Keeping AI Honest in Education: Identifying GPT-generated text

Arend Groot Bleumink CONTACT@EDUKADOAI.COM
Co-Founder / CTO Edukado AI

Aaron Shikhule
Co-Founder / COO Edukado AI

Abstract

The recent introduction of ChatGPT, a conversational AI built on top of the Generative Pre-trained architecture, has drawn significant attention from the general public in the field of AI and its possible applications. In this study, we introduce a transformer-based model that can predict if a GPT model, including the latest ChatGPT model, wrote a given sentence or text. During the evaluation of our model, the team at Edukado AI achieved an accuracy of 99.7% in identifying a mix of human and AI-written content.

The Chatbot, released by OpenAI, has since significantly impacted various industries, including customer service and support, content creation, marketing and sales, and education. It is estimated that within just two months after launch, the Chatbot reached 100 million monthly active users in January 2023 (Hu, 2023). In a recent survey (Westfall, 2023), it was found that 89% of students have used the platform to help with homework, and 48% of students admitted to using it for a quiz or at-home test. At the same time, a staggering 52% of students have used it to write an essay.

Our findings highlight the potential of detecting texts generated by GPT models with high accuracy and low False Positive rate. The impact of AI on education cannot be underestimated if there is no way to detect improper or undocumented use in writing essays, doing research, and performing exams. However, there are also ethical concerns about relying solely on such detectors, as a false classification can impact an individual. Further research is needed to enhance the reliability of the model and possibly additional methods of confirming the authenticity of the submitted text.

Keywords: ChatGPT, conversational AI, Generative Pre-trained architecture, transformer-based model, Natural Language Processing, GPT-3, OpenAI, Large Language Models, detection of AI-generated texts, plagiarism, education.

1. Introduction

Generative Pretrained Transformer (GPT) models have been shaking up the field of Natural Language Processing since its introduction by OpenAI in 2018. The original GPT model (Radford et al., 2018) proposed a transfer learning approach where the model was pre-trained on a large corpus of data from the internet. Later the same pre-trained model was fine-tuned for specific tasks, including classification and language generation. Fast forward two years and GPT-3 was released in June 2020 with

© A. Groot Bleumink & A. Shikhule.

a much-improved model. The improvements included a significant increase in the number of parameters and were trained on much more diverse training data. Due to the improvements, the model was able to outperform the previous GPT models in text generation, question-answering, and text classification.

The GPT models are so-called Large Language Models (LLMs), systems that are able to understand and generate text. Besides GPT, other LLM systems are available, including BERT, PaLM, Sparrow, RoBERTa, and others. These systems are trained on a large corpus of text data found on the internet, books, journals, and other sources. Each LLM has its advantages and disadvantages, and mostly a different underlying technology.

1.1. Impact on Education

Previous research into the reasons for plagiarism within higher education (Šprajc et al., 2017) demonstrated that the main reasons for committing plagiarism include ease of copying and ease of access to materials and new technologies.

In relation to new technologies, a student survey conducted by our team at Edukado AI in early December 2022, revealed that university students were likely to use AI writing tools as a result of having to meet tight assessment deadlines. They also expressed that they found assessments very time consuming and also highlighted writing skill difficulties. Majority of students from this cohort acknowledged that they interpreted the use of AI writing tools as cheating and would be deterred from using them if a detection method became available.

Other research into the impact of ChatGPT on education was conducted, which touched on the idea of reverting back to physical closed-book exams where the students write by hand (Rudolph et al.). The researchers suggested that instead of reverting back to that, higher education institutions should focus on graduate employability rather than students cramming information for their exams, only to forget most of it later.

1.1.1. Positive impact

While the impact of ChatGPT is not yet known, we can imagine what impact ChatGPT can have on students. In the research on plagiarism (Šprajc et al., 2017), an important reason to plagiarize included a poor explanation by the teacher. Chatbots like ChatGPT are able to explain certain questions students might have about certain subjects or concepts or give an alternative way of supporting students during their studies. Furthermore, such systems would be able to give students feedback on their submitted work or during the process. Additionally, it would be able to be a ”sparring” partner during research and experiments due to the sheer knowledge that is within such LLMs.

1.1.2. Negative impact

As we mentioned above, there are multiple potential positive impacts of large language models like ChatGPT, but there are also concerns about authenticity and academic integrity with AI generative models applied within education. The models could allow students to quickly generate answers or essays without the need for research. Additionally, the current models have been demonstrated to fabricate facts and provide inaccurate or biased answers to certain topics. It can also cause an over-reliance on technology, which might prevent students from developing critical thinking and problem-solving skills.

2. Methodology

2.1. Transformer-based model

In 2017 a team of researchers from Google Brain, Google Research, and the University of Toronto introduced the concept of transformer-based models (Vaswani et al., 2017). Transformer-based models are a type of deep learning neural network that processes input data in parallel rather than in a sequential matter like RNNs and LSTM models. A few reasons why transformer-based models were introduced include the training and inference time; since transformers are able to process an entire input sequence in parallel, the training and inference time drastically decreases. Additionally, transformers use a special mechanism called self-attention that weighs the importance of each input element for the current prediction. The output of that self-attention layer is used to compute the final output. This method has been proven to increase the effectiveness in modeling relationships between elements in the sentence, primarily in Natural Language Processing tasks.

