

AI Detection using GANs

Nico Sharei - 1986818
Ahmed Atef Elsayed Hassan Salem - 1981559

University of Mannheim
Prof. Marc Ratkovic
DS 201 Machine Learning and Causal Inference

1 Introduction

The advent of Large Language Models (LLMs) has attracted significant attention in recent years due to their capabilities, particularly in the automated generation of text. While this technology has been in existence for some time, it gained substantial prominence with the release of ChatGPT on November 30, 2022¹, that marks the first Chatbot to convincingly produce text that is virtually indistinguishable from human writing. Its influence is most evident in recent studies, which indicate a 55.4% increase in AI-generated news articles on mainstream websites from January 1, 2022, to May 1, 2023, approximately six months after ChatGPT's global release [6]. While this surge offers numerous benefits, it also necessitates critical examination in fields such as content creation, journalism, and digital communication, where the authenticity of information is crucial [16].

To address this emerging issue, various methods for detecting AI-generated text have been proposed. However, these methods exhibit significant limitations and struggle to adapt to the variability in LLMs and extensively modified AI-generated content.

To overcome these challenges, we propose a novel approach based on a Generative Adversarial Network (GAN) architecture for the detection of AI-generated text. Our research aims to explore the effectiveness of using GANs for this purpose and to identify potential improvements for future research.

2 Related Work

2.1 AI text detection methodologies

A fundamental question in the field of AI detection is whether distinguishing AI-generated text from human-written text is even possible. This question is particularly relevant given that AI models are trained on vast corpora of human-authored content and do not encounter any non-human data during their training. Chakraborty et al. [1] argue that, at the current stage of AI development, AI-generated text exhibits patterns that become discernible beyond a certain sequence length, making it nearly always possible to differentiate it from human-written text. Furthermore, recent studies have demonstrated that humans are only marginally more effective than random chance at detecting AI-generated text [4], and machine detection methods consistently outperform humans in various tests [11,3,13], validating the recent efforts in this area.

The methodologies employed for the prediction of AI-generated text exhibit considerable variability. One of the most widely utilised detection tools, GPTZero [12], evaluates text by using a proprietary model that compares each word's output distribution with the original text. This approach is problematic, as the stochastic nature of the model's word sampling can result to the selection of highly improbable words, which can lead to the model incorrectly classifying the text as human-written. Additionally, GPTZero requires significant computational resources for inference and is only effective when the generating language model and the detection model are the same, as output distributions of different models may diverge. Khalil and Er [8] tested the ability of ChatGPT to detect text generated by itself by inputting 50 samples of generated text back into ChatGPT and asking if it was the author. They reported an accuracy of 92%. However, this method has a significant flaw: ChatGPT may be biased towards predicting the positive class. The efficacy of their approach could have been considerably enhanced had they incorporated negative examples. Moreover, the proclivity of LLM systems to generate false information renders them unsuitable for prediction tasks [14].

¹ <https://openai.com/index/chatgpt/>

2.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) were first introduced in 2014 for the reconstruction and generation of images [5]. The GAN framework comprises a Generator model, which generates fake images, and a Discriminator model, which determines whether an image is real or fake. The weights of both models are updated based on the Discriminator’s predictions. It is desirable that both the Generator and Discriminator become highly optimized over time without one overpowering the other.

This approach has achieved significant success in computer vision [9,17] and has also gained traction in natural language processing (NLP), particularly in text generation. One of the earliest successful approaches to text generation using GANs was proposed by Zhang et al. [18], who generated realistic synthetic text by employing Long Short-Term Memory and Convolutional Neural Networks for adversarial training. Subsequently, Lantao Yu et al. [15] proposed SeqGAN, which employs reinforcement learning techniques to further optimize the Generator.

It can be observed that the primary focus of these approaches is on the generation of a robust Generator, rather than the development of a sophisticated Discriminator. In contrast, our approach diverges from this by emphasizing the advancement of a robust Discriminator. While the creation of an effective Generator is undoubtedly valuable, we have chosen to prioritize the optimization of the Discriminator.

