Is it Possible to Distinguish Between AI and Human-Generated Text Using Watermarking Techniques?

Sigurður Haukur Birgisson

23rd of October, 2024

Introduction

AI-generated text is becoming increasingly difficult to distinguish from human-generated text as large language models (LLMs) continue to improve rapidly. This improvement has led OpenAI, a leading AI research company, to discontinue access to its AI detector due to its low accuracy (Kirchner, Ahmad, Aaronson, & Leike, 2023). Detecting AI-generated text is challenging because such text can be edited to appear more human-like. Additionally, training detection models for non-English languages is difficult due to limited training data. Lastly, neural network-based methods have a high rate of false positives (i.e., human-generated text incorrectly flagged as AI-generated) for inputs outside their training data (Kirchner et al., 2023).

One proposed technique to distinguish between AI and human-generated text is watermarking. Watermarking is a statistical technique that embeds a unique identifier within text by altering the language model's token sampling. This watermarking causes the generative model to favor certain tokens over others, which can be used to identify the model that generated the text.

This raises the question: is it possible to distinguish between AI and human-generated text using watermarking techniques?

This essay argues that watermarking is a reliable method for distinguishing between human and AI-generated text. It can be applied across languages and text types, even extending to other media types such as images, videos, and sound. Crucially, watermarking can be embedded within a model's parameters, making it nearly impossible to remove without drastically reducing the model's performance.

Body

Watermarking is a reliable method for distinguishing between AI and human-generated text. Large language models are trained on extensive datasets to predict the next token in a sequence of tokens (where a token can be a character, subword, or word). Given an input, language models output the most likely token by sampling from a probability distribution. By modifying the token sampling's probability distribution, the model can be guided to generate predictable tokens, which effectively watermarks the text (Dathathri et al., 2024).

This is illustrated in Figure 1, where the input is converted into a sequence of tokens. The model generates a probability distribution for the next token in the sequence, which can be adjusted (in this example, by increasing the probability of three specific tokens). This affects the model's output and effectively watermarks the media, whether text, image, or sound. Of course, this adjustment must be applied to each language model in a manner that does not impact its performance.

