Language Testing in Asia

RESEARCH Open Access

Using AI-based detectors to control AI-assisted plagiarism in ESL writing: "The Terminator Versus the Machines"

Karim Ibrahim[1*]

*Correspondence:
Ibrahim.K@gust.edu.kw

[1] Gulf University for Science and Technology, Block 5, Building 1, Mubarak Al-Abdullah Area, West Mishref, Kuwait

Abstract

The release of ChatGPT marked the beginning of a new era of AI-assisted plagiarism that disrupts traditional assessment practices in ESL composition. In the face of this challenge, educators are left with little guidance in controlling AI-assisted plagiarism, especially when conventional methods fail to detect AI-generated texts. One approach to managing AI-assisted plagiarism is using fine-tuned AI classifiers, such as RoBERTa, to identify machine-generated texts; however, the reliability of this approach is yet to be established. To address the challenge of AI-assisted plagiarism in ESL contexts, the present cross-disciplinary descriptive study examined the potential of two RoBERTa-based classifiers to control AI-assisted plagiarism on a dataset of 240 human-written and ChatGPT-generated essays. Data analysis revealed that both platforms could identify AI-generated texts, but their detection accuracy was inconsistent across the dataset.

Keywords: Generative AI, AI-assisted plagiarism, Language assessment, Academic integrity, ESL composition

Introduction

In November 2022, the world was taken aback by the public release of OpenAI's ChatGPT, an interactive artificial intelligence chatbot powered by GPT-3.5 (generative pre-trained transformative model). GPT-3.5 is a natural language processing (NLP) model (OpenAI, n.d.-a) that can respond to users' requests as a conversation agent by performing a variety of NLP tasks, including translating texts, generating texts, summarizing content, gathering information, telling stories, composing music, or developing computer code (Gao et al., 2023; Khail & Er, 2023; Cotton et al., 2023; Lund & Wang, 2023; Yeadon et al., 2022). Along with its ability to perform a variety of cognitive tasks (Susnjak, 2022), ChatGPT's impressive ability to generate coherent, intelligible, human-like language (Cotton et al., 2023) has attracted 100 million users in 2 months, making it the fastest-growing internet app ever created (Hu, 2023) with an average of 1.8 billion visits per month (Carr, 2023). OpenAI's GPT-3.5 is not the first large language model (LLM), but it represents a quantum leap in language modeling performance that supersedes competing natural language models currently available (Brown et al., 2020; Chen et al., 2023; Ouyang et al., 2022). This leap has produced a model with an unprecedented ability to emulate human cognitive abilities concerning NLP, complex problem-solving, and critical thinking (Solaiman et al., 2019; Lund & Wang, 2023; Radford, 2018; Radford, 2019). Thus, the release of ChatGPT marked a revolutionary transformation in the NLP capabilities of artificial intelligence and the dawn of a new era of automation of what until recently were believed to be uniquely human capabilities.

Despite the tremendous potential implications of ChatGPT technology for different fields, the evolutionary development of ChatGPT into a brilliant and expansive technology capable of simulating human intelligence and able to learn and encompass new expertise seamlessly (e.g., Bommarito II & Katz, 2022) has sent shockwaves in every field, shifting how users view AI-technologies (Haque et al., 2022). Ironically, similar to Schwarzenegger's *Terminator* movie franchise, where humans panicked upon realizing that AI can give machines the power to revolt against the human rule, academics and thinkers in every field are wondering about the potential misuses of the technology and how they could affect every facet of modern life from swaying public opinion in politics to transforming the face of the job market in economics and everything in between (e.g., Zeller et al., 2019). One of the seminal concerns about ChatGPT (and generative language models in general) is the potential misuse of its generative abilities for plagiarizing in higher education (Haque et al., 2022). Specifically, with the ability of ChatGPT to generate coherent texts (Cotton et al., 2023) and respond to critical questions (Susnjak, 2022), there is a growing fear that it could be widely utilized by college students to complete essay assignments (Francke & Benett, 2019; King & ChatGPT, 2023; Yeadon et al., 2022). Thus, the disruptive and sudden rise of ChatGPT as a powerful generative technology marks a new age of AI-assisted plagiarism that constitutes an existential threat to higher education as it brings into question the reliability of its assessment practices and could devaluate the college degree (Cotton et al., 2023). The threat of AI-assisted plagiarism is more eminent in ESL composition contexts where educators are challenged by the wide prevalence of plagiarism (Pecorari & Petric, 2014) either because ESL students do not fully grasp the Western concept of plagiarism (Ali et al., 2012) or because they draw on it as a resource to compensate for their limited L2 proficiency (Waltzer & Dahl, 2023). And, in the face of this existential threat, ESL educators have very little guidance or reliable resources in detecting AI-generated texts (e.g., Mitchell, 2022), especially since human raters have difficulty distinguishing between AI-generated and human-written texts (Brown et al., 2020; Hovy, 2016; Ippolito et al, 2019; Solaiman et al., 2019) and traditional plagiarism detection software fail to detect AI-generated texts (Khalil & Er, 2023).

With the rising fear of AI-generated texts' proliferation, a growing body of research in NLP has explored different approaches to detecting AI-generated texts (e.g., Bakhtin et al., 2019), including watermarking AI-generated texts (e.g., Kirchenbauer et al., 2023), detecting traces of language modeling in generated texts (e.g., Tay et al., 2020), and training AI models to discriminate between AI-generated and human-written texts (e.g., Fangi et al., 2021). One of the promising detection approaches that can help ESL educators control AI-assisted plagiarism is training an AI-based classifier (i.e., a program that classifies texts based on specific criteria) to distinguish between human-written and AI-generated texts (Jawahar et al., 2020). This way, AI-based classifiers can use the power of AI to identify AI-generated text and aid educators in controlling AI-assisted plagiarism. Early research in this field suggests that fine-tuned classifiers (Solaiman et al., 2019; Tay et al., 2020; Mitchell et al., 2023), which are complex classifiers based on generative models and trained further on text detection, can achieve higher accuracy in detecting AI-generated texts compared to other classifiers (e.g., Fangi et al., 2021). Specifically, fine-tuned classifiers based on RoBERTA model (Liu et al., 2019), a fine-tuned AI detector trained on the output of existing large language models (e.g., GPT-2), has demonstrated higher detection accuracy compared to other classifiers (Solaiman et al., 2019; Jawahar et al., 2020; Fagni et al., 2021). However, these findings are preliminary, and further research on the model and parameters of its use is needed to develop reliable AI classifiers. Alongside NLP research, several plagiarism-detection companies and language modeling companies responded to the growing fears of AI-assisted plagiarism by developing online AI text detection platforms; however, most of these commercial platforms are experimental, and their providers suggest that their results are not always accurate. For instance, Turnitin Feedback Studio warns against using the results of its AI detector in making decisions about academic integrity. Turnitin reported that in-house experiments on their AI detector found that essays flagged as potentially AI-generated with a less than 40% probability are likely to be false positives (Turnitin, 2023). In addition, most commercially available AI-detection platforms are powered by simple detection models (i.e., simple classifiers), which NLP research suggests are of limited reliability (e.g., Solaiman et al., 2019). The questionable reliability of commercial detectors limits their applicability in controlling plagiarism, as decisions about academic integrity need to be based on solid grounds. The fact that available research on AI-text classifiers is still maturing and commercial detectors are still experimental leaves L2 educators with almost no viable resource to face the challenge of AI-assisted plagiarism. This predicament underlines the pressing need for cross-disciplinary research in L2 assessment that draws on NLP research on AI-text generation and detection to evaluate classification approaches and commercial detection platforms to identify viable and reliable AI-detection approaches that can help L2 educators control AI-assisted plagiarism.

To address the challenges of AI-assisted plagiarism in ESL writing contexts and explore this uncharted research territory, the present cross-disciplinary descriptive study examined the potential of AI-based classifiers to detect AI-assisted plagiarism in ESL composition classes. Specifically, the researcher evaluated the performance of RoBERTa-based AI text detection platforms: *GPT-2 Output Detector Demo* and *Crossplag Detector*. These platforms were selected because they are (a) based on a robust detection model supported by empirical research (i.e., RoBERTa) and (b) are publicly available to educators for free. Both detectors were trained on GPT-2 data (OpenAI, n.d.-b). The study will use each detector separately to evaluate a data set of AI-generated and human-written texts and then compare the performance of both detectors. It is hoped that this line of research can shed light on (a) the effectiveness of publicly available AI-based detectors in controlling AI-assisted plagiarism, (b) the effectiveness of RoBERTa model in AI text classification (in realistic contexts with longer, more sophisticated, and more diverse texts), and (c) the reliability of detectors trained on earlier versions of GPT in the face of more recent and larger versions of the model (given that both detectors were trained on GPT-2 and are used to detect the outputs of GPT-3.5). For discussion, *AI-*, *machine-generated*, and *AI-generated texts* will be used interchangeably to refer to texts generated by an AI language model. This paper will start with an overview of early research on the implications of ChatGPT for plagiarism and a survey of different approaches to detecting machine-generated texts. After that, the methodology and research questions of the study will be presented. Next, the study's findings will be presented, analyzed, and discussed in light of current research. Finally, the study will conclude with implications, limitations, and directions for future research.

Literature review

ChatGPT

ChatGPT is an interactive artificial intelligence chatbot (i.e., an internet-based computer program that can interact verbally with users) released by *OpenAI* on November 30, 2022 (OpenAI, n.d.-a). ChatGPT has gained massive popularity among internet users worldwide due to its unique ability to generate intelligible and coherent natural language texts that are difficult to distinguish from human-written ones. The chatbot can perform different language processing tasks, including answering complex questions, generating articles, poems, stories, or computer code; and reading, summarizing, paraphrasing, editing, or translating texts. The chatbot is powered by the Generative Pre-trained Transformer model (GPT-3.5), a 175B parameter auto-regressive natural language processing model (Radford et al., 2018).[1] As a generative language model (Pavlik, 2023), GPT is a machine-learning neural network trained to analyze and interpret natural language texts and model the linguistic system of a language (MacNeil et al., 2022). After training, generative models can transfer their understanding of natural language systems to new language processing contexts and generate coherent natural language texts (Pan et al., 2010). A generative model's capacity is measured by the number of parameters (i.e., modeling criteria acquired in training) used to model a language and generate texts. GPT is not the only generative language model (other models include Megatron-Turing Natural Language Generation by *Nvidia*, BARD Model by *Google*, and Chinchilla and Gopher AI by *DeepMind*). Still, the reason for GPT's outstanding performance and immense popularity lies in how the model was developed through iterative cycles of innovative machine-learning training.

