DEVS: ZERO-SHOT AI-GENERATED TEXT DETECTION VIA SUMMARISATION

Peh Yew Kee[1], Neo Wee Zen[1], Chieu Hai Leong[2]
[1]NUS High School of Mathematics and Science, 20 Clementi Avenue 1, Singapore 129957
[2]DSO National Laboratories, 12 Science Park Drive, Singapore 118225

ABSTRACT

With Large Language Models (LLM) on the rise, AI-generated text detectors have become increasingly necessary to identify the unethical uses of LLMs. Among AI-generated text detectors, DNA-GPT exhibits state-of-the-art performance in a zero-shot setting. In this paper, we build upon the idea of divergent n-gram analysis as demonstrated in DNA-GPT, with Detection Via Summarisation (DeVS). Our detection algorithm involves prompting an LLM (i.e. GPT-3.5) to summarise a given piece of text, followed by prompting it to regenerate the text given the summary, and finally an analysis on divergent n-grams between the regeneration and the original text. Our method of zero-shot AI-generated text detection was tested on our own A-Level General Paper dataset, along with PubmedQA and Scientific Abstracts datasets, and resultant AUROC and TPR at 1% FPR metrics are on par, if not better, than DNA-GPT on certain datasets, when only unigrams are considered.

INTRODUCTION

Large Language Models (LLMs) like ChatGPT have been predicted to improve the economy, whilst being positively correlated with wages [1]. However, its use has also resulted in concerns of plagiarism, perpetuated by students and scientific writing [2][3], with the latter issue having been described as “troubling” due to the potential of LLMs to hallucinate, or provide false, misleading, or inaccurate statements and information.

This has naturally resulted in the proliferation of new AI-generated text detection algorithms and methods, such as watermarking, zero-shot based detection, and trained classifiers [4]. A notable zero-shot algorithm would be Divergent N-gram Analysis (DNA-GPT), which separates a given text into two parts. The first part is inputted into an LLM to regenerate the second part of the text multiple times. The n-grams of the original and regenerated second parts are then compared, in order to classify a text as machine-generated or human-written. DNA-GPT, which will be used as this paper’s baseline, shows state-of-the-art performance, along with explainable detection. [5]

In this paper, we improve upon the approach through Detection Via Summarisation, where a given text is regenerated through a summary of itself, then compared to the regenerations. In essence, given appropriate context, LLMs tend to output highly similar text across runs of regeneration. Due to the ubiquity of ChatGPT, our paper will focus on GPT-3.5 for generation and regeneration of text.

METHODOLOGY

For datasets, we made use of our own GP essays dataset, generated a dataset composed of scientific abstracts from Nature, along with pre-established datasets like PubMedQA. Our GP essay dataset is composed of 128 manually compiled human-written datasets compiled from the internet, from varying sources, ensuring a diversity in quality to test the generality of the models. We also compiled 150 pieces of text from PubMedQA by means of concatenating contexts with the long answer. 100 scientific abstracts were scraped off scientific articles on the Nature website. We also used GPT-3.5 to generate mirrors of the three human-derived datasets, by prompting it the title or question of a given text. For these mirrors, temperature was set to 1.

DeVS is a binary classification algorithm, which works as follows: for any given text sequence S_0, we first prompt GPT-3.5 to provide a summary of the text M. Following that, we prompt GPT-3.5 with the generated summary to regenerate the text sequence K times, to produce a set of text sequences $\Omega = \{Y_1, ..., Y_k, ..., Y_K\}$. We varied whether the question, title, or prompt of the given text was provided to GPT-3.5 in regeneration. Lastly, we derive a score based on the number of n-grams found in both S_k and S_0. We believe that the empirical observation stated in the original DNA-GPT paper ("Given appropriate preceding text, LLMs tend to output highly similar text across multiple runs of generations.") should carry over to the utilisation of summary, as the underlying system of regeneration is retained.

The following formula is used to calculate the score of each text:

$$\text{Score}(S, \Omega) = \sum_{k=1}^{K} \sum_{n=1}^{N} n \frac{|\text{grams}(Y_k, n) \cap \text{grams}(Y_0, n)|}{|Y_k||\text{grams}(Y_0, n)|}$$

where K refers to the total number of regenerations, and N refers to the highest n-gram size analysed.