2.2. AICheatCheck

In the first week of 2023, the Edukado AI team released "AICheatCheck," a web-based AI detection tool that had been in development since November 2022. Before the launch of ChatGPT, the team knew this technology would be disruptive to the education sector. AICheatCheck's model was developed in-house and has been constantly updated and refined in the weeks following launch. On the 7^{th} of January, the team released their first model version to the public via a web interface.

In the weeks after, our team at Edukado AI noticed certain limitations in the current model and set out to work on more robust approaches to tackle this problem. We built a deep learning model able to distinguish AI-generated content using transformers by looking at patterns in the data. Specifically, it extracts certain characteristics of a sentence or group of sentences and uses that to predict if it was AI-generated or not; these characteristics include sentence structure, word choice, fluency, and many others. After that process, it combines the data of each of the sequences of text and computes a prediction based on those.

2.3. Data and Evaluation

The model's training was done with a combination of ChatGPT, GPT-3, and human-written texts. The model was trained on about 50,000 human versus GPT-generated text examples among an extensive collection of domains and education levels. We produced outputs in GPT-3 and ChatGPT to match any education level between Middle School and the Ph.D. level to ensure the model is used for different writing styles. The data was split into a training set (70%), a test set (20%), and a validation set (10%).

We matched the AI generated text against the human text that used publicly available data on different subjects and education levels. The data was balanced and preprocessed to ensure that the model was not overfitting specific sentence structures and punctuation, among other text features. The publicly available data, which was part of the evaluation data set, was presented by researchers (Guo et al.) examining patterns and linguistic analysis of ChatGPT-generated answers.

In the end, our model computes a binary classification score, determining whether text is AI-generated or is human-generated. A binary classification score assigns either 0 (for human-generated text) or 1 (for AI-generated text) to each text. We also include confidence levels so that some nuance can be applied during the outcome evaluation. Additionally, we computed the precision and recall, calculated by dividing the total number of positive predictions by the number of true positive predictions. The recall measures the model's ability to find all AI-generated texts in the data set. It can be calculated as the number of true positive predictions divided by the number of AI-generated texts in that data set.

Metric	Value
Accuracy	99.73%
Precision	99.90%
Recall	99.61%
False Negative	0.188%
False Positive	0.046%

Table 1: Validation Metrics

3. Results

The AICheatCheck model was evaluated using the accuracy, precision, recall, and False Positive rate. The most important metrics for us were the False Positive rate and the overall accuracy, as we want to minimize the negative impact of falsely classified texts.

The model achieved 99.7% accuracy after multiple repeated experiments (Table 1). When conducting these experiments, we monitored the False Positive rate and adopted the model with the lowest False Positives. That False Positive rate is 0.046%, so about 1 in 2200, which was much better than expected. A precision of 99.90% and a recall of 99.61% was also achieved.

To achieve our model with 99.7% accuracy and the other validation metrics, we conducted hundreds of experiments testing various models, data sources, parameters, and other variables. The model consists of over 100 million parameters which were fine-tuned for our purpose, the classification of GPT-generated text.

The model's results showcase that Edukado AI has built the first AI-generated text detection solution with the highest accuracy and lowest False Positive rate to date. In future research, we intend to enlarge the entire data set, train the model on new AI models focused on text generation, and minimize the False Positive rate further.

4. Conclusion

This technical paper aims to provide educators and the public with a better understanding of the AICheatCheck model, which aims to distinguish AI-generated text from human-generated text. Also, the model aims to classify text regardless of different education levels, subjects, and fluency in English.

The AICheatCheck model achieved 99.7% accuracy after multiple repeated experiments and a False Positive rate of 0.046% (1 in 2200). Edukado AI is developing an Application Programming Interface (API) and Learning Tools Interoperability (LTI) to make their latest model version available for educators and other interested parties to use.

References

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arxiv:2301.07597*.

Krystal Hu. ChatGPT sets record for fastest-growing user base - analyst note, 2 2023. URL https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-anal

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Jürgen Rudolph, Samson Tan, and Shannon Tan. Chatgpt: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6(1).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

Chris Westfall. Educators Battle Plagiarism As 89OpenAI's ChatGPT For Homework, 1 2023. URL https://www.forbes.com/sites/chriswestfall/2023/01/28/educators-battle-plagiarism-as-89-of-students-admit-to-using-open-ais-chatgpt-for-homewo?sh=5f029382750d.

Polona Šprajc, Marko Urh, Janja Jerebic, Dragan Trivan, and Eva Jereb. Reasons for plagiarism in higher education. *Organizacija*, 50(1):33–45, 2017. doi: doi:10.1515/orga-2017-0002. URL https://doi.org/10.1515/orga-2017-0002.

Appendix A. Published Metrics

The published metrics on the model can be found at the following URL:

https://api.wandb.ai/links/arend/wog1yr8o

This page provides an in-depth analysis of the model's performance and its various metrics. We encourage the reader to refer to this page for a more comprehensive understanding of the model's capabilities and limitations.