The most common deep-learning approach used to train a language model is supervised training, which involves training a model on large quantities of labeled data (i.e., tagged texts) to perform task-specific linguistic functions (Radford, 2019). Supervised training involves training a machine learning system on a dataset of examples exhibiting correct behavior on a specific linguistic task so that the system imitates that behavior on similar tasks. However, the use of single-task training (i.e., examples exhibiting correct behavior on one task) on single classes of datasets (i.e., types of texts) has limited the generalizability of these systems (i.e., they have problems generalizing correct behaviors to different kinds of texts or domains), which impacts the performance of a language model (Radford, 2019). The alternative to supervised training is unsupervised

[1] OpenAI released a newer version of their model, GPT-4, however, their popular and widely accessible platform ChatGPT is still powered by GPT-3.5 and GPT-4 is only limited to *ChatGPT plus*, which is a paid service that is only available to select users.

training, which entails training a model on a vast dataset of unlabeled data (i.e., text samples), allowing the model to detect the underlying structures and patterns of a language system from the data (Lee et al., 2018); however, unsupervised training required vast quantities of data, which makes this type of training expensive and time-consuming (Radford, 2018). To overcome the challenges of unsupervised training and the limitations of supervised training, OpenAI developed an innovative semi-supervised approach that combines unsupervised pretraining (i.e., initial training of a language model) and supervised fine-tuning (i.e., more specific further training of a pre-trained model) that they used to train their GPT model (OpenAI, n.d.-a). OpenAI used *Common Crawl* to train the model on a dataset of a trillion words of internet texts (175 billion parameters) in supervised training (Brown, 2020) and then fine-tuned the model using reinforcement learning from human feedback approaches (Christiano et al., 2017) to improve output alignment and consistency with users' intent, along with improving task-agnostic performance (i.e., modeling a language across different functions and task) and transfer learning capabilities (Ouyang et al., 2022). As a result of this innovative hybrid training approach, ChatGPT can generate humanlike texts that are difficult to distinguish from human-written ones. This powerful ability brings into question the potential misuse of the technology to facilitate plagiarism.

AI-assisted plagiarism

The impressive ability of ChatGPT to engage in critical thinking and generate coherent, fluent texts that are difficult to distinguish from human-written ones has raised serious concerns about potential misuses of the technology to plagiarize on college assessments as students can submit machine-generated texts as their work (Francke & Bennet, 2019; Haque et al., 2022; Yeadon et al., 2022; Cotton et al., 2023; Susnjak, 2022; King & ChatGPT, 2023; Khalil & Er, 2023; Gao et al., 2023). As Yeldon et al. (2022) put it:

> *"We may be at the beginning of an AI revolution. In order to facilitate authentic assessment it is vital that we are aware of the capabilities of this technology and its ramifications on the way that credited work is assessed. In the present case, it is hard to avoid the conclusion that non-invigilated assessments based on short-form essay questions are already no longer fit for purpose; they are simply too vulnerable to current AI text-completion technologies, which can produce creditable content cheaply and quickly" (p.11-12).*

Not only can AI-assisted plagiarism undermine higher education and devaluate college degrees by misrepresenting students' educational performance on course assessments (Cotton et al., 2023), but they are also more challenging to identify using traditional plagiarism detection approaches adopted by higher education institutions since they constitute original texts that were not reproduced and do not get flagged on traditional plagiarism detection systems (Khalil & Er, 2023). These severe concerns have led to some preliminary investigations into the potential of ChatGPT to facilitate AI-assisted plagiarism.

One of the earlier attempts to probe into the potential misuses of ChatGPT to violate academic integrity was Francke and Bennett's (2019) case study investigating the potential impacts of GPT-2 model (OpenAI, n.d.-b) on the proliferation of plagiarism in higher education. Using focus-group interviews, the researchers gathered data from two groups of academics on the quality of ChatGPT-generated texts and their potential to spread plagiarism in higher education. Data analysis revealed that AI can generate intelligent texts that simulate human writing and increase the proliferation of plagiarism in higher education. Similarly, but with a broader focus on internet users' reaction to generative AI, Haque et al. (2022) conducted a sentiment analysis of 1732 Tweets from early adopters of ChatGPT to explore the public response to the new technology. Using a mixed-methods design and topic modeling of discussions, data analysis revealed that early adopters of the technology perceived ChatGPT to be a disruptive technology that is likely to transform traditional education (both in positive and negative ways) and expressed serious concerns that students can misuse it to plagiarize on homework assignments. Also, a few studies have examined ChatGPT's perspectives on the implications of generative AI for academic integrity. For example, Cotton et al. (2023) explored ChatGPT's views on the opportunities and challenges it poses to higher education. They reported that one of the critical challenges the system acknowledged was the possibility that students could misuse ChatGPT's ability to generate coherent texts to commit plagiarism. Similarly, King and ChatGPT (2023) inquired about ChatGPT's opinion on college students' potential misuse of ChatGPT to cheat on essay assignments and reported that the platform supported the possibility that students can feed their assignments to ChatGPT as prompts and present generated responses as their original work.

A few studies have also explored the potential of generative AI to facilitate plagiarism by investigating its ability to engage in critical thinking and generate coherent academic essays. For instance, Susnjak (2022) explored the ability of ChatGPT to engage in critical thinking by evaluating its output in response to critical thinking questions. Using the universal intellectual standards (Paul, 2005) to measure logical reasoning and critical thinking, the researcher analyzed the model's responses for clarity, accuracy, precision, depth, breadth, logic, persuasiveness, and originality. Data analysis revealed that ChatGPT can understand the context of a prompt, engage in critical thinking, and generate logical and coherent responses, demonstrating depth and breadth of knowledge. The researchers concluded that ChatGPT can reason critically and express its ideas in logical, coherent prose that is indistinguishable from human-written texts and, thus, poses a severe concern for academic integrity.

Similarly, but with an explicit focus on essay composition skills, Yeadon et al. (2022) investigated the threat of AI-generated essays to academic integrity by examining the quality of academic papers generated by ChatGPT. Using an essay-question prompt from a Physics course, the researchers gathered ten AI-generated scripts and had them marked independently by five human raters. The samples achieved an average score of 71%, close to the average student score in the course. The researcher inferred from the results that students in the lowest third of the cohort would have a solid motive to use ChatGPT to complete their assignments. In another study, Bommarito and Katz (2022) explored the ability of ChatGPT to complete tasks that require depth of knowledge and complex semantic understanding by having the system take the MCQ section of the BAR exam. The study reported that ChatGPT achieved 50.3% on the practice exam and concluded that ChatGPT exhibited an advanced ability to understand complex questions and gain depth of knowledge in domain-specific tasks.

Even though available research on AI-assisted plagiarism is scarce in quantity, broad in scope, and exploratory in nature, it indicates that ChatGPT (and generative AI) could pose a serious threat to academic integrity and calls into question available means of detecting AI-generated texts as fundamental resources to controlling AI-assisted plagiarism.

AI-generated text detection

In response to the growing concerns about potential misuses of generative AI, such as the mass-generation of web texts (e.g., tweets) to spread misinformation or manipulate public opinion (e.g., Zeller et al., 2019), NLP researchers trained text classification algorithms to classify texts as either human-written or machine-generated as a means to detect machine-generated texts (Solaiman et al., 2019). Classifiers are trained to identify the characteristics of a machine-generated text based on their semantic understanding of a text (Ippolito et al., 2019). Machine-generated text detection classifiers fall into three categories based on their underlying detection mechanism and training approach: (1) simple classifiers, (2) zero-shot classifiers, and (3) fine-tuned classifiers (Jawahar et al., 2020).

Simple classifiers

Simple classifiers are basic machine-learning algorithms trained from scratch to discriminate between human-written and machine-generated texts (Solaiman et al., 2019). A basic form of this classifier is the "bag-of-words," which is typically a logistic regression model trained on large sets of machine-generated and human-written texts to identify subtle stylistic or linguistic differences between the two categories (Jawahar et al., 2020). Solaiman et al. (2019) experimented with training a simple classifier using logistic regression to discriminate between GPT-2 output and internet articles across different model sizes and sampling methods. The researchers found that detection accuracy varied depending on the generative model's size, as texts generated by larger models were more challenging to detect. Another form of simple classifiers is machine-configuration detectors, which distinguish machine-generated texts by identifying distinct traces of the modeling choices used to generate a text (e.g., model size, decoding method, or model architecture) and use these traces to determine the origin of a text (Jawahar et al., 2020). For instance, Tay et al. (2020) empirically investigated the potential use of textual artifacts that appear in machine-generated texts due to modeling and sampling choices to identify their generative language model. Using texts generated by GROVER (Zeller et al., 2019) language model in different modeling conditions (e.g., model size) and the CNN/Dailymail news corpus (as human-written texts), the researchers trained and tested a classifier model to distinguish between texts based on modeling choices' artifacts. The experiments revealed that modeling choices left artifacts in the generated texts that could be used to predict modeling choices based on the generated texts alone and could be the basis of a new approach to machine configuration detection. These studies suggest that large language models challenge simple classifiers trained from scratch. Still, they can be used to identify traces of a language model in a generated text, which can contribute to future innovations in AI-text detection.

Zero-shot classifiers

Zero-shot classifiers employ a pre-trained generative model (e.g., GPT-2) to identify texts generated by itself or comparable models without additional supervised training (Jawahar et al., 2020). An example of a zero-shot classifier is Gehrmann et al. (2019)'s Giant Language Model Test Room (GLTR), which utilizes BERT (Devlin et al., 2018) and GPT-2 (Radford et al., 2019) models to discriminate between machine-generated and human-written texts based on the probability of word distribution in these models (i.e., the likelihood that the distribution of words in a text was generated by a given model). The researchers tested GLTR empirically with machine-generated and human-written articles from different sources. The analysis revealed that GLTR offered frequency and ranking information that improved human raters' accuracy in detecting machine-generated texts. Similarly, Solaiman et al. (2019) used a zero-shot detector based on GPT-2 model to assess text origin based on total log probability (i.e., GPT-2 estimation of the likelihood that word distribution in a text is similar to its own generated text). However, they found that it performed poorly compared to a simple classifier. Similarly, Mitchell et al. (2023) developed a zero-shot classifier based on GPT-3, DetectGPT, that detects GPT-generated texts based on their log probability. The researchers conducted several experiments comparing the performance of DetectGPT to previous zero-shot and two fine-tuned classifiers. The experiments revealed that DetectGPT outperformed previous zero-shot classifiers, but underperformed fine-tuned classifiers on general domain topics. These studies indicate that zero-shot classifiers perform poorly in AI text detection but can offer data that aid human-rating of AI-texts.

Fine-tuned classifiers

These classifiers are pre-trained generative models that are fine-tuned for text classification with supervised training on detection examples. An example of fine-tuned classifiers is Liu et al. (2019)'s fine-tuning of BERT model (Delvin et al., 2019), RoBERTa (Robustly optimized BERT approach), which uses a modified training approach involving longer training cycles, bigger training batches, and longer data sequences. The researchers trained RoBERTa with over 160 GB of text and found that RoBERTa's optimized training methods, especially data size and variety, resulted in performance improvement compared to BERT. Similarly, Zeller et al. (2019) developed a linear classifier with the GROVER model to detect its own generated fake news articles. The researchers compared GROVER fine-tuned classifier to several fine-tuned classifiers: BERT, GPT-2, and FastText. The results revealed that the fine-tuned GROVER classifier outperformed other fine-tuned classifiers in detecting its own generated texts. Similarly, Solaiman et al. (2019) fine-tuned RoBERTa classifier to detect GPT-2-generated texts and found that it could detect GPT-2-generated texts with a higher accuracy than a GPT-2-based detector. Similarly, Jawahar et al. (2020) experimented with using RoBERTa to distinguish between GPT-2-generated and human-written Amazon reviews. They found that RoBERTa detector needs training with several thousands of examples to achieve high accuracy in detection. In another study, Fagni et al. (2021) used a deepfake and human-written tweets database to compare the performance of simple, zero-shot, and fine-tuned classifiers. The researchers classified 25,572 tweets using a simple Bag-of-word classifier, BERT model (as a zero-shot classifier), and several fine-tuned classifiers:

XLNet. RoBERTa, BERT, and DistilBERT. The experiments revealed that fine-tuned classifiers outperformed zero-shot and simple classifiers and that RoBERTa achieved the highest accuracy, outperforming other fine-tuned classifiers by a wide margin. These studies revealed that fine-tuned classifiers, especially RoBERTa model, can outperform other classifiers in AI-text detection.