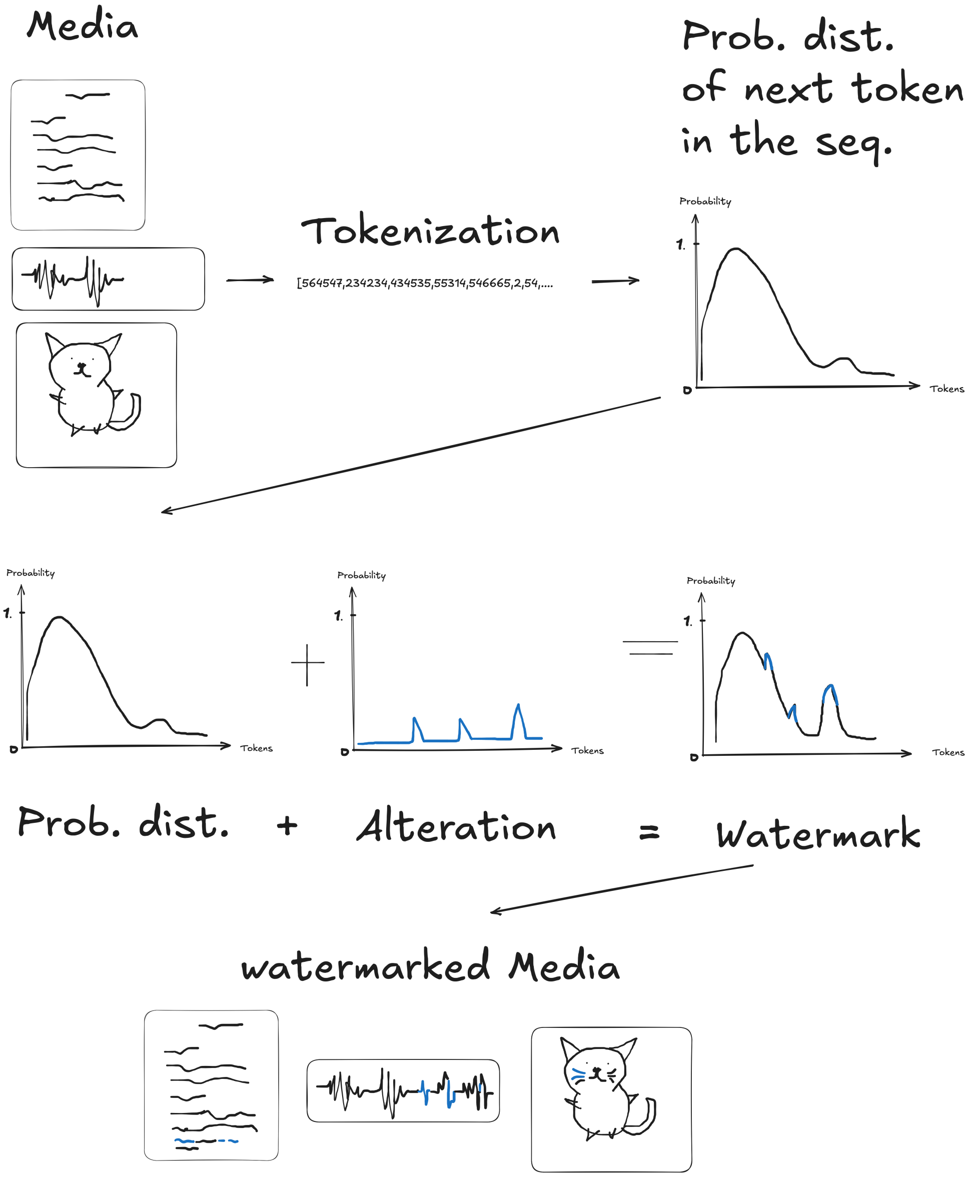

Figure 1: Probability distribution of token sampling and how it can be altered to watermark text

Watermarking is applicable across all languages and types of text, even extending to images. Because watermarking works by altering the randomness of token sampling, it is not language-dependent and can be applied universally. This universality makes watermarking a highly favorable method for distinguishing between AI and human-generated text. Different language models can be trained on different languages while using the same watermarking method. In fact, watermarking is not even media-dependent, as similar methods can be applied to images, videos, and sound (DeepMind, 2024).

A counter-argument suggests that watermarking only works with closed-source models, as bad actors have significant motivation to remove any watermarks from open-source models. However, watermarking can be integrated down to the parameter level, meaning that altering the parameters to remove the watermark would severely degrade model performance. Recent research by Christ, Gunn, Malkin, and Raykova (2024) found that "the strongest attack we consider requires deteriorating text quality to zero out of 100 to bring the detection rate to 50%." Text quality was computed by having Mistral-Instruct-7B, a large language model, assign a score out of 100. This significant performance degradation would render the model practically unusable. Removing the watermark without affecting performance would require retraining the model from scratch, a costly and time-consuming process.

Conclusion

As language models continue to advance, knowing the origin of text will become increasingly important to prevent the misuse of AI in propaganda campaigns and to uphold academic integrity. Watermarking is a reliable method for distinguishing between AI and human-generated text and is applicable across all languages and types of text. Watermarking can be integrated at the parameter level, making its removal impossible without significantly compromising performance.

This essay concludes that watermarking is a reliable and immediately applicable method for distinguishing between AI and human-generated text, offering a robust solution to the challenge of detecting AI-generated content.

References

Christ, M., Gunn, S., Malkin, T., & Raykova, M. (2024). *Provably robust watermarks for open-source language models.* Retrieved from https://arxiv.org/abs/2410.18861

Dathathri, S., See, A., Ghaisas, S., Huang, P.-S., McAdam, R., Welbl, J., ... Kohli, P. (2024). Scalable watermarking for identifying large language model outputs. *Nature*, *634*(8035), 818–823. Retrieved from https://doi.org/10.1038/s41586-024-08025-4 doi: 10.1038/s41586-024-08025-4

DeepMind. (2024). *Synthid: Deepmind's technology for identifying ai-generated content.* https://deepmind.google/technologies/synthid/. (Accessed: 2024-10-28)

Kirchner, J. H., Ahmad, L., Aaronson, S., & Leike, J. (2023, Jan). *New ai classifier for indicating ai-written text.* Retrieved from https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text