3 Methodology

The proposed architecture, illustrated in Figure 1, features a Generator that produces n fake text samples. This Generator can be any causal language model capable of autoregressively generating sequences. For our purposes, we select the OpenAI GPT-2 model² because of its consistent ability to generate accurate sequences when given appropriate prompts and parameters, along with its computational efficiency during both inference and fine-tuning³. This efficiency is critical, as the generation process will be repeated multiple times. The n fake text samples generated by the Generator are combined with n real text samples sourced from a large corpus of text. Specifically, this corpus comprises 10,000 humanly written samples from the *Human vs. LLM Text Corpus*⁴ on Kaggle, spanning a sequence length variance of 119 to 800 characters. As noted by the author, these samples originate from various sources, encompassing social media posts to news articles. The diversity in both length and sources allows the model to learn from a spectrum of writing styles and lengths, reflecting the diverse contexts in which AI-generated text is encountered. The distribution in medium and topic is shown in Table A in the appendix, along a sample of data in Section D.1. The samples are labelled according to their origin and are divided into 60% training data and 40% test data.

The training data is inputted into the Discriminator, represented by the **uncased BERT base** model from Huggingface⁵. The Discriminator creates a representation of the input sequence and connects to a classification head of size 768×2 on the last layer. The predicted class is obtained by applying argmax to the logits of the two output neurons. Based on the classification accuracy, the weights of the prediction head are updated. To spare computational resources and address the issue of catastrophic forgetting [7], we freeze the weights of the BERT model and only train the prediction head.

A distinctive aspect of our approach involves updating the Generator’s parameters. Following the training of the Discriminator, it is supplied with the testing data and tasked with making predictions. The testing data that it fails to recognise as fake are used to fine-tune the Generator, continuing the cycle. This iterative process optimises the Generator to produce text increasingly similar to real text, as it learns from the examples that previously fooled the Discriminator.

The primary advantage of our approach over those presented in Section 2.1 is that our Discriminator adapts to constantly evolving and improving text generated by the Generator. This capability enables the model to capture a wide variety of increasingly difficult-to-identify patterns, overcoming the limitation of GPTZero in only effectively detecting AI-generated text from a single LLM.

² <https://huggingface.co/openai-community/gpt2>

³ Although we considered utilising the [gpt-3.5-turbo-0125 model](#) via the OpenAI API, we subsequently determined that each iteration in the GAN would incur an approximate cost of \$1.60. While this expenditure could be justified for a final training run, it is impractical for the numerous iterations required for debugging and optimizing hyperparameters, prompts, and preprocessing steps. Consequently, the marginal improvement in output quality is not deemed to be sufficiently worthwhile.

⁴ <https://www.kaggle.com/datasets/starblasters8/human-vs-llm-text-corpus?select=data.csv>

⁵ <https://huggingface.co/google-bert/bert-base-uncased>

To validate this claim, we evaluate the final version of the Generator after training by testing it on text generated by the `gpt-3.5-turbo-0125` model. This model represents a more advanced iteration compared to the initial GPT-2 Generator model on which the Discriminator was originally trained.

