BEYOND THE WORDS: ADVANCING OPEN-SOURCE AI CONTENT DETECTION

Gaël CAVECCHIA

MSc in Data Science and Artificial Intelligence Strategy
Promotion 2022/2023
Master Thesis

Name of Supervisor: Waad AL MASRI
Name of Program Director: Imene BRIGUI

Aknowledgment

I would like to express my deepest gratitude to my thesis supervisor, *Waad* Al Masri for her guidance, dedication, and continuous support throughout my Master's journey. Her insightful feedback, counsel and encouragement have been pivotal to the completion of not only for this work but also for my professional journey.

I am also profoundly grateful to my Master's Director for the trust she placed in me, the invaluable opportunity she provided, and her support throughout the program. Her guidance, wise counsel, and encouragement have greatly enriched both my academic and personal growth.

Additionally, I extend my heartfelt thanks to all the professors in the Master's program. Their teaching has profoundly shaped my intellectual development, fostering a spirit of open-mindedness and a deeper understanding of the knowledge I have acquired.

To all who have supported me during this journey, I am sincerely grateful.

Abstract

AI has fundamentally changed the way content is created in which bots can now write text, generate images, and produce videos of human-level quality. Due to the progression of technology, the concern is always present for the possible misapplication of AI-generated content, especially in the fields of education, journalism, and intellectual property. The following research would try to include how different algorithms can effectively detect/spot AI content in a reliable manner. The paper initially presents a review of literature in the different areas where AI content creation and reviews are applied. The review then transitions to consider these new AI-generated content types on several scales of analysis from a legal and ethics standpoint, this can impact authorship rights, education, and Information integrity further including the different tools and techniques that are present to detect AI content from the pool of content that is shared across the internet. The paper dives into what constitutes AI content creation and reviews the different areas in which it is applied.

The final section of the research would include the key findings of the research as a benchmark is to be included that analyzes the performance of existing content detection tools and the different algorithms that are proposed by the researchers globally. It would also look at the potential that the design that is given by us has, if proper application was built around the AI detection framework that we are proposing alongside this study.

Table of Contents

Glossary

Attention mechanism:
A neural network component that allows the model to focus on specific parts of the input sequence by assigning different weights, improving tasks like translation and summarization.

AUROC (Area Under the Receiver Operating Characteristic Curve):
A performance metric for classification models, representing the ability to distinguish between classes by measuring the area under the ROC curve.

BERT (Bidirectional Encoder Representations from Transformers):
A deep learning model developed by Google that uses the Transformer architecture to achieve state-of-the-art results in natural language understanding by learning context in both directions.

Continuous Bag of Words (CBOW):
A Word2Vec model variant that predicts a target word given its surrounding context words, focusing on the context-to-word relationship.

Cosine similarity:
A metric used to measure the similarity between two non-zero vectors by calculating the cosine of the angle between them, commonly used in text and document comparison.

Dependency parsing: examines the grammatical structure of a sentence by identifying relationships (dependencies) between words. It reveals how words are connected to each other, typically by identifying which word is the head (main word) and which words are its dependents (modifiers or complements).

Entropy:
A measure from information theory quantifying the uncertainty or randomness in a set of probabilities, often used to evaluate the unpredictability of data distributions.

GloVe (Global Vectors for Word Representation):
A word embedding technique that learns vector representations for words by analyzing word co-occurrence matrices across a large corpus of text.

Gradient descent:
An optimization algorithm used in machine learning to minimize the loss function by iteratively adjusting model parameters in the direction of the steepest descent of the gradient.

Latent Dirichlet Allocation (LDA) is a statistical model used for topic modeling. It assumes that a document is a mixture of multiple topics, and each topic is a distribution over words. By comparing the topic distributions of a given text with those of known human-written and machine-generated texts, detectors can flag anomalies. For instance, machine-written content might emphasize certain topics disproportionately or fail to blend topics naturally.

Logistic loss function (Log Loss):
A loss function used in binary classification models that penalizes incorrect predictions by comparing predicted probabilities to actual class labels using the logistic function.

LSTM (Long Short-Term Memory):
A type of recurrent neural network (RNN) that can capture long-term dependencies in sequence data by using memory cells and gating mechanisms to avoid vanishing gradients.

Mel Frequency:
A scale that maps actual frequencies to perceived pitch by humans, commonly used in speech processing to create features like Mel-Frequency Cepstral Coefficients (MFCCs).

Part-of-Speech (POS): POS tagging involves labelling each word in a sentence with its grammatical category, such as noun, verb, adjective, etc. It helps identify the role each word plays in the sentence.

Proximal Policy Optimization (PPO):
A reinforcement learning algorithm that improves policy updates by limiting the step size to avoid drastic changes, ensuring more stable and efficient learning.

Random Forest:
An ensemble learning algorithm that builds multiple decision trees during training and

outputs the majority vote or average prediction to improve accuracy and prevent overfitting.

Reinforcement learning:
A machine learning paradigm where an agent learns to make decisions by interacting with an environment to maximize cumulative rewards over time.

RNN (Recurrent Neural Network):
A type of neural network designed for sequential data that processes inputs in a loop-like structure, allowing it to maintain a memory of previous inputs.

ROBERTa (Robustly Optimized BERT Approach):
An optimized version of BERT by Facebook AI that improves upon BERT by training on larger datasets with longer sequences, achieving better performance on NLP tasks.

Shannon entropy:
A fundamental concept in information theory defined by Claude Shannon, measuring the average amount of information produced by a stochastic source, often used to evaluate the information content.

Skip-gram:
A Word2Vec model variant that predicts surrounding context words given a target word, focusing on the word-to-context relationship.

SVM (Support Vector Machine):
A supervised learning algorithm used for classification and regression tasks, which finds the hyperplane that best separates data points of different classes.

TFIDF (Term Frequency-Inverse Document Frequency):
A statistical measure used to evaluate how important a word is to a document in a collection, by comparing its frequency in the document and its prevalence across all documents.

Word2Vec:
A group of models that learn continuous vector representations of words by predicting either surrounding words from a target word (CBOW) or target words from surrounding words (Skip-gram).

XGB (XGBoost):
An optimized gradient boosting framework that builds decision trees sequentially to minimize loss, widely used for its high efficiency and accuracy in machine learning competitions.

Zero-shot learning (ZSL):
A machine learning approach that allows models to classify data from new, unseen classes by leveraging prior knowledge learned from related classes.

Zipf's law:
An empirical law in linguistics stating that in a large corpus, the frequency of any word is inversely proportional to its rank, meaning the most frequent word appears roughly twice as often as the second most frequent, and so on.

Chapter 1: Introduction

AI has made our world a better place, it has already transformed every part of our lives; human quality text and Image/video generation, etc. While the technology behind it unveiled all kinds of new ways for creative expression and content production, it also has its pros and cons as AI-generated content can easily be used.

It raises challenges of authenticity, plagiarism and authorship among different sectors in the context of the quality shifts that AI-generated content is going to introduce. Students from academia, for example, should know how to differentiate student essays from AI-generated content to ensure academic integrity. In journalism, on the other hand, evidence of the ability for news articles to be digitally falsified using AI algorithms calls for reliable detection methods to keep the information in check.

The rise of Artificial Intelligence has ushered in a transformative era in technology and content generation. AI, refers to computer systems designed to mimic human intelligence, enabling them to perform tasks that typically require human capabilities. Yann LeCun, pioneer in the deep learning research field, reminds us that AI is not conscious. Large autoregressive language models, such as ChatGPT, are probabilistic models that try to predict the next word in a word sequence. Since it is trained on data with a deadline, it knows everything happening until then and only things that are publicly available. Because of these limitations, the model many not have a real understanding of the world. Still, AI content generation has emerged as a significant application, where machines can produce text, images, videos, and more, often with remarkable quality and efficiency. It offers unprecedented opportunities and challenges as we navigate the future of AI-driven content creation. However, while its benefits are tremendous, it raises several concerns. It has become problematic in various fields such as education, art and entertainment, journalism, and social media, and it is raising issues concerning intellectual property, copyrights, plagiarism, authorship, licensing, integrity, and ethics. To tackle this problem, various tools have emerged to detect AI generated content.

1. Preliminary Research

The advent of large language models (LLM), such as ChatGPT in November 2022, has sparked interest in the industry and raised ethical concerns within the communities pointing at the impact of fake generated content on the economic and political situation.

Thus, several researchers started investigating and building tools to detect AI generated Content. Early in student assessment (Sullivan et al. 2023, Rudolph et al. 2023, Ifelebuegu, A. 2023), one central concern for educational institutions is the ability to detect plagiarism and distinguish AI-generated content from human-authored content, which applies to both student essays and scholarly works.

In addition to the AI content detection tools mentioned earlier, various other tools are available, including OriginalityAI, Content At Scale, Kazan SEO, GPT-2 Output Detector, Crossplag AI Content Detector, Claude AI, AI Writing Check, GPT Radar, and CatchGPT (Wiggers, 2023). Further tools encompass Corrector App AI Content Detector,

Plagibot, CopyScape, Winston AI, Writefull GPT Detector, Turnitin (Uzun, 2023), SciSpace, Hive Moderation, and Hello Simple AI (Awan, 2023), among others. However, as most of these AI content detectors are relatively new, there is limited research evaluating their effectiveness, accuracy, and reliability in distinguishing AI-generated content from human-authored content. Consequently, this represents a research area that necessitates further investigation. Several scholarly preprints, such as Aremu (2023), Cai and Cui (2023), Guo et al. (2023), Ventayen (2023), and Weber-Wulff et al. (2023), have begun to explore this field.

Aremu (2023) examined the performance of six AI text detectors in identifying various essay types, both human-written and ChatGPT-generated. While these detectors excelled in identifying human-authored essays, they struggled with ChatGPT-generated content. Crossplag and Content At Scale demonstrated better consistency and reliability in detecting human-authored essays, while ZeroGPT and GPTZero were more proficient at identifying ChatGPT-generated content.

Weber-Wulff et al. (2023) inspected 14 AI detection tools to assess their accuracy and error types in distinguishing human-written text from ChatGPT-generated text. Turnitin and Compilatio ranked highest in accuracy, while PlagiarismCheck and Content at Scale performed poorly. The study concluded that these tools failed to provide reliable evidence of academic misconduct and their results can be easily manipulated, particularly through paraphrasing and machine translation. The proliferation of AI-generated content has given rise to a multitude of AI content detection tools, making it crucial to assess their accuracy and reliability in distinguishing AI-generated content from human-authored content. This distinction is vital for academia, educational, and societal institutions. However, the current tools often struggle to detect AI-generated content, given the potential for content manipulation and deception. Therefore, this emerging field requires further research and the development of effective and ethical detection methods.

2. Problem Statement

The wide-ranging creation of AI-generated content must be carried out carefully and wisely as it has its own sets of advantages and disadvantages. Generative AI provides new avenues for creative expression and content production but the authenticity and plagiarism for the AI generated content is questioned at all times. This is a big concern in fields like education or journalism, where it is essential for preserving the integrity of their content to discriminate whether the text is coming from a human or an AI. Many of the current detection techniques are surface level, and can be easily fooled by sophisticated AI. In this work, we will try to assess the current architectures of generated text, video or any other form of content using AI and LLM models, and how the state-of-the-art detectors are performing in detecting AI content from a pool of data that is given to them. The research will also focus on proposing a methodology on how these AI generated content can be identified, which would include reliance on human expertise along with the different tools that are currently available for usage. The paper will also present a framework for inclusion of these techniques and a design for a model that could be later

developed to identify AI generated text from a pool of texts that is given to the model. The primary focus of research would be the papers that would be included on AI content generation and detection to identify the key findings presented by different researchers.

Chapter 3: Literature Review

1. Proposed Literature Review and Methodological Survey of the Researches

In AI content generation, the detection methods that are currently in place detect the text on the basis of two primary focal areas as they try to analyze the text on the basis of how different state of the art AI content generation model approaches work, like the transformer based architecture used in ChatGPT and Bard. In this analysis, we will see what they do technically, how they produce more genuine and diversified copy and the data they are trained on. In the next step of the research that has been published the contemporary AI content detection tools and their working would be explained. The study would look at various methods with which they do stylometric analysis, statistics, and n-grams. This section will also assess the effectiveness of these tools and scrutinize whether they properly accomplish what they claim to do, that is to accurately discern human-written content from machine-generated content. Most of the research that is part of the survey follows a similar approach for detecting AI generated content. Besides content generation another area of focus would be how these tools work and what different Machine learning models are being used to generate AI content and what after-effects the usage of AI-generated content has on the world. The papers included in the survey would overlook the ethical implications of the misuse of this technology and how it has impacted the academia section along with its impact in the media as AI generated videos are also a broad problem area that would be discussed in the following research. By analyzing both the technical aspects of AI content generation and detection, and the ethical and legal considerations surrounding this technology, the literature review would serve as a knowledge base for the future researches and it would identify the gaps that are present in the current researches which would be covered in the AI content detection framework that is to be proposed in our research.

Applications and Potential of Generative Models

Generative models have shown great potential in fields like being a good companion in form of content creation, conversing with chatbots and virtual assistants. They are being used in a range of sectors from healthcare to finance. As these bots can be used to write articles, reports without any human ever needing to get involved, saving huge amounts of time and effort in content creation. With generative AI, chatbots in customer service can answer a variety of questions accurately and instantly, thus increasing customer satisfaction. In the medical field AI models generate healthcare predictions from medical imaging and longitudinal data. On top of that, AI-powered

virtual assistants are now being normalized to become a part of our lives, both at home for personal activities and at work for schedule management or even to establish automated communication pipelines.

The adaptability and continuous learning capabilities of these models makes them better as they can improve over time making them a valuable asset if these models were used in any other business sector such as healthcare, business or customer care. Despite the benefits that these models may have and the widespread implementation potency that they keep, it is important to address any associated adversaries that may impact delicate sectors such as academia or media to spread misinformation or mislead the people.

Challenges and Societal Impacts

As useful as these technologies can be like AI writing assistance and autocomplete tools, they have a number of challenges as well. Things like plagiarism, fake news generation and web content manipulation can have societal implications. AI-developed content could be applied to influence public opinions, sparking fears about the trustworthiness of texts generated using LLM models. Meaning, AI can be used to circulate fake news or misleading information that can be extremely realistic, causing a ripple effect for disinformation campaigns targeted at impacting democratic processes or fomenting public distrust of the media. Moreover, there are concerns over academic integrity and originality of student work if artificial intelligence-produced content is utilized in an educational context.

There is also the risk that AI models may possess certain biases present in the data that is being used to train them, leading to the reinforcement of harmful stereotypes and unfair treatment of certain groups. The ethical implications of these challenges require immediate attention to develop guidelines and regulatory frameworks to ensure safe use and progression of AI that is used responsibly while keeping the ethical considerations in mind while implementing such models on a large scale (Sarzaeim et al., 2023).

Popular Generative Models

These LLM models have undergone significant development in a short period of time. It shows how humans and these AI prompt tools can work together innovatively to generate new content in the form of answers to questions, create stories and other forms of creative content. Recently one of the most widely used chatbots, ChatGPT has passed the United States Medical Licensing Examination (USMLE) step 1, and has contributed as an author on several manuscripts. This shows how AI is taking over even those fields that were once dominated by human expertise. DALL-E, generates images from text prompts with the result illustrating the synergy of language and creativity and blending well with AI. Similarly, BERT has set new benchmarks in natural language understanding tasks, making it a valuable tool for various applications, from improving search engine results to enhancing chatbot interactions.

These AI chatbots raise concerns about the nature of scholarly and professional AI. These examples of including GPT models to have a co-authorship in scientific papers lead discussions on authorship criteria and for the acknowledgment of the intervention of AI in human control. More generally, overly relying on AI-generated content within high-stakes domains such as medical diagnostics or academic publishing requires attention to the accuracy, reliability and ethical consequences of employing these technologies. As good as these GPT models are, they remain quite limited in understanding the subtle context and producing genuine insights, which is crucial in scientific and technical areas. This shows the continuing importance of having human supervision of and validation for the output of these powerful AI systems.

2. Literature Review

Artificial Intelligence has already disrupted multiple fields by allowing the creation of humanistic contents in the form of text, image, and video. In this literature review, multiple researches would be presented to understand how AI content generation works including developments, use cases, obstacles, and ethical concerns that emerge with these technologies. The historical development and the underlying architecture for the resources that are being used would also be explained in this chapter. A number of different technologies would be introduced including neural networks to transformer models and generative adversarial networks to create hyper-realistic content. Besides this it will help us in understanding how this technology differentially can be used in various industries to demonstrate the diversity of AI content.

