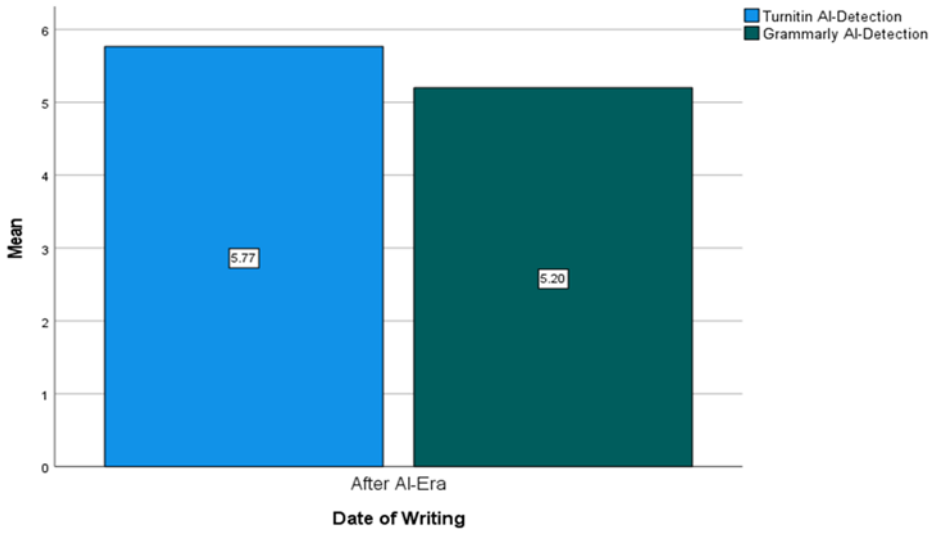

Figure 3. AI detection means for turnitin and grammarly on theses (2020-2025)

[Downloaded from mail.ijreeonline.com on 2026-10-30]

Table 4. Descriptive statistics for theses (2020-2025)

Tool	N	Mean%	Std. Deviation	Median	Range
Turnitin	30	5.77	13.90	0	0-62
Grammarly	30	5.20	12.58	0	0-55

A Wilcoxon Signed-Rank Test revealed that Turnitin generated statistically significant higher detection scores than Grammarly in this corpus, $Z = -2.00$, $p = .046$. However, the effect size was small ($r = .25$), suggesting that while Turnitin is more sensitive, the overall pattern of elevated scores is consistent across both tools for the post-AI theses. This critical finding highlights a clear divergence in the AI-detection profile between the pre-AI and post-AI corpora, presenting a fundamental interpretive challenge for academic integrity.

5. Discussion

This study provides one of the first empirical evaluations of AI detection tools specifically within the context of Iranian EFL academic writing, yielding two major findings with significant implications for academic integrity and equity.

5.1 The Divergent Efficacy of Detection Tools

The first key finding is the fundamental performance gap between Turnitin and Grammarly in identifying AI-generated text. Turnitin's near-perfect accuracy (99.33%) on AI-assisted essays aligns with its design purpose as a dedicated AI-writing detector and confirms its robustness as reported in other studies (e.g., Malik & Amjad, 2025). Its model, trained on a massive corpus of AI-generated text, appears highly effective at identifying the statistical "fingerprints" of LLM output.

In contrast, Grammarly's significant false negative rate (10%) reveals a critical limitation for its use as a primary integrity tool. This discrepancy likely stems from a difference in core design. Grammarly Premium is primarily a grammar and plagiarism checker; its AI-detection feature appears to be a secondary function not yet optimized to the same degree as Turnitin's dedicated model. This finding empirically validates concerns about the unreliability of many commercially available detectors (Chaka, 2024a; Weber-Wulff et al., 2023), creating a dangerous blind spot where AI-assisted work can evade detection.

5.2 The Algorithmic Bias Dilemma in L2 Writing

The most consequential finding of this study lies in the analysis of the thesis corpora. The clean, near-zero baseline established by the pre-AI era theses confirms that authentic Iranian EFL academic writing from that period is correctly classified as human. This baseline gives critical context to the elevated scores witnessed in the post-AI era corpus.

