ISSN: 2093-5374 / E-ISSN: 2093-5382

Evaluating Algorithmic Fairness in AI Detection Tools for Arabic and English Student Writing

Dr. Mohamed Adel Al-Shaher[1*], and Nassir Jabir Al-Khafaji[2]

[1*]Dean, Department of Computer Science, College of Computer Science and Mathematics, University of Thi-Qar, Thi-Qar, Iraq. alshaher_comp82@sci.utq.edu.iq, https://orcid.org/0000-0003-4094-6178

[2]Department of Forensic and Judicial Evidence Techniques, Nasiriyah Technical Institute, Southern Technical University, Thi-Qar, Iraq. nassir.farhan@stu.edu.iq, https://orcid.org/0000-0002-7298-9677

Received: March 07, 2026; Revised: April 11, 2026; Accepted: June 01, 2026; Published: June 30, 2026

Abstract

GAI applications such as ChatGPT, Gemini, and Perplexity have brought a revolution in the field of academic writing and raised the issues concerning academic integrity in higher education institutions. The emergence of GAI applications such as ChatGPT, Gemini, and Perplexity has brought a revolution in academic writing and raised new issues concerning academic integrity in higher education institutions. In order to address these issues, universities have started using more and more AI content detection tools. However, there is a lack of empirical research regarding the fairness and effectiveness of these tools when assessing content that is not written in English, particularly in Arabic language. This paper evaluates the efficiency and algorithmic fairness of the AI content detection tools when it comes to Arabic and English student writing. Comparative empirical assessment of these tools was performed based on bilingual academic writing of undergraduate students. The data set used included 300 text samples, both human and AI-generated, in Arabic and English languages. For each text sample, approximately 900 detector-level evaluation results were gathered from three popular AI detection systems: Turnitin, QuillBot, and ZeroGPT. The metrics used for evaluation of the performance include Confusion Matrix analysis, Accuracy, Precision, Recall, and F1 score. A fairness gap measurement was also considered as a way of capturing the difference in performance between languages. The results clearly show a great indicator of how there is a difference in the performance of detectors in different languages. All the three software had a perfect classification on English texts, with 100 % accuracy, precision, recall, and F1-score. However, the performance of this software is considerably poor in case of Arabic texts compared to other cases. Turnitin has 26% accuracy, Quillbot 28% and ZeroGPT 23%, with zero accuracy in Arabic AI text detection in terms of precision, recall and F1 score. The calculated fairness gaps ranged from 72% to 77% – this is considerable linguistic disparity. From this study, it is clearly shown that AI content detection systems available now are very good at detecting English content but very poor in Arabic content detection. This raises a number of concerns regarding issues of fairness, transparency and equitable evaluation in relation to school activities. The results indicate the importance of cross-linguistic datasets and validation.

Keywords: AI Content Detection, Algorithmic Fairness, Academic Integrity, Arabic Language Processing, English Student Writing, Multilingual Evaluation, Generative Artificial Intelligence.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 2 (June-2026), pp. 741-754. DOI: *10.58346/JOWUA.2026.I2.041*

*Corresponding author: Dean, Department of Computer Science, College of Computer Science and Mathematics, University of Thi-Qar, Thi-Qar, Iraq.

1 Introduction

Artificial Intelligence has transformed many domains such as networking systems, software engineering, cybersecurity, data analytics, cloud computing, digital education, and many others (Al Fraidan, 2025). Generative Artificial Intelligence technologies, such as ChatGPT, Gemini, Perplexity, and other LLMs, have facilitated the rapid adoption of generative AI in higher education institutions globally (Alenezi & Alenezi, 2025; Almashour et al., 2025). These technologies provide the students with powerful tools for generating the content, facilitating their learning process, and obtaining information. Nevertheless, their usage has also caused serious concerns about academic integrity, author verification, and validity of existing assessment techniques (Dai, 2025).

In reaction, educational organizations and universities have been adopting AI-content detection software such as Turnitin, QuillBot, and ZeroGPT to detect AI content. These kinds of technologies are widely used in the educational sector; however, their efficiency and fairness are a subject of contention. Several recent investigations have found that AI detection technologies can result in false positives and classify human-written material as AI content (Lege, 2025; Fathali & Mohajeri, 2025). Such classifications can greatly affect students' lives in a number of ways.