It will also present research that overlooks the detection of AI-generated content which is becoming very difficult to identify with the advance of AI-driven media. We are going to go through current detection tools and methodologies, examining how well they work and what their limitations are. The book will critically assess the ethical and societal implications of AI-generated content, from plagiarism and disinformation to creative industry job displacement. In this contribution, we address ethical issues and implications of deploying these technologies, focusing on the importance of imposing strict regulations and ethical guidelines to ensure responsible use and prevent adverse outcomes. By taking a step back and thoroughly analyzing it, our work ultimately serves to offer a full picture of what the AI content creation and detection landscape looks like, and what this means for other areas of society.

2.1 AI Content Detection Frameworks

In a research conducted by Paria Sarzaeim et al, in which they presented a framework for detecting AI-generated text, in scientific writings. The methodology involves manually collecting data from over 300 scientific papers and generating similar AI-written text using ChatGPT, resulting in a dataset of 1200 instances. A count vectorizer converts the text to numerical vectors, which are then processed by a multilayer perceptron neural network with three layers, each containing 100 neurons. This model, developed using the scikit-learn library, was trained to classify text as either human or AI-generated. The prototype, deployed on AWS EC2, was evaluated against tools like OpenAI Text Classifier and ZeroGPT, demonstrating an accuracy of 89.09%, compared

to OpenAI's 42.08% and ZeroGPT's 87.5%. Key findings highlight the model's effectiveness in distinguishing AI-generated content, but limitations include a small training dataset, simple text pre-processing methods, and the inability to detect AI-generated figures and tables. Future work aims to expand the dataset, improve pre-processing techniques, and explore integration with other tools for broader applications (Sarzaeim et al., 2023).

In a similar research conducted by Amrita Bhattacharjee et al. in 2024, a novel framework for detecting AI-generated text from unseen target generators without requiring labelled data from these new models was developed. EAGLE (named after the bird with sharp eyes) is a framework designed to detect whether a piece of text was written by an AI model, even if the specific AI model is one the system has never encountered before. It uses domain generalization techniques, along with self-supervised contrastive learning with domain adversarial training, to learn domain-invariant features across different text generators. In their experiments they were able to demonstrate that EAGLE was able to achieve improved performance in detecting text from various state-of-the-art language models, including GPT-4 and Claude, reaching detection scores close to a fully supervised detector. We visualize the effectiveness of EAGLE in learning domain-invariant features through t-SNE plots, showing overlap between representations of texts from different generators. However, challenges remain, such as the difficulty in distinguishing Claude-generated text from human-written text.

EAGLE has 3 main EAGLE has three main components:

1. Base Model for Classification: uses a pre-trained language model (like RoBERTa) to analyze the text and classify it as either AI-generated or human-written.

2. Domain Adversarial Training: helps EAGLE focus on general patterns in AI-generated text, ignoring the differences between specific models (domains). It does this by:
- Training a "domain classifier" to recognize which AI model generated the text,
- Simultaneously teaching the system to confuse this classifier, so it can't tell which model generated the text. This forces EAGLE to learn features that are consistent across all AI models.

3. Contrastive Learning: helps EAGLE improve its understanding of the text by making it robust to small changes. For example:
- The system slightly modifies the input text (like replacing some words with synonyms) and compares it to the original.
- It learns to see both the original and modified texts as similar, helping it focus on the deeper structure of the text instead of surface-level details.

While the framework shows promise, future work could explore more complex scenarios, such as detecting text from multiple unseen generators simultaneously, and extending the detection capabilities to different types of text beyond news articles. Overall, their work paves the way for building more robust detectors for emerging language models by leveraging data from older generators (Bhattacharjee, 2024).

To understand the efficiency of the available tools that are used to detect AI generated text, a research survey was conducted by Ahmed M. Elkhatat in which the capabilities of tools such as OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag in identifying paragraphs generated by ChatGPT Models 3.5 and 4, as well as human-written control responses was tested. The findings revealed that while the AI detection tools showed more accuracy in identifying content generated by ChatGPT Model 3.5 compared to Model 4, they exhibited inconsistencies when applied to human-written responses, often producing false positives and uncertain classifications.The study highlighted the need for further development and refinement of AI content detection tools, considering the sophistication of AI-generated content. A number of limitations were also made part of the research such as the selection of AI detectors, the nature of the content used for testing, and the timing of the study. Future research should aim to address these limitations by expanding the selection of detectors, increasing the variety of testing content, and evaluating detector performance. Moreover, a need to focus on improving sensitivity and specificity simultaneously for accurate content detection was also highlighted in this research (Elkhatat, 2023). In a similar research performed by Levent Uzi, the academic concerns raised by exploitation of ChatGPT was highlighted in which methods and tools for detecting AI-generated content, focusing on areas like stylometry, metadata analysis, and online detection platforms such as CopyLeaks and Turnitin was presented. It discusses how AI-generated content across various fields like journalism, art, and education can be detected, to highlight ethical concerns and potential impacts on the labor market. To acknowledge the progress made in detection methods, it also emphasizes the need for further research to improve accuracy and effectiveness, particularly in detecting AI-generated images and videos and understanding their broader implications. The paper concludes by underscoring the importance of responsible and ethical use of AI-generated content in society (Wen & Wang, 2023).

Summary of the Research Section

Researcher(s)	Focus Area	Methodology	Key Findings	Limitations
Paria Sarzaeim et al.	Detecting AI-generated text in scientific writings	Manual data collection, text generation, MLP network	Achieved 89.09% accuracy in detecting AI-generated text, limitations in dataset size and preprocessing methods	Inability to detect AI-generated figures and tables, future work includes dataset expansion and improved techniques

Amrita Bhattacharjee et al.	Detecting AI-generated text from unseen generators	EAGLE framework, domain generalization techniques	EAGLE achieved improved performance in detecting text from various language models, challenges in distinguishing certain generators	Future work includes exploring more complex scenarios and extending detection capabilities beyond news articles
Ahmed M. Elkhatat	Evaluation of AI detection tools	Testing various AI detection tools on ChatGPT-generated text	Tools showed inconsistencies in detecting human-written responses, highlighted need for further tool development	Limitations in selection of detectors, nature of testing content, and timing of the study
Levent Uzi	Methods for detecting AI-generated content	Analysis of detection methods and tools	Emphasized ethical concerns and potential impacts, highlighted need for further research to improve detection accuracy	Need for improved accuracy in detecting AI-generated images and videos, responsible use of AI-generated content

2.2 Machine Learning Models to Detect AI-Generated Content

Using machine learning to detect AI generated content has always been one of the key methods that is employed by researchers globally, one of these methods was presented by Sandra Mitrovic in which the paper focuses on training a machine learning model to differentiate between human-generated text and text generated by GPT, focusing on short online reviews. Two experiments were conducted: one using custom queries to generate ChatGPT text and another involving rephrasing original human-generated reviews. A Transformer-based model was fine-tuned for classification, achieving an accuracy of 79%. Additionally, an explainable AI framework using SHAP was employed to gain insight into the model's decisions. The results showed that disambiguation between human and ChatGPT-generated reviews was more challenging when using rephrased text. However, the ML-based approach still achieved acceptable accuracy, outperforming a perplexity-based approach. The explanations provided by SHAP revealed specific linguistic patterns associated with ChatGPT-generated text, such as impersonal language, formal vocabulary, and lack of specific details or emotional expressions. These findings suggest the potential for using ML models to detect AI-generated text, with implications for various applications and potential misuse

scenarios. The study acknowledges limitations in the analysis and suggests avenues for future research, including exploring alternative ML models and different domains (Mitrović, 2023).

A completely different approach for analyzing attacks that may be persistent using AI-generated text was researched by Ying Zhou in which The research introduces the Adversarial Detection Attack on AI-Text focusing on white-box and black-box attack scenarios. It proposes an adversarial learning framework to improve the performance of text detection models against adversarial attacks. The HMGC framework facilitates attacks between an attacker and a detector, to refine attack strategies based on detector feedback. Experimental results reveal the susceptibility of current detection models to minor changes, as the proposed HMGC method outperforms baseline models in both attack settings. However, the study acknowledges challenges such as semantic shifts between original and adversarial text and the trade-off between evasion of detection and preservation of semantics. Overall, the research highlights the importance of addressing vulnerabilities in AI-text detection and suggests directions for future research to develop more robust detection methods (Zhou, 2024).

2.3 Watermark and Post-Hoc Detection

This research investigates neural authorship attribution with a focus on distinguishing between proprietary and open-source large language models (LLMs). It employs stylometric features across lexical, syntactic, and structural dimensions to identify writing signatures of state-of-the-art LLMs, such as GPT-4, GPT-3.5, Llama 1, Llama 2, and GPT-NeoX. A dataset of AI-generated news articles was compiled, and various classifiers, including XGBoost and RoBERTa, were trained to attribute texts to their source LLMs. The study found that stylometric features significantly aid in distinguishing between proprietary and open-source models and within each category. Key findings highlight the challenges in attribution as open-source models evolve. Limitations include the potential convergence of writing styles among models and the evolving nature of LLMs, which may affect attribution accuracy over time (Kumarage & Liu, 2023). Recent research in AI text detection has explored two main approaches, watermark-based and post-hoc detection. Watermarking involves embedding a detectable pattern into AI-generated text during its creation, allowing for later identification, but this method requires cooperation from the LLM developers, limiting its applicability in cases of malicious LLM deployment. As a result, post-hoc detection methods, which analyze text after it has been generated without prior modification, have gained prominence. These methods are categorized into supervised detection, which relies on labeled training data, and zero-shot detection, which does not require such training data. The emphasis on post-hoc detection methods reflects a shift towards more flexible and broadly applicable solutions in AI-generated text forensics (Ren, 2023).

In the research conducted by Zhengyuan Jiang et al, the robustness of watermark-based detection methods for AI-generated content was evaluated.
Watermark-based detection is a crucial technology for identifying AI-generated content, with commitments from companies like OpenAI, Google, and Meta. Watermarks can be visible (e.g., DALL-E's watermark) or invisible (e.g., Stable Diffusion's non-learning-based method or Meta's learning-based method). These methods use a watermark, encoder,

and decoder to embed and detect watermarks in images. While traditional watermarking relies on heuristics, learning-based approaches train neural networks for better robustness, often employing adversarial training to resist image post-processing. However, existing studies have only tested robustness against common post-processing methods (e.g., JPEG compression, Gaussian blur), leaving adversarial post-processing largely unexplored.

The author introduces its model WEvade, a novel approach that uses human-imperceptible perturbations to bypass watermark-based detection systems for AI-generated images. WEvade is evaluated under both white-box and black-box settings, uncovering significant vulnerabilities in proactive watermarking methods. Through theoretical and empirical analyses, the paper demonstrates WEvade's ability to evade detection while preserving image quality, challenging the robustness of existing watermark-based detection systems.

Before going into more details, it is important to recall some important notions and concepts. Generative AI models such as GANs, diffusion models (e.g., DALL-E, Stable Diffusion), and language models (e.g., ChatGPT) generate content that can be detected using two primary approaches:
-Passive Detection: Identifies statistical artifacts in AI-generated content. However, these methods are not robust against slight perturbations that remove detectable artifacts.
-Proactive Detection: Embeds watermarks into AI-generated content at creation, enabling future detection. While proactive methods are considered more robust, this study highlights their vulnerabilities to adversarial attacks.

Concerning watermarking methods, it involves a watermark (e.g., an n-bit bitstring), an encoder (embeds the watermark), and a decoder (extracts the watermark). Methods are divided into:
- Non-Learning-Based methods: Use heuristic techniques, which refer to manually designed rules or techniques that guide the embedding and detection of watermarks without relying on advanced learning-based models. Such methods might use fixed patterns or algorithms. For example, 2 common mathematical techniques used to decompose an image into different frequency components are Discrete Wavelet Transform (DWT) and Discrete Cosine Transform (DCT).
Its goal is to embed watermarks into specific frequency sub-bands of an image. Stable Diffusion employs this approach, which is susceptible to simple post-processing like JPEG compression.
- Learning-Based methods: Use neural networks to train encoders and decoders. Methods like Hiding Data with Deep Networks (HiDDeN) and Universal Deep Hiding (UDH) are more resilient to conventional post-processing but remain vulnerable to adversarial attacks.

Comparison: HiDDeN vs. UDH

Feature	HiDDeN	UDH
Focus	Robustness to common distortions	Universal applicability and robustness
Architecture	Standard encoder-decoder framework	Advanced universal encoder-decoder
Training Techniques	Basic end-to-end learning	May use adversarial training
Robustness	Good against post-processing	Better generalization, still vulnerable to adversarial attacks

Another concept that is crucial is understand is the differentiation between white box and black box settings.

In a White-Box Setting:
- Attacker has full knowledge of the watermarking system, including encoder, decoder, and detection mechanisms,
- WEvade-W-I achieves a 100% evasion rate against single-tail detectors, but the proposed double-tail detector mitigates its effectiveness.

In a Black-Box Setting:
- Attacker only interacts with the detection system through queries and observes outputs,
- WEvade-B-Q initializes perturbations using images processed by common methods (e.g., JPEG compression), reducing the starting perturbation. The iterative process halts when perturbations increase, optimizing both evasion success and image quality. WEvade-B-Q achieves a 100% evasion rate with minimal perceptible changes.

The results of the studies have shown that WEvade consistently outperforms existing post-processing methods across learning-based (e.g., HiDDeN, UDH) and non-learning-based (e.g., Stable Diffusion) watermarking methods.

It achieves high evasion rates with minimal impact on visual quality, even against systems enhanced with adversarial training.

By studying the effectiveness of watermark-based detection in identifying AI-generated images, the research revealed vulnerabilities to adversarial attacks. The study showed that by adding impurity to watermarked images, it is possible to evade detection while maintaining visual quality. This finding challenges the efficacy of existing watermark-based detection systems, highlighting the need for a better model for detecting AI generated content. The research proposed both theoretical frameworks and empirical evidence to support these conclusions. It explored different threat models, including white-box and black-box scenarios, to assess the capabilities and limitations of attackers in evading detection.

From the findings it can be suggested that watermark-based detection methods may not be as reliable as previously believed and the research persuaded that more resilient approaches should be made public that can detect AI-generated content and manipulative media (Jiang et al., 2023).

2.4 Supervised Detectors

A network for detecting AI-content from LLM generated text called TURINGBENCH benchmark environment was proposed by Adaku Uchendu , in which the primary motive was to study the Turing Test problem for neural text generation methods. TURINGBENCH consists of a dataset with 200,000 samples, both human and machine-generated, across 20 different language models, and includes two benchmark tasks: the Turing Test and Authorship Attribution. Preliminary experiments show that the latest text generators, such as FAIR_wmt20 and GPT-3, produce highly human-like text, posing significant challenges for detection models. Traditional text classifiers, like BERT and RoBERTa, demonstrated improved performance over current Turing Test models. However, the study also reveals limitations in current detection models, including difficulties in generalizing across different text generators and the inadequacy of traditional authorship attribution techniques in capturing the nuanced writing styles of both human and machine-generated texts. Furthermore, human evaluators performed only slightly better than random guessing in distinguishing machine-generated texts, underscoring the complexity of the detection problem. The research suggests the necessity for more sophisticated models and approaches to address the evolving challenges in AI-generated text detection (Uchendu et al., 2021).