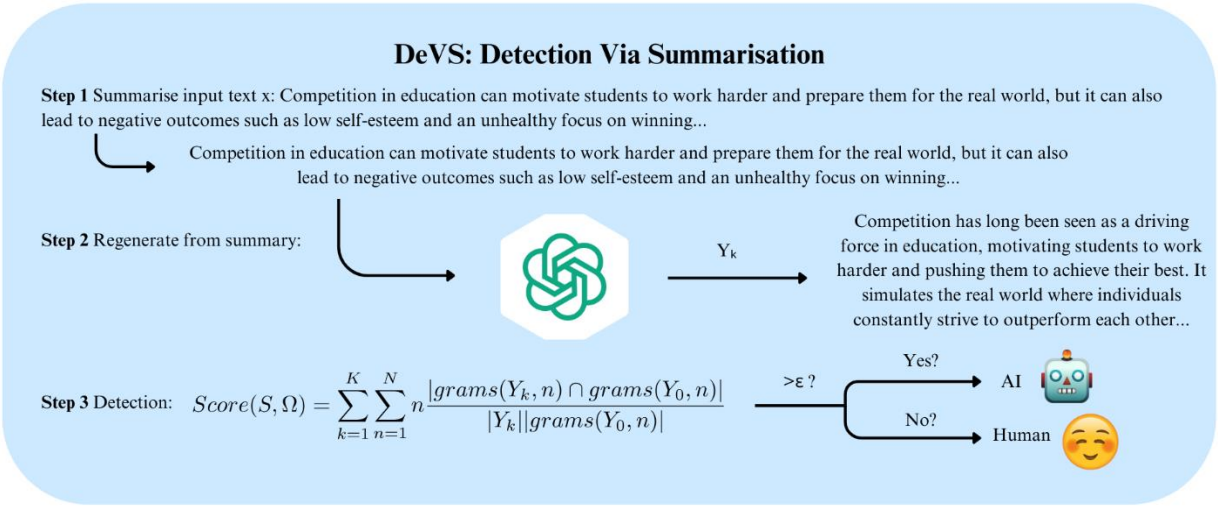


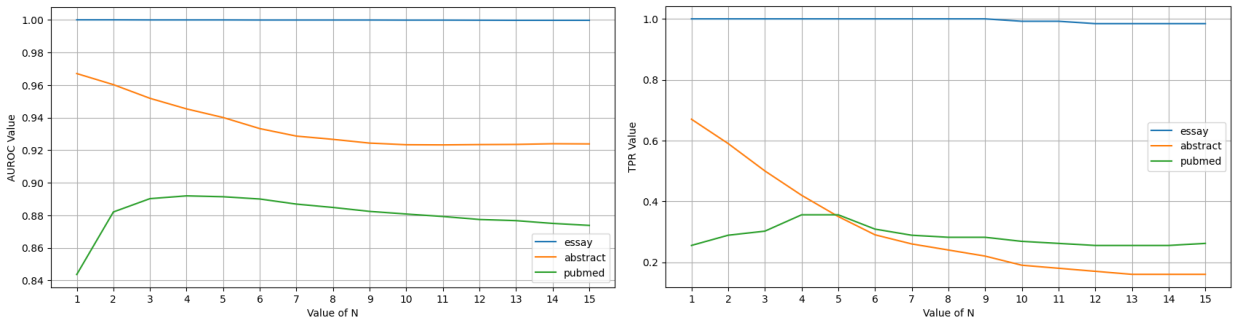
Figure 1: Diagram of algorithm of Detection Via Summarisation.