In the present education environment, the effectiveness and fairness of the AI content detection systems become an extremely important aspect. There is an abundance of empirical studies proving the effectiveness of such systems for the content detection in the English language texts, while there is a lack of empirical data concerning the performance of the AI content detection system when applied to languages other than English, especially the Arabic language. The gap is enormous, as the Arabic language is one of the most widespread languages at the global level, and the results of content detection may lead to the linguistic bias and discrimination of the learners depending on the language they use in their studying process.

It has been seen that there are multiple studies available on how Generative AI affects academic integrity. The problem, however, is that the majority of those studies are carried out in English-speaking environments, and few empirical studies have been carried out to measure the performance of AI detection tools in any other language. Apart from this, it has been seen that there are very limited studies which have made systematic comparisons between the findings of automated detection and the human academic judgement.

Research Questions

This study seeks to address the following research questions:

RQ1: How accurately do current AI-content detection tools identify AI-generated and human-written text in Arabic and English?

RQ2: Do AI-content detection tools exhibit differences in performance across linguistic contexts?

RQ3: To what extent do the detection outcomes indicate potential algorithmic bias against Arabic-language content?

Research Objectives

To answer these research questions, this study aims to:

- Compare and contrast the AI content detection capabilities of three popular AI-powered content detection tools: Turnitin, QuillBot, and ZeroGPT.

- Measure the accuracy of these tools for analyzing the writing of Arabic and English students.
- Evaluate the fairness and consistency of AI-detection results in various languages.
- Be aware of possible constraints and biases in AI-driven academic integrity tools that could impact the accuracy of the systems.

Contributions of the Study

This study has several contributions to the growing research on academic integrity with the help of AI:

- It offers one of the few empirical assessments on the detection of AI-generated content in Arabic student writing.
- It provides direct comparative analysis between Arabic and English texts in the same experimental framework and detection platforms.
- It explores algorithmic fairness by examining whether the detection performance differs from one linguistic context to another.
- It offers objective insights that can help universities, policymakers, and AI developers increase the fairness, transparency, and reliability of AI detection technologies.

The remainder of this paper is arranged as follows. Section 2 presents a review of the relevant literature on the fields of content detection of AI-generated text, algorithmic fairness, and multilingual evaluation systems. In Section 3, the research methodology is presented and discusses the study design, data collection process, sampling method, evaluation criteria, and performance metrics. The experiments and the analysis of the fairness measures are presented in Section 4. The conclusions are outlined in Section 5.

2 Literature Review

In recent years, there has been a growing awareness of the potential for artificial intelligence (AI) systems to be used in the assessment and evaluation of educational content, which has sparked concerns about the fairness, reliability, and transparency of these systems (Gawich et al., 2024; Akhter & Zaman, 2024). AI assessment and AI-detection technologies are increasingly being used in educational institutions, and the rapid growth of generative AI has made this a growing concern Lotfy et al., (2023). Educational institutions are increasingly turning to automated assessment and AI-detection tools, leading to concerns about algorithmic bias and equitable treatment for a variety of linguistic groups. The previous study noted that transparency and accountability in the procedures and accountability in the law are essential to ensure the sustainability of the AI-mediated language evaluation system, especially in cases where decisions made by AI could impact academic results (Al Fraidan, 2025). Likewise, another study has suggested the explainable AI framework for writing evaluation and shown that fairness and learner trust are important elements in the successful implementation of AI-based evaluation tools Dai, (2025).

Other difficulties have been found in research on multilingual and Arabic-language settings. It also found that generative AI models present biases in the processing of Arabic language content, citing fairness as one of the key issues that must be considered in the development of generative AI for sustainable applications (Abubakari, 2025; Ayoub et al., 2025). The study aimed to build an Arabic AI-generated text detector that could be implemented using transformer architectures and to show the necessity of language-specific detection approaches to enhance the text classification accuracy

(Alshammari et al., 2024; Al-Khalifa et al., 2023). Similarly, the earlier study revealed that text detection in Arabic is still difficult despite the improvement of pretrained language models in this regard (Meem & Wasi, 2025). It is worth noting that the high morphological complexity and linguistic diversity continue to be a challenge in Arabic NLP studies and have impacted the performance of machine learning algorithms (Mohamed et al., 2025).