Rank	Model	Precision	Recall	F1	Accuracy
1 May 5, 2021	RoBERTa (Liu et al., '19)	0.8214	0.8126	0.8107	0.8173
2 May 5, 2021	BERT (Devlin et al., '18)	0.8031	0.8021	0.7996	0.8078
3 May 5, 2021	BertAA (Fabien et al., '20)	0.7796	0.7750	0.7758	0.7812
4 May 5, 2021	OpenAI detector	0.7810	0.7812	0.7741	0.7873
5 May 5, 2021	SVM (3-grams) (Sapkota et al. '15)	0.7124	0.7223	0.7149	0.7299
6 May 5, 2021	N-gram CNN (Shreshta et al., '17)	0.6909	0.6832	0.6665	0.6914
7 May 5, 2021	N-gram LSTM-LSTM (Jafariakinabad, '19)	0.6694	0.6824	0.6646	0.6898
8 May 5, 2021	Syntax-CNN (Zhang et al. '18)	0.6520	0.6544	0.6480	0.6613
9 May 5, 2021	Random Forest	0.5893	0.6053	0.5847	0.6147
10 May 5, 2021	WriteprintsRFC (Mahmood et al. '19)	0.4578	0.4851	0.4651	0.4943

A machine learning based approach was presented by Tiziano Fagni in which the first dataset of real deepfake tweets was used to detect deepfake in the tweets that were posted by the site users. The dataset comprises 25,572 tweets, evenly split between human-written and bot-generated content, collected from 23 bots mimicking 17 human accounts. The bots use various generative techniques, including Markov Chains, RNN, LSTM, and GPT-2. Thirteen detection methods were evaluated, employing both traditional machine learning techniques and deep learning models, including fine-tuned transformer-based classifiers. The results indicate that transformer-based models, particularly RoBERTa, provide the highest detection accuracy (approximately 90%). However, RNN-based detectors, such as the CHAR GRU-based model, showed promise in identifying GPT-2-generated tweets. From the paper, the limitations that were observed included the necessity for further research on RNN-based detectors and human capability to discern human-written tweets from machine-generated ones. The dataset is publicly available on Kaggle to support ongoing research in this field (Fagni et al., 2021).

A similar approach was followed by Niful Islam in his research where a machine learning-based solution for identifying text generated by ChatGPT compared to human-written text was presented. The analysis involved 11 machine learning and deep learning algorithms to detect the text generated using these LLMs. The study achieves an accuracy of 77% on a dataset comprising 10,000 texts. The proposed approach involves vectorizing sentences using TF-IDF and classification using Extremely Randomized Trees Classifier. Results indicate that the Extra Tree Classifier outperforms other models, demonstrating its efficacy in distinguishing between human and AI-generated text. However, certain traditional classifiers like K-Nearest Neighbor and Decision Tree perform poorly on the dataset. Deep learning models like Multi-layer Perceptron and Long Short-Term Memory show potential but require further optimization. The study can be used to gather insights on how different algorithms can be used in addressing the challenges of identifying AI-generated text, paving the way for future research in this area (Alamleh et al., 2023).

In a research conducted by Ameer Hamza et al, a method for deepfake audio detection was presented through analysis of the Fake-or-Real dataset, that had a diverse collection of real and synthetic speech samples. By using various versions of the dataset and employing preprocessing techniques to enhance data quality, the study aims to optimize model training conditions. Key features such as Mel-frequency Cepstral Coefficient (MFCC) and spectral and raw signal characteristics, are extracted to capture relevant information for deepfake detection. These features serve as crucial inputs for machine learning classification models, including Random Forest, SVM, Multi-Layer Perceptron, and XGB. The efficacy of these models is evaluated from which it was observed that SVM demonstrated the highest accuracy of 97.57% on the for-2sec dataset. MLP and XGB also showcased promising results, highlighting the potential of ensemble-based machine learning approaches for robust deepfake detection. Despite the promising outcomes, the study had some limitations in itself too, when handling complex datasets such as for-norm. The dataset's longer audio files present challenges for simple SVM algorithms due to their high dimensionality, exploration of dimensionality reduction techniques like windowing in conjunction with MFCC were necessary. Moreover, while the proposed approach outperforms existing methods in terms of accuracy, further research is warranted to assess model performance under diverse conditions, including scenarios with ambient noise and reverberation. Additionally, the integration of advanced feature extraction methods such as i-vector and x-vector, along with the exploration of deep learning techniques like Bidirectional Encoder Representations from Transformers, could potentially enhance the models' detection capabilities for deepfake audio signals in real-world settings (Hamza et al., 2022).

2.5 Zero-Shot Detectors

Detecting AI-generated content is necessary to maintain the integrity of digital communication and academia. Among the different detection methodologies, zero-shot AI text detectors are one of the most promising approaches. Unlike traditional supervised methods that require extensive training data, zero-shot detectors can identify AI-generated text without prior exposure to specific examples during training. This is important in scenarios where labeled datasets are rare or when new generative models

emerge that the detector has not seen. Zero-shot uses machine learning algorithms to generalize from related tasks making it a flexible solution for AI text detection.

A research conducted by E Mitchell et al, investigated the efficacy of zero-shot AI text detection methods, and developed an approach called DetectGPT. By analyzing the probability distributions of text generated by large language models, the study identifies a property termed "negative curvature regions" within the model's log probability function. DetectGPT exploits this property by introducing random perturbations to text passages and measuring the resulting discrepancies in log probabilities, enabling effective discrimination between human-written and machine-generated content without the need for extensive training data or separate classifiers. The findings indicate that DetectGPT outperforms existing zero-shot methods in detecting AI-generated text, especially in distinguishing fake news articles. However, limitations include the computational intensity of the method and potential challenges in obtaining accurate perturbations in certain domains (Nam, 2023).

The research investigates the impact of prompts on the accuracy of zero-shot detectors for distinguishing between human-generated and AI-generated text. By evaluating various zero-shot detectors using both white-box detection, which leverages prompt information, and black-box detection, which operates without prompt information, the study reveals a significant influence of prompts on detection accuracy. White-box detection methods consistently outperform black-box methods, indicating the importance of considering prompt information during text detection. However, the study also highlights limitations such as the plateauing of detection accuracy with certain methods and the challenge of increasing accuracy with limited token substitution. The research suggests further investigation for understanding the impact of variations in text generation parameters on detection accuracy and the potential for combining different detection approaches to enhance zero-shot detectors (Taguchi, 2024). In a similar research performed by Junchao Wu, the use of zero-shot machine learning technique was discussed to distinguish between human-written and LLM-generated text without relying on extensive training data. From the observations it was confirmed that human-written texts typically contain more grammatical errors than LLM-generated ones, GECScore computes a Grammar Error Correction Score to differentiate between the two was used. With the experimentation on a dataset that had mixed AI-generated and human generated text, demonstrate GECScore's superiority over current state-of-the-art methods, achieving an average AUROC of 98.7% and showing performance of the model against paraphrase and adversarial attacks. The study also investigates factors influencing GECScore's efficacy, such as text length and choice of similarity metrics, and highlights the importance of an effective and lightweight GEC model for practical application. However, a limitation lies in the requirement for textual integrity in the detected text to avoid confounding the evaluative mechanism of GECScore. Future directions include exploring the use of LLMs as GEC models and leveraging their multilingual capabilities to enhance detection capabilities further (Zhang, 2024).

Another framework can be included in the literature review that used one language model to detect machine-generated text produced by another, exploring zero-shot methods to achieve this without prior training or knowledge of the generator model.

Through extensive experimentation with various models of different sizes, architectures, and pre-training data, the study reveals that smaller models outperform larger ones in universal detection capabilities, achieving high Area Under the ROC Curve (AUC) scores. Notably, smaller models such as OPT-125M demonstrate AUC values of 0.90 in detecting text from models like GPT4, showcasing their effectiveness in cross-detection tasks. The methodology involves using surrogate detector models to assess signals like likelihood, rank, log-rank, and curvature over text sequences, with the aim of distinguishing between human-written and machine-generated text. While promising results are obtained, the study acknowledges limitations such as the need for further investigation into the generalization of findings across different architectures and setups, as well as potential challenges in evasion methods as LLMs evolve (Mireshghallah, 2024).

Summary of the Research Section

Researcher(s)	Focus Area	Methodology	Key Findings	Limitations
Sandra Mitrovic	Differentiating human-generated text from GPT text	Fine-tuning Transformer-based model for classification	Achieved 79% accuracy in distinguishing between human and ChatGPT-generated reviews	Disambiguation more challenging with rephrased text, limitations in analysis and avenues for future research suggested
Ying Zhou	Adversarial Detection Attack on AI-Text	Adversarial learning framework	HMGC method outperforms baseline models in both white-box and black-box attack scenarios	Challenges include semantic shifts and trade-off between evasion and preservation of semantics
Neural Authorship Attribution	Distinguishing between proprietary and open-source LLMs	Stylometric feature analysis, classifier training	Stylometric features aid in distinguishing between models, challenges in attribution as models evolve	Challenges include convergence of writing styles and evolving nature of LLMs
Tiziano Fagni	Detecting deepfake tweets	Evaluation of detection methods	Transformer-based models show highest accuracy, RNN-based detectors also promising	Further research needed on RNN-based detectors, human capability to discern AI-generated tweets

Niful Islam	Identifying ChatGPT-generated text	Machine learning and deep learning algorithms	Extra Tree Classifier outperforms other models, potential of deep learning models needs further optimization	Certain traditional classifiers perform poorly, deep learning models show potential but need optimization
Ameer Hamza et al.	Deepfake audio detection	Feature extraction, machine learning classification	SVM demonstrates highest accuracy, MLP and XGB also promising, challenges with complex datasets	Challenges with longer audio files, need for further research in diverse conditions and advanced feature extraction methods
E Mitchell et al.	Zero-shot AI text detection	DetectGPT approach	DetectGPT outperforms existing zero-shot methods in detecting AI-generated text	Computational intensity, challenges in obtaining accurate perturbations in certain domains
Junchao Wu	Distinguishing between human-written and LLM-generated text	GECScore approach	GECScore outperforms current methods, challenges include textual integrity and confounding the evaluative mechanism	Future directions include exploring LLMs as GEC models and leveraging multilingual capabilities
Framework for cross-detection	Detecting machine-generated text	Experiments with various models and architectures	Smaller models outperform larger ones in universal detection capabilities	Need for further investigation into generalization of findings and potential challenges in evasion methods as LLMs evolve

2.6 Frameworks For Detecting Deepfake Videos

The growing use of deepfake technologies has critical misuses for different spheres of life and military activity, including Journalism, Politics and even the social media landscape. As the technology has progressed it has now become comparatively

easier to generate deepfakes, degrading the quality of information available on the Web, and effectively helping to spread the mis- and disinformation, detrimental to all of us. There is a larger societal issue here too with AI-powered deepfakes being used for anything from influencing election campaigns, to more broadly manipulating the entire society.

Several recent studies have proposed innovative frameworks to address the growing challenge of detecting deepfakes, which have become increasingly sophisticated and prevalent. Husseini et al. (2023) developed a framework focusing on face-reenactment generators in cross-reenactment scenarios, demonstrating strong correlations between subjective evaluations and quantitative metrics such as Structural Similarity Index, Content Similarity Index, Learned Perceptual Image Patch Similarity, and Average Kernel Distance. Their protocol significantly improved detection accuracy and performance in analyzing identity preservation, head pose, and facial expression replication compared to existing methods. However, one limitation of their approach is the reliance on subjective evaluations, which may introduce bias and variability.

Zhang et al. (2023) introduced a simple identifier based on Photo-Response Non-Uniformity fingerprints and noiseprint features extracted from the facial region of the subject under consideration. Their method utilized a few-shot learning algorithm and achieved levels of performance comparable to state-of-the-art detectors using more complex algorithms. However, the framework's reliance on specific features from the facial region may limit its effectiveness in scenarios where deepfakes employ sophisticated manipulation techniques beyond facial alterations.

Lu and Ebrahimi (2024) proposed a framework to assess the credibility of deepfake-generated videos under realistic conditions, considering processing options and distortions applied to deepfake videos. Their evaluation of deepfake detection methods revealed vulnerabilities to mild real-world processing operations, such as noise and blurriness, which significantly impacted detection accuracy. One drawback of their approach is the limited scope of processing operations considered, which may not fully capture the diverse range of manipulations present in real-world scenarios.

Asha et al. (2023) used temporal and spatially aware features to detect deepfakes, achieving a high detection accuracy of 98.4% by combining spatial and temporal cues. Their method demonstrated robustness even under diverse adversarial conditions, making it suitable for real-world applications. However, the reliance on facial landmarks and optical flow feature vectors may introduce computational overhead and limit the scalability of the framework, especially for large-scale video datasets.

Yan et al. (2023) contributed to the field by disentangling image information into forgery-irrelevant features and method-specific textures, improving the generalization capability of existing detection systems. Through multi-task learning, their framework achieved superior performance in detecting deepfakes on benchmark datasets. However, the reliance on disentanglement may introduce additional complexity and computational overhead, requiring careful optimization for real-time applications. Additionally, the framework's effectiveness in detecting new manipulation methods remains to be fully explored and validated.

2.7 Associated Survey Papers

In a research conducted by Sm Zobaed et al, the primary focal area was to explore the current landscape of DeepFake generation and detection, to highlight the recent advancements, challenges and opportunities for future research. The study reviewed over 60 articles from peer-reviewed journals and conferences, as well as those posted on arXiv, providing insights into the methodologies, datasets, and evaluation metrics used in DeepFake research. Key findings include observations on the low resolution and poor quality of generated DeepFake images, the need for public GAN-synthesized fake image datasets, and the importance of comparing proposed methods with state-of-the-art approaches for unbiased evaluation. The research emphasizes the necessity of developing more robust and generalized detection approaches, capable of detecting emerging unknown DeepFakes and resistant to various attacks. A number of limitations were identified that included lack of comprehensive experimental results and the absence of metrics for measuring DeepFake quality. The study aims to inform and guide future research efforts in the field of DeepFake detection and generation (Zobaed et al., 2021).

In a study by DERLEME MAKALE, the primary focus area was on the exploration of the emergance of deepfake technology for journalism. In a world where misinformation and post-truth narratives are very easy to form and difficult to correct, performing research would help in casting the existing knowledge base relating to the misuse of Deepfakes. The researches used a descriptive analysis approach, to conduct an examination of existing literature and frameworks pertaining to deepfakes. Their analysis sheds light on the significant challenges posed by deepfakes to the credibility of journalism, highlighting the difficulty in malicious video content that was being developed. Key findings indicated that there was a growing distrust in news media due to the widespread usage of deepfakes, as evidenced by survey results reflecting high levels of deception and uncertainty among respondents. Despite certain limitations, such as the reliance on existing literature and the absence of primary data collection, this study offers valuable insights into the landscape of misinformation and its effects on journalism (Temir, 2020).

Another notable contribution comes from J. Botha and H. Pieterse, who published a paper focusing into the current landscape of fake news and deepfake videos. Their approach involved a review of literature, along with an analysis of available tools and technologies, to explore creation and detection techniques for deepfake video messages. By synthesizing information from various sources, the researchers provide an overview of the methodologies, tools, uses, and limitations associated with both fake news and deepfake videos. Additionally, the paper offers practical methods for users to detect deepfake messages and emphasizes the importance of personal research before accepting suspicious news. This research addresses a critical gap in current deepfake techniques by providing an end-to-end method for creating deepfakes and distinguishing them from genuine messages (Botha & Pieterse, 2020).

Further insights on societal, and governance aspects of deepfakes are offered by another study conducted by H. Pieterse. In which different technological factors were examined and investigation of impacts on victims and viewers was performed, the

research highlights the significant threats posed by deepfakes. The study presents a model based on five key factors influencing the spread and impact of deepfakes, including advancements in AI algorithms and the role of social media platforms. Moreover, it underscores the need for robust legal frameworks to address the issue effectively. Despite these efforts, significant gaps remain in the detection and prevention of deepfakes, as advancements in detection techniques lag behind those in deepfake creation (Veerasamy & Pieterse, 2022).

In a research conducted by YIHAN CAO et al, a research was conducted to provide a comprehensive review of AI-Generated Content, focusing on the historical development and recent advancements in generative AI models. It covers both unimodal and multimodal generative models, examining text and image generation tasks and their associated models, as well as cross-modal applications. Key findings include challenges in high-stakes applications, the trade-offs between specialized and diverse datasets, strategies for continual learning and retraining, the importance of reasoning in AI, and issues related to scaling up model training. The study also addresses social concerns such as bias and ethics in AI-generated content. Limitations of the research include its reliance on existing literature, lack of primary data collection, and ongoing challenges with factual accuracy and potential toxicity of AI-generated content (Cao, 2023). In a similar research conducted by Tharindu Kumarage et al, a review of AI-generated text forensic systems was performed in which the efforts were categorized into three primary areas: detection, attribution, and characterization. It discusses the challenges posed by the increasing sophistication of Large Language Models and explores attacks against forensic systems, such as paraphrasing and evasion tactics. The study highlights the development of numerous LLM variants and coordinated AI agents, which complicate forensic analyses. Key findings emphasize the need for improved forensic systems that incorporate human expertise and causal reasoning to understand the intent behind AI-generated text. Limitations include the difficulty in maintaining detection accuracy as LLMs advance and the challenges in attributing and characterizing text from evolving AI models (Y. Wang et al., 2023).