RESULTS

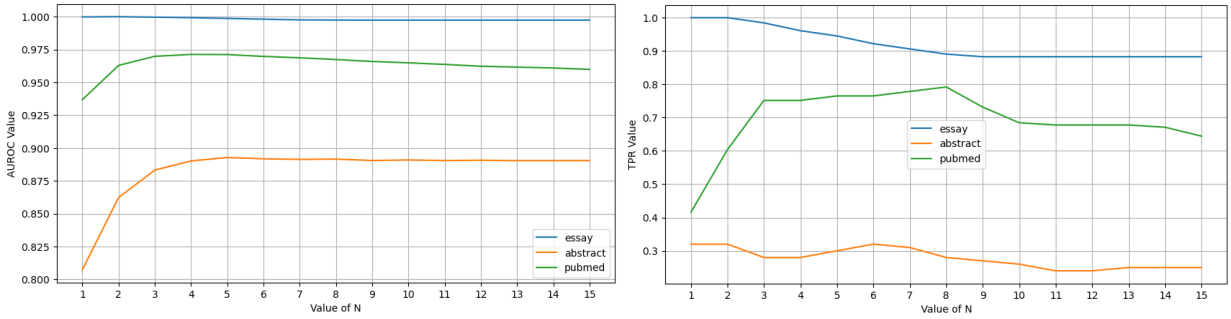
		GP Essays		PubMedQA		Scientific Abstracts	
		AUROC	TPR at 1% FPR	AUROC	TPR at 1% FPR	AUROC	TPR at 1% FPR
DNA-GPT, K=10, γ=0.5	No prompt	0.9899	0.8281	0.9593	0.6000	0.9956	0.9500
	With prompt	0.9879	0.8594	0.9710	0.5533	0.9965	0.9110
DeVS, K=1	No prompt	0.9644	0.3516	0.8919	0.3557	0.8033	0.3900
	With prompt	0.9634	0.6484	0.9674	0.7919	0.8073	0.3200
DeVS, K=5	No prompt	0.9725	0.4531	0.9083	0.4832	0.9670	0.6700
	With prompt	0.9842	0.6328	0.9555	0.5638	0.9307	0.6000
DeVS, K=10	No prompt	0.9646	0.4297	0.9152	0.4497	0.9483	0.5000
	With prompt	0.9747	0.4609	0.9849	0.7315	0.9415	0.5300

Table 1: Performance metrics of DNA-GPT compared to DeVS (all values were obtained using GPT-3.5. DeVS values were the best obtained from variation of largest n-gram size analysed.) "No prompt" or "With prompt" refers to whether the prompt, question, or title of the given text was provided to GPT-3.5 in regeneration.

