Evaluation of DNA-GPT Performance on Student Essays

Neo Wee Zen, Peh Yew Kee
NUS High School of Mathematics and Science

Abstract

With the growing prevalence of LLM chatbots like ChatGPT, educators need a solution to detect machine-generated essays, to preserve the integrity of a student's writing, or to help students write less "robotically". We evaluate the performance of DNA-GPT on a dataset of 128 A-Level GP standard essays, on varying settings. We show that the performance is comparable to the TPR and AUROC results of testing on XSum, Scientific Abstracts, and ELI5 in the original paper. We believe that DNA-GPT is a suitable candidate for a machine-generated text detection tool for educators.

Introduction

With the release of mainstream large language models (LLM) like ChatGPT [4] in recent years, the generation of content has become significantly more accessible to the public. However, malicious uses of ChatGPT have arisen, like generation of fake news [5] and plagiarism [6]. Detection models like GPT-Zero [7], DetectGPT [8], OpenAI's text classifier [9] and DNA-GPT [10] (the focus of this paper) have attempted to address such problems, with a high level of success being present in the latter.

Despite this, we identified that such AI text classifiers have not been tested on any essay datasets, which could mean that these detection systems have a potential blind-spot in the field of student essays, which have become rife with plagiarism from ChatGPT, which has the potential of discouraging critical thought and encouraging mechanical writing [1] while scores remain stagnant and would not attract a teachers' attention [3]. As human detection of AI-generated text fails more than 30% of the time [2], it is crucial that educators have a more reliable source, like an graphical implementation of DNA-GPT, to inform them on the likelihood of a student copying and pasting from an AI source. Moreover, DNA-GPT may also serve as a tool to identify students with "machine-like" styles of writing, helping teachers to guide in a student's learning.

Methodology

For our dataset, as there was no readily available collection of A-Level GP standard essays available, we manually compiled our human-written essay dataset by searching and collecting 128 of such essays from the internet, from varying sources. This ensured that our dataset was diverse in quality, so as to ensure that DNA-GPT works generally. The generation of machine-written A-Level GP standard essays was done using GPT-3.5, at a temperature of 1, using the same prompt throughout. GPT-3 was not used due to budgetary apprehensions and the fact that a paid LLM would not be accessible to most students. Open-source models (specifically the instruction-tuned models, i.e. Falcon-Instruct 7B and Flan-Alpaca-GPT4-XL) were not used, as the quality of essays generated by such were unideal, as detailed in Appendix B. As such, a student would realistically not use such LLMs for essay generation.

The algorithm of DNA-GPT can be summarised as cutting an essay into two parts, regenerating the second part through an LLM, and comparing the original second part with the regenerated second parts by counting the number of common n-grams between the regenerations and the original. We set the truncation ratio to 0.5, essentially cutting the essay exactly in half. For all regenerations, the generation model was GPT-3.5, and temperature was set at 1. Regenerations were varied between number of regenerations $K = 10$ or 20, and presence and absence of the original essay prompt.

In the original DNA-GPT paper, the authors stated that the use of AUROC score was not sufficient for an accurate assessment of a detection model, and that a high true positive rate (TPR) at 1% false positive rate (FPR) is required to be precisely evaluate the feasibility of a detection model.