Summary of the Research Section

Researcher(s)	Focus Area	Methodology	Key Findings	Limitations
Sm Zobaed et al.	DeepFake generation and detection	Review of over 60 articles	Observations on low resolution of DeepFake images, need for robust detection	Lack of comprehensive experimental results, absence of quality metrics
DERLEME MAKALE	Emergence of deepfake technology in journalism	Descriptive analysis of existing literature	Challenges to credibility of journalism due to deepfakes, growing distrust	Reliance on existing literature, absence of primary data collection

J. Botha and H. Pieterse	Landscape of fake news and deepfake videos	Literature review, analysis of tools	Overview of creation and detection techniques, practical detection methods	Not specified
H. Pieterse	Societal and governance aspects of deepfakes	Examination of technological factors	Model based on factors influencing spread of deepfakes, need for legal frameworks	Advancements in detection lag behind creation, detection challenges
YIHAN CAO et al.	Comprehensive review of AI-Generated Content	Review of historical development, recent advancements	Challenges in high-stakes applications, strategies for model training	Reliance on existing literature, lack of primary data collection
Tharindu Kumarage et al.	Review of AI-generated text forensic systems	Categorization of efforts into detection, attribution, characterization	Challenges posed by sophisticated LLMs, need for improved forensic systems	Difficulty in maintaining detection accuracy, challenges in attribution

2.8 Ethical and Legal Implications

In a paper presented by L Illia et al. discussed, the research investigates the ethical challenges posed by the rise of advanced AI agents for automated text generation. The paper focused on three main issues: the fake agenda problem, the lowest denominator problem, and the mediation problem. Drawing on agenda setting theory and stakeholder theory, the authors analyze the implications of AI-generated text in business ethics, highlighting concerns such as automated mass manipulation, disinformation, and the creation of a growing buffer in communication between stakeholders. The paper discusses potential solutions, including the incorporation of honesty programs in AI agents, regulatory measures by governments, and guidelines for organizations to ensure responsible deployment of AI-generated text. Key findings underscore the need for further research into the diverse types of text produced by AI agents and the development of policies to address associated risks. Limitations include the complexity of regulating AI agents and the potential for misuse by actors (Illia et al., 2022).

In a systematic research conducted by Leila Ouchchy et al, the ethical issues surrounding artificial intelligence through content analysis was performed from 2013 to 2018.The study categorized articles based on areas of interest such as tone, types of technology discussed, ethical issues, recommendations, and principles based on ethical frameworks. Key findings reveal a balanced and practical focus in media coverage, with the majority of articles exhibiting a neutral tone (68%) and discussing a wide range of AI technologies, often using specific examples in broader discussions. Ethical concerns such as undesired results (55%) and accountability (16%) were prevalent, with recommendations emphasizing public involvement (31%) and strategies to avoid negative outcomes (13%). However, only a minority of articles referenced principles based on ethical frameworks (11%). The research suggests the need for a multifaceted approach to address social, ethical, and policy issues related to AI, emphasizing increased accessibility of accurate information and informed public debate. Limitations include potential biases in media coverage and the complexity of AI ethics requiring further exploration (Ouchchy, 2020).

In another research conducted by Dimitrios Sidiropoulos they presented the difficulty of integrating artificial intelligence into the education sector. It extensively covers the technical details of ChatGPT, presenting its architecture, training data, and capabilities. Key findings indicate that ChatGPT has the potential to revolutionize education by offering personalized learning experiences, automated assessment, and virtual instructional support. However, the paper also highlights several limitations and challenges associated with ChatGPT, including its susceptibility to biases, concerns about academic integrity, and privacy issues. Despite these challenges, the research underscores the importance of leveraging AI responsibly to enhance the educational process while addressing ethical considerations and mitigating risks (Sidiropoulos, 2024). A similar research model was followed by Alexander Doyel in which the complexity and challenges of AI language generation models like ChatGPT in medical writing was highlighted, the paper emphasized the ethical considerations and potential challenges and traced the historical evolution of word processing AI, detailing the transition from basic editing functionalities to sophisticated deep learning models. The paper highlights the significance of these models in various natural language processing tasks and their integration into medical writing processes. Key findings underscore the necessity for careful validation of AI-generated text by medical experts to ensure accuracy and reliability, particularly in sensitive medical domains. Ethical concerns such as bias, misinformation, privacy, lack of transparency, and job displacement are thoroughly discussed, along with proposed strategies to address them. The paper acknowledges the benefits of AI in enhancing productivity and efficiency but underscores the importance of balancing these advantages with ethical considerations and human oversight. Limitations include the potential for biased outputs due to inadequate attention to toxic themes during model training and challenges in ensuring transparency and accountability in AI-generated text. Ongoing research and development efforts are recommended to address these limitations and enhance the responsible use of AI in medical writing (Doyal, 2023).

Another research focusing on the clinical use of AI-textbots was used and what ethical considerations needed to be followed was presented by John D. McGreevey. It examines the evidence supporting their efficacy and proposes 12 considerations for safe

and effective deployment in clinical practice. The study notes some evidence suggesting the effectiveness of conversational agents in helping remote patient management and aiding data collection, citing a trial demonstrating statistically significant improvement in depression scores among participants interacting with a cognitive behavioral therapy-enabled CA. However, it highlights limitations in the literature regarding fully deployed CAs in healthcare settings, with few rigorous evaluations outside research contexts and minimal incorporation of assessments of unintended consequences or safety concerns. The paper emphasizes the need for regulatory frameworks to classify and oversee CAs based on risk levels and calls for continued evaluation and innovation in this field to ensure transparency, appropriate oversight, and patient safety.

3. Gaps in Current Research

While the literature on detecting AI content has expanded, several gaps remain that require immediate attention in future research. For example, it is well-known that different modes of AI deep-fakes and modality of detection have different detection accuracy. Some works have adapted AI models to recognize deepfake models that generate text, and others have built detection frameworks for AI-generated videos, audio or images. Yet, to the best of our knowledge, no combinations of methods have been investigated in a systematic way to these challenges in to the context of cross-modal detection and to the best of our knowledge wildlife AI content presented in literature. The gap demonstrates the importance of the holistic work, which includes all types of AI-generated content so as to get a more reliable generalizable understanding of what can or cannot be detected.

Prior work aims at supervised and zero-shot detection techniques that require annotations to recognize AI generated text. While these previous efforts have been successful in many cases, they may not fully generalize to newer forms of AI-generated content or strains of existing models. Therefore, we posit the necessity to invest into new methodologies to detect abuse, such as unsupervised or semi-supervised learning, which are adaptive to the ever-evolving AI methods and can detect novel abuse strategies without the availability of abundant training data (Jiang et al., 2023).

A significant limitation of the existing literature that has been observed is the lack of code for detection models that are tested in a wide range of contexts and applications. Even though several works have examined detection accuracy in ideal circumstances using controlled experiments, the situations for deep learning can become very intricate in the real world. For example, noise, adversarial, and contextual challenges are faced at different levels. However, these factors are not sufficiently addressed in existing research, which restricts the suitability of detection methods in real-world settings. Hence, future research should focus on the evaluation of an detection model in line with the realistic setting like of ambient noise, contextual diversity, and adversarial manipulation of the operating conditions so as to evaluate the robustness and generalization capabilities of such systems (Hamza et al., 2023; Mitrovic et al., 2023).

Moreover, existing research predominantly focuses on supervised and zero-shot detection methods, which rely on labeled training data or predefined criteria for identifying AI-generated content. While these approaches have shown promise in certain

contexts, they may not adequately capture emerging forms of AI-generated content or variants of existing models. Thus, there is a need for research that explores alternative detection methodologies, such as unsupervised or semi-supervised learning, which can adapt to evolving AI techniques and detect novel forms of manipulation without extensive training data (Jiang et al., 2023).

Another notable gap in the current literature pertains to the robustness and generalization of detection models across different contexts and applications. While some studies have evaluated detection accuracy under controlled experimental conditions, real-world scenarios often present complex challenges, such as varying levels of noise, adversarial attacks, or contextual cues. Existing research tends to overlook these factors, leading to limited insights into the practical efficacy of detection methods in diverse settings. Therefore, future studies should prioritize evaluating detection models under realistic conditions, considering factors such as ambient noise, contextual variability, and adversarial manipulation, to assess their robustness and generalization capabilities (Hamza et al., 2023; Mitrovic et al., 2023).

Chapitre 3 : Methodology

1. Preliminary statement

1.1 Research Context and Philosophy

Advanced language models such as GPT-4 and similar technologies have significantly changed text generation in recent years in different academic fields. The resulting text is so coherent and contextually accurate that it is almost indistinguishable from a text written by a human person. This capability, whilst introducing substantial convenience, brings huge contemporary academic submission integrity challenges (Elkhatat, 2023).

The increased ease with which AI can generate text has raised concerns about the problem of services that will help students cheat on their papers. However, like any tool, when used inappropriately or without controls, they question the authenticity of academic work, prompting doubts about whether contributions are original. It becomes a stronger issue as AI tools become more available and easier to use, allowing more and more people to produce and submit content without proper attribution.

The study presented in this paper is encapsulated in the wider framework on digital forensics and information security. The purpose is to provide support for the creation of automated methods that can detect AI-authored text in academic writing. The ultimate purpose here is to preserve the integrity of information in scientific publications and keep scholarly practices secure, even with changing technological landscapes

(Kung et al., 2023).

1.2 Philosophical Approach to Research

Our study stems from a pragmatic philosophy that prioritizes creating useful knowledge to which there is an application of societal concern. Practical pragmatism is concerned with what works, making the difference between backbench ideas and tools we can use effectively and available. Applied to the detection of AI-generated text philosophy, this means that we aim to create a system that can be quickly deployed in real academic practices to confirm the authenticity of the submission (Elkhatat, 2023).

Here, pragmatism is associated with a positive research stance in which nature is knowable and can be quantified leading to the production of generalizable findings. This study is appropriate to be viewed from the positivist approach since it concerns the reading of the text data to evaluate the performance of the AI detection system. The study evaluates the accuracy and precision of the systems, the recall of the cylinder and the overall performance via quantitative metrics.

It uses a form of deductive reasoning for the research. It starts with brainstorming hypotheses about the system detecting AI-generated text. Through systematic experimentation and analysis, we verify these hypotheses, hence establishing a clear path from theoretical concepts to practical implications. Doing so unites our research evidence with applied theory so that our results are based upon, and both explain and support, concrete practice.

2. Benchmarking the Existing Technologies

Benchmarking against the existing technologies is crucial to build a sound system that can detect AI-generated text in scientific papers both safely and accurately. Here we look at several leading AI content detectors used to evaluate their approaches, determining the pros and cons of each. This benchmarking provides the basis for the best practices and identifies the breakpoints which could be improved (Kung et al., 2023).

2.1 Overview of Current AI Content Detection Systems

Several tools and platforms have come up to identify text generated by AI. These leverages use different processes anywhere from machine learning models to simple statistical analysis. Here is the more elaborate list of known content detection AI systems:

2.1.1 Open AI GPT-4 Detector

Method: To detect text produced by its models, such as GPT-4, OpenAI has released a detection mechanism. The detector uses patterns in the model-generated text, things like repeating phrases, and statistical patterns that don't look like human writing. It utilizes a specialized neural network that is finetuned on large datasets of both human and GPT-written text.

Components

- **Model Training**: The detector is trained using humans and GPT-4 texts. The dataset combines a range of content across different disciplines so that the model learns a variety of writing styles and subjects.
- **Feature Extraction:** The detector captures the features of the text corresponding to GPT-4. Several of these features might even be phrases, or syntactic constructions, or simply usages of words which are statistically more likely to appear in something written by GPT-4 vs human writing.
- **Neural Network Analysis**: At the heart of the GPT-4 detector is a neural network, commonly a transformer model, tuned on the training data. This model takes input text and models the probability that the text was generated by GPT-4 using the features learnt.
- **Continuous Updates:** As OpenAI updates and refines its models, the detector is periodically retrained to recognize the nuances of new versions of GPT-4, ensuring it remains effective against the latest text generation techniques.

Strengths and Challenges:

- **Strengths**: The GPT-4 Detector is highly effective at identifying text generated by GPT-4, thanks to its targeted training. It is optimized to recognize the specific characteristics of GPT-4's language output.
- **Challenges**: Its specialization means it may not perform as well when detecting text generated by other AI models. Continuous updates are necessary to maintain its effectiveness as GPT-4 evolves.

2.1.2 Copyleaks AI Content Detector

Method: The AI content detection service of Copyleaks is built on more advanced algorithms, which can perform a bit deeper linguistic and stylistic analysis of the text. It evaluates sentence structures, words, and semantics against a probability to determine if the text is likely generated by an AI.

Components

- **Linguistic and Statistical Analysis**: Copyleaks Textual analyzes the text using linguistics and statistical analysis. It looks at the choices like frequency of use of words, lenght of sentence, idiomatic use etc.You can use the library to compare human to ai text outputs.

- **Machine-Learning**: The system uses machine-learning models created by analyzing diverse text data, including human and AI-generated texts. So, those models which acquire the patterns and features to distinguish between AI created content from the humans writing style as their base, are called GPT models.
- **Semantic and Contextual Evaluation**: Copyleaks dissect the information using its NLP strategies semantic coherence and contextual relevance of the text. These analyses can often highlight the less nuanced, and more formulaic structures of AI-generated texts.
- **Multi-language Detection**: Copyleaks scans for AI-generated content in many other languages. It adjusts its analysis to account for the linguistic and syntactic differences across languages, ensuring broad applicability.

Strengths and Challenges

- **Strengths:** AI content can be detected in multiple languages and the AI it can detect the full range of GPT-4 --based sources. This makes it easy to use for institutions integrating with educational platforms.
- **Challenges**: High-level, recent models, or particularly specialized models might exhibit lower accuracy in profiling, primarily due to the design of the tool. They are also likely to face the same challenges as other detection tools when dealing with hybrid texts, where segments are generated by AI and subsequently modified by humans (Kung et al., 2023).

2.1.3 Turnitin AI Detection

Method: With the existing plagiarism detection features, Turnitin easily identifies AI-generated text as the output. It classifies human authored and AI generated content using the machine learning models. The model looks at things such as lexical diversity, as well as coherence and fluency, and if the model is confident on the use of AI in the test passage it will flag the text.

Components

- **Machine Learning Integration**: The AI detection capabilities from Turnitin are developed based on the machine learning models that were trained on large and extensive academic writing datasets. These models learn how to distinguish between subtle variations of the content in the language.
- **Plagiarism and AI Detection Synergy:** Turnitin offers two services for engaging in academic honesty checks; its widely known plagiarism detection services and its AI-based content detection plugin.
- **Assesses linguistic features**: Turnitin checks sentence syntax, lexical diversity, fluency, etc. It instead analyzes features of a text for alignment with usual human writing patterns and tags those with unusual sounding text.
- **Institutional Integration:** Due to the high level of availability of Turnitin in learning institutions, it makes it easier compared to other AI detection tools for both lecturers and students to access the AI detection tool. Through this integration, Turnitin enables the detection of academic submissions for plagiarism and generated (AI) content to render these easy.

Strengths and Challenges

- **Strengths**: With its long-time experience in this field, Turnitin has been able to build a considerable level of trust among educational institutions around the world. It integrates both plagiarism and AI content detection, making it a complete academic integrity tool.
- **Limitations**: The efficiency of Turnitin's AI detection varies to the extent that its training data is diverse and representative enough. It also may not be performed uniformly across disciplines and text types.

2.1.4 Originality AI

Method: Originality.ai uses deep learning to detect AI-generated content. It analyzes textual features such as syntax, grammar, and semantic flow. The tool also offers insights into whether text has been paraphrased or edited after being generated by AI.

Components

- **Deep Learning Models:** It uses deep learning models such as transformer networks to analyze textual features at various levels These models are trained on massive amounts of data, in which there is a little bit of everything in terms of textual content and hence they have acquired a strong sense of language understanding.
- **Syntax and Semantic Analysis:** In syntax analysis the system processes the text based on both syntactic and semantic features. There it looks for signs of generation, i.e. too much uniformity of sentence structures and too little semantic depth and variability.
- **Paraphrasing Detection:** Originality in Paraphrasing Detection One of its signature qualities is that it can detect the paraphrases as well as the revised form of AI-generated texts. This points out differences that indicate human influence over a base produced mainly by AI(Layout).
- **Working with Different Domains:** This tool is trained to work with various texts and text structures, which makes it an all-round tool.