This brief survey of the literature on AI-based classifiers suggests that fine-tuned classifiers, especially RoBERTa-based classifiers, constitute the most promising resources for detecting machine-generated texts; however, research has barely scratched the surface of this novice area, and further research is needed to gain a deeper understanding of the strengths and limitations of RoBERTa-based classifiers.

In addition, the only access that L2 educators have to AI classifiers is through commercial platforms claiming the ability to detect AI-generated content, including *GPTZero*, *Originality.AI*, *ZeroGPT*, *AI Writing Indicator* (by Turnitin), and *Writefull.* However, most of these platforms are still in their beta version (i.e., drawing on user interaction as training to improve their performance); thus, they cannot provide a reliable measure of text originality that constitutes valid grounds for making assessment decisions in higher education. For instance, OpenAI has developed an AI Text Classifier based on a fine-tuned GPT model to predict the likelihood that a text is machine-generated, but they indicated that "the classifier is not always accurate; it can mislabel both AI-generated and human-written text" (OpenAI, n.d.-b). Moreover, despite the growing research on text classifiers, most of the available commercial AI-detection platforms rely on training simple classifiers in-house and do not refer to any theoretical or empirical foundations for their platforms, which brings the validity of their results into question.

One of the few publicly available detection platforms that adopted an empirically supported underlying classification mechanism is *GPT-2 Output Detector Demo*. This fine-tuned classifier was developed by training RoBERTa model (Liu et al., 2019) with the output of the 1.5 billion GPT-2 model (OpenAI, n.d.-b). This classifier generates a score representing the probability that the examined text is human-written. OpenAI claims that in their in-house testing of the classifier, it achieved a 95% accuracy in machine-generated text detection. However, they warn against using its results to make "allegations of academic misconduct" (Hugging Face). Another available platform based on a fine-tuned classifier is *Crossplag's AI Content Detector*. This classifier is also based on RoBERTa model and was fine-tuned using OpenAI's GPT-2 1.5 billion parameters dataset (Crossplag). Crossplag's classifier generates a score indicating the probability that the examined text is machine-generated. Like OpenAI, Crossplag suggests that their classifier can accurately detect machine-generated texts. However, they also suggest that their platform is still in beta and unavailable for institutions (Crossplag). So even though these detection platforms have research foundations (i.e., RoBERTa model), they still need further testing to ensure the reliability of their results and the validity of academic integrity decisions that can be made based on them. To date and the researcher's knowledge, only one study has examined the reliability of an AI-text detection platform. Gao et al. (2023) examined the reliability of *GPT-2 Output Detector Demo* by comparing its originality scores for 50 human-written and 50 ChatGPT-generated abstracts. The researchers suggested that the detector accurately discriminated between original and generated abstracts. However, the wide range of variability in the results and the small sample size warrant further investigation. To extend this critical line of research, improve our understanding of methods for controlling AI-assisted plagiarism, and guide L2 educators in using AI-detection platforms, the present study will investigate the performance of AI-based detection platforms. And, given the empirical support for RoBERTa model and in line with Gao et al. (2023) suggestion to explore different platforms, the present study will test and compare the performance of *GPT-2 Output Detector Demo* and Crossplag *AI Content Detector.*

Methodology

Design

Generally speaking, descriptive research is well-suited for investigating L2 teaching and assessment since controlled experimental research is challenging to achieve in classroom settings (Nassaji, 2015). Given that the present study is exploring a new area where previous research is lacking and understanding is limited, descriptive research approaches would be most appropriate (Fireman Kramer, 1985). Within a descriptive research paradigm, the researcher utilized a comparative research design that involves comparing the dependent variables in several groups without manipulating the independent variable (Baker, 2017). Specifically, the study examined the performance of two AI-detection platforms, GPT-2 Output Detector, and Crossplag Detector, by analyzing and comparing their originality scores for 120 human-written and 120 machine-generated essays. In this design, the text type (i.e., human-written or machine-generated) constitutes the independent variable. The originality scores, defined as a percentage reflecting the probability that a text is human-written, comprise the dependent variable. The study investigated the following research questions:

(a) How effective is GPT-2 Output Detector Demo in detecting machine-generated essays?
(b) How effective is Crossplag AI Content Detector in identifying machine-generated essays?
(c) What is the difference in classification effectiveness between GPT-2 Output Detector Demo and Crossplag Detector?
(d) What does the comparison of two RoBERTa-based detectors suggest about the robustness of AI-detection of AI-generated texts?

Data collection

For the human-written dataset, a convenience sample of 120 student essays was gathered from a research-argument project in an advanced first-year composition course offered at a major American university in Kuwait in Fall 2021, Spring 2022, and Summer 2022. The researcher gathered samples from previous semesters before ChatGPT was publicly available (in November 2022) to ensure that the essays collected were human-written and that the chances of sample contamination with AI-generated essays were minimal. As for the machine-generated dataset, the researcher asked ChatGPT to generate 120 essays using the January 2023 version (GPT-3.5). The researcher used the research argument project description as a prompt for ChatGPT to ensure that the machine-generated dataset is consistent with the human-written dataset in format and development specifications (e.g., length, number of paragraphs, number of references) and minimize the chances of stylistic inconsistencies that can trigger text detection. To make the machine-generated dataset more comparable to the human-written one, the researcher specified the topics for the machine-generated essays in the prompt and matched them to those of the human-written dataset. Generated essays were transformed into plain texts, excluding the titles and references to minimize stylistic inconsistencies that could flag machine-generated texts. Both datasets were fed into GPT-2 Output Demo Detector and Crossplag Detector to generate originality scores for each dataset.

Data analysis

To measure the effectiveness of each platform in detecting machine-generated texts, the originality scores for the human-written and machine-generated datasets were compared and analyzed for each detector separately using descriptive and inferential statistical methods. Descriptive data analysis from both platforms revealed that the data did not meet the normal distribution condition required for parametric inferential statistical analysis of variance (ANOVA or *T*-tests). So, non-parametric tests were used to analyze the differences between the two datasets based on grade ranking of scores (Lowie & Seton, 2013). Data were analyzed using IBM SPSS version 28. To compare the effects of text type on text originality scores, Mann–Whitney *U* tests were performed on the originality scores of both datasets for each platform. GPT-2 Output Detector Demo generates an originality score consistent with the study's definition of the construct of "originality." So, GPT-2 Outputs were used without modification. Crossplag, on the other hand, generates a score representing the probability that the examined essay was machine-generated (a 0% score means that the platform is confident that an essay is human-written). So, Crossplag originality scores were modified according to the formula *originality score = 100-Crossplag scores* to generate scores consistent with the study's operationalization of the construct of originality and comparable to GPT-2 output Detector Demo Demo scores. After the scores for both datasets were compared for each detector to assess its ability to identify machine-generated texts, a confusion matrix (Fawcett, 2006) was developed to get a more accurate estimate of each detector's accuracy. A confusion matrix is a predictive analysis typical in machine learning research that uses probability data to predict the precision and accuracy of a model (Deng et al., 2016). Confusion matrixes can offer a reasonable accuracy estimate as they determine the percentage of false positives and false negatives an algorithm makes. Finally, the originality scores for both data sets were compared between the two detectors to compare the performance of both detectors and shed some light on the effectiveness of RoBERTa-based detectors.

Results

GPT Output Detector Demo

To get a sense of originality scores' distribution for human-written and machine-generated datasets, the researcher conducted a descriptive statistical analysis of each set of scores before comparing the score distribution for the two sets graphically (see Figs. 1 and 2). In the human-written data set (Fig. 1), descriptive statistics revealed that data had a mean of 91.4% and a standard deviation of 24.06, but the frequency

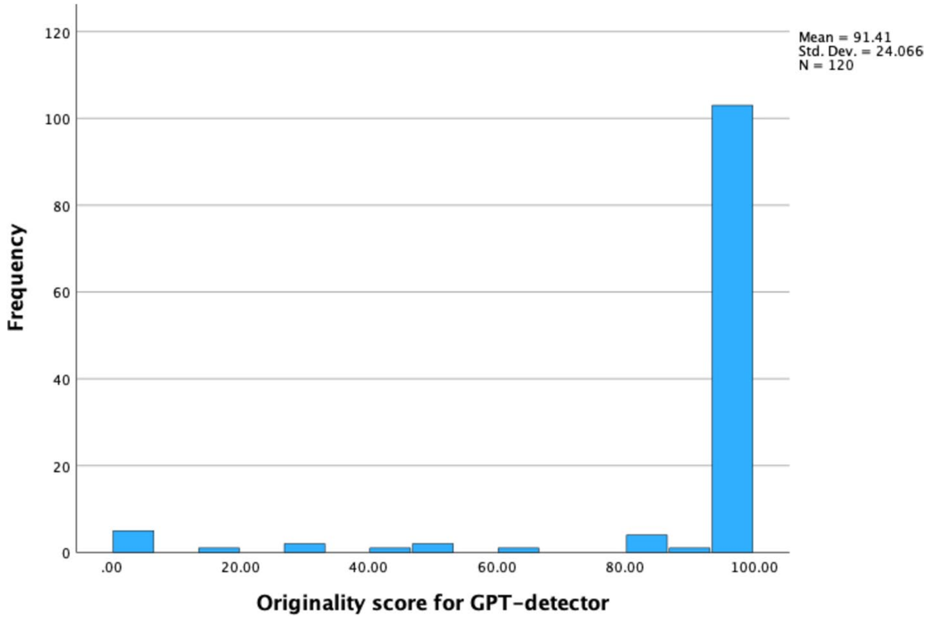

Fig. 1 Histogram representing the frequency distribution of originality scores for human-written texts in GPT-2 Output Detector Demo

