GPTZero vs. Text Tampering: The Battle that GPTZero Wins

David W. Brown
Elmhurst University, USA

Dean Jensen
Elmhurst University, USA

Abstract: The growth of Artificial Intelligence (AI) chatbots has created a great deal of discussion in the education community. While many have gravitated towards the ability of these bots to make learning more interactive, others have grave concerns that student created essays, long used as a means of assessing the subject comprehension of students, may be at risk. The bot's ability to quickly create high quality papers, sometimes complete with reference material, has led to concern that these programs will make students too reliant on their ability and not develop the critical thinking skills necessary to succeed. The rise in these applications has led to the need for the development of detection programs that are able to read the students submitted work and return an accurate estimation of if the paper is human or computer created. These detection programs use natural language processing's (NLP) ideas of perplexity, or randomness of the text, and burstiness, or the tendency for certain words and phrases to appear together, plus sophisticated algorithms to compare the essays to preexisting literature to generate an accurate estimation on the likely author of the paper. The use of these systems has been found to be highly effective in reducing plagiarism among students, however concerns have been raised about the limitations of these systems. False positives, false negatives, and cross language identification are three areas of concern amongst faculty and have led to reduced usage of the detection engines. Despite the limitations however, these systems are a valuable tool for educational institutions to maintain academic integrity and ensure that students are submitting original work.

Keywords: Natural Language Processing, Language Translation, Chatbot, Plagiarism

Citation: Brown, D.W. & Jensen, D. (2023). GPTZero vs. Text Tampering: The Battle that GPTZero Wins. In M. Shelley, V. Akerson, & M. Unal (Eds.), *Proceedings of IConSES 2023-- International Conference on Social and Education Sciences* (pp. 734-744), Las Vegas, NV, USA. ISTES Organization.

Introduction

Artificial Intelligence (AI) tools called chatbots, such as ChatGPT and Caktus AI, have been in the news since the release of ChatGPT by OpenAI in 2022. As faculty explored how these tools might be used by their students, plagiarism and academic integrity became an immediate concern. Academic integrity is a fundamental principle and is crucial for a functioning institution of higher learning. It is based on the values of honesty, trust, and respect requiring students and faculty to adhere to a set of ethical standards (International Center for

Academic Integrity, 2021). Plagiarism, or the intentional or unintentional act of using someone else's work or ideas without proper citation, is a serious ethical offense, as it undermines the integrity of academic work and learning (Chowdhury & Bhattacharyya, 2018). The effects of the emergence of generative AI are currently unknown, but many educators fear it may spell the end of essays as educational assignment (Eke, 2023).

The history of AI and chatbots can be traced back to the 1950's when scientists like Alan Turing began exploring the concept of intelligent machines and whether a computer program could communicate with people without the person realizing their partner was artificial. This question formed the basis for the Turing test, which is considered by many to be the generative idea of chatbots (Copeland, 2000). The first AI program called ELIZA was developed in 1966 to simulate a psychotherapist and while its ability to communicate was limited, it has been the source of inspiration for subsequent development. In the ensuing decades AI technology has continued to advance, leading to the creation of more sophisticated chatbots with the ability to understand and respond to complex requests. These include Siri developed by Apple in 2010, Watson developed by IBM in 2011, Cortana developed by Microsoft and Alexa by Amazon in 2014, and now ChatGPT and Caktus AI.

Chat Generative Pretrained Transformer (ChatGPT) is an AI chatbot that is specifically designed to generate human-like text in a conversational style (Cotton, Cotton, & Shipway, 2023).]. It is freely accessible, allowing the platform to attract millions of interactions and is based on a large language model (LLM) with over 175 billion parameters that was trained using Reinforcement Learning from Human Feedback (RLHF) based on a model in the GPT-3.5 series using Microsoft Azure AI supercomputing infrastructure (ChatGPT, 2023). This training was performed using over 40 terabytes of text, or close to 40 million books in an Amazon Kindle format (Khalil & Er, 2023). This allows it to use deep learning to perform a range of Natural Language Processing (NLP) tasks, such as translation, summarization, question answering, and text generation, with little to no task-specific training needed (Cotton, Cotton, & Shipway, 2023).