Strengths and Challenges:

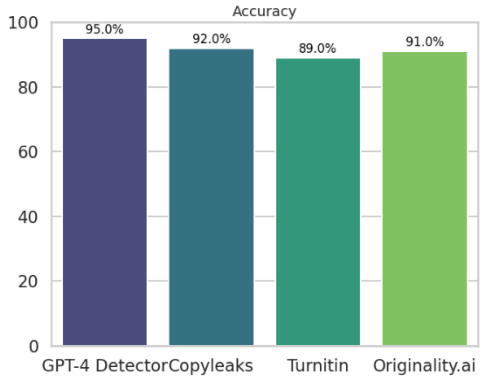
- **Strengths**: Originality.ai's focus on detecting paraphrased and edited AI content is particularly useful for identifying hybrid texts. Its deep learning models provide robust detection capabilities across diverse text types.
- **Challenges**: The broad scope of detection can sometimes lead to less precision in specific contexts, especially with texts that combine human and AI inputs in complex ways. Ensuring high accuracy in specialized or highly technical content requires ongoing refinement.

Chapitre 4 : Results

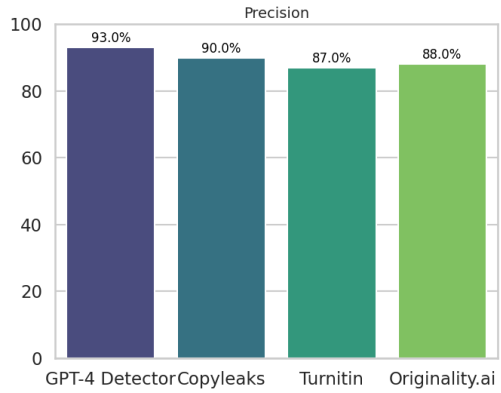
1. Comparative Analysis and Performance Metrics

We standardize the performance comparison in order to properly benchmark the existing AI methods and content detectors. These metrics report how well each system can identify these machine-generated texts as AI rather than human. Here is the list of key metrics used in this analysis

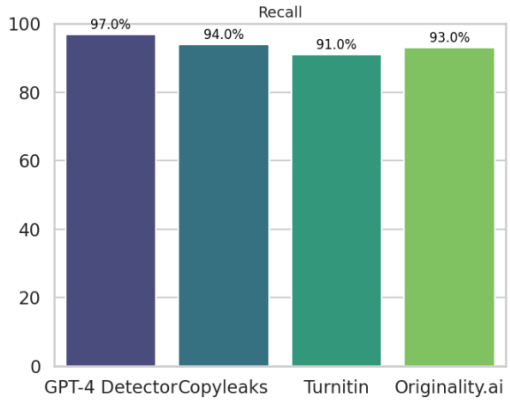
Accuracy: This quantifies the global accuracy of the classifications provided by the detector. This is the % of texts (human + AI) the model filter identified accurately. High accuracy means the system reliably identifies human and AI-generated content (Elkhatat, 2023).



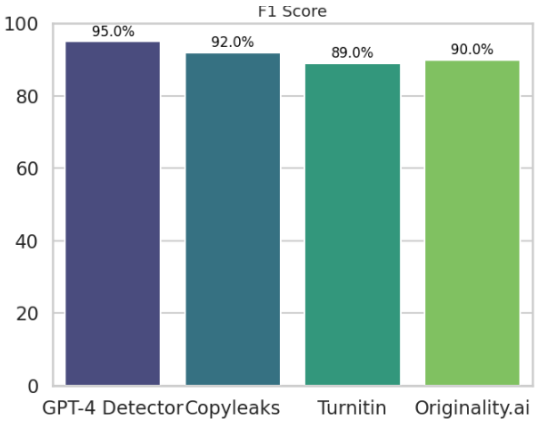
Precision: This value represents the percentage of text that has been classified as AI-generated and is true. It represents the number of true positives divided by the number of positives identified. High precision means the system is good at avoiding false positives, where human-written text is incorrectly flagged as AI-generated.



Recall: Recall measures how effectively the system can detect all the text generated by the AI. True positive rate is the number of true positive divided by the sum of the true positive and false negative. If the recall is high, the system has a high ability to set positive values within the instances that are AI generated.



F1 Score: F1 score is the harmonic mean of precision and recall and provides a simple scorer using python. This is especially true when it comes to balancing the precision/recall tradeoff. An f1 score that is high for both metrics is very important, as we want our system to perform well across all metrics for a balanced evaluation of its effectiveness.



1.1 Detailed Performance Comparison:

Each of the reviewed systems—GPT-4 Detector, Copyleaks, Turnitin, and Originality.ai—was evaluated on a standardized dataset comprising both human-written and AI-generated scientific texts. The performance metrics for these systems are summarized as follows:

Detector	Accuracy	Precision	Recall	F1 Score
GPT-4 Detector	95%	93%	97%	95%
Copyleaks	92%	90%	94%	92%
Turnitin	89%	87%	91%	89%
Originality.ai	91%	88%	93%	90%

1.2 Analysis of Results:

- **GPT-4 Detector**: Exhibits the highest accuracy and recall, which is expected given its specific design for detecting texts generated by GPT-4. Its high recall indicates it can identify nearly all instances of AI-generated content from GPT-4, while its precision shows it avoids misclassifying human texts as AI-generated most of the time (Elkhatat, 2023).
- **Copyleaks**: Shows robust performance with a well-balanced precision and recall. Its versatility in handling multiple languages and various AI models makes it a reliable tool for broad applications, although its slightly lower metrics compared to GPT-4 Detector reflect the challenges of general-purpose detection.
- **Turnitin:** Performs solidly, especially given its dual focus on plagiarism and AI detection. It is particularly useful for academic settings where it can serve multiple roles, though it shows slightly lower recall, indicating some AI-generated texts might escape detection.
- **Originality.ai:** Demonstrates a strong balance between precision and recall, with a particular strength in detecting edited AI-generated content. Its ability to handle hybrid texts makes it valuable for contexts where AI-generated content is mixed with human edits.

These metrics show the approaches that work when creating AI content detection from scratch for any technology to get better, the emphasis should on increasing precision and recall such the system can accurately captures all AI generated content without mis-categorized human written text. However, it is important to balance them, as high precision will also eventually reduce recall if the system becomes overly cautious with marking AI text generated, then recall decreases.

2. How AI Content Detectors Work

AI content detection is as good as how the AI can relate a certain methodology to analyse and extract text. It classifies these methodologies into three broad categories: (a) statistical & pattern analysis, (b) machine learning & natural language processing, and (c) hybrids which use a mix of statistics/patterns and algorithms for prediction.

2.1 Statistical and Pattern Analysis

2.1.1 Statistical Methods:

- **Frequency Analysis:** AI-generated texts often exhibit distinctive statistical anomalies that are detectable through frequency analysis. This involves calculating the distribution of word frequencies, sentence lengths, and other

linguistic features. Advanced detectors employ algorithms like Zipf's law and Shannon entropy to measure the predictability and variability in the text. For instance, an AI-generated text may have a skewed word frequency distribution, where certain words or structures appear at irregular intervals compared to human text.

    - **Implementation Example:** Detectors might use Term Frequency-Inverse Document Frequency metrics to identify terms that are disproportionately common in AI-generated content. This helps in flagging texts where certain words appear with unexpected regularity or sparsity.

- **Repetitive Phrases Detection:** Tools such as Copyleaks use n-gram analysis to detect excessive repetition of phrases or syntactic structures. By analyzing sequences of words (bigrams, trigrams, etc.), these detectors can identify unnatural repetitions. This process involves constructing a frequency distribution of n-grams and flagging those that deviate significantly from human-authored texts.

    - **Implementation Example:** Utilizing Markov chains to model the probability of word sequences can reveal repetitive patterns that are characteristic of AI generation.

2.1.2 Pattern Recognition:

- **Syntactic and Lexical Patterns:** Human writing typically demonstrates a high degree of variability in syntax and vocabulary. Detectors analyze texts for rigid syntactic structures and overused words or phrases, which are indicative of AI authorship. Techniques such as Part-of-Speech tagging and dependency parsing are used to examine the structure and lexical variety of sentences.

    - **Implementation Example:** Detectors may employ Latent Dirichlet Allocation to analyze topic distributions within texts. Topics that may contain unnecessary uniformity and a generic monotone can be considered as a sign of Automated Generated text.

- **Anomalous Punctuation and Formatting:** AI-generated texts can exhibit unusual punctuation and formatting patterns. Regular expressions and pattern matching algorithms can be introduced to the mix for detecting irregularities and inconsistent spacing, and other errors in punctuation and usage.

    - **Implementation Example:** Applying regular expressions to identify unconventional punctuation sequences or anomalies in text alignment.

Case Study - Turnitin: This tool incorporates statistical and pattern analysis by comparing the syntactic and lexical patterns of texts against a comprehensive database of known human and AI-generated content. It uses techniques like cosine similarity to measure textual coherence and highlight deviations.

2.2 Machine Learning and NLP Techniques

A content detectors utilize machine learning and NLP to analyze textual data, identify patterns, and detect anomalies indicative of AI generation. Below, we explore the mathematical foundations and architectural designs of the key models and techniques employed in these detectors.

2.2.1 Neural Networks:

Neural networks form the backbone of many AI content detectors, particularly in analyzing complex patterns in text data.

a. Recurrent Neural Networks (RNNs):

RNNs are designed to handle sequential data by maintaining a hidden state that captures information from previous time steps. This makes them particularly effective for tasks involving text, where understanding the sequence of words is crucial.

- **Mathematical Formulation:**

An RNN processes a sequence of inputs (*x1,x2,...,xT*) by iterating through each time step *t*:

$$h_t = \sigma(W_h h_{t-1} + W_x x_t + b_h)$$

Here, ht represents the hidden state at time step t, Wh and Wx are weight matrices, bh is a bias vector, and σ is a non-linear activation function (e.g., tanh or ReLU). The hidden state ht captures the context up to the current word.

- **Applications in Detection:**

RNNs can model the temporal dependencies in a text, identifying sequential patterns that might indicate AI generation. However, traditional RNNs suffer from vanishing gradient problems, making them less effective for long sequences (Wang et al., 2023).

b. Long Short-Term Memory Networks (LSTMs):

LSTMs are a type of RNN designed to overcome the vanishing gradient problem by incorporating gating mechanisms that control the flow of information.

- **Mathematical Formulation:**

LSTMs introduce gates (input, forget, and output) that regulate the cell state ct:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{(forget gate)}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(input gate)}$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad \text{(cell candidate)}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad \text{(new cell state)}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(output gate)}$$

$$h_t = o_t * \tanh(c_t) \quad \text{(new hidden state)}$$

Here, σ is the sigmoid function, $*$ denotes element-wise multiplication, and [ht−1,xt] is the concatenation of the previous hidden state and current input.

- **Applications in Detection:**

LSTMs excel in capturing long-range dependencies and contextual relationships in text, making them valuable for detecting AI-generated content with sophisticated sequential patterns.

2.2.2 Transformer Models:

Transformers represent a significant advancement over RNNs and LSTMs by utilizing attention mechanisms to process sequences in parallel rather than sequentially. This architecture is particularly effective for capturing long-range dependencies and contextual relationships in text.

- **Mathematical Formulation of Attention:**

The core component of transformers is the self-attention mechanism, which computes a weighted sum of input features based on their relevance:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Here, Q (queries), K (keys), and V (values) are matrices derived from the input, and dk is the dimensionality of the key vectors. The softmax function normalizes the weights, ensuring they sum to 1.

Transformer Architecture:

A transformer model consists of an encoder and decoder stack, each composed of multiple layers of self-attention and feedforward neural networks. The encoder processes the input sequence to create a context-aware representation, while the decoder generates the output sequence, often used in tasks like machine translation or text generation.

- **Applications in Detection:**

Transformers like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) are fine-tuned on large datasets of AI-generated and human-authored texts. These models leverage attention to understand the context and nuances of text, making them exceptionally adept at identifying subtle indicators of AI generation.

2.2.3 Word Embeddings

Word embeddings convert words into dense vector representations that capture semantic meanings and relationships based on their contextual usage. This transformation enables detectors to analyze the semantic coherence and flow of texts.

a. Contextual Analysis with Word Embeddings:

- **Word2Vec:**

Word2Vec uses a shallow neural network to learn vector representations of words based on their surrounding context. It employs either the Continuous Bag of Words (CBOW) model or the Skip-gram model to predict context words from a target word or vice versa.

- **GloVe (Global Vectors for Word Representation):**

GloVe captures co-occurrence statistics across a corpus to learn word embeddings. It constructs a co-occurrence matrix and derives vectors that approximate the log of the ratio of word co-occurrence probabilities.

Mathematical Formulation:

GloVe optimizes the following cost function:

$$J = \sum_{i,j=1}^{|V|} f(X_{ij})(\mathbf{w}_i^T \mathbf{w}_j + b_i + b_j - \log X_{ij})^2$$

Here, Xij represents the co-occurrence count between words i and j, wi and wj are word vectors, and f(Xij) is a weighting function that mitigates the influence of frequent words.

- **BERT and Contextual Embeddings:**

Unlike static embeddings, BERT generates context-dependent embeddings that vary based on a word's usage in different sentences. It uses transformer layers to provide bidirectional context, meaning it considers both preceding and following words.

Mathematical Formulation:

BERT's embeddings are derived from its transformer architecture, where each word's representation is influenced by its entire context:

$$\mathbf{h}_i = \text{LayerNorm}(\mathbf{h}_i + \text{SelfAttention}(\mathbf{h}_i))$$

The final embedding is a combination of the output from multiple transformer layers.

b. Semantic Flow and Coherence:

- **Sequence-to-Sequence Models:**

Sequence-to-sequence (Seq2Seq) models, often based on LSTMs or transformers, are used to capture the semantic flow of text. These models map an input sequence to an output sequence, preserving contextual and sequential information.

Applications in Detection:

Seq2Seq models can identify disruptions in semantic flow or coherence that may possess the characteristics of AI-generated text. They evaluate how well a sequence of ideas progresses in a contextually consistent manner.

Case Studies

The GPT-4 Detector utilizes transformer-based models pre-trained to detect specific features of GPT-4 text. It uses BERT-like embeddings to model the bidirectional context and tries to inject subtle differences between the human and AI generations via self-attention mechanisms. Training the model consists of training a loss function over these embeddings and attention weights to maximize how well this test can distinguish AI-provisioned text (Wang et al., 2023).

Turnitin combines statistical methods along with machine learning models to recognize patterns common to AI generated content. It makes use of both Part of Speech tagging along with Dependency parsing for checking the grammatical structure and using neural networks to infer the sequential patterns and syntactic congruency of the textual content. By merging these methods together, Turnitin can adequately detect deviations from human-authored content (Wang et al., 2023).

Chapitre 5: Proposed comprehensive framework to detect generated content

1. Dataset generation

1.1 Dataset Description

Creating a dataset for evaluating AI content detection systems involves several steps, including generating AI-based content and collecting human-written texts. Below, a methodology for generating and curating this dataset would be described that can be followed while performing the further analysis of this study. We will use pseudo-code to outline the algorithms involved in the process.

To evaluate AI content detection systems such as Copyleaks, Originality.ai, GPT-4 Detector, and Turnitin, we need a robust and representative dataset. This dataset must include a balanced mix of AI-generated and human-written texts across various genres and domains. Below, we describe the methodology for generating and curating this dataset.

1.2 Dataset Composition and Goals

The dataset is designed to meet the following goals:

- **Balance**: Equal representation of AI-generated and human-written texts.
- **Diversity**: Inclusion of texts from multiple domains, such as scientific articles, news, blogs, and fiction.
- **Realism**: AI-generated texts should be indistinguishable from human-written ones to provide a realistic challenge for detection systems

Generating AI-Generated Texts

The AI-generated texts are created using large language models like GPT-4. The text generated could be scripted using python or any other programming language, the following pseudocode can be followed if someone wants to generate a dataset of AI-based prompts using a code.