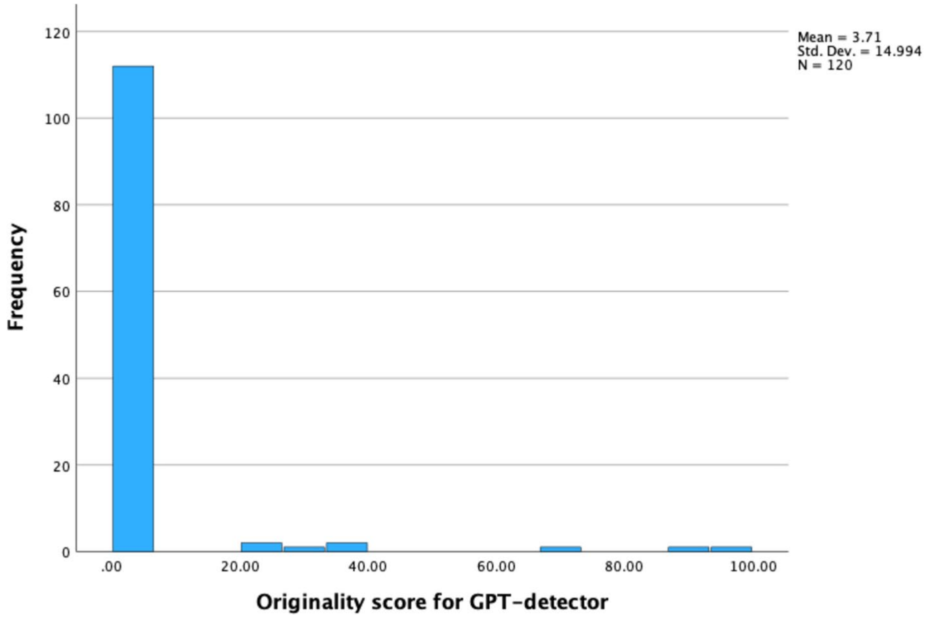

Fig. 2 Histogram representing the frequency distribution of originality scores for machine-generated texts in GPT-2 Output Detector Demo

distribution demonstrated that the data distribution violated the normal distribution assumption (skewness = −2.98; kurtosis = 7.77). The data distribution varied widely, with a range of 99.95% (minimum score = 0.03, maximum score = 99.98). The data was uni-modal with a median score of 99.98% and an interquartile range of

0.41% (25th percentile = 99.56%, 75th percentile = 99.98%). The researcher also used a Confusion Matrix analysis to assess the accuracy of the originality scores for the human-written essays by calculating the percentage of true and false positives that the detector generated. Using a 95% confidence interval for originality score accuracy (with a 5% margin of error), the researcher used 95% or above as an accurate estimate of human-written texts' originality to determine the percentage of false negatives (i.e., human-written essays that are mistakenly assumed to have low originality). Out of 120 essays in the human-written dataset, only 16 essays received an originality score below 95%, representing 13.3% of false negatives versus 86.7% of the data that were true positives.

Descriptive analysis of the machine-generated dataset (Fig. 2) revealed that data had a mean of 3.07% and a standard deviation of 14.99%, but the frequency distribution demonstrated that the data distribution violated the normal distribution assumption (skewness = 5.03; kurtosis = 26.57). The data distribution varied widely, with a range of 99.93% (minimum score = 0.02, maximum score = 99.93). The data was uni-modal with a median score of 0.02% and an interquartile range of 0.01% (25th percentile = 0.02%, 75th percentile = 0.03%). The researcher also used a Confusion Matrix analysis to assess the accuracy of the originality scores for the machine-generated dataset by calculating the percentage of true and false negatives that the detector generated. Using a 95% confidence interval for originality score accuracy (with a 5% margin of error), the researcher used 5% or less as an accurate measure of machine-generated texts' originality to determine the percentage of false positives (i.e., machine-generated texts assumed to have a degree of originality). Out of 120 essays in the machine-generated dataset, only ten essays received an originality score above 5%, representing 8.3% false positives versus 91.7% of the data that were true negatives.

To compare the scores of both data sets and explore the ability of the detector to discriminate between human-written and machine-generated essays, the researcher used descriptive statistical analysis to visualize the difference between the originality scores for both data sets (see Figs. 3 and 4). Firstly, the researcher charted a simple bar graph of mean originality scores for each data set to compare the two data sets and visualize the ability of the platform to classify essays (Fig. 3). The bar graph revealed that the scores were polarized between the two datasets.

The researcher also used SPSS to generate a histogram of originality scores classified by dataset (Fig. 4) to further explore the scores' distribution. The histogram demonstrated that the distribution of originality scores was mostly polarized as most human-written essays received an originality score in the 99th percentile, and most machine-generated essays received an originality score within the 1st percentile. However, the scores of some of the essays from both datasets varied widely and overlapped with the other dataset.

To measure the significance of the difference between the two datasets' mean scores, the researcher conducted a Mann–Whitney U test (Fig. 5), a non-parametric alternative to Independent-samples t-tests. The Mann–Whitney U test revealed a significant difference between the human-written and machine-generated essays' originality scores ($U = 163$, $P = 0.001$), with a mean rank of 61.86 for the machine-generated essay dataset and 179.14 for the human-written dataset.

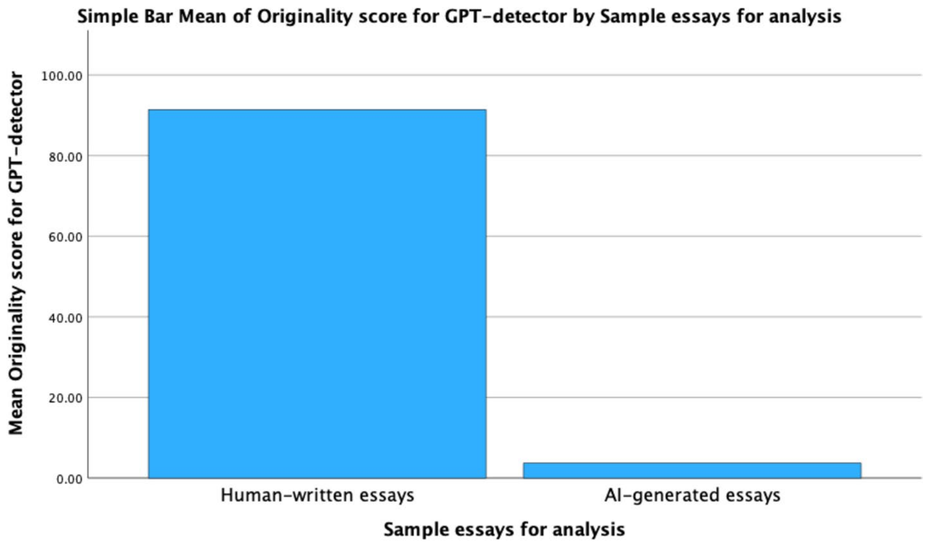

Fig. 3 Simple bar graph comparing the mean originality scores for human-written and machine-generated texts in GPT-2 Output Detector Demo

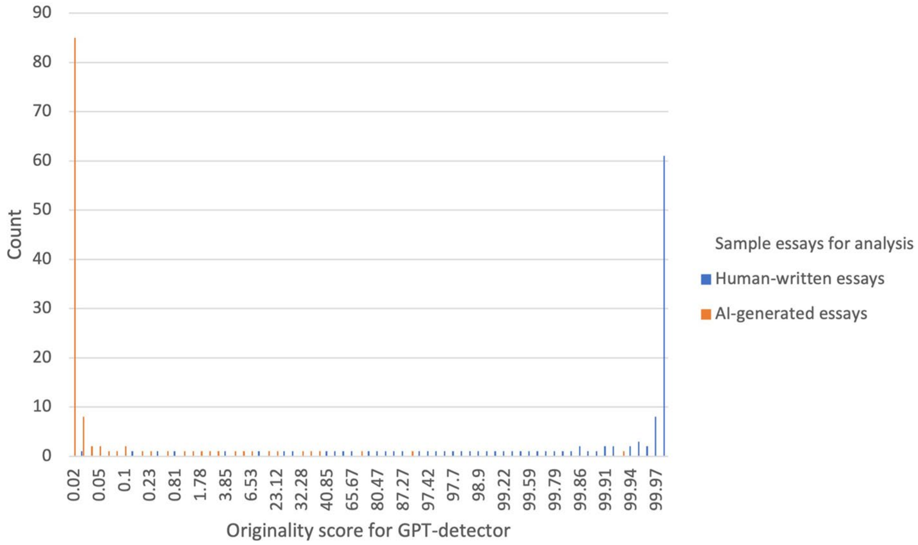

Fig. 4 Histogram comparing the mean originality scores for human-written and machine-generated texts in GPT-2 Output Detector Demo

Crossplag Detector

Following the same data analysis used for GPT-Output Detector Demo, the researcher conducted a descriptive statistical analysis of the originality scores of both datasets in Crossplag Detector separately before comparing them. In the human-written dataset (Fig. 6), descriptive statistical analysis revealed that the data had a mean originality score of 92.06% and a standard deviation of 21.43; however, the data distribution revealed a violation of the normal distribution assumption (skewness = − 3.57, kurtosis = 12.19). The data distribution varied considerably, with a range of 100% (min. = 0.001, max = 100). The data were unimodal with a median score of 99% and an interquartile range of 0.00002% (25th percentile = 99.00001%, 75th percentile = 99.00003%). Using a

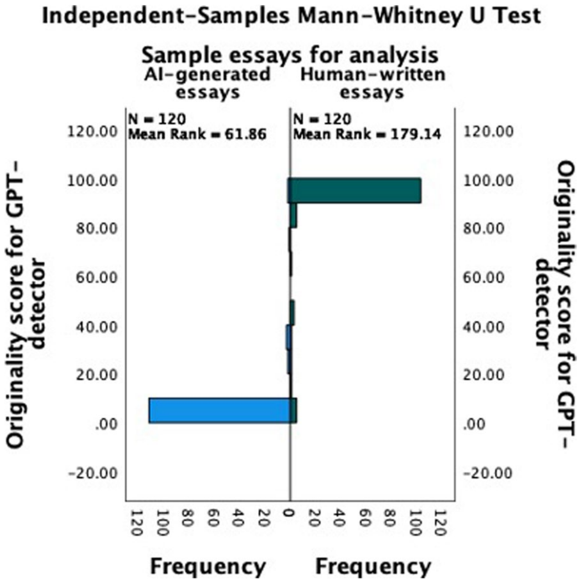

Fig. 5 Mann–Whitney *U* test comparing the mean originality scores for human-written and machine-generated texts in GPT-2 Output Detector Demo

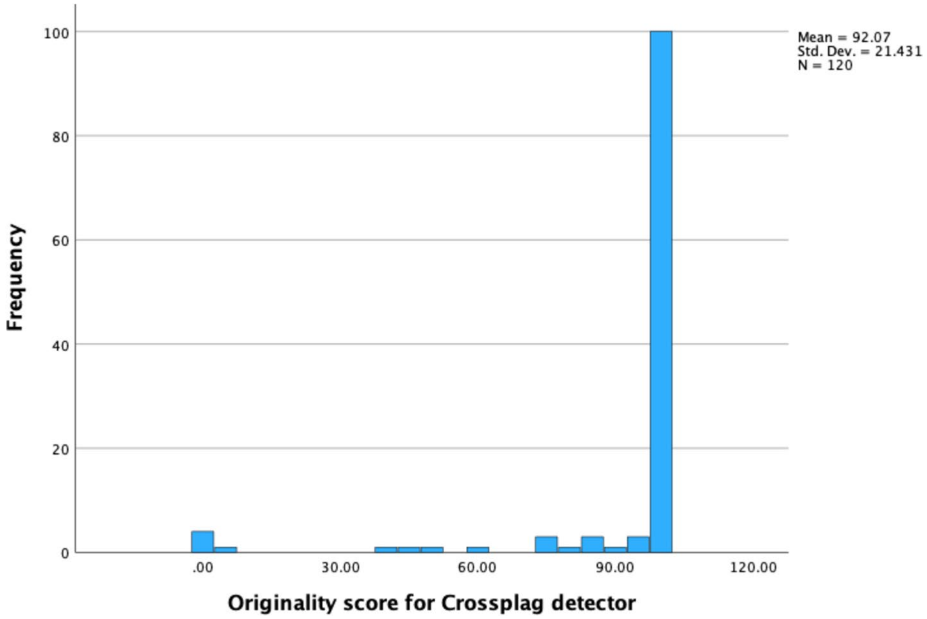

Fig. 6 Histogram representing the frequency distribution of originality scores for human-written texts in Crossplag Detector

95% confidence interval for originality score accuracy, the researcher used 95% or above to measure human-written texts' originality and estimate the percentage of false negatives. Out of 120 essays, only 18 essays received an originality score below 95%, representing 15% false negatives versus 85% of the data that were true positives.