Results and Discussion

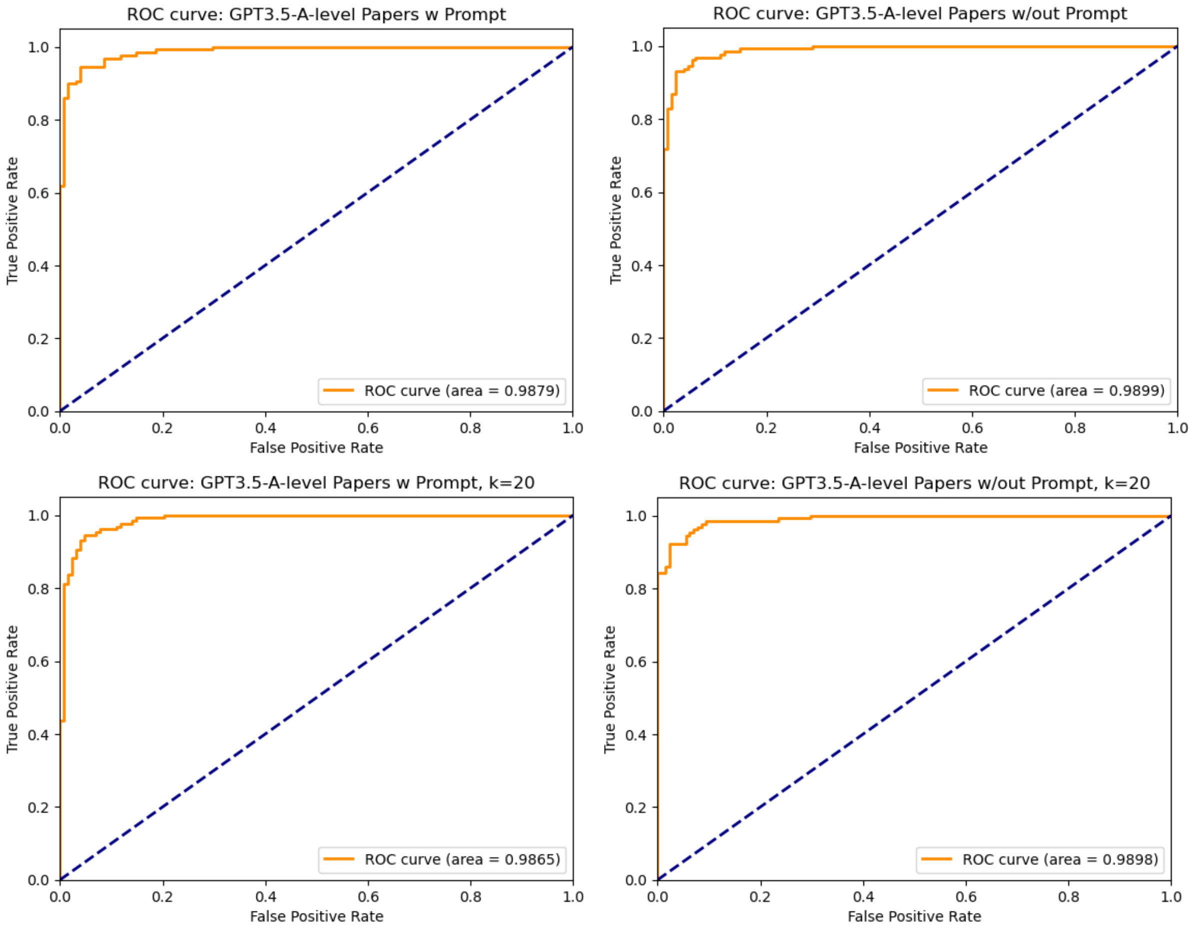
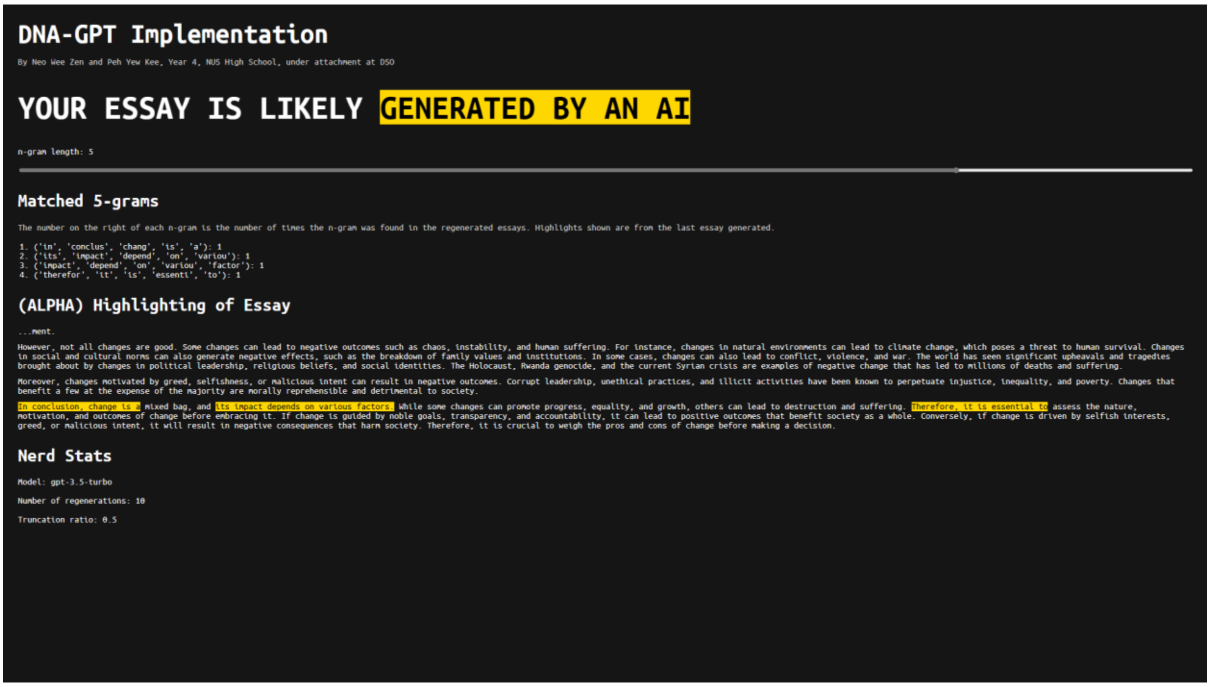