In contrast to ChatGPT, which was intended for generalized usage, Caktus AI was created by a group of AI educators, and designed to be a resource for students to clean up their own essays. Their process involved combining machine learning, natural language processing, and the processing power possible through cloud-based computing infrastructures. (Ju, et al., 2014).

Tools of this type are designed to create content from the data they are trained on when presented with a prompt. Using thousands of sources from the internet it will generate a response, often without further input from the user, that appears realistic. This ability has allowed them to become a popular choice among college and university students to generate academic essays for homework, which has increased the concerns of plagiarism (Khalil & Er, 2023). The coherent nature of this generated text makes it difficult to distinguish between the computer generated and the student's own writing, thus undermining the purpose of higher education, which is to challenge and educate students. When students use ChatGPT or Caktus AI to generate essays or other forms of written text and then pass them off as original works, they violates the core principles of academic integrity (Eke, 2023).

Background

Plagiarism Types

Plagiarism is the presentation of another's work or ideas as one's own without acknowledgement. Although it can appear in different forms, there are generally two types of plagiarism (1) textual plagiarism and (2) source code plagiarism. Textual plagiarism is more commonly seen in academic settings, and thus is the focus of this paper. The authors in (Chowdhury & Bhattacharyya, 2018) divide it into seven categories based on its form and application:

1. **Clone Plagiarism**: also known as deliberate copy/paste or identical copying and designates the situation where someone copies another work and presents it as their own with, or without, acknowledging the original source.
2. **Paraphrasing Plagiarism**: also known as hybrid or remix and refers to the use of another work presented in different ways simply by switching words, changing sentence constructs, and altering grammatical styles without citing the original source.
3. **Metaphor Plagiarism**: refers to someone using metaphors to present other ideas in better ways.
4. **Idea Plagiarism**: refers to someone borrowing an entire idea from other sources and claiming them as their own.
5. **Recycle Plagiarism**: also known as self-plagiarism, this occurs when someone borrows from their own previous documents without a proper citation.
6. **Illegal Source Plagiarism**: refers to someone citing references that are invalid.
7. **Retweet Plagiarism**: refers to someone citing the reference of proper sources; however, their presentation is very similar to the original contents wording, sentence structure, and/or grammatical usage.

Regardless of which form of plagiarism we are dealing with, it is a complicated and ethically difficult subject as it refers to the act of stealing and publishing another author's work under one's own name without crediting the original source (Mansoor & Al-Tamimi, 2022). Further complicating the issue, in this era of generative AI, the topic becomes more complicated and potentially morally ambiguous, as the originality of the content can be questioned. Machine generated content is the result of a computational process and not a deliberate act of copying or paraphrasing someone else's work. While it could be argued that this falls outside of these established categories of plagiarism, as the content was not stolen from another author, yet without proper attribution of the source it can still be defined as plagiarism as authors, even machines, must adequately be credited. No matter how it is defined, plagiarism is a type of academic deception that must be detected.

Plagiarism Detection

To counteract this issue, plagiarism detection tools have been developed to assist educators in identifying instances of plagiarism. These tools work by using advanced algorithms to scan and compare submitted written works against a database of existing texts, identifying similarities, and generating a report that illustrates potentially problematic areas. In this way, plagiarism detection has become an essential tool for maintaining academic honesty and ensuring the credibility of academic work. Textual plagiarism detection can occur between two same or two different natural languages. Based on the language homogeneity or heterogeneity of the documents being compared, the detection can be classified as either monolingual or cross-lingual plagiarism detection (Chowdhury & Bhattacharyya, 2018).