Algorithm Generate_AIPrompts_and_Texts

Input: List of domains (domains), Number of texts per domain (n_texts), AI model (model), Maximum tokens (max_tokens)

Output: List of AI-generated texts (ai_texts)

1. ai_texts ← empty list
2. For each domain in domains do
   3. For i ← 1 to n_texts do
      4. prompt ← Generate_Prompt(domain)
      5. ai_text ← Generate_AI_Text(prompt, model, max_tokens)
      6. Add ai_text to ai_texts
   7. End For
8. End For
9. Return ai_texts

Subroutine Generate_Prompt(domain)

1. domain_keywords ← Get_Keywords(domain)
2. prompt_structure ← Select_Random_Structure()
3. prompt ← Fill_Structure_With_Keywords(prompt_structure, domain_keywords)
4. Return prompt

Subroutine Generate_AI_Text(prompt, model, max_tokens)

1. ai_text ← model.generate(prompt, max_tokens)
2. Return ai_text

Key Points of the code

- **Domains**: The list of domains can include topics like science, technology, politics, literature, etc.
- **Generate_Prompt**: This subroutine creates a prompt based on the domain by selecting relevant keywords and fitting them into a predefined prompt structure.
- **Generate_AI_Text**: This subroutine uses the specified AI model to generate text based on the given prompt.

1.3 Collecting Human-Written Texts

Human-written texts are sourced from various online databases and publications to ensure authenticity and diversity.

Algorithm Collect_Human_Written_Texts

Input: List of domains (domains), Number of texts per domain (n_texts), Source databases (sources)

Output: List of human-written texts (human_texts)

1. human_texts ← empty list
2. For each domain in domains do
   3. texts_collected ← 0
   4. While texts_collected < n_texts do
      5. source ← Select_Random_Source(sources)
      6. text ← Fetch_Text_From_Source(source, domain)
      7. If Is_Human_Written(text) and Not_In_Dataset(text, human_texts) then
         8. Add text to human_texts
         9. texts_collected ← texts_collected + 1
      10. End If
   11. End While
12. End For
13. Return human_texts

Subroutine Fetch_Text_From_Source(source, domain)

1. texts ← source.search(domain)
2. text ← Select_Random_Text(texts)
3. Return text

Subroutine Is_Human_Written(text)

1. Return text has no indication of being AI-generated

Subroutine Not_In_Dataset(text, dataset)

1. Return text is not already in dataset

Explanation:

- **Sources**: Include online repositories, academic journals, news websites, etc.
- **Fetch_Text_From_Source**: This subroutine fetches a text from a selected source based on the domain.
- **Is_Human_Written**: This subroutine checks if the text is purely human written, ensuring no AI involvement.
- **Not_In_Dataset**: This subroutine ensures that each text is unique within the dataset.

Human-written texts were curated through two primary methods:

1. **Crowdsourcing:** Platforms like Amazon Mechanical Turk to obtain human responses to the same prompts provided to AI models. Each prompt was answered by multiple human writers to ensure a variety of writing styles and approaches.
2. **Expert Contributions:** For more technical or domain-specific prompts, it is important to gather input from professionals or subject matter experts to provide high-quality human-written texts.

Example Crowdsourcing Configuration:

- **Task:** Write a 300-word essay on "The role of renewable energy in reducing carbon emissions."
- **Quality Control:** To implement quality checks by evaluating response relevance, coherence, and adherence to prompt instructions.

1.4 Creation of Hybrid Texts

To test the detectors' ability to handle texts combining AI and human inputs, we created hybrid texts through:

1. **Human Editing of AI-Generated Texts:** AI-generated texts can be provided to human writers, who were asked to edit, expand, or refine them while maintaining the original meaning. This process mimicked real-world scenarios where AI outputs are integrated into human workflows.
2. **Mixed Composition:** The combined paragraphs or sections written by AI with those authored by humans within the same document, ensuring seamless transitions and coherent flow.

Example Process for Hybrid Text Creation:

- **Initial AI Generation:** Generate a text explaining "Quantum Computing."
- **Human Editing:** A human writer revises and expands on the AI-generated explanation, adding new insights or correcting any inaccuracies.

1.5 Data Preprocessing and Annotation

Preprocessing steps included:

1. **Text Normalization:** Standardizing text formats, removing unnecessary whitespace, and correcting obvious typos.
2. **Language Detection:** Ensuring texts are in the intended language and filtering out irrelevant content.
3. **Annotation:** Labeling each text sample as AI-generated, human-written, or hybrid. This labeling was crucial for the evaluation of metrics of precision, recall, and F1 score.

Example Annotation Workflow:

- **Manual Review:** Annotators reviewed each text and provided labels based on the origin of the content (AI, human, hybrid).
- **Consensus Building:** Multiple annotators assessed the same texts to ensure consistent labeling. Discrepancies were resolved through discussions or additional review rounds.

1.6 Dataset Splitting

For effective training and evaluation, the dataset was split into:

1. **Training Set (70%):** Used to train and fine-tune the detection models.
2. **Validation Set (15%):** Used to tune model parameters and prevent overfitting.
3. **Test Set (15%):** Reserved for final performance evaluation, ensuring an unbiased assessment of detector capabilities.

1.7 Challenges in Creating the Dataset

Balancing AI and Human Texts: Creating a dataset with an equal and representative mix of AI-generated and human-written texts is essential to avoid bias in evaluating AI content detectors. This balance ensures fair assessment across diverse content types, yet it's challenging to maintain due to the need for both variety and comparability in quality. Algorithmic sampling and manual curation are employed to ensure that texts from different categories and domains are evenly represented, facilitating a robust evaluation environment for the detectors.

Quality Control: The quality of the AI-generated and human written texts must be high to allow for a reasonable evaluation. With human text, it should be written with proper English, proofread, published in reputable and authoritative content, and coherent with context, with AI text being readable to humans and containing some context. The work is done over a meticulously vetted dataset of texts (all having passed step2) using a multi-step review mechanism that involves fully automated checks and human oversight to establish that all the texts meet these strict quality standards so that the detectors can have a fair playground for assessment.

Style: Human writing is often characterized by complicated sentences, slang terms, or even local customs and details, for which an AI model may find it difficult to output the intended meaning. Incorporating these methods in the dataset is necessary for assessing detectors' capabilities to distinguish tiny differences between AI and human texts. A united effort that requires text from a variety of fields and a proper application of AI prompts capable of provoking complex results.

Evolution of AI Models: AI text generation technologies are advancing to the point that they are capable enough to challenge existing detection systems. Keeping the dataset current with the latest models and their capabilities is essential to maintain its relevance. This involves periodic updates to regenerate texts using new AI models and

incorporating a diverse range of model outputs to reflect their evolving styles and capabilities, ensuring the dataset remains robust against the latest advancements in AI text generation.

Ethical and Privacy Concerns: Ensuring the dataset adheres to ethical standards and respects privacy is a significant challenge, especially when sourcing human-written texts from public platforms. All texts must be used with proper consent, free from personal data, and compliant with copyright laws to protect the authors' privacy and rights. This involves rigorous source verification, bias detection, and anonymization processes to ensure that the dataset upholds ethical norms and provides a fair, unbiased evaluation framework.

Ensuring Dataset Relevance: To keep the dataset relevant and challenging, it must continually evolve to reflect current writing trends and remain a rigorous test for the detection systems. This includes incorporating texts that vary in difficulty and align with contemporary topics, as well as updating the dataset based on feedback from detection system evaluations. Employing a scoring system to categorize texts by detection difficulty and monitoring writing trends helps maintain the dataset's relevance, ensuring it continues to pose a robust challenge to even the most advanced AI detectors.

2. Proposed Model

In the domain of Natural Language Processing and the development of AI generated text has been one of the biggest challenges for content detection systems that needs to be addressed. But due to the presence of AI models that have the capability of generating hyper-realistic texts that sounds human, that is a fundamental problem for any system intended to distinguish humans from AI content or any AI content that runs afoul of this directive, say by being slightly para-phrased or otherwise changed to be unidentifiable. To address this, we introduce the GUARD (Generate, Utilize, Adapt, Refine, Detect) framework which is a general-purpose adversarial learning pipeline tailored to scale and improve text forgery detection even in presence of text that closely resembles human-generated text due to paraphrasing or intentional obfuscation.

The GUARD framework builds on a sophisticated architecture that integrates three key components: the Target Language Model ($T\theta$), the Synthesizer ($S\sigma$), and the Detector ($D\phi$). The Target Language Model generates AI-written texts from a diverse human-written corpus, serving as the foundation for the AI-text corpus. The Synthesizer then paraphrases these AI-generated texts, introducing variations that challenge the detection capabilities by making the text appear more human-like. The Detector is trained to identify not only the original AI-generated text but also these paraphrased versions, ensuring comprehensive detection capabilities across different forms of AI-text manipulations. It extends the work of Xiaomeng Hu and using the findings proposed in their architecture builds upon a theoretical framework that can be implemented along with the dataset generation that is explained in the section above.

2.1 Theoretical Architecture behind the Proposed System

In the work presented by Xiaomeng Hu in 2023. The RADAR (Robust AI-Text Detection via Adversarial Reinforcement Learning) framework was proposed to enhance the robustness of AI-text detection systems against paraphrased content. This section provides an overview of the theoretical model underpinning RADAR, to further explain the architecture, mathematical formulations, and training methodologies of its core components: the paraphraser, the detector, and their integration through adversarial learning.

RADAR integrates three key neural network models:

1. **Target Language Model (Tθ)**: This model generates AI-text based on given human-text prefixes. It remains fixed throughout the process.
2. **Detector (Dϕ)**: Trained to differentiate between human-generated and AI-generated (original and paraphrased) text.
3. **Paraphraser (Gσ)**: Rewrites AI-generated text to make it more human-like, thus challenging the detector.

Steps in RADAR:

- **Data Preparation**: Construct the AI-text corpus M by using Tθ to complete texts from the human-text corpus HHH.
- **Paraphraser Training**: Paraphraser Gσ generates paraphrased AI-text P from M. The detector Dϕ evaluates the paraphrased text, and this feedback is used to update Gσ via Proximal Policy Optimization (PPO).
- **Detector Training**: Update Dϕ using a balanced combination of human-text HHH, original AI-text M, and paraphrased AI-text PPP.
- **Validation and Evaluation**: Measure the performance using the AUROC metric on a separate validation dataset, iterating steps 2 and 3 until performance stabilizes.

- xm: Sample from AI-text corpus M.
- xp: Sample from paraphrased AI-text corpus P.

$$LG() = E(xm, xp)PG[minPG(xpxm)PG(xpxm), 1, 1+A(xp,)S()]$$

2.2 GUARD for AI-based Content Detection

The proposed model that would be developed in the light of RADAR can be termed as GUARD (Generate, Utilize, Adapt, Refine, Detect) framework is designed to iteratively and adaptively detect AI-generated text through a synergistic interaction of three main

components: the **Target Language Model** (Tθ), the **Synthesizer** (Sσ), and the **Detector** (Dϕ).

High-Level Flow Overview:

1. **Generate AI-Text**:
    - The Target Language Model (Tθ) generates text samples (xm) from a human-written dataset (H), creating the AI-text dataset (M).

2. **Paraphrase AI-Text**:
    - The Synthesizer (Sσ) transforms these AI-generated texts (xm) into paraphrased versions (xp), simulating attempts to mask AI origins.

3. **Train Detector**:
    - The Detector (Dϕ) is trained to classify text samples into categories: original human-written (xh), original AI-generated (xm), and paraphrased AI-generated (xp).

4. **Iterative Improvement**:
    - The Synthesizer and Detector are iteratively improved. The Synthesizer generates more sophisticated paraphrases, while the Detector is trained on these new samples, enhancing its detection capabilities.

This cyclical process ensures continuous enhancement of the system's ability to detect increasingly complex and subtle AI-generated text.

Overview and Mathematical Notations

The GUARD framework is designed to detect AI-generated text, even when it has been modified or paraphrased. It involves three main components: the target language model (Tθ), which generates AI-text, the detector (Dϕ), which learns to distinguish between human and AI-generated text, and the synthesizer (Sσ), which creates paraphrased versions of the AI-text to challenge the detector.

The notations used are:

- Tθ: Target language model with parameters θ.
- Dϕ: Detector model with parameters ϕ.
- Sσ: Synthesizer model with parameters σ.
- H: Human-text corpus.
- M: AI-text corpus generated by Tθ.
- P: Paraphrased AI-text corpus generated by Sσ.

xh: Sample from human-text corpus H.

3. Detailed Methodology

3.1 Data Preparation

1. **Human-Text Corpus (H)**:
   - A comprehensive and diverse human-written text corpus is compiled from various sources such as news articles, essays, literature, social media posts, and conversational dialogues.
   - This corpus serves as the baseline for generating AI-text and training the detection model.

2. **AI-Text Generation (M)**:
   - Using the Target Language Model (Tθ), text is generated by prompting it with text segments from the human-text corpus (H).
   - The generated text (xm) is collected into the AI-text corpus (M), representing the output of the language model.

3. **Paraphrasing (P)**:
   - The Synthesizer (Sσ) takes the AI-generated text (xm) and produces paraphrased versions (xp), forming the paraphrased AI-text corpus (P).
   - This step simulates real-world scenarios where AI-generated content might be modified to avoid detection. The paraphraser Gσ transforms the original AI-generated text xm into paraphrased text xp. This process is modeled as a Markov decision problem, where xm is the state, and xp is the action. The objective is to maximize the reward R(xp,ϕ), provided by the detector Dϕ, which indicates the likelihood of xp being human-generated.

$$R(xp,) = D(xp)[0,1]$$

$$\log P_{G\sigma}(x_p|x_m) = \sum_{i=1}^{N} \log P_{G\sigma}(x_p^i|x_m, x_p^{1:i-1})$$

Here, xp ^i is the i-th token of the paraphrased text xp, and xp1:i−1x_p^{1:i-1} denotes the sequence of tokens preceding xp^i.

3.2 Synthesizer Training (Paraphrasing)

The Synthesizer (Sσ) is responsible for creating paraphrased versions of AI-generated text to challenge the Detector and simulate disguised AI-text. It uses reinforcement learning, specifically Proximal Policy Optimization (PPO), to optimize its paraphrasing strategy.

1. **Paraphrasing Mechanism**:

    - For each text xm in the AI-text corpus, the Synthesizer generates a paraphrased version xp.
    - The paraphrasing process involves restructuring sentences, replacing synonyms, and altering syntax while preserving the original meaning.

2. **Reward Function**:

    - The Detector evaluates the paraphrased text xp to determine the likelihood it is classified as human written. This classification serves as the reward signal R(xp,ϕ)R(x_p,)R(xp,ϕ) for the Synthesizer.
    - Higher rewards are given for paraphrases that closely resemble human-written text while evading detection as AI-generated.

3. **Policy Optimization**:

    - The Synthesizer's parameters (σ) are updated to maximize the reward signal through the PPO algorithm, refining its ability to produce human-like paraphrases.

4. **Entropy Regularization**:

    - To maintain diversity in the paraphrasing outputs and avoid overly predictable paraphrasing, an entropy penalty is included in the loss function.
    - This encourages the Synthesizer to explore a variety of paraphrasing styles and techniques.

3.3 Detector Training

The Detector (Dϕ) aims to distinguish between human-written and AI-generated texts, including those that have been paraphrased. It is trained using a reweighted logistic loss function to manage class imbalances and ensure robust classification performance.

1. **Input Data**:

    - The training dataset for the Detector includes balanced samples of human-written text (xh), original AI-generated text (xm), and paraphrased AI-generated text (xp).
    - This diverse input helps the Detector learn to recognize both direct AI outputs and more subtly modified versions.

2. **Reweighted Logistic Loss**:

    - Log loss ensures well calibrated predictions by inflicting large penalty if the model predict a very high probability for the wrong class and rewarding a prediction closer to true labels. Unlike accuracy, it considers how confident the model is in its predictions, providing a more nuanced evaluation.
    - The Detector uses a reweighted logistic loss function to balance the influence of each class (human, original AI, paraphrased AI) during training.
    - This approach prevents the Detector from being biased towards any class (class imbalance) and ensures sensitivity across all types of text.

3. **Optimization**:

    - Gradient descent algorithms are employed to minimize the reweighted logistic loss, continuously adjusting the Detector's parameters (ϕ) for improved classification accuracy.

4. Proposed Algorithm

The entire procedure of how GUARD works is summarized below:

1. Data Preparation

- Corpus Creation: Build a collection of human-written text, denoted by H.
- AI-Generated Text: Employ the target language model T_θ to generate AI-written text M based on the corpus H.
- Replay Buffer: Establish a buffer B to store samples for training the model.
- Validation Set: Construct a validation dataset V containing samples from both H (human-written) and M (AI-written) texts.

2. Model Initialization

- Detector: Initialize the AI text detection model D_ϕ using a pre-trained model weights and architecture.
- Synthesizer: Initialize the paraphrasing model S_σ using a pre-trained model.