Descriptive analysis of the machine-generated dataset originality scores (Fig. 7) revealed that data had a mean of 3.52% and a standard deviation of 15.40, but the

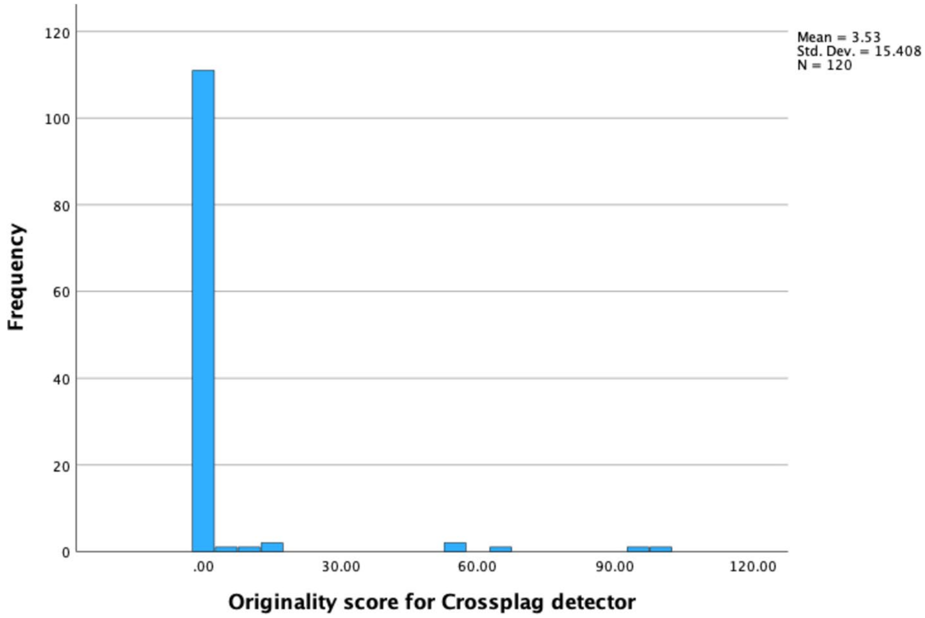

Fig. 7 Histogram representing the frequency distribution of originality scores for machine-generated texts in Crossplag Detector

frequency distribution demonstrated that the data violated the normal distribution assumption (skewness = 4.95; kurtosis = 24.78). The data distribution varied widely with a range of 99% (minimum score = 0.001, maximum score = 99). The data was unimodal with a median score of 0.00001% and an interquartile range of 0.00002% (25th percentile = 0.00001%, 75th percentile = 0.00003%). Using a 95% confidence interval for originality score accuracy, the researcher used 5% or less as an average measure of machine-generated texts' originality to estimate the percentage of false positives (i.e., machine-generated texts assumed to have a degree of originality). Out of 120 essays, only nine essays received an originality score above 5%, representing 7.5% false positives versus 92.5% of the data that were true negatives.

To compare the scores of both data sets and explore the ability of Crossplag Detector to discriminate between human-written and machine-generated essays, the researcher used descriptive statistical analysis to visualize the difference between the originality scores for both data sets. Firstly, the researcher charted a simple bar graph of mean originality scores for each data set to compare the two data sets and visualize the ability of the platform to classify essays (Fig. 8). The bar graph revealed that the scores were polarized between the two datasets.

The researcher also used SPSS to generate a histogram of originality scores classified by text type (Fig. 9). The histogram demonstrated that the distribution of originality scores was mainly polarized as the majority of human-generated essays received scores in the 99th percentile, and most machine-generated essays received scores within the 1st percentile. However, some of the scores for essays from both datasets varied widely and overlapped with the scores of the other dataset.

To measure the significance of the difference between the means scores of the two datasets, the researcher conducted a Mann–Whitney *U* test (Fig. 10). The Mann–Whitney *U*

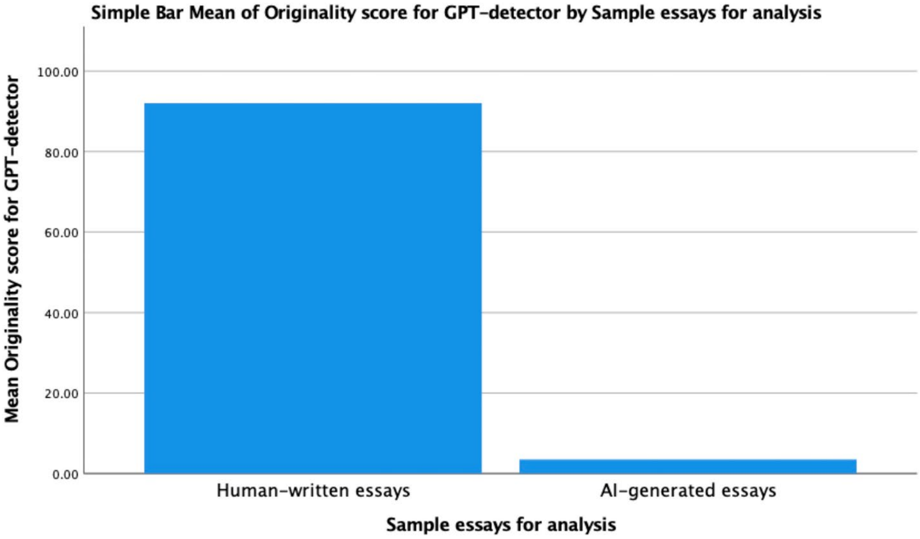

Fig. 8 Simple bar graph comparing the mean originality scores for human-written and machine-generated texts in Crossplag Detector

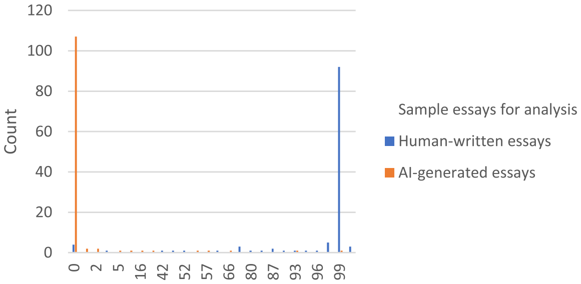

Fig. 9 Histogram comparing the mean originality scores for human-written and machine-generated texts in Crossplag Detector

test revealed a significant difference between the human-written and machine-generated essays' originality scores ($U = 363$, $P = 0.001$) with a mean rank of 63.53 for the machine-generated dataset and 177.47 for the human-written dataset.

GPT Output Detector Demo and Crossplag Detector Performance Comparison

To compare the performance of both platforms in discriminating between machine-generated and human-written texts, the researcher conducted several analyses comparing the performance of both detectors on each data set. First, s/he used a clustered simple bar graph to visualize the difference between originality scores' distributions for the machine-generated dataset across both platforms (Fig. 11). Visually, the originality scores distribution seemed consistent between the two platforms. However, Crossplag appeared to be more sensitive to machine-generated texts, as indicated by the higher concentration of scores close to 0% in the score distribution.

To assess the significance of the difference in score distribution, the researcher conducted a Mann–Whitney *U* test between the originality scores of the machine-generated dataset in both platforms (Fig. 12). The Mann–Whitney *U* test revealed a significant difference between Crossplag Detector and GPT-2 Output Demo

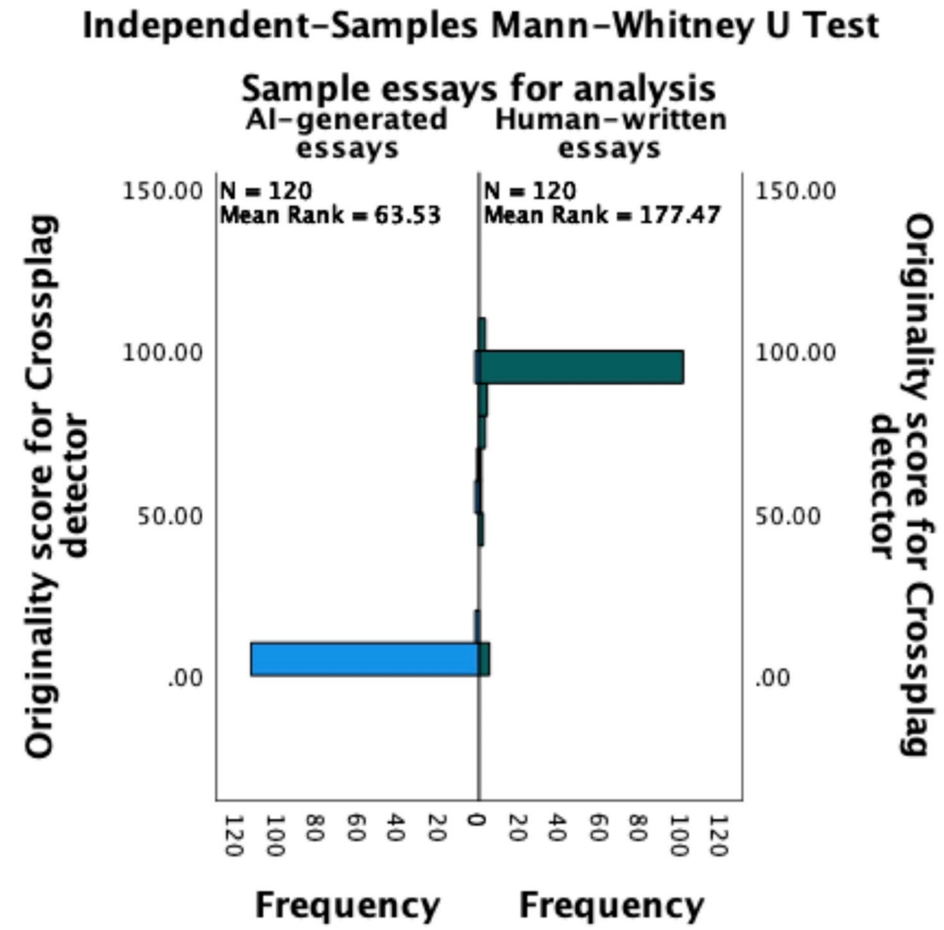

Fig. 10 Mann–Whitney *U* test comparing the mean originality scores for human-written and machine-generated texts in Crossplag Detector

originality scores in favor of the latter ($U=1459$, $P=0.001$) with a mean rank of 72.66 for Crossplag Detector and 168.34 for GPT-Output Detector Demo.

Similarly, the researcher compared the originality scores assigned by both platforms to the human-written dataset visually and nonparametrically. The researcher used a clustered simple bar graph to visualize the difference between originality scores' distributions for human-written texts in both platforms (Fig. 13). Visually, originality scores of GPT-2 Output Detector Demo appeared to be higher than those of Crossplag given that they appeared in high concentration in the 99.98% area while Crossplag scores were concentrated in the 99.00% area; however, the concentration of Crossplag Detector scores was higher than that of GPT-2 Output Detector Demo.