[Downloaded from mail.ijreeonline.com on 2026-10-30]

The presence of non-zero, and sometimes very high, AI-detection scores in contemporary theses presents a serious interpretive dilemma with two plausible explanations:

1. **Actual AI Use:** The scores reflect genuine, and likely undisclosed, use of AI writing assistance by students.
2. **Algorithmic Bias (False Positives):** The tools are misclassifying the inherent linguistic patterns of Iranian EFL writers as AI-generated.

While this study's design cannot definitively arbitrate between these two, the latter interpretation is powerfully supported by existing literature. Researchers have consistently warned that AI detectors exhibit bias against non-native English speakers, with false positive rates for EFL texts reaching up to 61.3% (Chaka, 2024b; Mizumoto et al., 2024). The linguistic profile of L2 writers, often characterized by grammatical simplicity, less lexical diversity, and a reliance on formulaic academic phrasing, may inadvertently mirror the "perfected" but ultimately predictable and uniform style of LLM output. Consequently, the authentic intellectual effort of language learners, marked by the very features of their developmental journey, is at risk of being algorithmically pathologized as dishonest.

6. Conclusion

This study set out to critically evaluate the efficacy and potential biases of two prominent AI-detection tools, Turnitin and Grammarly, within the specific and high-stakes context of Iranian EFL academic writing. Through a two-phase empirical investigation, the results paint a nuanced and cautionary picture of the current state of algorithmic authorship assessment. The central conclusion of this research is that while these tools offer a technological advantage, they are not infallible arbiters of integrity, and their uncritical application poses a significant risk of systematic misjudgment against second language writers.

The investigation yielded two paramount findings. First, it revealed a stark divergence in tool performance, with Turnitin demonstrating robust, near-perfect accuracy in detecting AI-generated text in shorter essays, while Grammarly exhibited a statistically significant false negative rate. This underscores that not all detection services are created equal; their underlying design and primary function critically impact their reliability for this new task. Second, and more critically, the analysis of longer theses revealed a fundamental dilemma: the significant elevation of AI-detection scores in contemporary (2020-2025) theses compared to the near-zero baseline of pre-AI (2010-2015) theses. This discrepancy forces a confrontation with two equally plausible interpretations, either it reflects actual, undisclosed AI use, or it signals a form of algorithmic bias where the authentic linguistic patterns of Iranian EFL writers are misclassified as AI-generated.

This potential for bias validates the core problem this study sought to address. The very features that characterize the diligent work of L2 learners, such as grammatical simplicity, controlled lexicon, and reliance on conventional academic phrasing, may inadvertently mirror the stylistically "perfected" yet ultimately predictable output of large language models. When

[Downloaded from mail.ijreeonline.com on 2026-10-30]

tools trained primarily on native-speaker corpora encounter these features, they risk pathologizing the legitimate writing process of language learners, potentially leading to false accusations with severe academic and personal consequences.

Therefore, the most critical imperative to emerge from this study is that AI-detection software must never serve as the sole or final evidence in adjudicating academic misconduct. Their outputs should be treated not as definitive proof, but as probabilistic data points that necessitate further, human-led investigation. A high AI-detection score should be an invitation for pedagogical dialogue, a chance to review drafting history, discuss the student's writing process, and understand the context of the work, rather than an immediate trigger for a punitive outcome.

In summary, this research underscores a pivotal moment for academic assessment. The path forward requires a hybrid, ethically-grounded approach. This involves designing AI-resistant assessments, developing clear institutional policies that prioritize human judgment, and compelling tool developers to mitigate bias through more diverse and inclusive training datasets. By balancing technological capabilities with a constructive commitment to supporting all students, particularly those navigating the complex challenges of writing in a second language, the academic community can uphold integrity without compromising equity.

6.1 Implications

The findings of this study carry significant implications for a range of stakeholders in higher education, from classroom instructors and institutional policymakers to the developers of the technologies themselves. The demonstrated risks of false negatives and potential algorithmic bias necessitate a proactive and nuanced shift in how academic integrity is maintained in the age of AI.

The inescapable presence of AI requires a move beyond simplistic prohibitions. Educators and curriculum designers must adopt a more strategic approach that integrates education about AI while safeguarding assessment integrity. Instead of a blanket "do not use" policy, educators should develop explicit modules on the ethical use of AI tools. This includes teaching students the distinction between using AI for brainstorming, editing, or summarizing (which may be permitted) versus using it to generate entire submissions (which typically constitutes misconduct).