In the case of cross-lingual plagiarism, detection methods are limited due to the difficulty in finding proximity between two text segments from different languages, e.g., English-to-Spanish or English-to-Japanese (Danilova, 2013). Conversely, monolingual plagiarism detection, which is the most common type, the detection deals with similar languages, e.g., English-to-English, and can be further subdivided based on the use of external references used during the detection process as either intrinsic or extrinsic plagiarism detection. Intrinsic detection analyzes the written style or uniqueness of the author and attempts to detect plagiarism based on own-conformity or deviation between the text segments requiring no external sources for detection. Extrinsic detection compares the submitted work against many other available relevant digital resources in databases or on the internet for its detection (Mansoor & Al Tamimi, 2022).

Extrinsic detection can be further divided into Source Retrieval where given a suspect document, a search engine is used to identify all plagiarized sources. Text Alignment instead seeks to identify all contiguous, possibly reused text passages between a given pair of documents (Ali & Taqa, 2022). The development of software detection systems has taken decades of research and has focused on developing sophisticated text-matching algorithms to identify plagiarism. Such systems include, but are not limited to: Turnitin, iThenticate, PlagAware, PlagScan, CheckForPlagiarism.net, and PlagiarismDetection.org.

These tools detect plagiarism from various perspectives, including Character Based, Vector Based, Syntax Based, Semantic Based, Fuzzy Based, Structural Based, Stylometric Based, Grammar Based, Classification and Cluster Based, and Citation Based. Many studies have tested their effectiveness in plagiarism detection; however, with the release of ChatGPT more sophisticated methods are required to detect the machine generated work, as its originality would not be represented within existing online repositories (Ali & Taqa, 2022), (Ali, Abdulla, & Snasel, 2011), (Mansoor & Al-Tamimi, 2022).

GPTZero

GPTZero is a relatively new classification model released in the wake of ChatGPT that attempts to predict whether a document was written by a LLM or a human. It provides predictions on a sentence, paragraph, or document, and was initially trained on a large and diverse corpus of human-written and AI-generated text with a focus on English prose (GPTZero, 2023). Its classifier returns a score that specifies the probability of the entire document being AI-generated.

The classifier has achieved an AUC, or Area Under the Curve, score of 0.98. The higher the AUC score, the better the AI program is at distinguishing between the two extremes, in our case, student created or plagiarized

(Bhandari, 2020). At a threshold of 0.65, 85% of AI documents are classified as being AI-generated and 99% of human documents are classified as human. At a threshold of 0.16, 96% of AI documents are classified as AI and 96% of human documents are classified as human. It is recommended that a threshold of 0.65 or higher is used to minimize the number of false positives.

GPTZero further utilizes perplexity and burstiness as indicators (Bowman, 2023). Perplexity is a measure of how well a statistical language model can predict a sequence of words given the preceding context and is a way to measure the quality of these predictions. The score is calculated as the inverse probability of the test set normalized by the number of words in the test set. The lower the perplexity score, the better the language model is at predicting the test set. Burstiness is a measure used to describe the distribution of words or phrases in text It refers to the phenomenon of certain words or phrases occurring in clusters, or bursts, within a particular context, rather than being evenly distributed throughout the text (He, Shen, Chen, Backes, & Zhang, 2023).

If GPTZero is perplexed by the text, then it has a high complexity and it is considered more likely to be human written. However, if the text is more familiar to GPTZero, because it has been trained on such data, then it will have a low complexity and therefore is more than likely to be AI-generated. Similarly, humans tend to write with greater burstiness, for example, with longer or more complex sentences alongside shorter ones, whereas AI sentences tend to be more uniform (Bowman, 2023).

Cross Language Translation

Language switching plagiarism is a type of source code plagiarism where the developer changes the programming language, or a program is written in one language and rewritten in another language and declared to be their own work (Chowdhury & Bhattacharyya, 2018). One way students might attempt to plagiarize work, or submit work created using a chatbot is through the use of Google Translate. Translate is a multilingual neural machine developed by Google to translate text and documents from one language into another (Google Translate, 2023). By changing the language of the generated text to a foreign language and then back into English, the student may hope to confuse detection methods being employed by the school. While GPTZero was not trained to identify AI-generated text that has been heavily modified after generation, this paper will examine whether GPTZero can detect generative AI content slightly modified through obfuscation using Google Translate.