To assess the significance of the difference in human-written dataset scores between both platforms, the researcher conducted a Mann–Whitney *U* test between the originality scores of human-written texts in both platforms (Fig. 14). The Mann–Whitney *U* test revealed a significant difference between Crossplag Detector and GPT-2 Output Detector Demo originality scores' in favor of the latter ($U=2979$, $P=0.001$) with a mean rank of 85.33 for Crossplag Detector and 155.68 for GPT-Output Detector demo.

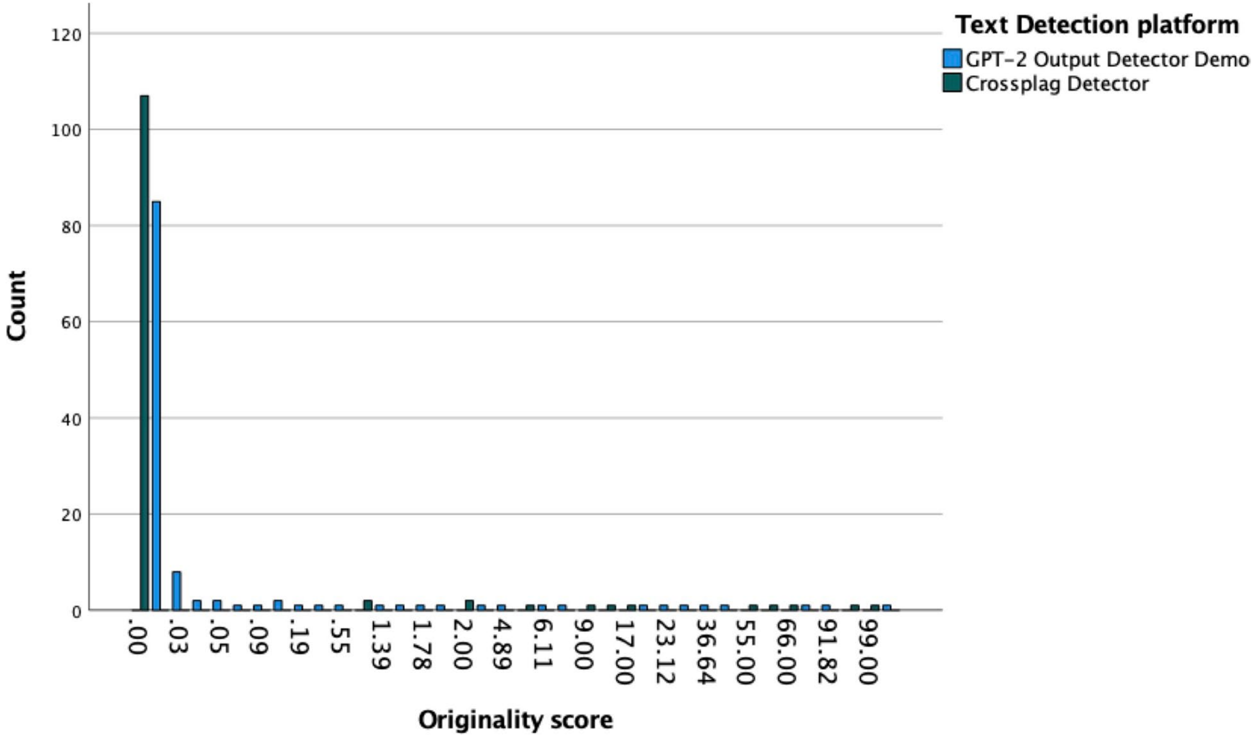

Fig. 11 Clustered simple bar graph comparing the mean originality scores for machine-generated texts across the two detectors

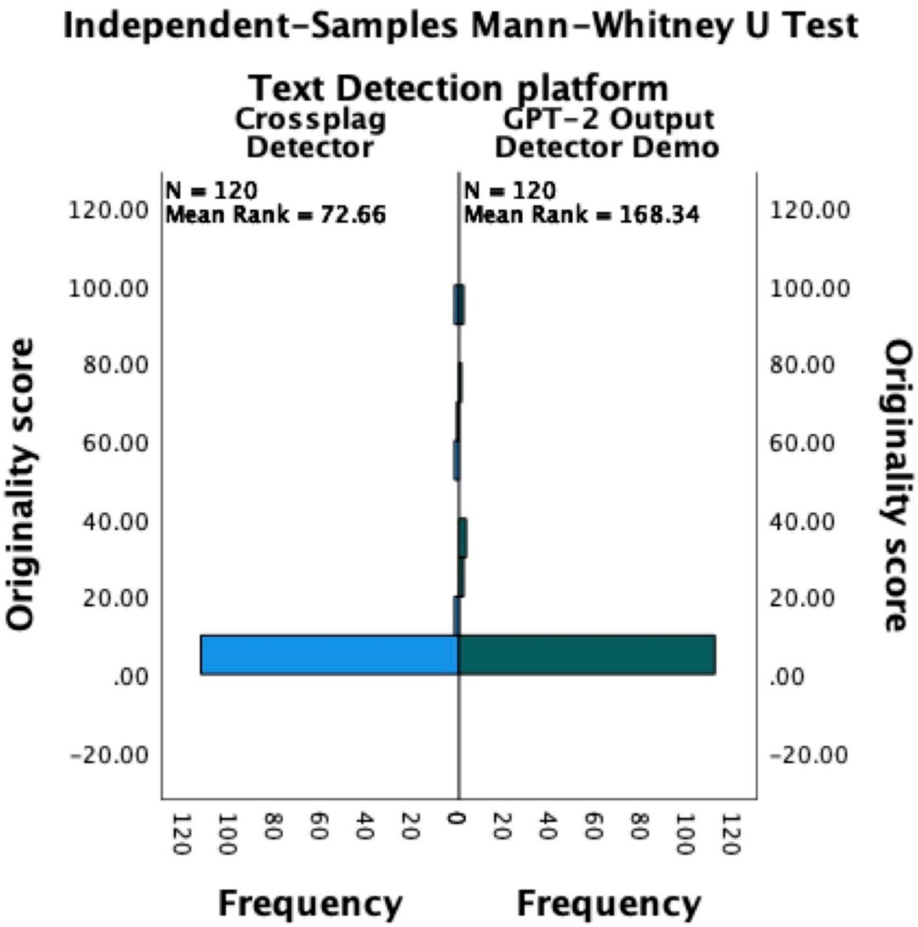

Fig. 12 Mann–Whitney *U* test comparing the mean originality scores for machine-generated texts across the two detectors

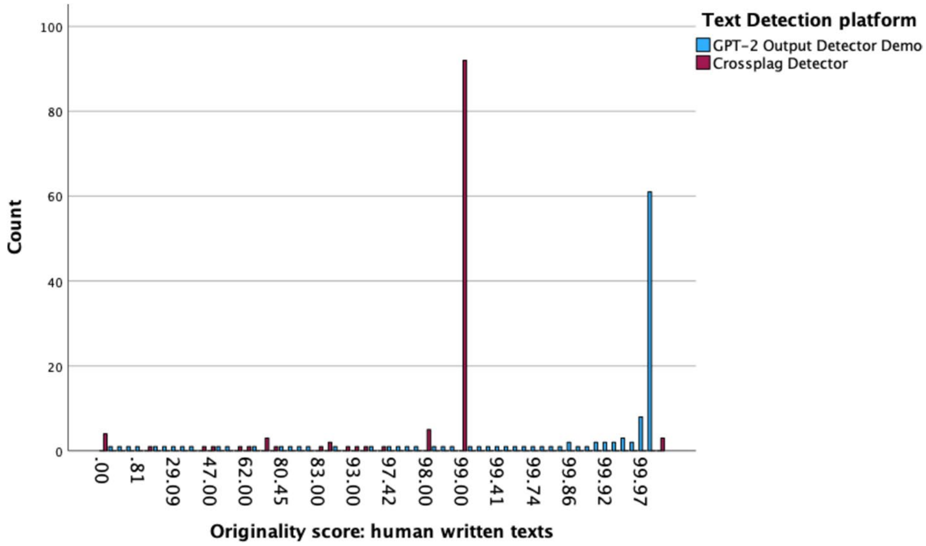

Fig. 13 Clustered simple bar graph comparing the mean originality scores for human-written texts across the two detectors

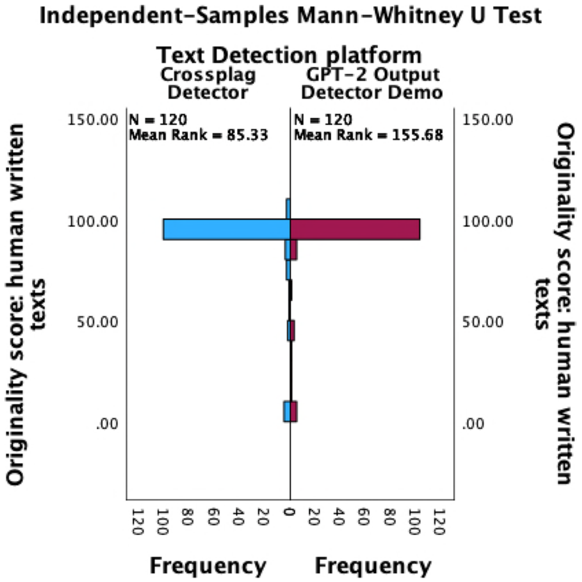

Fig. 14 Mann–Whitney *U* test comparing the mean originality scores for human-written texts across the two detectors

Discussion

RQ#1: How effective is GPT-2 Output Detector Demo in detecting machine-generated essays?

Data analysis revealed that GPT-2 Output Detector Demo could discriminate between human-written and machine-generated texts, but its detection accuracy was inconsistent across the datasets. Most of the originality scores the detector assigned to human-written texts were concentrated in the 99th percentile with a median score of 99.98% and a narrow interquartile range of 0.41%. Based on a 95% confidence interval, the detector identified 86.7% of human-written texts as highly original, but it mis-evaluated 13.3% as possibly machine-generated (i.e., false negatives). As for machine-generated texts, the detector flagged most of the samples as potentially AI-generated. Most originality scores were concentrated in the 1st percentile, with a median score of 0.02% and a narrow interquartile range of 0.01%. Based on a 95% confidence interval, the detector identified 91.7% of machine-generated texts as unoriginal, but it mislabeled 8.3% of the samples as partially original (i.e., false positives). A Mann–Whitney *U* test comparing the originality scores of both machine-generated and human-written texts indicated that the difference between the scores assigned to both groups was significant, with a mean rank of 61.86 for the machine-generated essay group and 179.14 for the human-written essay group. However, the degree of variation in the scores of mis-evaluated samples (false negatives and false positives) varied widely, indicating that the detector can confuse machine-generated and human-written texts over some parameters. Thus, on average GPT-2 Output Detector Demo could discriminate between human-written and machine-generated texts with an 89.2% accuracy rate.

RQ#2: How effective is Crossplag Detector in identifying machine-generated essays?