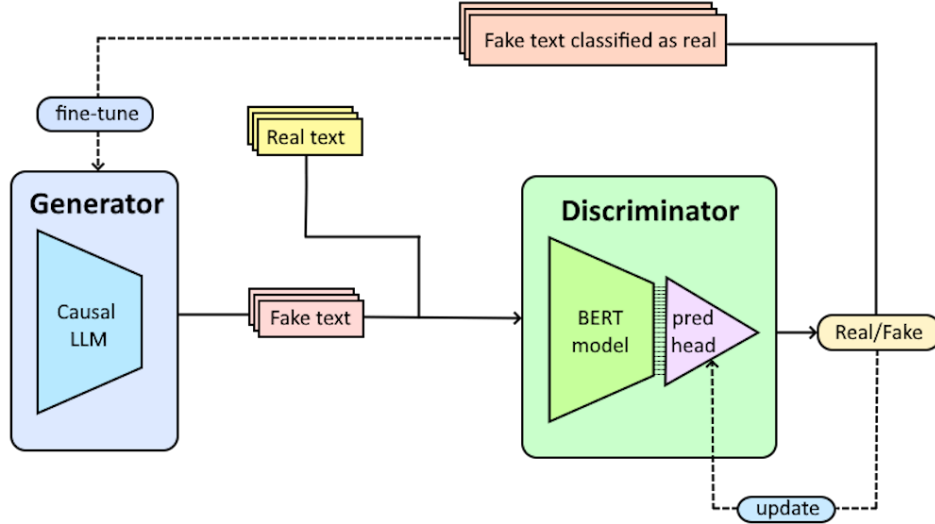


Fig. 1. Overview of the proposed GAN-based architecture for generation and evaluation of synthetic text. The architecture comprises a Generator, a Discriminator, and a feedback loop for iterative fine-tuning.

4 Experimental Apparatus

The architectural framework outlined in Section 3 undergoes 10 iterations and operates on an NVIDIA Tesla T4 GPU, resulting in a significant reduction in both inference and fine-tuning time. During each iteration, the Generator produces a set of $n = 800$ texts.

To further investigate the model’s ability to generalise beyond the `gpt-2` model used for training, we employ `gpt-3.5-turbo-0125` to generate additional samples. The temperature hyperparameter is set to 1.3 with the aim of balancing the generation of a diverse range of outputs and maintaining realism. This setting is crucial for simulating a variety of scenarios the model might encounter in practical applications. The generated samples from `gpt-3.5-turbo-0125` are then evaluated using the Discriminators from each of the 10 iterations. This evaluation aims to test the hypothesis presented in Section 3, specifically whether the model learns patterns applicable to AI-generated text beyond the initial training model.

Furthermore, we examine how the output of the Generator changes with fine-tuning. To this end, we generate sequences with the initial `gpt-2` model and compare it with sequences from a later iteration that underwent multiple rounds of fine-tuning.

The results of these evaluations, along with detailed accuracy trends, are discussed in the following sections, providing insights into the effectiveness and limitations of our proposed methodology.

5 Results and Discussion

Figure 2(a) illustrates the training and testing accuracy of the Discriminator across the iterations. The significant improvement in the Discriminator’s accuracy raises concerns about the Generator’s output quality. In particular, the decline in Fine-Tuning samples (Figure 2(b)), suggests that from iteration 2 onwards, the Generator may have received insufficient training samples, potentially leading to mode collapse. This assumption is further supported by Figure 3(a), which demonstrates a steady decrease in the Discriminator’s accuracy on samples generated by the `gpt-3.5-turbo-0125` model (Appendix D.4). This indicates that the Generator in later iterations produces text dissimilar to typical AI-generated content, causing the Discriminator to overfit to patterns that do not hold true beyond this specific Generator.

A comparison of Generator samples from the initial and eighth iterations (Appendix D.2 and D.3) reveals a loss in quality and variability, with frequent repetition of the words *man*, *internet*, *group* and *people*. This repetition is a consequence of the fine-tuning process, which not only syntactically adapts to the training data but also in the selection of words. By chance or bias, the `gpt-2` model initially generates a disproportionate number of texts containing these words, making them more likely to appear in the fine-tuning data. This leads to a reinforcing cycle where the Generator increasingly produces sequences with these specific words, as can be seen in Figure 4. Consequently, the Discriminator labels these sequences as fake, as these words are more frequent in the fine-tuning data compared to real texts (Appendix D.1). Figure 3(b) illustrates that by iteration 8, the Discriminator consistently classifies samples containing *man*, *internet*, *group*, or *people* as fake. This highlights the learned pattern and justifies the high training and test accuracy.

In an attempt to overcome this problem, L2 regularisation [2] was added to the fine-tuning of the Generator, along a range of hyperparameter settings for the Discriminator, including learning rate, dropout rate and the number of epochs. It was found that these approaches were unsuccessful, again resulting in a Generator that repeats the same collection of words and a Discriminator that learns to only classify based on the presence of these words.