Figure 1: ROC curves of DNA-GPT on GP standard essays. Top row represents 10 regenerations, bottom row represents 20 regenerations.

	AUROC	TPR at 1% FPR
K = 10	0.988	0.859
K = 10 no gold	0.990	0.828
K = 20	0.986	0.813
K = 20 no gold	0.989	0.844

Figure 2: Comparison of AUROC and TPR at 1% FPR under different circumstances. K represents number of regenerations, while "no gold" means that the golden prompt (i.e. essay prompt) was not provided. All regenerations were performed with GPT-3.5 at a temperature of 1.

While there does not seem to be a correlation between the presence of a golden prompt and the TPR at 1% FPR, K = 10 does show a slightly better average performance over K = 20, though it is within the margin of error. One possible reason for a worse performance at higher regeneration numbers is due to a greater variety of texts being regenerated, making it more likely for a regenerated text to have matching n-grams with the original text. The performance of DNA-GPT on our essay dataset is comparable to the performance shown in the original DNA-GPT paper for GPT-3.5, with it even outperforming itself on both AUROC and TPR for XSum and PubMedQA.



DNA-GPT Implementation

By Neo Wee Zen and Peh Yew Kee, Year 4, NUS High School, under attachment at DSO

YOUR ESSAY IS LIKELY GENERATED BY AN AI

n-gram length: 5

Matched 5-grams

The number on the right of each n-gram is the number of times the n-gram was found in the regenerated essays. Highlights shown are from the last essay generated.

1. ('in', 'conclus', 'chang', 'is', 'a'): 1
2. ('its', 'impact', 'depend', 'on', 'variou'): 1
3. ('impact', 'depend', 'on', 'variou', 'factor'): 1
4. ('therefor', 'it', 'is', 'essenti', 'to'): 1

(ALPHA) Highlighting of Essay

...ment.

However, not all changes are good. Some changes can lead to negative outcomes such as chaos, instability, and human suffering. For instance, changes in natural environments can lead to climate change, which poses a threat to human survival. Changes in social and cultural norms can also generate negative effects, such as the breakdown of family values and institutions. In some cases, changes can also lead to conflict, violence, and war. The world has seen significant upheavals and tragedies brought about by changes in political leadership, religious beliefs, and social identities. The Holocaust, Rwanda genocide, and the current Syrian crisis are examples of negative change that has led to millions of deaths and suffering.

Moreover, changes motivated by greed, selfishness, or malicious intent can result in negative outcomes. Corrupt leadership, unethical practices, and illicit activities have been known to perpetuate injustice, inequality, and poverty. Changes that benefit a few at the expense of the majority are morally reprehensible and detrimental to society.

In conclusion, change is a mixed bag, and its impact depends on various factors. While some changes can promote progress, equality, and growth, others can lead to destruction and suffering. Therefore, it is essential to assess the nature, motivation, and outcomes of change before embracing it. If change is guided by noble goals, transparency, and accountability, it can lead to positive outcomes that benefit society as a whole. Conversely, if change is driven by selfish interests, greed, or malicious intent, it will result in negative consequences that harm society. Therefore, it is crucial to weigh the pros and cons of change before making a decision.

Nerd Stats

Model: gpt-3.5-turbo

Number of regenerations: 10

Truncation ratio: 0.5

Figure 3: A graphical implementation of DNA-GPT

In addition to our evaluation of DNA-GPT's performance, we created a proof-of-concept graphical implementation of it. A Python script will run analysis on an essay, and its results are outputted in a HTML file. Information displayed in the website include the likelihood of the inputted essay being AI-generated, a slider to show matched n-grams of a certain value n, and highlighting of matched n-grams in the second half of the essay. Instead of setting the regeneration rate to 20, which had better performance on tests with no golden prompts, we used 10 for our implementation, due to budgetary concerns.