Methodology

This is a descriptive study that presents the results of cross-lingual plagiarism detection analysis on AI-generated content and cross-language translated by Google Translate. This study follows a quantitative analysis, where the outputs generated are analyzed and evaluated numerically based on the scores produced by the plagiarism detection tool GPTZero. Below we explain in more detail the process for data collection, plagiarism detection, and further analysis.

Data Generation and Collection

To gather a representative sample of data, the authors suggested 6 topics dealing with ethical issues in computer science for scoring by GPTZero. For each topic, 5 anonymized human versions were combined with ChatGPT and Caktus AI generated example essays. In addition, a cross language translation version of the ChatGPT and Caktus AI version of the paper was converted back into English and left in a foreign language, and versions of the ChatGPT essay were fed back into Caktus' AI's "Improve" functionality (Table 5). This led to a total of 84 essays to be evaluated. The different versions of AI generated text were created as a means of showing the minimum many students would do to try and disguise the work being created mechanically, rather than from their own work.

Cross Language Translation

Essays generated by both ChatGPT and Caktus AI were input into Google Translate and converted from English to French to German to Danish to Māori to Russian and finally back to English. In addition, a copy of the essay was left converted into French to test the Cross Language capabilities of the GPTZero engine.

Analysis

The results of the plagiarism detection were analyzed to determine the originality and uniqueness of the AI-generated essays. The analysis is descriptive following quantitative measures of perplexity and burstiness scores.

Table 5. Input Essay Distribution

Paper Source	Quantity
Anonymized essays from students	5
Anonymized essay improved by Caktus AI	1
ChatGPT Generator	1
ChatGPT Improved by Caktus AI	1
ChatGPT Foreign Language	1
ChatGPT converted to English	1
Caktus AI's standard essay generator	1
Caktus AI's guided essay generator	1
Caktus AI Foreign Language	1
Caktus AI converted to English	1

Results

Initial Findings

After uploading the essays to GPTZero, the perplexity and burstiness scores (Figure 27) were recorded along with the engine's interpretation of the scores. GPTZero uses a six-point Likert scale to provide an easily understandable result to its computations ranging from "Likely to be entirely written by a human" to "Likely to be entirely written by an AI".