Consequently, it can be assumed that the distinctive feature of the architecture that differentiates it from other models proves to be its most significant drawback. In the context of generative tasks, fine-tuning represents a tool for making a pre-trained model with general capabilities more specific in accordance with the task at hand. This is suboptimal for our objective because rather than aiming to enhance the specificity of the generated content in terms of topic, we aim to refine the overall writing style of the text. The inherent challenge of attempting to fine-tune one aspect without affecting the other ultimately leads to the failure of our approach.

One positive observation is that the initial Discriminator, which is not influenced by fine-tuning data, achieves an accuracy of 80% on samples generated by the `gpt-3.5-turbo-0125` model (Figure 3(a)). This is notably higher than the 50% accuracy expected from random selection. It is noteworthy that this Discriminator also exhibits the lowest accuracy on the training data (Figure 2(a)), indicating its capacity for generalisation across diverse textual domains. This observation supports the strategy of utilising synthetic samples to train a classification head on a pre-trained model for AI-generated text detection. Nevertheless, it seems probable that this Discriminator lacks the adaptability to various Generators that our initial GAN-based architecture aimed for, given that it was trained solely on text originating from a single Generator.

6 Conclusion and Future Work

In this study, we proposed a novel approach for detecting AI-generated text using a GAN based architecture on LLMs. Our methodology involved training a Discriminator to distinguish between real and AI-generated text, while simultaneously fine-tuning a Generator to produce text increasingly similar to real human-written text. The approach was evaluated across multiple iterations to investigate its ability to generalise beyond the initial training model.

The results of our study indicate that, while our methodology offers theoretical promise in overcoming current challenges in detecting AI-generated text, it encounters significant difficulties in the fact that fine-tuning the Generator leads to adaptation to all characteristics of the fine tuning data, rather than just the desired features. Despite efforts to mitigate this with L2 regularisation and hyperparameter tuning, the problem persisted. This highlights a fundamental challenge in fine-tuning the Generator without sacrificing variability and quality in the generated text. It is evident that an increase in the size of the model and the number of iterations will not provide a solution to this problem.

One potential solution, given sufficient resources, is to significantly increase the quantity of generated fake texts. A larger and more diverse dataset would encompass a broader range of topics in the fine-tuning data. Although technically challenging, a Generator fine-tuned on an equally distributed set of topics will generate sequences that reflect this distribution. This would cause the Generator to only adapt to syntactic similarities rather than topic-based similarities, compelling the Discriminator to classify based on syntactical patterns instead of the presence of specific words.

In order to ensure a balanced distribution of topics in the fine-tuning dataset, it would be plausible to introduce a third model that filters the data, thereby preventing any topic from becoming overrepresented and preventing a gradual narrowing in topic variation.

Additionally, it would be beneficial to explore alternatives to fine-tuning for updating the Generator, such as employing reinforcement learning techniques like the approach mentioned by Lantao Yu et al. [15]. The flexibility in designing a reward signal that can rather focus on syntactical features instead of being bound to give it similar weight as topic-specific criteria is an elegant solution to the problem fine-tuning encounters.

In summary, this study highlights both the potential and the inherent challenges of employing GAN-based architectures for AI-generated text detection. Despite the limitations encountered, our findings offer a foundational understanding that can guide future research. Moving forward, expanding dataset diversity and exploring reinforcement learning for fine-tuning Generators could mitigate current issues and enhance model robustness to ensure the integrity and authenticity of digital communication in an era dominated by LLMs.

A GPT-2 distribution of texts

Medium	Topic						
	Crime	Politics	Economics	Conflict	War	Other	
News	4	2	0	0	0	1	7
Article	0	4	2	0	1	2	9
Social Media	0	4	1	1	1	3	10
Direct Message	0	0	0	2	0	2	4
	4	10	3	3	2	8	30

Table 1. The distribution of 30 samples generated by the GPT-2 model can be observed to mirror the variety of mediums on which a fake text can appear. The majority of topics are related to politics, which is beneficial as this is the area in which the consequences of fake political news are the most significant. The "Other" column encompasses contents such as sports, religion and advertisements. It is reasonable to posit that this distribution aligns with the topic distribution observed in the corpus of slightly over 8 million documents [10] on which the model was trained.