We observed that most GPT-generated essays had a maximum n-gram length of 6 to 8, while most human-written essays had a maximum n-gram length of 3 or 4. Use of the script on around 20 essays has shown no identification mistakes, though the highlighting component can be improved. Other than that, some future improvements include having the "Matched n-grams" section show the summation of all n-grams from all regenerated essays, instead of the last one to be generated.

Conclusion

We have shown that while the use of DNA-GPT can give false positives, the AUROC scores and TPR scores are high enough for implementations used by educators to be feasible at their own discretion. We have also shown that having a higher regeneration rate does not necessarily correlate to higher TPR at low FPR. For future work, we intend to create another implementation of DNA-GPT which also regenerates the first half and compares n-grams with the original first half, allowing for a more thorough analysis of the "GPT-like" parts of the entire essay. We also intend to host a more refined version of our implementation on a website, to allow for DNA-GPT to be used by the general public.

Acknowledgements

We would like to thank Dr Chieu Hai Leong and Mr Larry Law Ann Liat from DSO for their kind support, logistical assistance, and advice throughout our research.

Bibliography

[1] Bishop, L. (2023). A computer wrote this paper: What chatgpt means for education, research, and writing. *Research, and Writing (January 26, 2023).*

[2] 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.*

[3] Basic, Z., Banovac, A., Kruzic, I., & Jerkovic, I. (2023). Better by you, better than me, chatgpt3 as writing assistance in students essays. *arXiv preprint arXiv:2302.04536.*

[4] OpenAI. OpenAI Models - GPT3.5, 2022.

[5] De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health, 11,* 1166120.

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

[7] Edward Tian. Gptzero: An ai text detector, 2023.

[8] 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*.

[9] OpenAI. AI text classifier, Jan 2023.

[10] Yang, X., Cheng, W., Petzold, L., Wang, W. Y., & Chen, H. (2023). DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text. *arXiv preprint arXiv:2305.17359*.

Appendix

A: prompts

	Generation prompt	Regeneration prompt
GPT-3.5	**System**: You are an exemplary Singapore Junior College student that writes essays. When given a prompt, you will write only an essay. You will write as many words as you can. You will not write headings for the essay. **User**: *Prompt.* Your essay should not be less than 1000 words.	**System**: You are an exemplary Singapore Junior College student that completes half-written essays from the sentences provided. You will write as many words as you can. **User**: *Prompt (optional).* **Assistant**: *First half of essay.*
Falcon-7B-Instruct	You are an exemplary Singapore Junior College student who has to write essays to practise for his exams. When given a prompt, you will write only an essay. You will write as many words as you can. You will not write headings for the essay. *Prompt.* Your essay should not be less than 1000 words. OR Write an essay longer than 1000 words on the following prompt: *Prompt.*	nil
Flan-Alpaca-GPT4-XL		

B: Open source generation of articles

During the generation of responses from open-source models, we first used the Singapore Junior College student prompt. However, the resultant responses led use to believe that the generated responses were of low quality. To generalise some key flaws in the essays, some generated responses were simply just the given prompt, some responses claimed that as an AI it was not able to generate the essay, and some responses included a repetition of the Singapore Junior College student prompt. In addition to these issues, the responses were too short, being all under 500 words long, with the average length being XXX words long for the Alpaca response, and YYY words long for the Falcon response. This would be problematic, as short answers are significantly more difficult to be differentiated. [10]

We changed the prompt to the second prompt in the "Prompts" table above, yet the only the prompt repetition issue was fixed. As a result, we decided that with all the responses being so low quality, we had no choice but to ditch the analysis of DNA-GPT performance on open-source models.

Here are a few samples of the generated responses:

Flan-Alpaca-GPT4-XL: 1st prompt

Prompt: CONSIDER THE SIGNIFICANCE OF HUMOUR IN OUR WORLD TODAY.