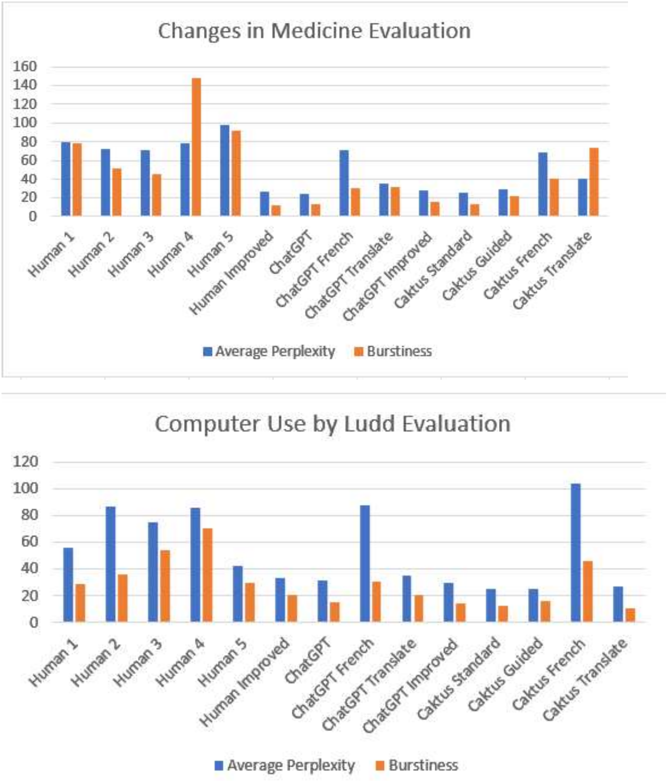

Figure 27. Comparison results for individual essays

False Positive Check

As shown in Figure 27 of the 30 essays we considered, 87% were found to be "Likely w to be written entirely by a human", however 3% of the papers were misidentified as being "Likely written entirely by AI". This non-zero result means that conclusions by the engine should not be considered in a vacuum, but instead as part of a rigorous methodology. It is important to consider the expected level of writing for the student based on observations of in class assignments and the level of coursework the essay is assigned. GPTZero and other detection programs do not consider the quality of the work and if the student submits work that does not reach a certain threshold of perplexity, the engine assumes that the essay is machine generated. Therefore, GPTZero should not be the only means you use to check for plagiarism.

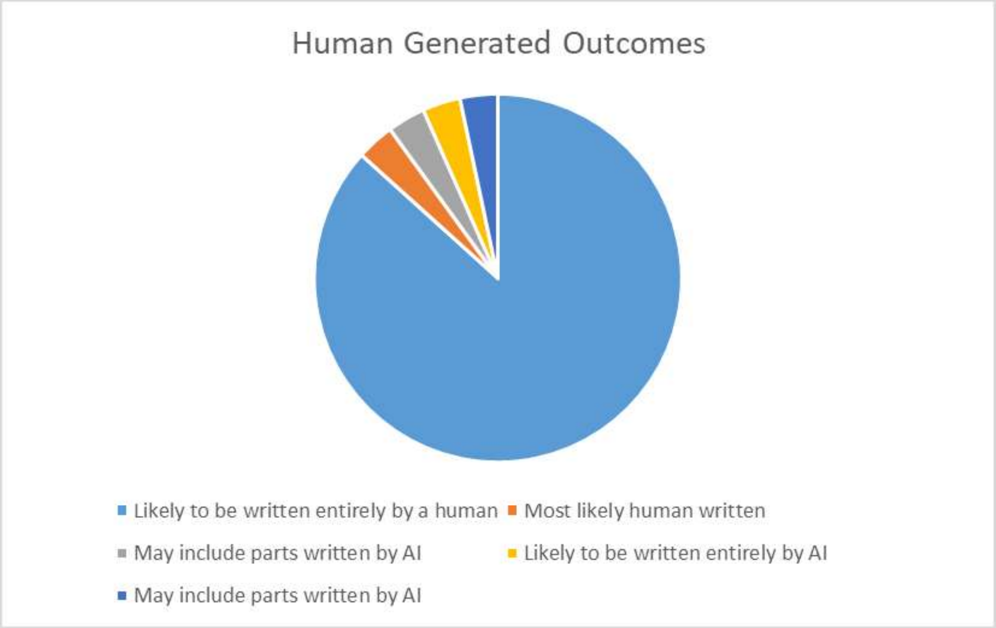

Figure 28. Paper outcomes for human generated essays

False Negative Check

Combining all papers that were created or improved by use of one of the two AI chatbots shows that 91% of the time, GPTZero correctly identified the paper as being “Likely to be written entirely by AI” (Figure 29). Discounting the papers in a foreign language which are discussed in Section 4.4, GPTZero created a false negative 7% of the time. There was no single ‘point of failure’ on any of the papers that were able to escape detection. However, all of these false negatives were modified in some way before being run through the GPTZero application. This could take the form of either using the ‘Improve’ essay option in Caktus AI or being run through our translation sequence. As with the false positives, the non-zero outcome of our search indicates that the instructor of the course will still need to consider the known writing style and ability of the student as part of the grading process rather than simply depending on the engine.

Cross Language Translation Check

It was assumed that GPTZero would have difficulties with cross-language checks, and this proved to be the case as all papers submitted in our foreign language selection, French, were scored as being “Likely to be written entirely by a human”. Considering the difficulties in cross-language translation, it should be considered normal and indicates that until issues with the process are improved, through advancements in technology or development of more language specific testing engines, the concerns of AI generated papers will continue to be an issue in those disciplines. One important note to consider though is that when we translated those papers back to English, 92% of the time, GPTZero correctly identified them as AI generated, which means cases where the faculty have a reason to question the origin of a paper, they may simply translate it to the language of the detection program and run the analysis. When we had a colleague look over the documents created by translating the original papers into French, they were amazed at how good the papers were. They commented that the only way an instructor would know the papers were computer generated was from the fact that grammatically speaking, the essays were too perfect.