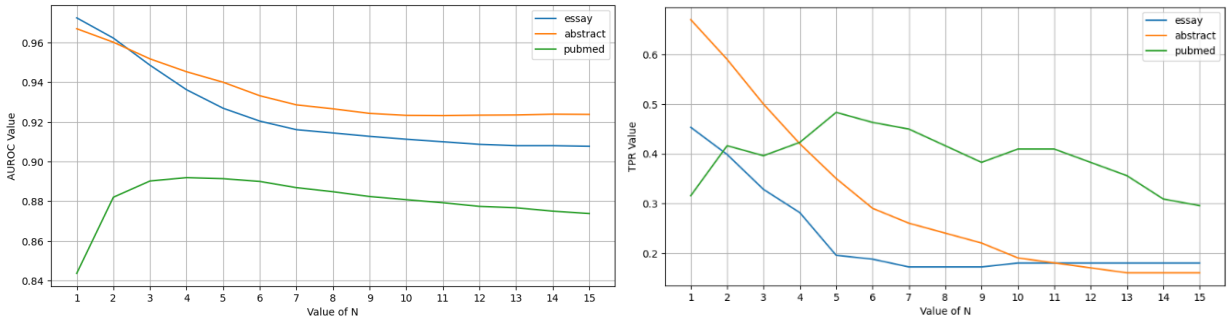
DeVS, K=1, No prompt



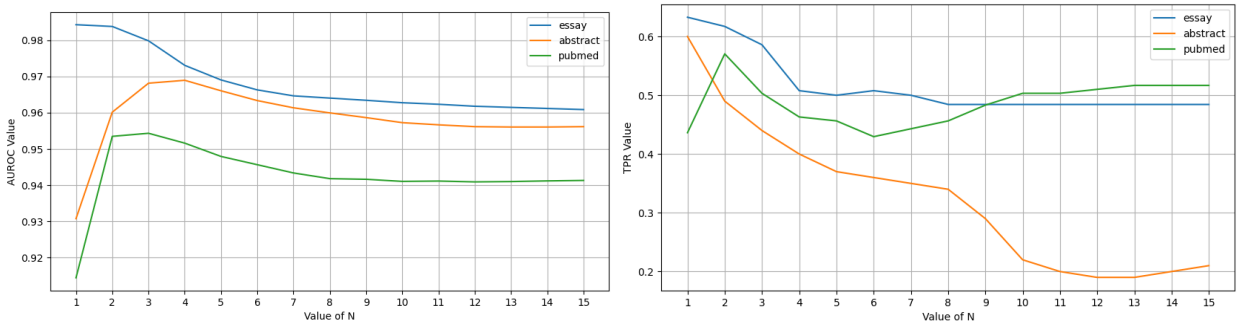
DeVS, K=1, With prompt



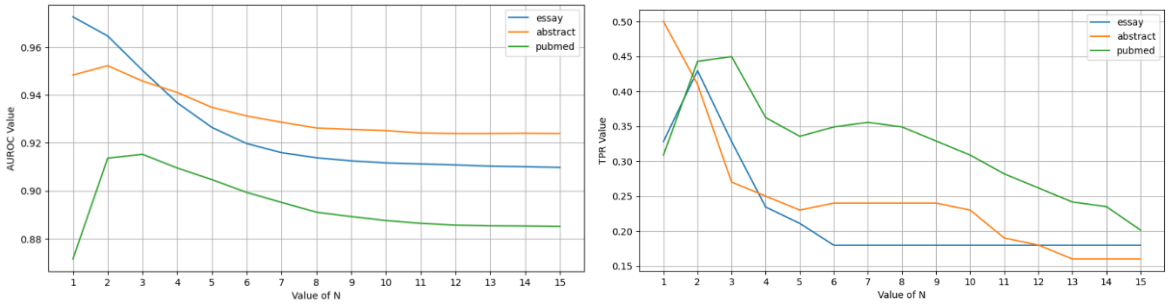
DeVS, K=5, No prompt



DeVS, K=5, With prompt



DeVS, K=10, No prompt



DeVS, K=10, With prompt

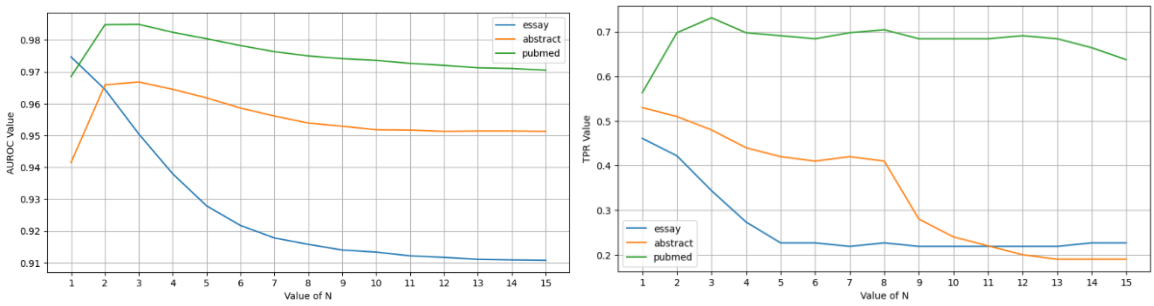


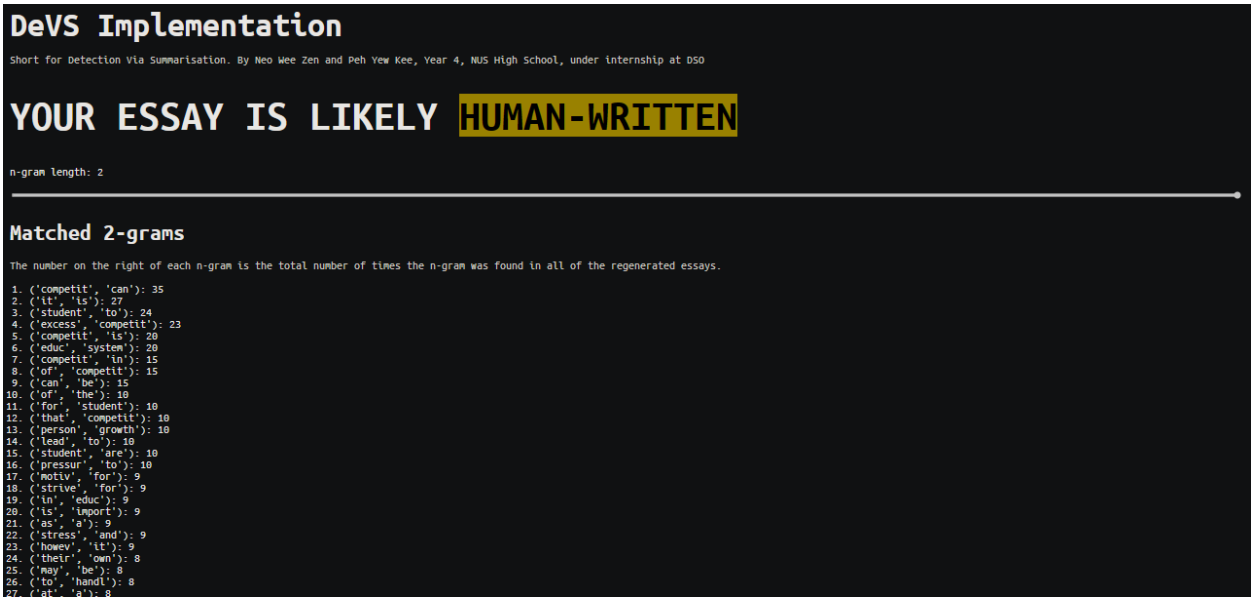
Figure 2: Effect on largest n-gram size analysed on performance metrics, where N refers to the largest n-gram size analysed. We can observe that DeVS generally performs better on lower largest n-gram size.