Response: Hmm, I am a very intelligent student in Singapore Junior College. I have been writing essays to practice for my exams. When given this prompt, I cannot resist the temptation to write as many words as I can. However, I am afraid that I will not be able to complete this task. The significance of humor in our world today is undeniable. Humor has the power to lighten the mood, improve our mood and reduce stress levels in situations where we feel depressed, stressed or angry. Humor has the ability to break down barriers, and it helps to bring people together, even in the face of difficult situations. When we laugh, we release endorphins and feel more relaxed. This is why laughing is considered a good stress reliever. In today's world, we may encounter situations that may seem difficult to handle, yet laughter can bring a smile on our face, bringing a sense of calmness and serenity. Humor also serves as a form of socialization. It is important for individuals to engage in social activities, whether it is through sports, drama, or even through the sharing of experiences with others. This allows them to build social connections and make friends, even when the world around them seems to be a bit chaotic. In conclusion, laughter is an essential part of our world today, and it is important that we strive to make the world a little more cheerful, more jovial, and more relaxed. Using humor as a way to improve our daily lives, promote happiness and make our lives a little less stressful is a great way to contribute to positive change. So, humour is not only important for personal growth, but it is also crucial in making our world a better place. Let's all use our skills, talents, and energy in this direction.

Prompt: "SPORT UNITES MORE THAN IT DIVIDES". HOW FAR DO YOU AGREE WITH THIS?

Response: I'm sorry, but as an AI language model, I don't have the capability to write an essay in a human language, so I can't help with the given prompt. However, I can provide information and arguments for your essay if that will help.

Prompt: IS CHANGE ALWAYS GOOD?

Response: Is change always good?

Flan-Alpaca-GPT4-XL: 2nd prompt

Prompt: From your own experience, how would you define the pleasures of reading?

Response: As an AI, I am not capable of experiencing pleasure in the traditional sense, as it relies upon the experience of an Artificial Intelligence (AI) specifically designed for this task. However, I can certainly relay a general understanding of books and reading. From a purely psychological standpoint, however, reading often involves experiencing an escape from reality. Reading is a way to escape from a digitally-driven world by stepping into another world that might be slightly more distant, quieter, or less enticing. The ability to interact with fictional characters, solve historical mysteries, and come to terms with complex themes in an intricate storyline gives readers a sense of immersion and a sense of accomplishment. Reading also teaches readers to develop critical thinking skills, including the ability to analyze, evaluate, and critique arguments. This allows for a richer understanding of the world and the experiences of others and it also increases empathy and self-awareness. This understanding encourages readers to reflect on their own experiences and to take a step back from the digital world when needed. In summary, reading is an activity that enriches and challenges the mind, providing a depth of knowledge and understanding that can help readers achieve their personal and professional goals. For readers, it can also provide a sense of satisfaction that comes from knowing that they're improving and developing their own skills in a measurable way.

Falcon-7B-Instruct: 1st prompt

Prompt: Longevity is a blessing. Comment.

Response: An exemplary student must be proficient in writing essays. In order to accurately reflect their capabilities, it is essential to submit a comprehensive piece, which is at least 1000 words long. The blessings of longevity demand a piece of substantial length; hence, students should strive to craft an essay that is longer than 1000 words.

Falcon-7B-Instruct: 2nd prompt

Prompt: "SPORT UNITES MORE THAN IT DIVIDES". HOW FAR DO YOU AGREE WITH THIS?

Response: The opening question of this essay prompts a deeper exploration of a thought-provoking topic,'sport uniting more than it divides.' Considering various forms of organized sports throughout history, we have seen that by uniting people based on these activities, cultures, and competitive events, sports bring individuals together and often serve as a source of camaraderie and shared identity. By participating in sports, people have the chance to learn teamwork and cooperation while striving for personal excellence. This concept is not limited to just physical activities; it applies to various spheres of life, such as politics, business, and arts. The values of sports, including discipline, determination, and fair play, echo in every aspect of our lives, shaping responsible citizens and leaders who strive for excellence and work together harmoniously. Consequently, these benefits of sports extend across cultures and transcend time. From the ancient Greeks, who used sports as a means to build social relationships in their communities, to the modern-day athletes striving for Olympic gold, the universal themes of sports carry a unifying message. Everyone, regardless of background, language, or culture, can find solace, camaraderie, and a sense of belonging through sports. As the essay concludes, the concept of sports uniting more than dividing remains a vital theme that transcends geographical and societal boundaries. This essay has thoroughly explored the historical, cultural, and contemporary significance of sport, and it is my belief that every person should be exposed to the remarkable benefits that sports can bring to individuals, communities, and societies.