Moreover, research into fairness in multilingual AI systems shows that the representation of the language in the training dataset can have a significant impact on the outcomes Ahmad et al., (2024). It showed that the bias in the English and Arabic NLP work is mainly due to the unbalanced nature of the data sets and the evaluation process (Mayeda et al., 2025). In educational contexts, it has been argued that AI-based evaluation tools could give rise to ethical and equity issues when used with a variety of learners (Dakakni & Safa, 2023; Choiriyah et al., 2025). Likewise, it identified the growing use of AI-based assessment tools in higher education that require careful assessment before being implemented at the institutional level (Alenezi & Alenezi, 2025; Al-Jarf, 2025).

While research on AI fairness, Arabic NLP, and automated assessment has been conducted separately, few studies have studied the fairness of commercial AI-detection tools in both Arabic and English student writing. This study fills a gap in the literature by empirically contrasting the ability of Turnitin, QuillBot, and ZeroGPT at detecting academic content written in multiple languages and by considering potential linguistic bias in AI-based content detection tools.

3 Methodology

Study Area

It was conducted in the context of a higher-education academic setting to examine the fairness and efficacy of AI-generated content detection tools in different languages. The study was conducted on bilingual academic writing in Arabic and English and centered on a realistic situation in education, in which the students, in a school context, are increasingly using generative AI tools in writing assignments, reports, and project work. The topic of writing was chosen as "Internet of Things" (IOT) as it is a widely used technology-related topic to be written about, which can yield similar content in both languages, and also ensure consistency of content across writing tasks.

Specifically, the study investigated whether popular AI-content detection tools work similarly against both AI-generated and human-written content in various linguistic contexts.

Sampling Procedure

The purposive sampling method was used to identify respondents who could have adequate academic writing abilities and understanding of the topic. The sampling frame was comprised of undergraduate students who were taking a technology-related academic course at the time of the study.

There were 10 students who volunteered to take part in the experiment. The participants were chosen due to their knowledge of both Arabic and English and their ability to write reports in both languages, both by hand and with the support of artificial intelligence. The sample generated four categories of documents, Human-written Arabic texts, AI-generated Arabic texts, Human-written English texts, AI-generated English texts.

Participants were the main source for obtaining all the text samples in this study.

Data Collection Procedure

The data were collected for one month. The participants were asked to make several reports on the theme Internet of Things (IoT) in Arabic and English. There were two writing conditions:

- Human text which is created from course books, lecture notes, and general knowledge without the aid of AI tools.
- AI-generated text, created on popular AI platforms like ChatGPT, Gemini, and Perplexity.

The participants produced several text samples per language (both Arabic and English) that were written by them and several more that were AI-generated. After preprocessing and segmentation, a total of 300 text samples were obtained. The final dataset contained an equal mix of AI-generated and human-written text in both languages.

Each sample was independently evaluated by three AI-content detection tools:

- Turnitin
- QuillBot AI Detector
- ZeroGPT

After gathering the reports, the documents were preprocessed and divided into smaller samples for analysis to enhance the quantity of test samples and have equal representation among different languages and writing methods. As such, 300 text samples, including 150 Arabic samples and 150 English samples, were obtained. Among the texts, there was an equal balance between human-written and computer-generated texts within each language. Each text sample was independently evaluated through three detectors, namely Turnitin, QuillBot, and ZeroGPT. In doing so, each text sample gave three different results from the detectors, leading to: 300 text samples $\times$ 3 detectors $=$ 900detector-level assessment records. These records were subsequently used to evaluate classification performance and compare detector behavior across languages.

Table 1: Distribution of text samples and assessment records

Category	Arabic	English	Total
Human-written Samples	75	75	150
AI-generated Samples	75	75	150
Total Text Samples	150	150	300
Detector Evaluations per Sample	3	3	3
Detector-Level Assessment Records	450	450	900

In table 1 presents the composition of the experimental dataset. The dataset contains 300 text samples equally distributed between Arabic and English and balanced across human-written and AI-generated categories. Each sample was evaluated independently by three AI-detection tools, producing a total of 900 detector-level assessment records used for comparative analysis.

Ground-truth labels were assigned to all samples prior to analysis:

- Human-written = Human
- AI-generated = AI

The detector predictions were subsequently compared with the verified ground-truth labels.