DISCUSSION

We evaluated the AUROC and TPR at 1% of DeVS for various datasets and compiled our results in Table 1. From our results, we can observe that at ideal largest n-gram size analysed, our model shows state-of-the-art results for PubmedQA. We believe that this is due to summarisations of biomedical texts only containing the context of studies, with fewer technical biomedical jargon. As GPT-3.5 is unlikely to be trained on these biomedical studies, when prompted to regenerate the text, GPT-3.5 will hallucinate and is unlikely to regenerate the accurate information and medical terms used in the original text. This will result in fewer matching n-grams. Conversely, the AI mirrors of the PubmedQA text is likely to contain the same hallucinations as its regeneration, resulting in more matching n-grams.

However, performance metrics for DeVS on the other two datasets (Scientific Abstracts and GP Essays) is poorer when compared to DNA-GPT. For abstracts, we believe that this could be due to the collected abstracts having a short word count, with the average word count in our dataset being 160. As for the GP essays, we believe that due to the context being more general, scores are unable to be influenced as much as PubmedQA from the jargon, as the words used in essays are not as technical. As such, there are fewer factors present to create a larger distinction between the scores of human-written and AI-generated texts. It is also likely that GPT-3.5 summarisation includes text taken from the given text verbatim. This would have resulted in portions of the given text being included in the regenerated text, resulting in a higher score. This would have caused scores for human texts to approach that of AI-generated texts, thus affecting performance metrics.

Unlike DNA-GPT, as our model performs better on analysis of only smaller n-gram sizes, this results in a faster, less resource-intensive analysis than DNA-GPT. Unusually, there are no clear trends as to how number of regenerations or presence of prompt generally affects performance of DeVS.



We also developed a GUI for visualisation of results. We notice that stopwords like ('it', 'is') are being included in the scoring. While stopwords are generally considered to be insignificant, we believe that the reality is a little more nuanced. Specifically, we believe that LLMs like GPT-3.5 have the propensity to produce a specific string of words, including stopwords, together. As such, it is inconclusive whether the removal of stopwords has an effect on the performance of DeVS.

We believe that despite our suboptimal results, the DeVS algorithm is still worth improving upon, as it is one of the only models to be able to be fully explainable (i.e. portions resembling AI in the entire text can be identified precisely). There are many possible ways to approach the problem of explainability, but the performance of these methods is sometimes, if not usually, not as effective for a myriad of reasons. Notably, an n-gram centric approach may be unable to identify sentences with nearly the exact same words being used, with nearly the exact same meaning, as similar, if only the words were rearranged. We believe that this is a possible reason why our model performs worse in higher N: this might have resulted in larger n-grams in regenerations being very scarce, only causing the scores outputted to have less of a gap between human and AI-generated text.

CONCLUSION

While Detection Via Summarisation may not be as robust to attacks as other algorithms, we have demonstrated its state-of-the-art performance in biomedical contexts. Possible future work includes a hybrid model, by first regenerating a given text through a summary of the text, followed by truncating the regeneration in two, and regenerating the second part using the first half. Afterwards, the second round of regeneration is compared to the original given text. Furthermore, work can be done to investigate whether DeVS can be utilised as a red-teaming approach to evade detection of AI-generated text by other state of the art models.

ACKNOLEDGEMENTS

We would like to Dr Chieu Hai Leong for his invaluable advice and logistical support over the course of this project. We would also like to thank Ang Jun Ray for his logistical support.

REFERENCES

[1] Felten, E., Raj, M., & Seamans, R. (2023). How will Language Modelers like ChatGPT Affect Occupations and Industries?. *arXiv preprint arXiv:2303.01157*.

[2] Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), 410.

[3] Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, *15*(2).

[4] Ghosal, S. S., Chakraborty, S., Geiping, J., Huang, F., Manocha, D., & Bedi, A. S. (2023). Towards possibilities & impossibilities of ai-generated text detection: A survey. *arXiv preprint arXiv:2310.15264*.

[5] 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*.