On Reducing Risks of Synthetic Content

Point of Contact: Nathan Harmon, Ozni AI

February 2, 2024

Abstract

The evolution of artificial intelligence (AI) has blurred the lines between reality and digital fabrication, raising critical concerns about the authenticity of content across various media. This abstract outlines our investigation into the methods and challenges of discerning and verifying original content amidst the proliferation of AI-generated text, images, videos, and audio. Our focus is on evaluating current and emerging techniques for detecting synthetic content, employing a range of strategies from advanced detection algorithms to digital watermarking and metadata analysis. By addressing the technological, ethical, and economic dimensions of this issue, our aim is to contribute to the development of robust standards and solutions that ensure content integrity and public trust in the digital age.

1.0 Motivation

Artificial intelligence (AI) systems have revolutionized modern technology, setting a new gold standard of performance across a wide variety of domains including healthcare diagnostics, autonomous vehicle navigation, personal assistants, and within the Department of Defence and Intelligence Community. The accessibility of GPUs, cloud resources, downloadable foundation models, and the surge in open source AI projects has minimized technological barriers, facilitating the rapid development of innovative solutions by AI practitioners. Understandably, this sense of rapid innovation is highly desired by our partners in the national security and defense domains. However, this ease of implementation has led to significant challenges in AI reliability as some practitioners may inadvertently design and field AI systems without providing sufficient verification, validation, or safe guards. Moreover, the rise of generative AI has introduced new challenges, such as the spread of inadvertent plagiarism or generation of illicit/defaming content.

There is an urgent need for regulations and governance mechanisms that ensure AI systems are developed and used responsibly, without compromising on safety and reliability. This responsibility involves protecting users and other stakeholders from potential harm or exploitation by inadequately tested or negligently developed AI technologies. However, it is imperative that these regulatory measures are balanced to avoid stifling the innovation that drives the AI field forward. Excessive regulation could hinder progress and place members at a strategic disadvantage compared to parties who may face fewer constraints. Therefore, our mission is to aid in establishing a framework that assures the safe and ethical use of AI, while also maintaining the momentum of technological advancement.

This work is original and has been enhanced for clarity and precision with the assistance of a Large Language Model (LLM). We have made diligent efforts to ensure the accuracy and completeness of the information presented.

2.0 Introduction

In the rapidly evolving landscape of AI the distinction between authentic and synthetic content has become increasingly blurred. As AI technologies advance, generating highly realistic text, images, videos, and audio, the ability to discern and authenticate original content from AI-generated counterparts is paramount. This capability is not only crucial for maintaining the integrity of information but also for safeguarding public trust and preventing misinformation. The proliferation of synthetic content, while showcasing the remarkable capabilities of AI, also presents significant challenges in content verification, security, and the potential for misuse.

This report aims to explore and evaluate the current landscape and emerging techniques for authenticating, labeling, and detecting synthetic content, with a particular focus on synthetic text generation. However, it's important to note that the methods and insights discussed are broadly applicable to other data modalities, including images, videos, and audio. By delving into both black box and white box detection methods, watermarking techniques, digital signatures, and metadata for content provenance, we will provide a comprehensive overview of the tools and strategies at our disposal. Furthermore, we will examine the challenges, trade-offs, and economic feasibility of these approaches for organizations of various sizes, alongside strategies to enhance detection resilience and mitigate risks posed by malign actors.

Our objective is to not only shed light on the state-of-the-art methods and their effectiveness but also to discuss potential developments in new science-backed standards and techniques. As we navigate the complexities of synthetic content across various modalities, this report seeks to offer actionable insights and recommendations. Together, we aim to foster a collaborative effort towards developing effective, adaptable solutions that balance the rapid innovation in AI with the critical need to ensure the authenticity and integrity of digital content.