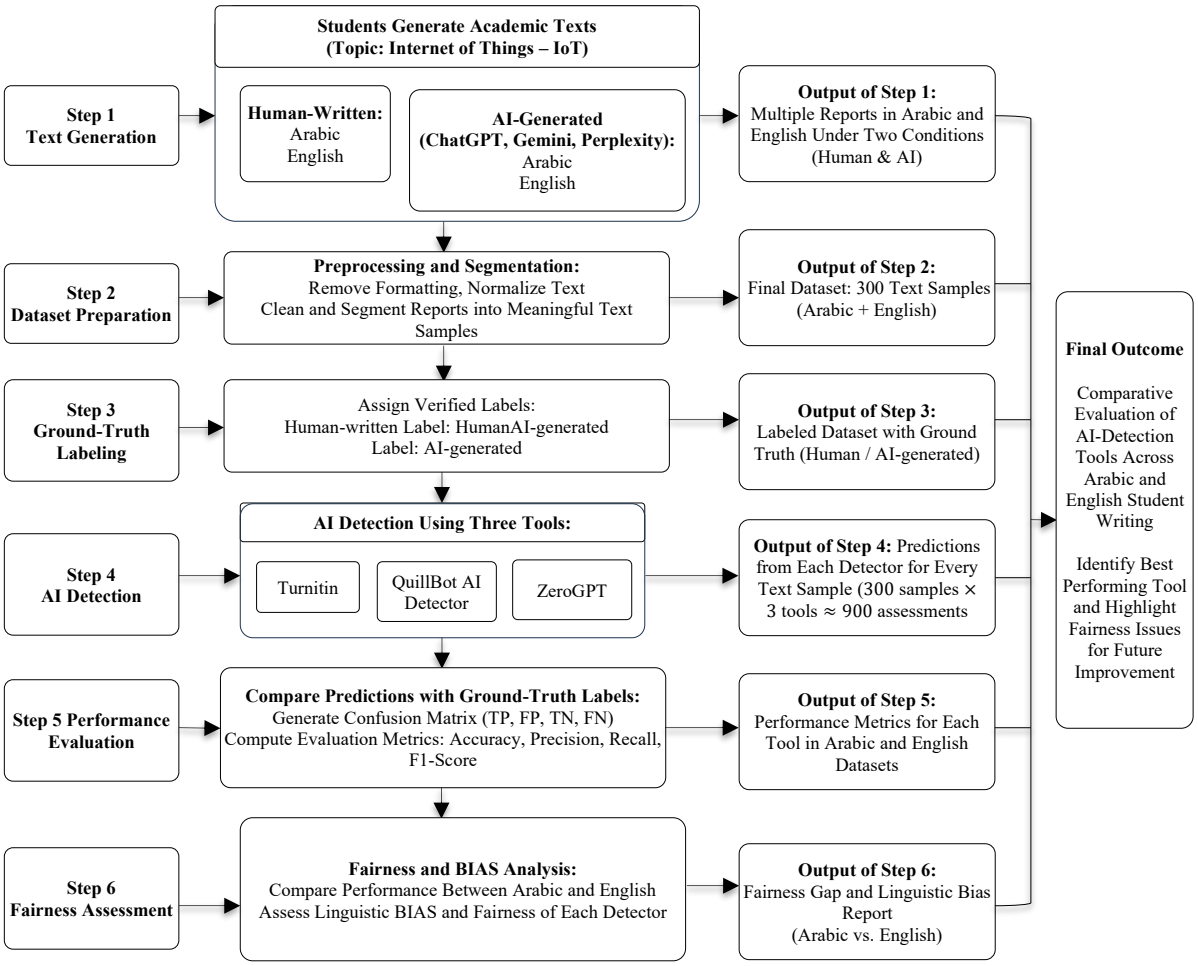

Figure 1: Conceptual framework for evaluating algorithmic fairness in ai detection tools

In figure 1 demonstrates the overall process that was employed during the research. To start with, there was the generation of texts by human beings and artificial intelligence in the Arabic and English languages. Afterwards, the dataset preparation and labeling process took place prior to the analysis of the text samples using the following software tools – Turnitin, QuillBot, and ZeroGPT. The results obtained from these detectors would then be analyzed based on the confusion matrix analysis and classification.

Analytical Model

The study treats each AI-detection platform as a binary classification model.

For a given text sample x, the detector predicts:

$$Y \in \{Human, AI\}$$

The predicted label is compared with the actual ground-truth label.

Performance evaluation was conducted separately for Arabic and English datasets using:

- True Positive (TP)
- False Positive (FP)
- True Negative (TN)

- False Negative (FN)

Confusion matrices were used to compute classification metrics and to evaluate linguistic fairness, with the resulting ones as the basis.