To mitigate the temptation and ease of AI misuse, assignments should be redesigned to be inherently more difficult for AI to complete authentically. Effective strategies include:

- Process-Oriented Tasks: Incorporating multiple drafts, reflective memos, annotated bibliographies, and in-class writing exercises that document the student's authentic writing journey.

[Downloaded from mail.ijreeonline.com on 2026-10-30]

- Personalized and Context-Specific Prompts: Designing assignments that require students to connect course material to their personal experiences, local community issues, or specific class discussions.
- Oral Defense (Viva Voce): Requiring students to orally explain and defend their written work, which quickly reveals their genuine understanding of the content and its development.

Moreover, universities must lead by developing clear, fair, and transparent policies that reflect the new technological reality. Institutions must create and widely communicate specific guidelines that define acceptable and unacceptable AI use. These policies should outline the procedures for investigating suspected AI misuse and the associated penalties, ensuring due process for students.

In addition, key recommendation is to require students to include a statement of AI use in their submissions. This practice, aligned with guidelines from bodies like the Committee on Publication Ethics (COPE), promotes transparency and helps instructors contextualize AI-detection results. A sample statement could be: “During the preparation of this work, the author(s) used [Name of AI Tool] to [Purpose of Use, e.g., 'for brainstorming initial ideas' or 'for language editing'] in [Section(s) of the document]. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the final work.”

Crucially, institutional policy must stipulate that AI-detection scores cannot be the sole evidence for an academic integrity violation. A high score should trigger a structured, human-led investigation that includes interviewing the student, reviewing their drafting history and outlines, and comparing the work to their previous writing.

The potential for false positives against non-native writers, as strongly suggested by this study, represents a critical flaw that developers have an ethical obligation to address. Developers must actively work to eliminate bias by training and continuously testing their models on diverse, global datasets that include a wide variety of non-native English writing styles. This will help ensure that legitimate L2 writing patterns are not mischaracterized as artificial. Developers should also provide greater transparency on how their tools work, their known limitations, and potential error rates for different demographics. They should also offer comprehensive training resources for instructors on how to interpret results correctly and avoid misapplication.

The unique challenges faced by EFL students are now compounded by the risk of being falsely flagged by AI detectors. Academic support services must be equipped to address this new challenge. This includes not only traditional support for grammar and structure but also guidance on how to develop a strong, authentic academic voice that is less likely to be flagged by flawed algorithms. The fear of being falsely accused can be paralyzing for students. Institutions must proactively communicate their fair investigation protocols to reassure students

[Downloaded from mail.ijreeonline.com on 2026-10-30]

that their authentic work will be protected. Empowering students with knowledge about these tools and processes is crucial for maintaining trust in the academic system.

In conclusion, navigating the new terrain of AI in academia requires a collective and thoughtful response. By embracing pedagogical innovation, enacting fair policies, demanding more equitable technology, and providing robust student support, educational stakeholders can harness the benefits of AI while upholding integrity and protecting the most vulnerable members of the academic community.

Author Contributions

Conceptualization, E.E. and A.F.; methodology, M.A.D.B. and E.E.; software, M.A.D.B.; validation, M.A.D.B. and E.E.; formal analysis, M.A.D.B.; investigation, M.A.D.B.; resources, E.E.; data curation, M.A.D.B.; writing, original draft preparation, M.A.D.B.; writing, review and editing, E.E. and A.F.; visualization, M.A.D.B.; supervision, E.E. and A.F.; project administration, E.E.; funding acquisition, N/A All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

Not applicable.

Ethical considerations

Not applicable.

Funding

The authors received no financial support for the research, authorship, or publication of this manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

Ahmadi, A. (2014). Plagiarism in the academic context: A study of Iranian EFL learners. *Research Ethics, 10*(3), 151–168. https://doi.org/10.1177/1747016114523781

Amirzhanov, A., Turan, C., & Makhmutova, A. (2025). Plagiarism types and detection methods: A systematic review. *Frontiers in Computer Science, 3*(2), 32–43.