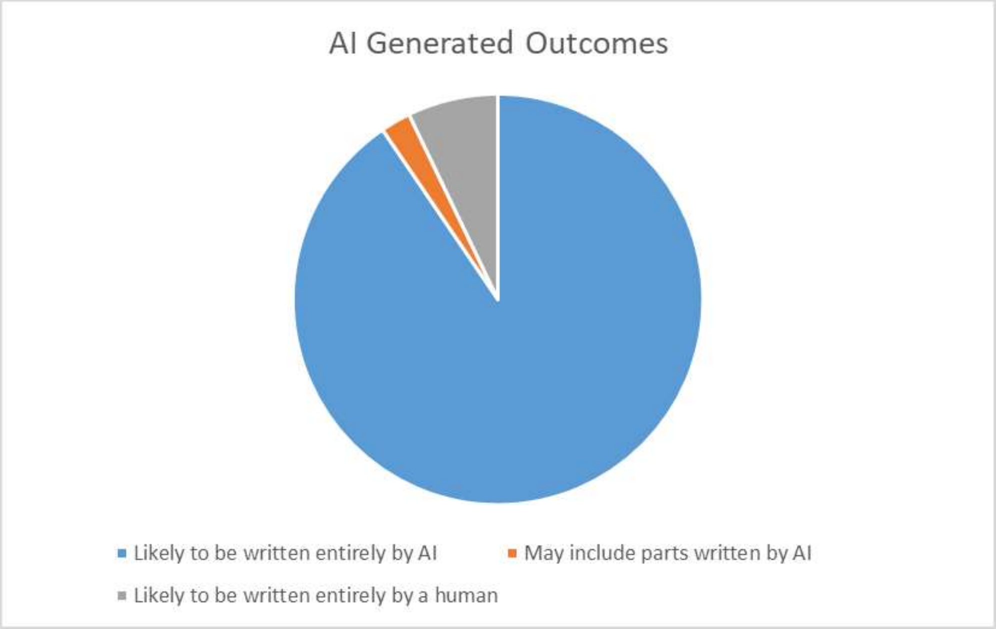

Figure 29:Paper outcomes for AI generated essays

Future Work and Discussion

Even though our numbers of false positives and negatives were higher than anticipated, we do not find that this invalidates the claim of a .98 AUC with the GPTZero engine. Several papers authored by students show as being partially written by an AI, but except for one instance, the scores did not rise to the level of passing into being more likely than not to be AI generated. Our false negatives, withholding the cross-language checks, fall directly into the anticipated range.

In addition, GPTZero has so far managed to stay ahead of common means of students to obfuscate the AI generated text through language conversion. Would paraphrasing of the supplied text allow the student to pass the checks provided by detection algorithms like GPTZero? Our initial research says that it would, but at this point it becomes difficult to control and becomes no different from other forms of plagiarism. As with the false positives, being an observant instructor and knowing the writing ability of students is truly the only way to eliminate all forms of cheating. In the future, the creation of a larger pool of human and AI generated material is the next step in this process, as it will allow us to see if our numbers of false positives and negatives improve. In addition, the modification of the GPTZero engine to accept papers in foreign languages to eliminate those false negatives will be explored.

Study Limitations

The methodology incurred several limitations. First, the study is limited to 500 word essays, as that appears to be the maximum capability of the free version of ChatGPT. The length of essay assignments in undergraduate college and university courses vary depending on the institution, department, and course level but typically range between 1500-5000 words (McCombes, 2019). The relatively short length of the AI-generated essays could have an impact on their perplexity and burstiness. Second, the results of our study are dependent on the accuracy of GPTZero in its classification and plagiarism detection. Third, the sample size of 84 human and AI-generated essays used in this study may not be sufficient to generalize for further implications. A larger sample size, e.g., >1000 essays, may be necessary to increase the reliability of the results.