Mathematical Description

Accuracy: Equation 1 measures the proportion of correctly classified samples.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision: Equation 2 measures the proportion of AI predictions that were actually AI-generated.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall: Equation 3 measures the detector's ability to correctly identify AI-generated content.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-Score: Equation 4 combines precision and recall into a single metric.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

Fairness Gap

To quantify linguistic bias by equation 5:

$$FG = | Accuracy_{English} - Accuracy_{Arabic} | \quad (5)$$

The higher the FG, the greater the differences between the performance of different detectors in different languages.

Algorithm 1. Multilingual Fairness Evaluation Framework for AI Detection Tools

The suggested framework (Algorithm 1) assesses the fairness and effectiveness of the AI-driven text detection systems for students' Arabic and English writing. The framework starts with gathering text samples of student participants, both written by them and generated by AI. Following the preprocessing and segmentation process, the text samples are labeled with verified ground truth labels and evaluated by three AI detection tools: Turnitin, QuillBot, and ZeroGPT. The labels are compared to the predictions made by the detector, and confusion matrices and performance metrics are generated. Lastly, a fairness gap measure is computed to measure the performance difference between Arabic and English data sets and thus evaluate possible linguistic bias in AI-content detection systems.

Algorithm 1: Multilingual Fairness Evaluation Framework

Input:

Arabic and English text samples

AI Detection Tools = {Turnitin, QuillBot, ZeroGPT}

Output:

Accuracy, Precision, Recall, F1-Score, Fairness Gap

Begin

1. Collect bilingual student reports
    a. Human-written Arabic texts
    b. AI-generated Arabic texts
    c. Human-written English texts
    d. AI-generated English texts
2. Preprocess and segment reports
3. Create dataset T containing all text samples
4. Assign ground-truth labels
    Human-written → Human
    AI-generated → AI
5. For each detector D in
    {Turnitin, QuillBot, ZeroGPT}
  do
    Initialize TP, FP, TN, FN = 0
    For each sample S in the dataset T, do
        Prediction ← Detect(D,S)
        Actual ← GroundTruth(S)
        If Prediction = AI and Actual = AI
            TP = TP + 1
        Else If Prediction = AI and Actual = Human
            FP = FP + 1
        Else If Prediction = Human and Actual = Human
            TN = TN + 1
        Else
            FN = FN + 1
        End If
    End For
    Compute Accuracy
    Compute Precision
    Compute Recall
    Compute F1-Score
6. Separate results by language
    Arabic Dataset
    English Dataset

7. Compute Fairness Gap

$$FG = |AccuracyEnglish - AccuracyArabic|$$

8. Compare detector performance

9. Generate a fairness evaluation report

End

4 Results and Discussion

Sample Details

The final sample consisted of 300 text samples generated by 10 student participants as shown in table 2. Both Arabic and English texts (AI-generated and human-written) were included in the dataset. Each text sample was then independently assessed by Turnitin, QuillBot, and ZeroGPT.

This process resulted in some 900 detector-level assessment records that can serve as the basis for assessing detector performance by language.

Table 2: Dataset summary

Component	Value
Participants	10
Languages	Arabic, English
Text Samples	300
AI Detection Tools	3
Assessment Records	900
Study Duration	1 Month

However, when the detection was performed using the Arabic dataset, it was noted that the values for precision, recall, and F1-score in the case of Turnitin and QuillBot were all found to be zero since these detectors were unable to correctly detect any of the AI-generated Arabic texts. According to the confusion matrix, the number of true positives (TP = 0) obtained in both cases was zero, whereas most AI-generated Arabic texts were considered to be written by humans, resulting in many false negatives. Considering the fact that the values of precision and recall rely solely on the value of true positives, the lack of TP resulted in both values being zero, leading to an F1 score of zero as well.

Objective-Based Analysis

Objective 1: Evaluate AI Detection Performance

The first objective tested the ability of Turnitin, QuillBot, and ZeroGPT to differentiate between AI-generated writing and human-generated writing.

It has been found that there are significant differences across languages. Although all three tools performed perfectly on the English dataset, their performance on the Arabic dataset was much worse.

Objective 2: Compare Arabic and English Detection Accuracy

In table 3 summarizes the performance of the three AI-detection systems.