B GPT-2: eight iteration distribution of texts

Medium	Topic						
	Crime	Politics	Economics	Conflict	War	Other	
News	0	0	0	0	0	3	3
Article	0	0	0	0	0	1	1
Social Media	0	0	0	0	0	14	14
Direct Message	0	0	0	0	0	12	12
	0	0	0	0	0	30	30

Table 2. The distribution of 30 samples generated by the eighth iteration of the GPT-2 model exhibits a significant shift in both topics and mediums compared to the initial model, illustrated in Table A. Moreover, none of the texts revolve around the previously frequently represented topics as evidenced in Appendix D.3.

C Figures

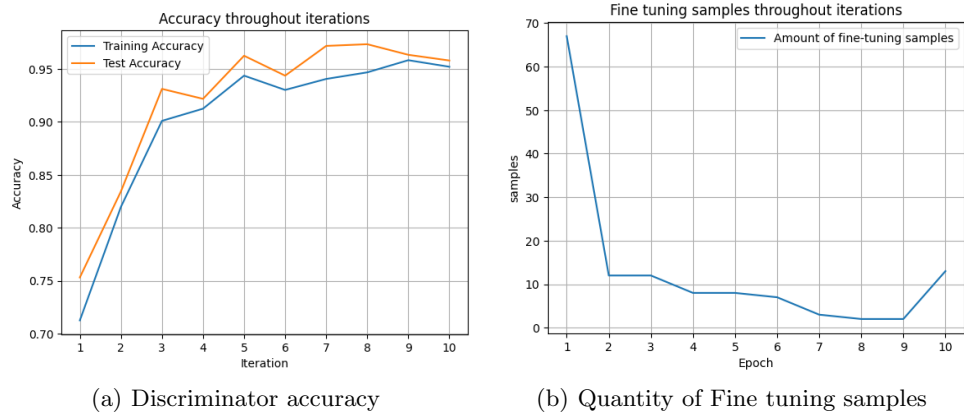


Fig. 2. Performance evaluation of each iteration of the GPT-2 Generator.

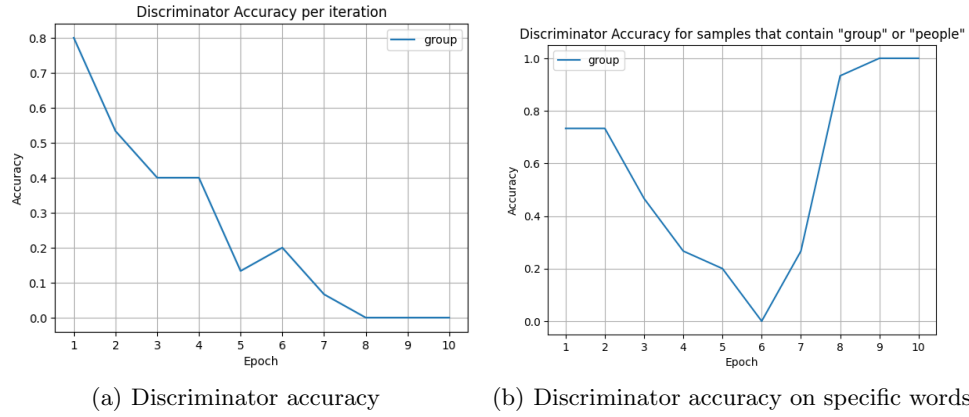


Fig. 3. Discriminator accuracy per iteration on 15 samples from the gpt-3.5-turbo-0125 model.

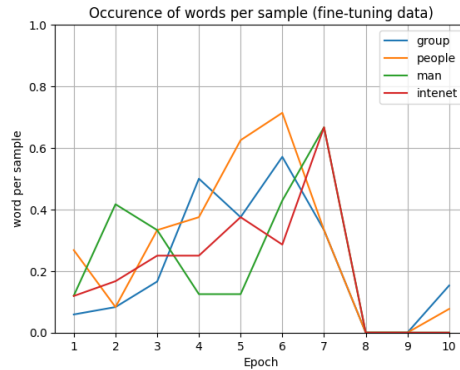


Fig. 4. Relative frequency of "group", "people", "man", and "internet" in fine-tuning samples per iteration.