Conclusions

AI improvements and the growth of chatbots have had a chilling effect on the use of essays as a means of judging comprehension and understanding among students around the world. Apprehension from some faculty arose that they would be unable to distinguish between original work performed by their students from that created by entering the topic of the paper into a chatbot and hitting a button, especially with faculty who do not see themselves as technically proficient. As we have seen in this study, however, these concerns seem to, for now at least, be overblown. Limitations to the length of an essay generated and the ability of engines such as GPTZero to correctly distinguish the source of the provided text means that essays can still be part of the educational experience.

References

Adamopoulou, E., & Moussiades, L. (2020, December 15). Chatbots: History, technology, and applications. *Machine Learning with Applications, 2*.

Ali, A., & Taqa, A. Y. (2022). Analytical Study of Traditional and Intelligent Textual Plagiarism Detection Approaches. *Journal of Education and Science*, 8-25.

Ali, A., Abdulla, H., & Snasel, V. (2011). Overview and comparison of plagiarism detection tools. *Dateso*, 161-172.

Bhandari, A. (2020, June 16). *Guide to AUC ROC Curve in Machine Learning: What is Specificity*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Bowman, E. (2023, January 9). *A college student created an app that can tell whether AI wrote an essay*. Retrieved from NPR: https://www.npr.org/2023/01/09/1147549845/gptzero-ai-chatgpt-edward-tian-plagiarism

Caktus AIAI. (2023). Retrieved from Caktus AI: https://www.Caktus.AI

ChatGPT. (2023). Retrieved from ChatGPT: https://help.openai.com/en/collections/3742473-chatgpt

Chowdhury, H. A., & Bhattacharyya, D. K. (2018). Plagiarism: Taxonomy, tools, and detection techniques. *arXiv preprint arXiv:1801.06323*.

Copeland, B. J. (2000). The Turing test. *Minds and Machines*, 519-539.

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*, 1-12.

Danilova, V. (2013). Cross-language plagiarism detection methods. *Proceedings of the Student Research Workshop associated with RANLP*, (pp. 51-57).

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., . . . Wright, R. (2023). So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*.

Eke, D. O. (2023). ChatGPT and the Rise of Generative AI: Threat to Academic Integrity? *Journal of Responsible Technology*.

Ferrero, J., Besacier, L., Schwab, D., & Agnes, F. (2017). Deep investigation of cross-language plagiarism detection methods. *arXiv preprint arXiv:1705.08828*.

Google Translate. (2023). Retrieved from Google Translate: https://translate.google.com

GPTZero. (2023). Retrieved from GPTZero: https://gptzero.me

He, X., Shen, X., Chen, Z., Backes, M., & Zhang, Y. (2023). MGTBench: Benchmarking Machine-Generated Text Detection. *arXiv preprint arXiv:2303.14822*.

International Center for Academic Integrity. (2021). *Fundamental values of academic integrity.* Albany, NY: International Center for Academic Integrity.

Ju, J., Yao, X., Yang, S., Wang, L., Sun, R., & Chen, E. (2014). An Intelligent Personal Assistant for Task Management and Recommendation. *Advanced Functional Materials*, 301-309.

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

Mansoor, M. N., & Al Tamimi, M. (2022). Plagiarism Detection System in Scientific Publication Using LSTM Networks. *International Journal on Technical and Physical Problems of Engineering (IJTPE)*, 17-24.

Mansoor, M. N., & Al-Tamimi, M. S. (2022). Computer-based plagiarism detection techniques: A comparative study. *International Journal of Nonlinear Analysis and Applications*, 3599-3611.

McCombes, S. (2019, January 28). *How Long is an Essay? Guidelines for Different Types of Essay*. Retrieved from Scribbr: https://www.scribbr.com/academic-essay/length