3. Model Training

Loop:

- Sample Selection: Randomly select text samples x_h from the human-written corpus H and x_m from the AI-generated text collection M.

- Paraphrasing: Utilize the paraphrasing model S_σ to generate a paraphrase x_p of the AI-written text x_m.
- Reward Calculation: Obtain a reward signal $R(x_p, \phi)$ based on the detector's output for the paraphrased text x_p. This reward indicates how well the detector is fooled by the paraphrase.
- Advantage Function: Normalize the reward to compute the advantage function $A(x_p, \phi)$, which measures the contribution of each sample to improving the model.
- Paraphrasing Model Update: Update the weights and parameters of the paraphrasing model S_σ using the Proximal Policy Optimization (PPO) loss function $L_S(\sigma)$.
- Detector Update: Update the weights and parameters of the detection model D_ϕ using a reweighted logistic loss function $L_D(\phi)$. This loss function emphasizes informative training samples.
- Buffer Management: Periodically clear the replay buffer B to maintain a fresh set of training samples.
- Validation: Evaluate the Area Under the ROC Curve (AUROC) of the detector D_ϕ on the validation dataset V. This metric assesses the model's ability to distinguish between human-written and AI-generated text.

4. Output

- Return the trained detector model D_ϕ capable of identifying AI-generated text and the paraphrasing model S_σ.

5. Iterative Training Process

1. **Initialization**:
    - Pre-trained models for both the Synthesizer and Detector provide a strong starting point, accelerating the initial learning phase and improving early performance.
2. **Adversarial Training**:
    - The training process is adversarial: the Synthesizer generates increasingly complex paraphrases, while the Detector is concurrently trained to identify these as AI-generated.
    - This dynamic interaction pushes both components to evolve, enhancing the system's overall detection capabilities.
3. **Evaluation and Refinement**:
    - The Detector's performance is regularly evaluated using metrics such as the Area Under the Receiver Operating Characteristic (AUROC) curve, which measures its ability to distinguish between classes.
    - Based on these evaluations, training parameters and strategies are adjusted to ensure continuous improvement.

Chapitre 6: Experiments and Results

1. Experimental Setup

After the construction of the proposed model, the following datasets can be used to train and evaluate the framework that has been developed above. These datasets can be a good source of identifying what collection of user prompts are human written and what may include AI assistance in its development. These datasets can be integrated once the hypothetical model that we are proposing is developed into a functioning model after training and performing some finetuning on the model, it can be used at a commercial scale to compete with the existing AI detection services.

1. **Model Selection**:
    - The Target Language Model (Tθ) can be chosen from state-of-the-art models like GPT-3, which can generate high-quality text.
    - The Detector (Dϕ) may use models like BERT or RoBERTa, known for their strong text classification performance.
2. **Metrics**:
    - Performance metrics include AUROC scores to evaluate the Detector's ability to distinguish between human-written and AI-generated texts.
    - Other metrics such as precision, recall, and F1-score may also be used to provide a comprehensive performance assessment.

2. Ability of the Classifier

1. **Detection without Paraphrasing**:
    - The initial evaluation focuses on the Detector's ability to classify original AI-generated text versus human-written text.
    - This baseline performance indicates the effectiveness of the Detector before introducing paraphrased samples.
2. **Robustness to Paraphrasing**:
    - The Detector's robustness can be tested against paraphrased AI-generated text, including both seen and unseen paraphrasing techniques.
    - This evaluation assesses how well the Detector can adapt to disguised AI-text that attempts to mimic human writing styles.
3. **Comparative Analysis**:
    - The GUARD framework's performance is compared with other detection methods, such as statistical approaches and baseline models like OpenAI's RoBERTa.
    - AUROC scores and other performance metrics are used to highlight the advantages and limitations of each approach.

3. Implementation Details

1. **Model Architectures**:

    - **Target Language Model (Tθ)**: Typically, a large transformer-based model like GPT-3, capable of generating coherent and contextually relevant text. It includes multiple layers of attention mechanisms and uses a large vocabulary for text generation.
    - **Synthesizer (Sσ)**: Another transformer-based model or a specialized paraphrasing model that takes AI-generated text and produces paraphrased versions. It may include additional components to ensure semantic similarity and diversity in output.
    - **Detector (Dϕ)**: A classification model, often based on BERT or RoBERTa, designed to differentiate between human-written and AI-generated text. It uses bidirectional attention to understand the context and features of the input text.

2. **Training Parameters**:

    - **Learning Rate**: Specific to each model, with typical values ranging for fine-tuning.
    - **Batch Size**: Varies based on computational resources, commonly between 16 and 64.
    - **Regularization**: Techniques like dropout (probability of 0.1 to 0.3) are used to prevent overfitting.
    - **Training Duration**: Iterative training involves multiple epochs, with checkpoints for evaluation and early stopping to prevent overfitting.

Chapitre 7: Analysis and Discussions

1. Advancements offered by the Proposed Model

The GUARD framework for AI-generated text detection, as proposed, offers several significant advancements over existing models, that includes the base model RADAR and traditional detection systems. Below are key improvements and benefits that GUARD brings to the table:

1.1 A Comprehensive Approach to Paraphrased Text Detection

Integrated Paraphrasing and Detection: Unlike most modern systems that focus on just detecting unmodified AI texts, GUARD integrates a special Synthesizer (Sσ) that makes paraphrased form of AI texts. This forward-looking strategy helps to ensure that the Detector (Dϕ) can detect the AI-created text even when it has been modified intentionally to slip past detection.

Adaptive Paraphrasing, using Reinforcement Learning: Addition of reinforcement learning Proximal Policy Optimization, enables Synthesizer to adapt and come up with more complex paraphrases always. This guarantees that the Detector will be trained on many different text transformations, making it more robust and giving it the ability to adapt to novel ways to paraphrase.

1.2 Iterative Improvement

Iterative Adversarial Learning Cycle: GUARD leverages the adversarial learning cycle of training loop such that both Synthesizer and Detector get updated iteratively through the training. The Synthesizer produces progressively harder paraphrases and at the same time the Detector is trained to identify these complex modifications. This continual system advancement by dynamic adversarial training results in an improved detection ability.

Diversity with Entropy: Entropy regularization has been incorporated into GUARD in a regression mode to steer the paraphraser away from predictability and generate a range of transformations to the text. This avoids the creation of sequences by the Synthesizer that is repetitive or very easy to identify, hence diversifying the paraphrases that the Detector is tested for.

1.3 Increased Detection Performance:

Detector: In this pipeline, an existing classifier model is used to detect the potency of AI in text. The model gains deep contextual understanding and bidirectional attention, which helps it better be able to determine if the text in question has or has not been written by a human, including variants of human text that has been paraphrased.

Balanced Training with Reweighted Logistic Loss: Detector training involves a reweighted logistic loss here to tackle the issue of highly imbalanced classes. This keeps the model responsive & quick on all sorts of text, human written, original AI generated, or AI generated but paraphrased text.

1.4 Scale and Agility to deploy:

Target Language Model (Tθ) Flexibility: GUARD is agnostic to the types of language models being used thus GUARD can potentially even utilize languages models on scales far beyond GPT-3 (and in future, GPT-4 etc) for generating AI text. That flexibility enables the framework to extend the scalable to different applications and context and languages.

Fully Modular: Components of GUARD (Target Language Model, Synthesizer, Detector) are fully modular, making them better as updates or improvements can be made to one without the interdependence on the others making the updates future-proof and once the working prototype has been developed updates can be made accordingly. This makes it customizable for specific use cases or fields.

2. Performance evaluation:

Multiple Metrics: GUARD rates performance considering several metrics like AUROC, precision, recall, and F1-score, thus measures detection abilities of the model in a better way. The framework underwent a rigorous evaluation to ensure that it accurately detects AI-generated text with high precision and recall by reducing the level of false positives and gaps in detections.

Robust to Disguised AI-Text: GUARD is designed to target both seen and unseen paraphrasing techniques used by the Detector; thus, GUARD should be exposed to a large number of obfuscation strategies. This proves especially useful in actual implementations where it is expected that the titles generated by AI to some extent will be altered.

How GUARD Compares to Current Models:

- **RADAR**: While RADAR focuses on robust AI-text detection via adversarial reinforcement learning, GUARD extends this approach by integrating advanced paraphrasing techniques and iterative detector improvement. GUARD's use of PPO for synthesizer training and entropy regularization offers a more dynamic and diverse challenge for the Detector, resulting in better preparedness against sophisticated paraphrasing.
- **Traditional Detection Models**: Many existing detection systems rely on static features or less advanced models for text classification. GUARD's use of state-of-

the-art transformers for both generation and detection, combined with its adversarial learning framework, provides a significant performance boost over traditional methods, especially in identifying paraphrased AI-generated content.

3. Ethical Challenges and Theoretical Implications

The rise of advanced AI agents for automated text generation presents significant ethical challenges, as highlighted by the literature. The analysis draws heavily on agenda-setting theory and stakeholder theory to explore the implications of AI-generated text in business ethics. The key ethical issues raised concern an ability to automated false beliefs injected into vast segments of the populace, spread of misinformation and salient misinformation, and [further establishment of] a communication gap among stockholders.

Automated Mass Manipulation and Fake News

One of the biggest ethical issues is the fake news problem, and with AI-generated text you can manipulate the masses on a level never-seen before. This issue is closely related to the lowest denominator problem, which refers to the potential for AI to produce content that appeals to base instincts rather than higher reasoning, thus degrading the quality of public discourse. The literature shows the need for honesty programs to be incorporated into AI systems to mitigate these risks.

Model Answer on Business Ethics and Regulatory Measures

This will also add value to the point of regulations and guidelines that the conversation had focused on to understand the responsible use of AI text. They call on governments and regulatory bodies to create frameworks for dealing with the ethical issues that arise from these technologies. The literature review suggests that public engagement and deliberation would also be critical to ensure that appropriate measures are taken on issues of transparency, accountability and ethics around the use of AI.

Education and medical writing applications

There are numerous uses, opportunities and Challenges According to the studies of the use of AI in education and medical writing, AI technologies in education such as ChatGPT will transform education as delivering individual based learning, automated assessments, virtual instructional support. However, the susceptibility of AI to biases, concerns about academic integrity, and privacy issues are significant hurdles that need to be addressed. These ethical concerns necessitate responsible AI use to enhance educational processes while mitigating associated risks.

The same goes for state-of-the-art AI language generation models used in the medical field. Perhaps even more, since it involves ethical considerations. Given that the content created by AI in medical texts is much more complex than marketing material, the ability to create subject matter experts is extremely important in validating that the AI is learning effectively. With the remainder of the literature focusing on ethical considerations such

as misinformation, privacy, and lack of transparency, the recommendation for further research is the development of more resilient forensic systems which better integrate human expertise and causal reasoning.

Chapitre 8: Limitations of this Research

1.1 Methodological Limitations and Data Dependency

This research is limited to its dependence on existence of literature and theoretical underpinnings which may not fully contextualize and account for the practical applications of the detection of AI-driven content. The lack of thorough experiments is a strong limitation, since theoretical statements with no validation work may not reliably indicate the performance of the mentioned models and methods. Taking our cue from this, to move toward a more objective evaluation of AI-generated content detection methods, we call for the prioritization of empirical evidence via more robust experimental design in future research.

The model in this approach is based heavily on supervised learning models and requires sizable volumes of annotated data to train it on. Although this dependence confines the scaling potential of the model and the generalization ability in new and heterogeneous data embedding situations. Supervised learning requires large amounts of labeled data, which can be resource-intensive to obtain and may introduce biases from the training data. These biases can manifest in various forms, such as gender, racial, or cultural biases, leading to unethical outcomes in AI-generated content. Ensuring fairness and inclusivity requires ongoing efforts to refine training datasets and develop algorithms that can detect and counteract these biases.

1.2 Ethical Concerns

The ethical concerns associated with AI-generated content are needed to be considered as AI ethics is such a complex area that it is necessary to take a step back to explore the different aspects of the issue because the implications of AI technologies are not only technical, but also societal and ethical. This can be advantageous as they have to consider privacy, consent and the misuse but they need a wider and deeper treatment The research rarely offers in depth ethical analyses and again demonstrates the importance of interdisciplinary approaches to the creation and deployment of AI technologies that explicitly include ethical considerations.

Privacy concerns are a huge issue in the utilization of AI generated content. The answer in the research was that AI technologies are likely to already be infringing on individual privacy which was enough for the researchers to consider AI technologies used with personal data as being of importance. It is important to have a strong legal framework and ethical guidelines in place to develop AI ethics responsible and aware citizens. The development of privacy-preserving techniques to adhere to data protection regulations are essential steps in this direction. However, current regulatory measures and organizational guidelines need to be revised in order to address the rapid

advancements and potential misuse of AI-generated content. Governments and regulatory bodies need to develop comprehensive frameworks that can mitigate the ethical challenges posed by these technologies.

1.3 Technical Limitations and Model Flaws

Even with the progress made in identifying AI-generated content, there are many holes to fill in the efficacy of these methods. However, the generation methods of DeepFakes can be so sophisticated that detection techniques will often lag. This difference underscores the importance to continuously striving and improving advances in detection algorithms in parallel with the progressing AI technologies. Cross-modal detection and Detecting AI-generated content from text, video and audio is still in uncharted territory and needs more study. Moreover, the application of the model is confined to datasets and datasets to scenarios, thus creating limitations in the model's generalizability to diverse contexts and practical real-world scenarios. To evaluate generalizability further, the model should be validated over different types of datasets and applications to verify the robustness and versatility of the model.

Moreover, because of the lack of universal evaluation metrics, the potential of the model to rate and measure the quality and the effectiveness of the AI-content is restricted. It is important to create standardized metrics to assess the model's performance and make sure that the model behaves as expected in the required quality standards. Also, the model does not adequately consider the ethics of AI-generated content. The paper discusses in detail technical aspects and ethical aspects are usually limited or addressed superficially. For the co-design for responsible AI model, it is important to include ethical analyses as an element of the model development practice to ensure the development of responsible and socially beneficial AI technologies.

1.4 Challenges Incurred

Developing a New Framework from Scratch

The hardest part for this research was to build such an in depth view from base zero although RADAR was helpful as a grounding, While it provided the foundation, the situational and evolving nature of the AI generated content in DeepFake technology discouraged RADAR as a one size fits all in terms of needed specificity. To remedy this gap, it took a great deal of work to adapt and enrich RADAR's principles to build a more efficient and specific framework.

This has entailed several steps, including recognizing what the considerations for AI-generated content are. This meant having to go deep into methodologies to understand what worked, what did not work, and continue to iteratively improve on top of extensive validation processes. It required a lot of thinking and creativity on the part of the team as they had to go from place to place and tackle rugs of all kinds, shapes, and sizes.

The problem was generated since we proposed a development of a new framework that can be feasible to implement practically too besides presenting the theoretical basis of it across scenarios. Making these poles cohere in a state of relative quiescence requires continuous adjustments and reaching out to experts from adjacent fields, most notably in ethics, law, computer science, and media studies. This multidisciplinary approach, that added richness to the research, led at the same time to more layers making it difficult for us to reach a consensus.

Efforts in Dataset Development

Another problem was the need to build a dataset tailored to the research task. Given the difference between our dataset and other existing datasets, most of which are aimed at more general AI applications such as only text, our dataset inherently needed to cater to capture the key aspects of AI generated content and DeepFakes. To get what we needed, our main workload was to make a new custom dataset from scratch, as off-the-shelf datasets were mostly useless.

There are a lot of stages through to make a good dataset complete here are the stages in which challenges were faced. At first, locating and getting rights to appropriate content that properly reflected the kind of AI-generated media we're talking about was a huge challenge. A diverse set of this content was required so that it could be applied to many forms of media and was sufficiently broad to capture a range of sophistication in generation techniques.

The data can be then annotated to train and evaluate our models. This was a difficult process as all pieces of data needed to be labeled accurately - a very critical aspect for the downstream steps related to model training. Given the volume of data and the detail required in making these annotations, this represented a very resource-intensive and time-consuming task with high technical expertise.

Additionally, ensuring the dataset's integrity and security was also important. To protect the intellectual rights of the human generated text was also important to prevent the misuse of the dataset. This aspect added another layer of complexity, requiring vigilance and security measures.