Data analysis demonstrated that Crossplag Detector could differentiate between human-written and machine-generated texts, but detection accuracy was not maintained across the dataset. Similar to GPT-2 Output Detector Demo, Crossplag identified most of the human-written texts by assigning them originality scores in the 99th percentile with a median score of 99% and a tiny interquartile range of 0.00002%; however, the detector had a few lapses in judgment. Based on a 95% confidence interval, Crossplag identified 85% of human-written texts as highly original, but it mis-evaluated 15% as possibly machine-generated (i.e., false negatives). Regarding the machine-generated dataset, similar to GPT-2 Output Detector Demo, Crossplag flagged most machine-generated texts by assigning them originality scores close to 0% with a median score of 0.00001% and interquartile range of 0.00002%. However, it also misjudged a small portion of the dataset. Using a 95% confidence interval, the detector flagged 92.5% of machine-generated texts as unoriginal, but it mislabeled 7.5% of the samples as partially original (i.e., false positives). A Mann–Whitney *U* test comparing the originality scores of human-written and machine-generated datasets demonstrated a significant difference, with a mean rank of 63.53 for the machine-generated dataset and 177.47 for the human-written dataset. However, the degree of variation in the scores of mis-evaluated samples (false negatives and false positives) varied widely, indicating that the detector can confuse machine-generated and human-written texts over some parameters. In conclusion, Crossplag could classify texts as machine-generated or human-written with an average 88.75% accuracy rate.

RQ#3: What is the difference in classification effectiveness between GPT-2 Output Detector Demo and Crossplag Detector?

GPT-2 Output Detector Demo and Crossplag Detector achieved comparable classification accuracy, yet both detectors misjudged human-written and machine-generated texts over some parameters. While GPT-2 Output Detector Demo had an overall 89.2% average detection accuracy, Crossplag had an overall average detection accuracy of 88.75%. The detection accuracy for both platforms appeared to be similar, as the slight differences could be justified by random variability. Further investigation of detection accuracy between the two detectors based on text type (i.e., human-written or machine-generated) confirmed that both detectors had comparable levels of detection accuracy. It also demonstrated that each one was more sensitive to a specific text type. A Mann–Whitney *U* test comparing the originality scores assigned by both platforms for machine-generated texts indicated that GPT-2 Output Detector Demo scores were significantly higher than those of Crossplag with a mean rank of 72.66 for Crossplag Detector and a mean rank of 158.34 for GPT-2 Output Detector Demo. This finding indicates that Crossplag is more sensitive to the configurations of machine-generated texts and can detect them more accurately.

Conversely, another Mann–Whitney *U* test comparing the originality scores for human-written texts between both platforms revealed that originality scores for human-written texts were significantly higher for GPT-2 Output Detector Demo than for Crossplag Detector with a mean rank of 155.68 for GPT-2 Output Detector Demo and 85.33 for Crossplag Detector. This finding suggests that GPT-2 Output Detector is more sensitive to the configurations of human-written texts and less likely to produce false positives (i.e., misjudging human-written texts as unoriginal/machine-generated). Data analysis further revealed that both platforms misjudged the same text samples, either in the human-written or machine-generated dataset, suggesting that both platforms had similar challenges with specific configurations of text design that caused them to produce false negatives and false positives of comparable originality scores for the same essays. This finding is not unexpected given that both platforms share the same training data, GPT-2 model, and underlying detection mechanism, RoBERTa.

RQ#4: What does the comparison of two RoBERTa-based detectors suggest about the robustness of AI detection of AI-generated texts?

Data analysis demonstrated that RoBERTa-based classifiers could detect AI-generated texts with an average accuracy rate of 89% and underlined some vital considerations regarding the design and training of classifiers/detectors. Both experiments demonstrated that GPT-2-trained RoBERTa-based detectors could detect ChatGPT-generated texts. However, their detection accuracy was inconsistent across the dataset (see detailed explanations in RQ#1 and RQ#2). These results are consistent with Gao et al. (2023) findings about the effectiveness of GPT-2 Output Detector Demo in detecting ChatGPT-generated abstracts and OpenAI's in-house experiments on using GPT-2-based detector to detect its generation (OpenAI, n.d.-b). The difference in accuracy rates compared to OpenAI's study and the inconsistency of results in the current study are not unexpected, given the vast difference between GPT-2 model that was used to train the detector and GPT-3.5 model that powers ChatGPT. The two models vary in their training approaches and capacities (i.e., model size). While GPT-2 is a 1.5 billion parameter model (Solaiman et al., 2019) trained using in-context learning to achieve unsupervised multitask learning (Radford, 2019), GPT-3.5 is a 175 billion parameter model (OpenAI, n.d.-a) trained using a combination of unsupervised machine learning and reinforcement learning from human feedback approaches to produce outputs aligned with users' intents (Ouyang et al., 2022). Also, given that GPT-3.5 (the model that powers ChatGPT) is much bigger than GPT-2 that was used to train both detectors, these findings support OpenAI's conclusion that the outputs of large models are more challenging to detect and that higher detection accuracy requires training on large models' outputs (OpenAI, n.d.-b). These findings also support Solaiman et al. (2019)'s conclusion that classifiers' detection accuracy is deeply impacted by generative model size and that texts generated by larger models are more challenging to detect. Secondly, the relatively high accuracy of GPT-2-trained detectors in identifying the generations of a more developed version of GPT suggests that machine detection is more effective when language models detect their generations. This inference is consistent with Mitchell et al.'s (2023) conclusion that detection performance drops when the detection model differs from the generation model and that classifiers seem to be most effective in detecting their own generations. The high degree of accuracy of machine detection (89%), despite dealing with the generations of a more advanced version of the model, suggests that machine-generated texts have distinctive characteristics that machine classifiers can identify and use to detect AI-generated texts. This suggestion is consistent with Tay et al. (2020) finding that modeling choices leave traceable artifacts in the generated texts that could be used to identify the generative model and machine-generated texts. In addition, the high effectiveness of RoBERTa-based detectors supports the conclusions of Liu et al. (2019) and Solaiman et al. (2019) that the optimized training of RoBERTa model results in significant performance improvements compared to other detection models.

In summary, data analysis and interpretation revealed that both RoBERTa-based, GPT-trained detectors had a comparable performance and could discriminate between human-written and machine-generated texts with an average 89% accuracy. It is worth noting that each detector was more sensitive to a particular text type as GPT-2 Output Detector Demo detected machine-generated texts more accurately, while Crossplag Detector was more sensitive to human-written texts. These findings suggest that the output of larger models is more challenging to detect and that classifiers are more effective in detecting their own generations.

Conclusion

In this project, the researcher explored the potential of AI-based classifiers designed to detect AI-generated text as plagiarism detection tools that can help educators control the potential misuse of ChatGPT as a resource for AI-assisted plagiarism. Specifically, the researcher tested the performance of two GPT-2-trained, RoBERTa-based machine-detection classifiers in discriminating between human-written and AI-generated essays: GPT2-Output Detector Demo and Crossplag AI Content Detector. The purpose of the study has been to (a) assess the effectiveness of available AI-detection platforms to help L2 educators identify reliable platforms they can use to control AI-assisted plagiarism; (b) offer an in-depth investigation of this new, disruptive, and unexplored research territory to guide educators in dealing with AI-assisted plagiarism; and (c) evaluate underlying classifier type, design, and training mechanisms to contribute to the development of more robust machine-detection classifiers. To this end, the present descriptive study involved two experiments comparing the performance of two RoBERTA-based AI detectors on classifying a random set of texts as human-written or machine-generated. Each experiment involved comparing the originality scores assigned by a detector to 120 human-written essays and 120 ChatGPT-generated essays. Using Mann–Whitney *U* tests and Confusion Matrixes, the researcher compared the average score ranking of human-written and machine-generated essays for both detectors. Also, the researcher used Mann–Whitney *U* tests to compare the originality scores of both detectors for each text type. The study's findings revealed that GPT-2 Output Detector Demo and Crossplag AI Content Detector achieved high accuracy in discriminating between machine-generated and human-written texts with an average accuracy of 89% despite being trained on a smaller and earlier version of GPT. Also, the findings indicated that GPT-2 Output Detector Demo was more sensitive to human-written texts and could detect them more accurately. At the same time, Crossplag Detector was more sensitive to machine-generated texts and had fewer false negatives. These findings demonstrate that AI-based classifiers could offer viable resources for detecting machine-generated texts and controlling AI-assisted plagiarism, even if they need further development to achieve higher detection accuracy. So, ironically, it seems that using the powerful capabilities of AI against itself can offer a viable resource for detecting AI texts, similar to the

Terminator movie franchise when humans realized that a terminator machine would be a feasible protection against the machines' attempts to assassinate the resistance leader.

Practical implications

The present study has several practical implications. First, as GPT Output and Crossplag detectors could identify AI texts with a relatively high degree of accuracy, L2 educators can rely on them to investigate potential cases of AI-assisted plagiarism. To overcome the limitations of each detector, it might be helpful to run a flagged text in one detector and then into another to double-check the results or upload the text to GLTR to get a visual report that can support initial results from an AI-based detector. Second, since available classifiers can misevaluate human-written texts as machine-generated (i.e., false positives), L2 educators should not base their decisions about academic integrity solely on detectors' analysis and use other traditional approaches to ensure the academic integrity of students' work. For instance, educators can interview students about the content of their work or analyze the linguistic complexity of flagged work compared to students' in-class work. Third, since fine-tuned classifiers, especially RoBERTa-based classifiers, appear to be relatively effective in detecting AI-assisted plagiarism, major plagiarism detection platforms should integrate such classifiers into their systems rather than train their baseline classifiers from scratch. Fourth, as the findings demonstrate that model size substantially impacts the detection accuracy of classifiers, plagiarism detection platforms should train their detection classifiers on datasets from larger language models to improve their detection accuracy.

Limitations and future research directions

The present study is not without limitations. One of the study's limitations is that the sample of human-generated essays was collected using convenience sampling and cannot be generalized to the population of ESL writing students. Also, the sample size was small, reducing the samples' representativeness. Therefore, future studies can use larger sample sizes and collect the human sample using random sampling. Another limitation of the study is that the study compared only two detectors that are based on similar training protocols and model architecture. Perhaps future research could compare several detectors that use different training approaches and are based on different model architectures to compare the effects of these factors on detection accuracy. Another study limitation is that these AI essay samples were gathered in January 2023. Accordingly, the quality of essays generated could have improved slightly by the time of publication of the study due to the maturation of the AI system (neural network) due to continuous practice through interaction with users. Finally, the study's results might not represent future generations of GPT. However, this is not likely to be the case for the recently released GPT-4 since it is based on the same model architecture and machine learning approaches as GPT-3.5. Its main improvements to its predecessor appear to be in processing multimodal context and cognitive processing (OpenAI, 2023), which are likely to have minor effects on the modeling processes of text generation that AI detectors use to identify AI texts (which is supported by the findings of the present study where classifiers trained on GPT-2 were able to detect GPT-3.5 outputs).

This study's findings also suggest several future research directions. First, given that both detectors could identify texts generated by a later version of the LLM with relatively high accuracy, it is fair to assume that fine-tuning these detectors using GPT-3.5 data will likely boost detection accuracy. Thus, future research should investigate the effects of training classifiers on larger models on their detection accuracy. Second, since both RoBERTa-based detectors achieved relatively high and comparable levels of detection accuracy, it sounds reasonable to assume that RoBERTa model offers a robust architecture for classifier models. Therefore, future research should investigate the impact of different architecture, training, and modeling configurations on the detection accuracy of RoBERTa-based classifiers to optimize their performance further. Finally, given the anticipated vast proliferation and unprecedented disruptive impact of ChatGPT (and large language models in general) in ESL composition contexts, applied linguists should develop a new area of research in the field of intelligent CALL that explores the implications of generative AI for L2 learning and assessment, especially by drawing on recent advancement in NLP research to guide educators' efforts to control AI-assisted plagiarism and integrate generative AI literacies in ESL classes. The present study is hoped to be a preliminary step in this direction.