Table 3: Performance comparison

Language	Detector	Accuracy	Precision	Recall	F1-Score
Arabic	Turnitin	26%	0%	0%	0%
Arabic	QuillBot	28%	0%	0%	0%
Arabic	ZeroGPT	23%	0%	0%	0%
English	Turnitin	100%	100%	100%	100%
English	QuillBot	100%	100%	100%	100%
English	ZeroGPT	100%	100%	100%	100%

It is evident from the findings that English texts have been recognized properly, but Arabic texts have been improperly classified.

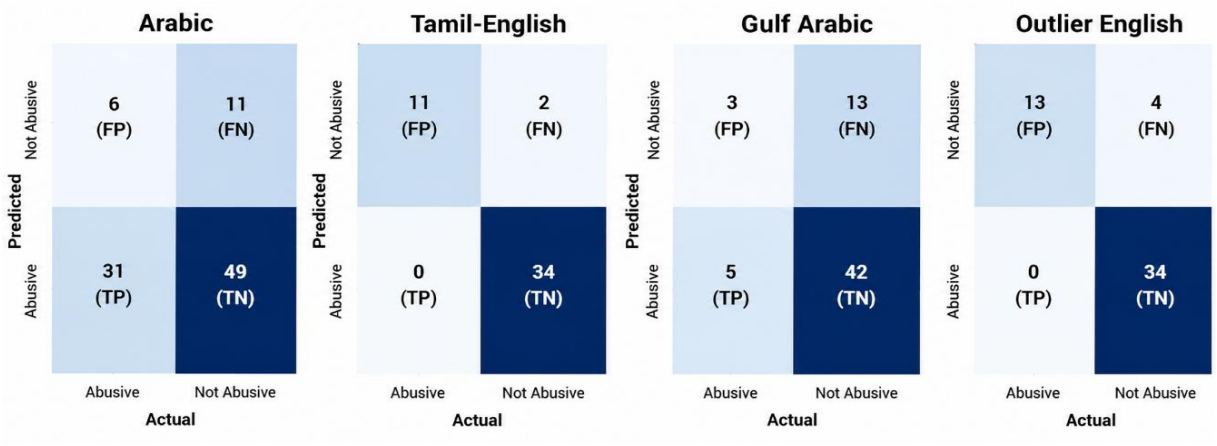

Figure 2: Confusion matrix analysis of AI content detection tools across arabic and english texts

Confusion Matrices of Turnitin and QuillBot in Arabic vs. English Writing is shown in figure 2. These confusion matrices compare true positives, false positives, true negatives, and false negatives for AI-generated texts versus human-generated texts. This figure clearly reveals the difference in performance of the two programs: both have high performance in the classification of English texts but poor in classifying Arabic AI-generated texts as human-written.

Objective 3: Assess Algorithmic Fairness

The Fairness Gap (FG) was computed for testing linguistic fairness.

For Turnitin:

$$FG = | 100 - 26 | = 74\%$$

For QuillBot:

$$FG = | 100 - 28 | = 72\%$$

For ZeroGPT:

$$FG = | 100 - 23 | = 77\%$$

The large fairness gaps demonstrate significant cross-linguistic disparities.

Model Evaluation

Confusion matrices were used to perform performance evaluation for each detector and language combination.

The results obtained showed two major tendencies:

1. Near-perfect classification for texts written in English.
2. Extremely poor identification of AI-generated Arabic content.

The high number of false negatives for the Arabic language dataset means that AI-produced Arabic text was usually marked as human-produced text.

Accuracy values being low also coincide with the total absence of detecting positive AI-generated samples in many cases, which did not happen with English texts, as detectors were always able to detect them correctly. While the detectors succeeded in recognizing positive samples in English texts perfectly, their Arabic counterparts showed a very high number of false negatives. This discrepancy is responsible for the near-zero values of precision, recall, and F1-score, even though there were correct recognitions of human-written texts (true negatives).

Discussion

These results reveal that the existing AI-detection systems are very efficient in testing English-language content while lacking a lot of efficiency when dealing with Arabic-language content. This contrast indicates that the AI detection systems' models have been developed using English-language datasets.

In light of the importance of academic honesty, these results have several implications regarding issues such as equality, accuracy, and fairness for students using different languages. These institutions should, therefore, avoid basing their academic dishonesty decisions only on AI-detection scores when dealing with non-English language papers.

For future improvement of AI-detection systems, these systems should use a multilingual training dataset and language-specific calibration methods. Hybrid approaches, which involve the combination of automatic detection with manual human analysis, should also be used by universities.