Chaka, C. (2024a). Accuracy pecking order - How 30 AI detectors stack up in detecting generative artificial intelligence content in university English L1 and English L2 student essays. *Journal of Applied Learning & Teaching,* 7(1), Article 33. https://doi.org/10.37074/jalt.2024.7.1.33

Chaka, C. (2024b). 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 & Teaching,* 7(2), Article 34. https://doi.org/10.37074/jalt.2024.7.2.34

Grammarly, Inc. (2023). *Grammarly Premium* [Computer software].

[Downloaded from mail.ijreeonline.com on 2026-10-30]

Hamdani, M., & Abid, N. (2023). Indonesian EFL students' writing challenges: A qualitative meta-analysis. *Journal of Language and Linguistic Studies, 19*(1), 1–20. https://doi.org/10.17263/jlls.19.1.1

Hosseini, M., & Ebadi, S. (2025). Exploring Iranian EFL learners' and instructors' perceptions of generative AI in academic writing: Ethical considerations and technological equity. *Computers and Education Open, 6,* Article 100215. https://doi.org/10.1016/j.caeo.2024.100215

Hyland, K. (2007). *Genre and academic writing in the disciplines.* University of Michigan Press.

Kouam, A. W. F. (2024). What is plagiarism in academic writing and how to avoid it: Guiding novice researchers. *International Journal of Research Publications, 144*(1), 283–286. https://doi.org/10.47119/IJRP1001441220246104

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Proceedings of the National Academy of Sciences, 120*(23), Article e2300837120. https://doi.org/10.1073/pnas.2300837120

Malik, M. A., & Amjad, A. I. (2025). AI vs AI: How effective are Turnitin, ZeroGPT, GPTZero, and Writer AI in detecting text generated by ChatGPT, Perplexity, and Gemini? *Journal of Applied Learning & Teaching, 8*(1), 1–5. https://doi.org/10.37074/jalt.2025.8.1.9

Mizumoto, A., Yasuda, S., & Tamura, Y. (2024). Identifying ChatGPT-generated texts in EFL students' writing: Through comparative analysis of linguistic fingerprints. *Applied Corpus Linguistics, 4*(3), Article 100106. https://doi.org/10.1016/j.acorp.2024.100106

OpenAI. (2023). *ChatGPT (GPT-4 version)* [Large language model]. https://chat.openai.com

Pennycook, A. (1996). Borrowing others' words: Text, ownership, memory, and plagiarism. *TESOL Quarterly, 30*(2), 201–230. https://doi.org/10.2307/3588141

Rahman, M. H., Danno, B. L., Mola, D. W., Islam, M. S., Andrabi, S. M. H., Reza, M. N., & Shaw, D. (2025). Analysis of similarity index between iThenticate and Ouriginal plagiarism detection software: A comprehensive study. *Indonesian Journal of Electrical Engineering and Computer Science, 37*(3), 2086–2094. https://doi.org/10.11591/ijeecs.v37.i3.pp2086-2094

Rahman, M. H., Islam, M. S., Andrabi, S. M. H., Sharma, J. P., & Reza, M. N. (2024). Comparing the similarity index across iThenticate, Ouriginal, and Turnitin plagiarism detection software. *Journal of Physical Education and Sport, 24*(2), 419–424. https://doi.org/10.7752/jpes.2024.02051

Rezanejad, A., & Rezaei, S. (2013). Academic dishonesty at universities: The case of plagiarism among Iranian language students. *Journal of Academic Ethics, 11*(4), 275–295. https://doi.org/10.1007/s10805-013-9193-8

Saeedi, Z., & Baleghizadeh, S. (2026). Navigating generative AI tools in second language postgraduate research writing: Affordances, algorithmic scrutiny, and pedagogical responses. *Iranian Journal of Language Teaching Research, 14*(1), 45–63.

Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills* (2nd ed.). University of Michigan Press.

Turnitin, LLC. (2023). *Turnitin similarity checker and AI writing detection* [Computer software]. https://www.turnitin.com

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,* Article 26. https://doi.org/10.1007/s40979-023-00146-z

[Downloaded from mail.ijreeonline.com on 2026-10-30]

Yuan, Y., Li, H., & Sawaengdist, A. (2024). The impact of ChatGPT on learners in English academic writing: Opportunities and challenges in education. *Language Learning in Higher Education, 14*(1), 1–15. https://doi.org/10.1515/cercles-2023-0006

[Downloaded from mail.ijreeonline.com on 2026-10-30]