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

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 (LLM) leads to plagiarism from students, academia[1][2]
- LLMs can hallucinate (provide false, inaccurate statements and information)
- Solution: Detection Via Summarisation, a novel approach for AI text detection, where given text is regenerated through summary of itself, then compared to regenerations.

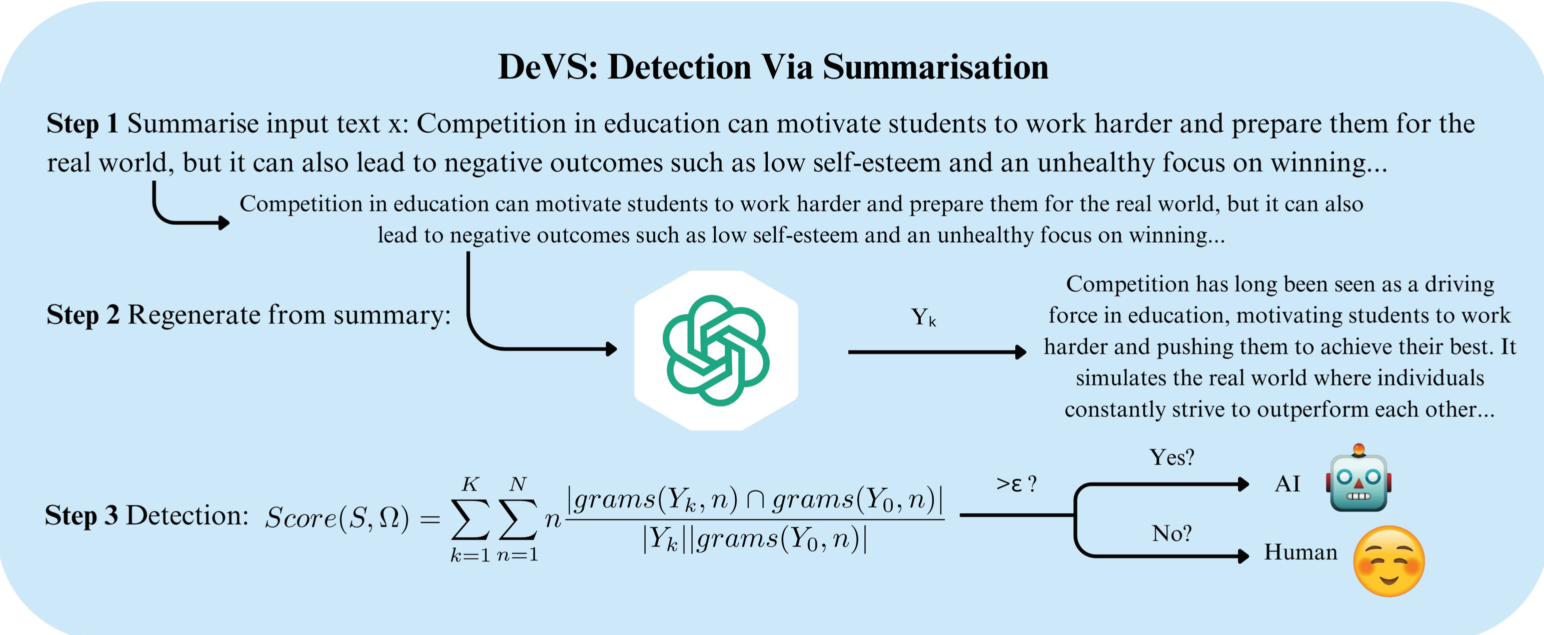

Figure 1: Effect on largest n-gram size analysed on performance metrics, where N refers to the largest n-gram size analysed. We can observe that DeVS generally performs better on lower largest n-gram size.

METHODOLOGY

Datasets: PubmedQA[3], GP Essays, Scientific Abstracts from Nature

Algorithm: Given text sequence Y_0…

1. Prompt GPT-3.5 for summary;
2. Prompt GPT-3.5 with the same summary to regenerate the text sequence K times, produce text sequences $\Omega = \{Y_1, ..., Y_k, ..., Y_K\}$
   - Vary whether question, title, or prompt of the given text was provided in regeneration.
3. Derive score based on the number of n-grams found in both Y_k and Y_0
   - $\text{Score}(S,\Omega) = \sum_{k=1}^{K}\sum_{n=1}^{N} n\frac{|\text{grams}(Y_k,n)\cap \text{grams}(Y_0,n)|}{|Y_k||\text{grams}(Y_0,n)|}$ where K refers to the total number of regenerations, and N refers to the highest n-gram size analysed.

RESULTS

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.

		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

RESULTS (CON'T)