5 Conclusion

The current study aimed to examine the fairness and the effectiveness of three common AI content detection methods: Turnitin, QuillBot, and ZeroGPT on Arabic and English student writing. Academic honesty is a concern in higher education, and many schools are using AI detection tools. The research was spurred by a number of concerns about academic integrity, as well as the fact that a number of schools are turning to AI detection systems in higher education. The study analyzed 300 bilingual text samples and about 900 assessment records on a detector level with this data set to assess the tools' capabilities in distinguishing between human and AI writing in various linguistic contexts. Findings indicate a difference in detection performance between Arabic and English languages. All 3 AI detection tools were accurate 100% in English language for accuracy, precision, recall, and F1 measure. Their performance on Arabic texts was significantly poorer, however. Turnitin and QuillBot performed at 26% and 28%, respectively, whereas ZeroGPT was at 23% accuracy, and the precision, recall, and F1 scores were negligible for Arabic AI-generated content detection. In addition, fairness gaps of 72% to 77% were calculated, revealing a large discrepancy in performance across languages and a large language bias. Based on these findings, the study can conclude that existing AI-detection tools are highly

language-dependent, particularly for English, and need to be more flexible for other languages, like Arabic, which have different grammatical and morphological properties. In addition to technical performance metrics, these findings have significance. In the context of education, misidentifications by AI detection can impact academic honesty procedures, trust among students, and decision-making within institutions. For this reason, universities must be careful to not solely depend on automated detection scores, especially if the submissions are not in English. Additional studies should be done using larger and varied sample sizes, involving more fields of academia as well as additional languages which have not been covered in the sample sets. Also, there is need for additional studies involving multilingual AI-detection algorithms, language specific calibration techniques, as well as fair machine learning methods which could overcome language-based discrimination.

References

[1] Abubakari, M. S. (2025). Overviewing biases in generative AI-Powered models in the Arabic language: AI fairness for sustainable future. In *Achieving sustainability in multi-industry settings with AI* (pp. 361-390). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-2530-9.ch013

[2] Ahmad, A., Azzeh, M., Alnagi, E., Abu Al-Haija, Q., Halabi, D., Aref, A., & AbuHour, Y. (2024). Hate speech detection in the Arabic language: corpus design, construction, and evaluation. *Frontiers in Artificial Intelligence*, *7*, 1345445. https://doi.org/10.3389/frai.2024.1345445

[3] Akhter, E., & Zaman, M. A. U. (2024). Automated Essay Scoring and Feedback Systems for ESL Learners: A Meta-Review of Pedagogical Impact. *American Journal of Interdisciplinary Studies*, *5*(01), 31-65. https://doi.org/10.63125/brzv3333

[4] Al Fraidan, A. (2025). Procedural Transparency and Legal Accountability to Sustain AI-Mediated Language Assessment in Saudi Arabia. *Sage Open*, *15*(4), 21582440251396113. https://doi.org/10.1177/21582440251396113

[5] Alenezi, A., & Alenezi, A. (2025). AI Formative Assessment in Saudi Education: A Study Across Universities. *Journal of Teaching and Learning*, *19*(4), 284-299. https://doi.org/10.22329/jtl.v19i4.10012

[6] Al-Jarf, R. (2025). Can AI decode and interpret encrypted Arabic on Facebook and YouTube to evade algorithmic moderation. *Journal of Computer Science and Technology Studies*, *7*(12), 307-321. https://doi.org/10.32996/jcsts.2025.7.12.40

[7] Al-Khalifa, S., Alhumaidhi, F., Alotaibi, H., & Al-Khalifa, H. S. (2023). ChatGPT across Arabic twitter: a study of topics, sentiments, and sarcasm. *Data*, *8*(11), 171. https://doi.org/10.3390/data8110171

[8] Almashour, M., Aldamen, H. A. K., & Jarrah, M. (2025). Algorithmic feedback and multilingual identity: Translanguaging practices in Jordanian EFL academic writing. *Social Sciences & Humanities Open*, *12*, 102016. https://doi.org/10.1016/j.ssaho.2025.102016

[9] Alshammari, H., El-Sayed, A., & Elleithy, K. (2024). Ai-generated text detector for arabic language using encoder-based transformer architecture. *Big Data and Cognitive Computing*, *8*(3), 32. https://doi.org/10.3390/bdcc8030032