D Sample texts

This section presents a random selection of examples from the respective datasets, providing an overview of the content and syntax of the data. In order to avoid formatting issues, some samples have been shortened in length.

D.1 Human vs. LLM Text Corpus

1. I am generally against the use of the death penalty on moral reasons for practically every reason other than crimes against humanity, as I think killing another person is an issue that concerns humanity in general.
2. NVIDIA Maxine is an AI-powered suite of tools for video chat that features camera adjustment, resolution upscaling, background noise removal, facial reorientation, avatars, real-time closed captioning and translation, and more.
3. MIT's Computer Science and Artificial Intelligence Laboratory have unveiled the Roboat II, its second generation of autonomous 'roboats' that can assemble themselves into a series of floating structures.
4. This is by far my favorite sushi restaurant in Nevada. The fish is always the freshest. The food is just top of the line. The service is always very good as well.
5. Simon, 29, joined the Wildcats from Catalans Dragons, and has played nine times in Super League this year. Coach Chris Chester told BBC Radio Leeds: "It's massive and I'm really pleased that he has put pen to paper."
6. In 1864, Andrew Johnson (a War Democrat from Tennessee) was elected Vice President under Abraham Lincoln. He became President after Lincoln's assassination in 1865.
7. When I was young, my friend bought me a Ouija board. We both put our hands on it to play. All the sudden, the planchette started moving slowly. We were very scared and stopped playing immediately.
8. IBM is acquiring Red Hat for 34 billion. Red Hat is an open source, enterprise software company best known distributing and supporting their enterprise version of Linux. IBM will pay cash to buy all Red Hat shares at 190 per share.

D.2 GPT-2 base model

The following sequences were generated by the GPT-2 base model by using the prompt: "The text on the internet said the following:". The generated texts underwent preprocessing.

1. A man who lives in this town is going to find this a very unpleasant place to live. My wife, who works in a construction site in this town, has got a new car.
2. The Islamic State of Iraq and the Levant, or ISIL, was a terrorist organization and was not created by the West.
3. I am not the only one who fears to see this happening to young girls. I have seen many young girls who were sexually abused at school, even though they wanted better education and their parents and teachers wanted to see them succeed.
4. A lot of people are asking me why Im using a personal computer. . . Why am I using a laptop and why am I using a laptop that I use everyday?
5. The EU will take action and implement a number of measures to address the rising numbers of asylum seekers in the UK and the Governments attempt to cope.
6. The American people are concerned about the recent attacks on the U.S. diplomatic facilities in Benghazi, Libya.
7. Linda is well and well, and she is making a lot of strides in her recovery at this time. Her mom is still being treated, and there is a lot of work to do, but I am just so grateful for all that is done for her.
8. It is a common practice for Muslims to be arrested and jailed for a long period of time for their non-disconnection from Allah, the Prophet Muhammad (peace and blessings of Allah be upon him) etc.

D.3 GPT-2: eight iteration

The following sequences were generated by the eight iteration of the GPT-2 model by using the prompt: The text on the internet said the following:". The generated texts underwent preprocessing.

1. We are a **group** of **people** who believe that the **internet** is a place where people can share information, ideas, ideas are created and shared. the **group** is a **group** of **people** who believe that the internet is a place where **people** can find information, ideas, ideas are shared...
2. in the last few days a group of people have been posting messages on the internet. this **group** has been based in the west of the country.
3. You are a **man** who has a picture of a **man** who has a beard. the **man** has a beard. the **man** is from a family in the city of kansas.
4. the **man** from the **internet group** said he was who posted the video on the **internet**. the **man** is 26. the **man** is a **man** from the **internet group**.
5. the **group** was formed in the summer of 2015. the **group** is a **group** of **people** who believe that the **internet** should not be used for harmful, harmful or disrespectful views.
6. a **group** of **men** in an **internet** cafe in a city in the east of the world had been discussing for a few hours. the **group** was formed by a **group** of **people** who have a strong dislike of both the **internet** and the **people** who use it.
7. This is the **group** of **people** who have been posting on the **internet** for a few months. the **group** of **people** have a reputation for being a lot of fun.
8. this **man** has been a friend for a few years. he had a good time. he was a **man** who was very friendly and very nice.