Abbreviations

AI	Artificial intelligence
ANOVA	Analysis of variance
BERT	Bidirectional Encoder Representations
ESL	English as a second language
GLTR	Giant Language Model Test Room
GPT	Generative Pre-trained Transformative m = Model
LLM	Large Language Model
MCQ	Multiple Choice Question
NLP	Natural Language Processing
RoBERTA	Robustly optimized BERT Approach
SPSS	Statistical Package for the Social Sciences

Acknowledgements
The author would like to thank the journal editor and reviewers for their feedback and support. The study was partially funding by a research grant from Gulf University for Science and Technology that covered the article processing charges.

Author's contributions
The paper was single authored by the corresponding author.

Funding
The present study did not receive funding from any entity.

Availability of data and materials
Data are available upon reasonable request.

Declarations

Competing interests
The author declares no competing interests.

Received: 1 August 2023 Accepted: 20 September 2023
Published online: 16 October 2023

References

Ali, W. Z. W., Ismail, H., & Cheat, T. T. (2012). Plagiarism: To what extent is it understood? *Procedia - Social and Behavioral Sciences,59*(2012), 604–611. https://doi.org/10.1016/j.sbspro.2012.09.320

Baker, C. (2017). Quantitative research designs: Experimental, quasi-experimental, and descriptive. In H. Hall and L. Roussel (Eds.), *Evidence-based practice: An integrative approach to research, administration, and practice*, (2nd Ed., pp.155–183). Jones & Bartlett Learning.

Bakhtin, A., Gross, S., Ott, M., Deng, Y., Ranzato, M. A., & Szlam, A. (2019). Real or fake? learning to discriminate machine from human generated text. *arXiv preprint. arXiv:1906.03351*. https://doi.org/10.48550/arXiv.1906.03351

Bommarito II, M., & Katz, D. M. (2022). GPT Takes the Bar Exam. *arXiv preprint. arXiv:2212.14402*. https://doi.org/10.48550/arXiv.2212.14402

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H Lin (Eds.), *Advances in neural information processing systems: Vol. 33*, (pp.1877–1901). ISBN: 9781713829546. Retrieved June 30, 2023, from https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html?utm_medium=email&utm_source=transaction

Carr, D. (2023, June 14). As ChatGPT Growth Flattened in May, Google Bard Rose 187%. *Similarweb*. Retrieved June 30, 2023 from https://www.similarweb.com/blog/insights/ai-news/chatgpt-bard/

Chen, X., Ye, J., Zu, C., Xu, N., Zheng, R., Peng, M., ... & Huang, X. (2023). How Robust is GPT-3.5 to Predecessors? A Comprehensive Study on Language Understanding Tasks. *arXiv preprint* arXiv:2303.00293. *https://doi.org/10.48550/arXiv.2303.00293*

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. *Advances in Neural Information Processing Systems,30*, 4299–4307. ISBN: 978151086096.

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and Cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*. https://doi.org/10.1080/14703297.2023.2190148

Crossplag (n.d.). AI Content Detector. Retrieved June 30, 2023 from https://app.crossplag.com/

Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences,340–341*, 250–261. https://doi.org/10.1016/j.ins.2016.01.033

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805. *https://doi.org/10.48550/arXiv.1810.04805*

Hugging Face (n.d.). RoBERTa-base-OpenAI-detector. Retrieved June 30, 2023 from https://huggingface.co/roberta-base-openai-detector

Fagni, T., Falchi, F., Gambini, M., Martella, A., & Tesconi, M. (2021). TweepFake: About detecting deepfake tweets. *Plos one*, *16*(5). https://doi.org/10.1371/journal.pone.0251415

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fireman Kramer, R. (1985). A overview of descriptive research. *Journal of the Association of Pediatric Oncology Nurses,2*(2), 41–45.

Francke, E., & Bennett, A. (2019). The Potential Influence of Artificial Intelligence on Plagiarism: A Higher Education Perspective. In P. Griffiths and M. N. Kabir (Eds.), *European Conference on the Impact of Artificial Intelligence and Robotics (ECIAIR 2019)* (pp. 131–140). Academic Conferences and Publishing Limited. DOI: https://doi.org/10.34190/ECIAIR.19.043

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *Njp Digital Medicine*, article 75. https://doi.org/10.1038/s41746-023-00819-6

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical detection and visualization of generated text. *arXiv preprint* arXiv:1906.04043. *https://doi.org/10.48550/arXiv.1906.04043*

Haque, M. U., Dharmadasa, I., Sworna, Z. T., Rajapakse, R. N., & Ahmad, H. (2022). "I think this is the most disruptive technology": Exploring Sentiments of ChatGPT Early Adopters using Twitter Data. *arXiv preprint* arXiv:2212.05856. *https://doi.org/10.48550/arXiv.2212.05856*

Hovy, D. (2016). The enemy in your own camp: How well can we detect statistically-generated fake reviews–an adversarial study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 351–356). Retrieved June 30, 2023 from https://aclanthology.org/P16-2057.pdf

Hu, K. (2023). ChatGPT sets record for fastest-growing user base-analyst note. *Reuters*, February 2, 2023. Retrieved June 30, 2023 from https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2019). Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint* arXiv:1911.00650. *https://doi.org/10.48550/arXiv.1911.00650*

Jawahar, G., Abdul-Mageed, M., & Lakshmanan, L. V. (2020). Automatic detection of machine generated text: A critical survey. *arXiv preprint* arXiv:2011.01314. *https://doi.org/10.48550/arXiv.2011.01314*

Johnson, A. (2023). ChatGPT in Schools: Here's Where It's banned-And How IT Could Potentially Help Students. *Forbes*, January 18, 2023. Retrieved June 30, 2023 from https://www.forbes.com/sites/ariannajohnson/2023/01/18/chatgpt-in-schools-heres-where-its-banned-and-how-it-could-potentially-help-students/?sh=3a758f366e2c

Khalil, M., & Er, E. (2023). Will ChatGPT get you caught? Rethinking of plagiarism detection. *arXiv preprint* arXiv:2302.04335. *https://doi.org/10.48550/arXiv.2302.04335*

King, M. R., & ChatGPT. (2023). Editorial: A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cellular and Molecular Bioengineering*, *16*(1), 1–2.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. *arXiv preprint* arXiv:2301.10226. *https://doi.org/10.48550/arXiv.2301.10226*

Lee, C., Panda, P., Srinivasan, G., & Roy, K. (2018). Training deep spiking convolutional neural networks with STDP-based unsupervised pretraining followed by supervised fine-tuning. *Frontiers in Neuroscience, 12*, article 435. https://doi.org/10.3389/fnins.2018.00435

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint* arXiv:1907.11692. *https://doi.org/10.48550/arXiv.1907.11692*

Lowie, W., & Seton, B. (2013). *Essential statistics for applied linguistics*. Bloomsbury Publishing.

Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library Hi Tech News*. ISSN: 0741–9058.

MacNeil, S., Tran, A., Mogil, D., Bernstein, S., Ross, E., & Huang, Z. (2022). Generating diverse code explanations using the GPT-3 large language model. *Proceedings of the ACM Conference on International Computing Education Research,2*, 37–39. https://doi.org/10.1145/3501709.3544280

Mitchell, A. (2022) Professor catches student cheating with ChatGPT: 'I feel abject terror.' *New York Post*, December 26, 2022. Retrieved June 30, 2023 from https://nypost.com/2022/12/26/students-using-chatgpt-to-cheat-professor-warns/

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. *arXiv preprint* arXiv:2301.11305. https://doi.org/10.48550/arXiv.2301.11305.

Nassaji, H. (2015). Qualitative and descriptive research: Data type versus data analysis. *Language Teaching Research,19*(2), 129–132. https://doi.org/10.1177/1362168815572747

OpenAI (n.d.-a) Documentation. Retrieved June 30, 2023 from https://platform.openai.com/docs/chatgpt-education

OpenAI (n.d.-b) GPT-2:1.5B release. Retrieved June 30, 2023 from https://openai.com/research/gpt-2-1-5b-release

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (Eds.), *Advances in Neural Information Processing Systems*: *Vol. 35*, (pp. 27730–27744). ISBN: 9781713871088.

Pan, W., Xiang, E., Liu, N., & Yang, Q. (2010). Transfer learning in collaborative filtering for sparsity reduction. In W. Pan, E. Xiang, N. Liu, and Q. Yiang (Eds.), *Proceedings of the AAAI conference on artificial intelligence* (Vol. 24, No. 1, pp. 230–235). https://doi.org/10.1609/aaai.v24i1.7578

Paul, R. (2005). The state of critical thinking today. *New Directions for Community Colleges,2005*(130), 27–38. https://doi.org/10.1002/cc.193

Pavlik, J. V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism and Mass Communication Educator,78*(1), 84–93. https://doi.org/10.1177/10776958221149577

Pecorari, D., & Petrić, B. (2014). Plagiarism in second-language writing. *Language Teaching,47*(3), 269–302. https://doi.org/10.1017/S0261444814000056

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pretraining. Retrieved June 30, 2023 from https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog,1*(8), 9.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., … & Wang, J. (2019). Release strategies and the social impacts of language models. *arXiv preprint* arXiv:1908.09203. *https://doi.org/10.48550/arXiv.1908.09203*

Susnjak, T. (2022). ChatGPT: The End of Online Exam Integrity? *arXiv preprint* arXiv:2212.09292. *https://doi.org/10.48550/arXiv.2212.09292*

Tay, Y., Bahri, D., Zheng, C., Brunk, C., Metzler, D., & Tomkins, A. (2020). Reverse engineering configurations of neural text generation models. *arXiv preprint* arXiv:2004.06201. *https://doi.org/10.48550/arXiv.2004.06201*

Turnitin (2023, March 16). *Understanding the false positive rate for sentences of our AI writing detection capability.* Turnitin. https://www.turnitin.com/blog/understanding-false-positives-within-our-ai-writing-detection-capabilities

Waltzer, T., & Dahl, A. (2023). Why do students cheat? Perceptions, evaluations, and motivations. *Ethics and Behavior,33*(2), 130–150. https://doi.org/10.1080/10508422.2022.2026775

Yang, M. (2023). New York City Schools ban AI chatbot that writes essays and answers prompts. *The Guardian*, January 6, 2023. Retrieved June 30, 2023 from https://www.theguardian.com/us-news/2023/jan/06/new-york-city-schools-ban-ai-chatbot-chatgpt

Yeadon, W., Inyang, O. O., Mizouri, A., Peach, A., & Testrow, C. (2022). The Death of the Short-Form Physics Essay in the Coming AI Revolution. *Physics Education, 58*(3). https://doi.org/10.1088/1361-6552/acc5cf

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche'-Buc, E.Fox, and R. Garnett (Eds.). *Advances in neural information processing systems,32*, 1–12. ISBN: 9781713807933.

Publisher's Note