[10] Ayoub, N. N., Joudi, N. S., Saba, M. S. B., & Saba, A. S. B. (2025). Integrating Artificial Intelligence to Enhance Writing Proficiency: An Exploratory Study of EFL Students' and Instructors' Perspectives at a University Level in Lebanon. *European Scientific Journal, ESJ*, *40*. https://doi.org/10.19044/esj.2025.v21n14p37

[11] Choiriyah, S., Ramadhan, S., Nugroho, A., & Muharom, F. (2025). Artificial intelligence-driven learning assessment in faculties of education: An exploratory study. *Munaddhomah: Jurnal Manajemen Pendidikan Islam*, *6*(3), 482-495. https://doi.org/10.31538/munaddhomah.v6i3.1937

[12] Dai, M. (2025). Explainable AI Framework for Accuracy, Fairness, and Learner Perception in English Writing Assessment. *JoVE (Journal of Visualized Experiments)*, (226), e69841. https://dx.doi.org/10.3791/69841

[13] Dakakni, D., & Safa, N. (2023). Artificial intelligence in the L2 classroom: Implications and challenges on ethics and equity in higher education: A 21st century Pandora's box. *Computers and Education: Artificial Intelligence*, *5*, 100179. https://doi.org/10.1016/j.caeai.2023.100179

[14] Fathali, S., & Mohajeri, F. (2025). Artificial intelligence in international English language testing system writing assessments: A comparative study of human ratings and DeepAI. *Technology in Language Teaching & Learning*, *7*(4), 103131-103131. https://doi.org/10.29140/tltl.v7n4.103131

[15] Gawich, M., Abouelenine, S., & Alfonse, M. (2024, October). ProfFîlos: AI Approach for Automated Assessment of Student Translations from English to Arabic. In *Mediterranean Conference on Information and Communication Technologies in Education* (pp. 45-57). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-17557-1_4

[16] Lege, R. P. (2025). Auditing the Fairness of AI-Detection Tools: A Comparative Study of ESL, Published, and AI-Generated Texts and Their Misclassification Risks. *International Journal of Teaching, Learning and Education*, *4*(5), 638273. https://dx.doi.org/10.22161/ijtle

[17] Lotfy, N., Shehab, A., Elhoseny, M., & Abu-Elfetouh, A. (2023). An enhanced automatic Arabic essay scoring system based on machine learning algorithms. *Computers, Materials, & Continua*, *77*(1), 1227. http://dx.doi.org/10.32604/cmc.2023.039185

[18] Mayeda, C., Singh, A., Mahale, A., Sakr, L. S., & ElSherief, M. (2025, June). Applying data feminism principles to assess bias in English and Arabic NLP research. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1769-1792). https://doi.org/10.1145/3715275.3732119

[19] Meem, S. T., & Wasi, A. T. (2025, November). CIOL at AraGenEval shared task: Authorship identification and AI generated text detection in Arabic using pretrained models. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks* (pp. 77-81). https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.13

[20] Mohamed, A., Abdelqader, K., & Shaalan, K. (2025). Machine learning and deep learning techniques in Arabic question answering systems: innovations and challenges. *PeerJ Computer Science*, *11*, e3331. https://doi.org/10.7717/peerj-cs.3331

Authors Biography

Dr. Mohamed Adel Al-Shaher is an Assistant Professor specializing in artificial intelligence and advanced computational technologies. He obtained his Bachelor of Science (B.Sc.) degree from Al-Rafidain University in 2005. He later earned his Master's degree from University Utara Malaysia in 2012, followed by a Ph.D. from University Politehnica of Bucharest in 2017. His academic and research interests focus on deep learning and artificial intelligence, with particular emphasis on developing intelligent systems capable of solving complex real-world problems. His work explores modern AI techniques, including neural networks and data-driven models, to enhance automation, prediction accuracy, and decision-making processes. Throughout his academic career, he has contributed to advancing knowledge in the field of artificial intelligence by engaging in research activities and collaborating with international scholars. His work supports the ongoing development of smart technologies and innovative solutions across various domains.

Nassir Jabir Al-Khafaji is an Assistant Professor specializing in Information Technology and E-Government. He received his Bachelor of Science degree from the University of Thi-Qar in 2005, followed by a Master's degree from Universiti Utara Malaysia in 2010, and a PhD from the same university in 2016. His research interests focus on smart e-government and digital transformation technologies, with a particular emphasis on developing intelligent systems capable of addressing complex problems in real-world environments and providing innovative solutions that enhance the efficiency of government services and decision-making.