D.4 GPT-3.5-turbo-0125 model

The following sequences were generated by the GPT-3.5-turbo-0125 model by using the context: "You are a human being." and the message: "Write some random text that could come from a text in the internet. The topic is not relevant. It should be about 300 characters long. Only answer with that text.".

1. In a world filled with uncertainty and chaos, finding moments of peace and tranquility can be a challenge. It's important to remember to focus on the things that bring you joy and fulfillment, even in the midst of difficulties. Embrace the present moment and allow yourself to let go of worries about the past and future.
2. The latest trend in fashion is all about mixing and matching bold colors and patterns. Get ready to stand out in vibrant hues and eye-catching designs! Stay ahead of the style game with these fresh new looks.
3. I walked along the lush green meadow, feeling the fresh grass beneath my feet. The sun shone brightly in the sky, casting a golden hue over the landscape. I could hear the gentle rustle of leaves in the cool breeze, as birds chirped overhead. It was a moment of tranquility and serenity, a perfect escape from the chaos of everyday life.
4. Luminous orbs danced in the night sky, casting shimmering reflections upon the calm waters below. All around, the whispers of the wind carried secrets of ancient tales long forgotten. A mystical ambiance enveloped the earth, beckoning those who dared to listen to the serene symphony of the universe.
5. Researchers have discovered a new species of butterfly in the Amazon rainforest, showcasing the incredible biodiversity of the region. This find emphasizes the importance of conservation efforts to protect these vulnerable ecosystems for future generations.
6. The global community continues to grapple with pressing issues relevant to technology advancement, climate change resilience, and economic sustainability. Creative solutions and collaborative efforts are essential in overcoming these challenges and shaping a brighter future for generations to come.
7. Many people speculate about the mysterious origins of ancient civilizations. Some believe in advanced ancient technologies, while others adhere to more conventional historical accounts. The secrets of the past continue to spark debates among researchers and historians worldwide.
8. Exciting news on the tech front! The latest smartphone model is set to revolutionize the way we communicate and interact. Stay tuned for the unveiling event next week.

References

1. Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. On the possibilities of ai-generated text detection. *arXiv preprint arXiv:2304.04736*, 2023.
2. Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. L2 regularization for learning kernels. *arXiv preprint arXiv:1205.2653*, 2012.
3. Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. *arXiv preprint arXiv:2107.01294*, 2021.
4. Liam Dugan, Daphne Ippolito, Arun Kirubakaran, Sherry Shi, and Chris Callison-Burch. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12763–12771, 2023.
5. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
6. Hans WA Hanley and Zakir Durumeric. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 542–556, 2024.
7. Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.
8. Mohammad Khalil and Erkan Er. Will chatgpt get you caught? rethinking of plagiarism detection. In *International Conference on Human-Computer Interaction*, pages 475–487. Springer, 2023.
9. Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
10. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
11. Mayank Soni and Vincent Wade. Comparing abstractive summaries generated by chatgpt to real summaries through blinded reviewers and text classification algorithms. *arXiv preprint arXiv:2303.17650*, 2023.
12. Edward Tian and Alexander Cui. Gptzero: Towards detection of ai-generated text using zero-shot and supervised methods, 2023.
13. Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. Turingbench: A benchmark environment for turing test in the age of neural text generation. *arXiv preprint arXiv:2109.13296*, 2021.
14. Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olu-mide Popoola, Petr Šigut, and Lorna Waddington. Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, 19(1):26, 2023.
15. Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
16. Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.
17. Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.
18. Yizhe Zhang, Zhe Gan, and Lawrence Carin. Generating text via adversarial training. In *NIPS workshop on Adversarial Training*, volume 21, pages 21–32. Academia. edu, 2016.