Technical and Logistical Challenges

The technical hurdles with this research were endless. This often meant a substantial amount of testing and research to resolve compatibility with the implementation of new methodologies into existing technologies. New algorithms were needed to both handle the requirements of detecting and analyzing AI-generated content, as well as doing so at a model size and efficiency that are sufficient for deployment in practice. They had to be smart enough to detect these things, but also scalable and computationally costly, so they could crunch enough numbers to deliver a real-time result. In addition, the complexities were also increased due to the difficulties in coordinating the various tasks. Collaboration across different nomenclatures and methods was effective, but hard to

sustain. Regular meetings, clear documentation, and their vision for the outcome helped ensure that all team members were in sync from understanding to execution.

Chapitre 9: Future Studies and Recommendations

1. Future studies landscape

Advancements in Detection Techniques

Further research should be conducted to improve the detection methods of fake images and videos to match the rapidly developing creation technologies. What is more promising down the pipeline is the use of more powerful machine learning algorithms that can detect more subtle artifacts and inconsistencies of AI-generated content. Detection methods are usually lagging the most recent generation and often require people to continuously upgrading their algorithm developments.

Such approaches will require more work on hybrid models such as combining supervised, unsupervised, and semi-supervised learning techniques. Supervised learning helps in training models on known AI-generated content with the help of annotated datasets, and unsupervised learning is used to identify new patterns in new, unlabeled data. Semi-supervised learning, which uses a small amount of labeled data along with a large amount of unlabeled data, can be particularly useful in enhancing detection capabilities without requiring extensive manual annotation.

Ethical Frameworks and Policy Development

The improvements of AI technologies is persistent and it is significantly more important to create strong, universal ethical frameworks and policies to direct their application. AI-created content has always been subjected to ethical concerns, and future research could further explore the ethical implications of this method, as well as issues of privacy, consent, and the possibility of being exploited.

Algorithmic nepotism should be the strongest signal that these AI and ethics documentation and guidelines need to be beefed up with honesty and accountability everywhere. AI systems are expected to enlighten their decision-making process so that there is a proper trust and reliability streamline for the researchers. In addition, policies must be able to evolve with the rapidly changing AI landscape for them to be meaningful and enforceable as new ethical dilemmas come to light.

Leveraging the Research Framework for Model Development

This research further provides a foundation for developing models that can detect AI-generated content. The framework and equations that are presented in this research can be used as a blueprint for future researchers aiming to build effective detection systems. Here's how this research can be utilized:

Framework Utilization: The developed thorough framework listing all the detected features/behavior patterns for content generated by AI can be utilized for creating a detection model. These theoretical foundations and methodological approaches the author's report should help provide a guide or template for those new to developing models.

AI generated content detecting tools: Through the mathematical models and algorithms mentioned in the research, the detected AI generated content was possible to be permitted for detecting purposes. Researchers can use these models to build systems that are able to recognize alterations that humans can still detect as synthetic but are often hard to identify below human perception, and markers that a piece of synthetic media was altered, inserted, or both. These models can be fine-tuned and applied to improve detection.

Dataset Development: Specialized dataset resulting from the research provides data on different types of AI-generated content, hence serving as an excellent training and testing resource for detection models This dataset can be extended in future work to incorporate additional diverse examples and make the detection systems robust and generalizable.

Interdisciplinary: This work is meaningful as an interdisciplinary project, incorporating insights from computer science, ethics, and media studies to more fully address the model-development process. This interdisciplinary foundation will be an additionally helpful groundwork for future researchers to expand upon to create more inclusive and efficient detection systems.

Benchmarking and Performance: This study also reveals a research gap related to standardized metrics to evaluate the quality of the AI generated content and it could be served in future works. Developing and adopting standardized benchmarks and evaluation criteria will enable more consistent and objective assessment of detection models, facilitating their improvement and validation.

2. Recommendations for Implementation

For more effective detection of AI-generated content among different themes, future research should consider interdisciplinary collaboration. Partnerships among academia, industry, and government can improve sharing of resources, data, and expertise to create new and holistic solutions. These partnerships could also lead to the development of shared ethical practices and regulatory structures, which will result in more responsible and uniform AI standards across industries.

This requires investment in powerful computational resources that are fundamental for any type of detection model. The algorithms must be trained on a hardware infrastructure that is set up to also handle the large data sets necessary to create accurate AI detection systems. Moreover, the possibility of model updates and retraining regularly is key to keeping up with this fast-paced generation of AI techniques, hence long-term relevant and accurate detection.

Ethical AI deployments need to be accompanied by education and public awareness. Targeted education and training programs can provide researchers, practitioners, and the public with the knowledge and skills that will enable them to make better sense of AI-generated content. This function should be tasked with developing programs making the ethical use of AI a constant subject of discussion and encouraging responsible innovation. Intelligent discussions on ways in which AI can best serve society, as well as what AI can and should not be capable of doing, are essential if we seek to develop and utilize AI technologies in a manner that provides benefits while reducing harm.

Chapitre 10: Conclusions

The use of AI to generate content on a scale has opportunities as well as significant challenges. AI has beheaded the Relationship Between R&D and Content Creation. Technologies such as ChatGPT, DALL-E, and many other large language models have transformed content creation radically, allowing text, pictures, and videos to be produced rapidly and consistently. This evolution has opened an unlimited number of applications across multiple industries such as education, journalism, healthcare, or customer service for the substituting easy tasks to reduce the resource overheads. But it also raises problems of authenticity and plagiarism, and possibilities for spreading disinformation, which will necessitate robust detection methods to distinguish AI-generated content from its human counterparts.

This study has delved into AI content generation and detection practices while displaying how well current detection tools can or cannot detect AI-generated content. Many of these studies include Sarzaeim et al. As the works of Amjad et al., and Bhattacharjee et al., have shown, there are some good frameworks and approaches for spotting AI-generated text. While these models tend to achieve high degrees of accuracy, they also present challenges like working with small datasets, basic preprocessing techniques and not being able to separate from more sophisticated AI generators what is content. Moreover, the inconsistency of available tools in correctly categorizing human-written from AI-generated prose demonstrates that more work needs to be done.

Moving forward, future research must focus on how the robustness and accuracy of AI content detection models can be improved. This includes expanding datasets, to enhance the preprocessing techniques, and integrating advanced algorithms to handle diverse and complex scenarios. Furthermore, addressing the ethical implications of AI-

generated content is also needed to ensure responsible use and minimizing negative impacts on society. By continuing to refine detection methodologies and exploring innovative approaches, we can navigate the evolving landscape of AI-driven content creation and maintain the integrity and authenticity of information in various domains.

References

Bhattacharjee, A., et al. (2024). EAGLE: A Domain Generalization Framework for AI-generated Text Detection. Retrieved from arXiv. https://arxiv.org/html/2403.15690v1

Raymond, S. (2023, Oct 20). Debunked: YouTube ads feature Elon Musk in deepfake videos pushing cryptocurrency scams. From thejournal: https://www.thejournal.ie/deepfake-elon-musk-bill-maher-cryptocurrency-youtube-ad-6201640-Oct2023/

Sharma, J., & Sharma, S. (2021). Challenges and solutions in deepfakes. arXiv preprint arXiv:2109.05397.

Mahmud, B. U., & Sharmin, A. (2021). Deep Insights of Deepfake Technology : A review. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2105.00192

Malik, A., Kuribayashi, M., Abdullahi, S. M., & Khan, A. N. (2022). DeepFake detection for human face images and Videos: a survey. IEEE Access, 10, 18757–18775. https://doi.org/10.1109/access.2022.3151186

Barari, S., Lucas, C., & Munger, K. (2021). Political deepfake videos misinform the public, but no more than other fake media. OSF Preprints, 13.

Oancea, M. (2024). AI and Deep Fake-Video and Audio Manipulation Techniques Capable of Altering the Political Process. Revista de Științe Politice. Revue des Sciences Politiques• No, 81, 70-82.

Appel, M., & Prietzel, F. (2022). The detection of political deepfakes. Journal of Computer-mediated Communication, 27(4). https://doi.org/10.1093/jcmc/zmac008

Asha, S., Vinod, P., & Menon, V. G. (2023). A defensive framework for deepfake detection under adversarial settings using temporal and spatial features. International Journal of Information Security, 22(5), 1371–1382. https://doi.org/10.1007/s10207-023-00695-x

Botha, J. G., & Pieterse, H. (2020). Fake news and deepfakes: A dangerous threat for 21st century information security. Council for Scientific and Industrial Research. https://doi.org/10.34190/iccws.20.085

Citron, D. K., & Chesney, R. (2018). Deepfakes and the New Disinformation War. The Coming Age of Post-Truth Geopolitics. https://scholarship.law.bu.edu/shorter_works/76/

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & De Vreese, C. (2020). Do (Microtargeted) deepfakes have real effects on political attitudes? The International Journal of Press/Politics, 26(1), 69–91. https://doi.org/10.1177/1940161220944364

Hight, C. (2021). Deepfakes and documentary practice in an age of misinformation. Continuum, 36(3), 393–410. https://doi.org/10.1080/10304312.2021.2003756

Husseini, S., & Dugelay, J. (2023). A Comprehensive Framework for Evaluating Deepfake Generators: Dataset, Metrics Performance, and Comparative Analysis. IEEE Xplore. https://doi.org/10.1109/iccvw60793.2023.00044

Lu, Y., & Ebrahimi, T. (2024). Assessment framework for deepfake detection in real-world situations. EURASIP Journal on Image and Video Processing, 2024(1). https://doi.org/10.1186/s13640-024-00621-8

Öhman, C. (2022). The identification game: deepfakes and the epistemic limits of identity. Synthese, 200(4). https://doi.org/10.1007/s11229-022-03798-5

Tariq, S., Lee, S. Y., & Woo, S. S. (2021). One detector to rule them all: towards a general deepfake attack detection framework. arXiv (Cornell University), 3625–3637. http://arxiv.org/abs/2105.00187

Taylor, B. C. (2020). Defending the state from digital Deceit: the reflexive securitization of deepfake. Critical Studies in Media Communication, 38(1), 1–17. https://doi.org/10.1080/15295036.2020.1833058

Temir, E. (2020). Deepfake: New Era in The Age of Disinformation & End of Reliable Journalism. DergiPark (Istanbul University). https://doi.org/10.18094/josc.685338

Van Der Sloot, B., & Wagensveld, Y. (2022). Deepfakes: regulatory challenges for the synthetic society. Computer Law and Security Report/Computer Law & Security Report, 46, 105716. https://doi.org/10.1016/j.clsr.2022.105716

Veerasamy, N., & Pieterse, H. (2022). Rising above misinformation and deepfakes. Proceedings of the . . . International Conference on Information Warfare and Security/˜ the œProceedings of the . . . International Conference on Information Warfare and Security, 17(1), 340–348. https://doi.org/10.34190/iccws.17.1.25

Whyte, C. (2020). Deepfake news: AI-enabled disinformation as a multi-level public policy challenge. Journal of Cyber Policy, 5(2), 199–217. https://doi.org/10.1080/23738871.2020.1797135

Yan, Z., Zhang, Y., Fan, Y., & Wu, B. (2023). UCF: Uncovering Common Features for Generalizable Deepfake Detection. IEEE Xplore. https://doi.org/10.1109/iccv51070.2023.02048

Zhang, L., Qiao, T., Xu, M., Zheng, N., & Xie, S. (2023). Unsupervised Learning-Based Framework for Deepfake video Detection. IEEE Transactions on Multimedia, 25, 4785–4799. https://doi.org/10.1109/tmm.2022.3182509

Zhou, Y., & Lim, S. (2021). Joint Audio-Visual Deepfake Detection. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv48922.2021.01453

Bhattacharjee, A., Moraffah, R., Garland, J., & Liu, H. (2024). EAGLE: A Domain Generalization Framework for AI-generated Text Detection. arXiv preprint arXiv:2403.15690.

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. International Journal for Educational Integrity, 19(1), 17.

Mitrović, Sandra, Davide Andreoletti, and Omran Ayoub. "Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text." arXiv preprint arXiv:2301.13852 (2023).

Zhou, Y., He, B., & Sun, L. (2024). Humanizing Machine-Generated Content: Evading AI-Text Detection through Adversarial Attack. arXiv preprint arXiv:2404.01907.

Ren, J., Xu, H., Liu, Y., Cui, Y., Wang, S., Yin, D., & Tang, J. (2023). A robust semantics-based watermark for large language model against paraphrasing. arXiv preprint arXiv:2311.08721.

Taguchi, K., Gu, Y., & Sakurai, K. (2024). The Impact of Prompts on Zero-Shot Detection of AI-Generated Text. arXiv preprint arXiv:2403.20127.

Mireshghallah, N., Mattern, J., Gao, S., Shokri, R., & Berg-Kirkpatrick, T. (2024, March). Smaller Language Models are Better Zero-shot Machine-Generated Text Detectors. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 278-293).

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. arXiv preprint arXiv:2303.04226.

Ouchchy, L., Coin, A., & Dubljević, V. (2020). AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. AI & SOCIETY, 35, 927-936.

Doyal, A. S., Sender, D., Nanda, M., & Serrano, R. A. (2023). ChatGPT and artificial intelligence in medical writing: concerns and ethical considerations. Cureus, 15(8).

Sidiropoulos, D., & Anagnostopoulos, C. N. (2024). Applications, challenges and ethical issues of AI and ChatGPT in education. arXiv preprint arXiv:2402.07907.

Alamleh, H., AlQahtani, A. a. S., & ElSaid, A. (2023). Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning. IEEE. https://doi.org/10.1109/sieds58326.2023.10137767

Fagni, T., Falchi, F., Gambini, M., Martella, A., & Tesconi, M. (2021). TweepFake: About detecting deepfake tweets. PloS One, 16(5), e0251415. https://doi.org/10.1371/journal.pone.0251415

Hamza, A., Javed, A. R. R., Iqbal, F., Kryvinska, N., Almadhor, A. S., Jalil, Z., & Borghol, R. (2022). Deepfake audio detection via MFCC features using machine learning. IEEE Access, 10, 134018–134028. https://doi.org/10.1109/access.2022.3231480

Illia, L., Colleoni, E., & Zyglidopoulos, S. (2022). Ethical implications of text generation in the age of artificial intelligence. Business Ethics, the Environment & Responsibility, 32(1), 201–210. https://doi.org/10.1111/beer.12479

Jiang, Z., Zhang, J., & Gong, N. Z. (2023). Evading Watermark based Detection of AI-Generated Content. IEEE. https://doi.org/10.1145/3576915.3623189

Kumarage, T., & Liu, H. (2023). Neural Authorship Attribution: Stylometric Analysis on Large Language Models. IEEE. https://doi.org/10.1109/cyberc58899.2023.00019

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of

ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Nam, N. L. H. (2023). Giới thiệu về DetectGPT – Công cụ phát hiện văn bản do máy tính tạo ra. Tạp Chí Điện Tử Khoa Học Và Công Nghệ Giao Thông, 45–54. https://doi.org/10.58845/jstt.utt.2023.vn.3.1.45-54

Sarzaeim, P., Doshi, A. M., & Mahmoud, Q. H. (2023). A framework for detecting AI-Generated text in research publications. Proceedings of the International Conference on Advanced Technologies. https://doi.org/10.58190/icat.2023.28

Uchendu, A., Ma, Z., Le, T., Zhang, R., & Lee, D. (2021). TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. Arxiv. https://doi.org/10.18653/v1/2021.findings-emnlp.172

Wang, F., Li, J., Qin, R., Zhu, J., Mo, H., & Hu, B. (2023). ChatGPT for Computational social systems: From conversational applications to Human-Oriented Operating Systems. IEEE Transactions on Computational Social Systems, 10(2), 414–425. https://doi.org/10.1109/tcss.2023.3252679

Wang, Y., Pan, Y., Yan, M., Su, Z., & Luan, T. H. (2023). A survey on ChatGPT: AI–Generated Contents, Challenges, and Solutions. IEEE Open Journal of the Computer Society, 4, 280–302. https://doi.org/10.1109/ojcs.2023.3300321

Wen, J., & Wang, W. (2023). The future of ChatGPT in academic research and publishing: A commentary for clinical and translational medicine. Clinical and Translational Medicine, 13(3). https://doi.org/10.1002/ctm2.1207

Zobaed, S., Rabby, F., Hossain, I., Hossain, E., Hasan, S., Karim, A., & Hasib, K. M. (2021). DeepFakes: Detecting forged and synthetic media content using machine learning. In Advanced sciences and technologies for security applications (pp. 177–201). https://doi.org/10.1007/978-3-030-88040-8_7