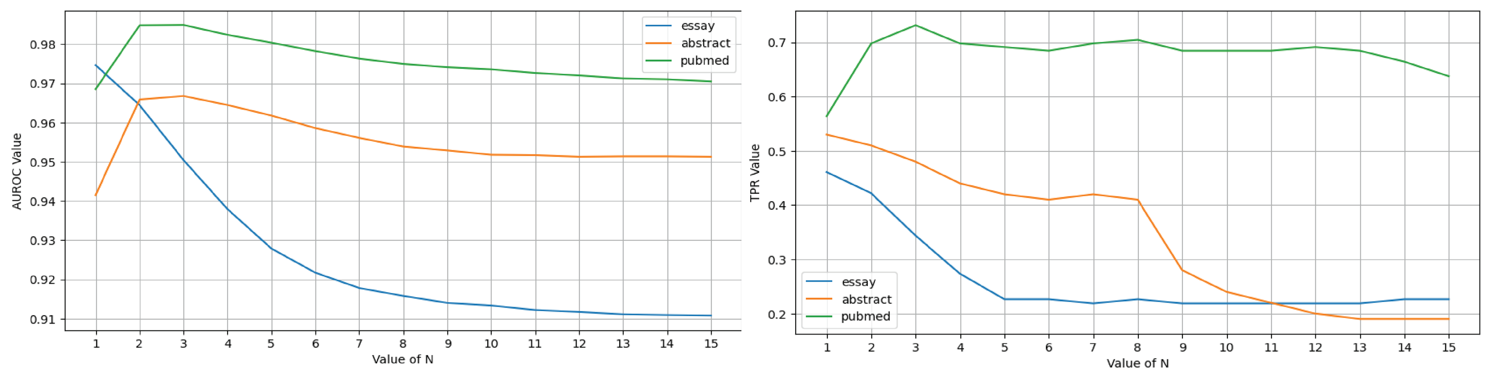

Figure 2: Effect on largest n-gram size analysed on DeVS, when number of regenerations is set to 10, and with prompt provided in regeneration, where N refers to the largest n-gram size analysed. We can observe that DeVS generally performs better on lower largest n-gram size.

- From Table 1: at ideal largest n-gram size analysed, DeVS shows state-of-the-art results for PubmedQA.
  - Likely due to GPT-3.5 hallucinating; unlikely to regenerate accurate information and medical terms used in original text, resulting in fewer matching n-grams.
- Performance metrics for DeVS on Scientific Abstracts and GP Essays poorer when compared to DNA-GPT.
  - Scientific Abstracts: likely due to short text length (average word count: ~160)
  - GP essays: likely due to the GPT-3.5 summarisation including text verbatim from Y_0, larger portions of the regenerated text will appear in the original text
- From Figure 2: unlike DNA-GPT, DeVS performs better on analysis of only smaller n-gram sizes
  - Can perform a faster, less resource-intensive analysis than DNA-GPT.
- No clear trends for how K or presence of prompt affects performance

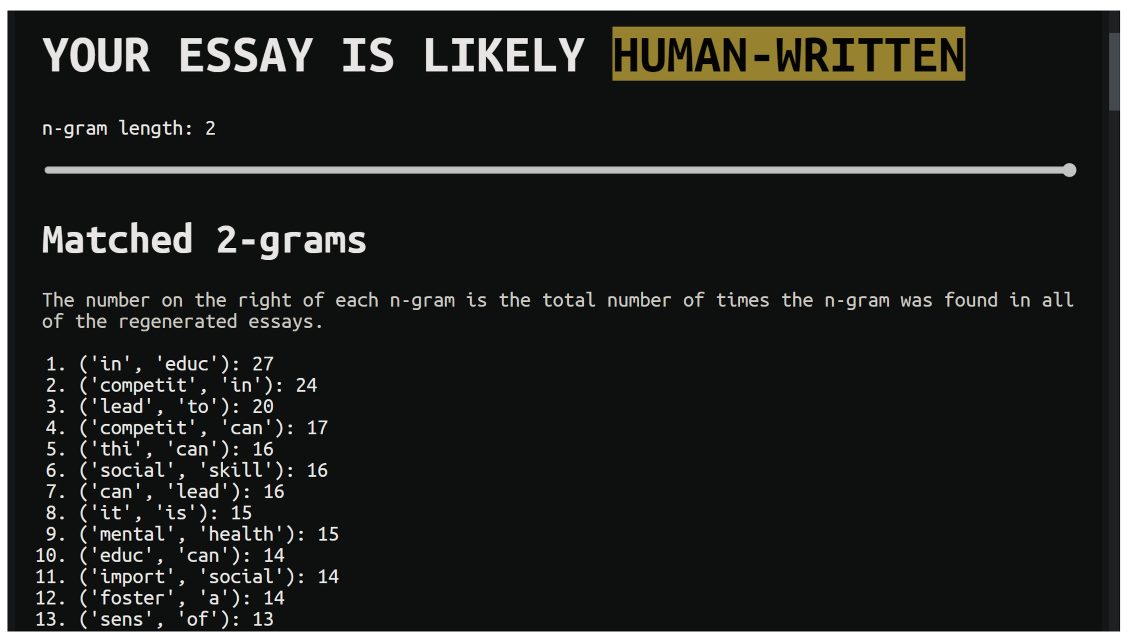

Figure 3: Graphical user interface (GUI) of DeVS, to provide a more accessible way of visualising the results.

- From Figure 3: Graphical user interface developed for visualisation of results
  - Allows others to visualize the way DeVS decides on the similarity of given text
- Despite suboptimal results, DeVS is still worth improving upon,
  - One of the only models to be able to be explainable on the entire given text
  - Fast and low-cost to run; suitable for large amounts of text that other methods may take unrealistic amounts of time to process.

CONCLUSION

- Demonstrated its state-of-the-art performance in biomedical contexts.
- Possible future work:
  - Hybrid model of DeVS and DNA-GPT
    - Regenerate given text through a summary of the text
    - Truncate regeneration in two, regenerate the second part using the first part
    - Second round of regeneration compared to original given text.
  - 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.
  - Increase robustness to attacks

ACKNOWLEDGEMENTS

We would like to thank Dr Chieu Hai Leong and Ang Jun Ray from DSO for their kind support, logistical assistance, and advice throughout our research.

REFERENCES

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

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

[3] Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146.*

Members:
Peh Yew Kee, NUS High School of Mathematics and Science
Neo Wee Zen, NUS High School of Mathematics and Science
Mentor:
Dr Chieu Hai Leong, DSO National Laboratories