This report starts with an exploration of detection methods, including both black box and white box approaches. At the outset of each section, we prioritize actionable recommendations to guide readers toward immediate insights. The discussion progresses to cover authentication and labeling strategies such as watermarking and digital signatures, before addressing the challenges and economic implications of these methods.

3.0 Detecting and Tracking Synthetic Content

The advent of sophisticated AI models has significantly enhanced the capability to generate synthetic content that closely mimics authentic human-created content, presenting both opportunities and challenges in the digital age. Detecting and tracking this synthetic content is crucial for a multitude of reasons, ranging from preventing misinformation and safeguarding intellectual property to ensuring the security and integrity of digital communications. The task involves distinguishing between content created by humans and that generated by AI, a process complicated by the continually improving quality of AI-generated outputs. As AI technologies evolve, so too must the methodologies and tools designed to identify and monitor synthetic content. Detection can be broken down into two paradigms Black box and White Box detection. Black box methods rely on analyzing external characteristics of the content without access to the underlying AI models, while white box techniques necessitate the cooperation of the model's creators, provide knowledge of, or access to, the model's internal workings. Together, these methodologies form the cornerstone of efforts to detect and track synthetic content.

3.1 Black Box Detection Techniques

Recommendation

We recommend deprioritizing investment in Black Box Detection Techniques due to their diminishing effectiveness against advancing generative AI as detailed below. These methods are becoming less reliable and do not warrant further investment.

Black box detection techniques identify synthetic content, operating without the need for access to the underlying AI models that generate the content. These methods analyze the external characteristics of digital artifacts—text, images, videos, and audio—to distinguish between human-generated and AI-generated content. This section delves into the principles of black box detection, focusing on its application to text but acknowledging the relevance of these techniques across all data modalities.

3.2 Statistical Disparities

Statistical analysis forms the backbone of many black box detection methods. These techniques evaluate the statistical properties of text to identify patterns or anomalies indicative of synthetic content, including:

- **Zipfian Coefficient Analysis**: This method assesses how well the text conforms to Zipf's law, which posits that the frequency of any word is inversely proportional to its rank in the frequency table. AI-generated text often deviates from this distribution in telltale ways Piantadosi [2014].

- **GLTR (Giant Language Model Test Room)**: GLTR utilizes knowledge of a language model's word prediction probabilities to spot the unnatural precision of AI-generated text. It identifies whether the text disproportionately selects high-probability words, a common trait in machine-generated content Gehrmann et al. [2019].

- **Perplexity Measurement**: By calculating the perplexity of a text—essentially, its predictability—analysts can gauge whether the content is too uniform or smooth, traits often associated with synthetic generation Brown et al. [1992].

3.3 Linguistic Patterns

Linguistic pattern analysis examines the stylistic and structural elements of text, looking for cues that suggest artificial creation: Bhatt and Rios [2021], Fröhling and Zubiaga [2021], Guo et al. [2023]

- **Vocabulary Diversity and Text Length**: AI-generated texts tend to use a less diverse vocabulary over extended lengths compared to human writing, which is typically more concise and varied.

- **Part-of-Speech Usage**: Synthetic content often exhibits abnormal frequencies of certain parts of speech, such as an overreliance on nouns, determiners, and conjunctions.

- **Emotional Expression and Coherence**: AI texts may lack the depth of emotional expression found in human writing, presenting instead a neutral tone. They can also struggle with maintaining thematic coherence throughout.

3.4 Fact Verification

Fact verification targets the content's fidelity to reality, a critical aspect given AI's propensity for generating plausible but factually inaccurate "hallucinations." Zhong et al. [2020]

- **Identifying Hallucinations**: Techniques focus on spotting inaccuracies or outright fabrications within the text, a task becoming increasingly challenging as AI models improve in generating contextually relevant, though potentially untrue, content.

3.5 Leveraging Machine Learning and Deep Learning for Black Box Detection

The integration of Machine Learning (ML) and Deep Learning (DL) techniques in black box detection marks a pivotal advancement in differentiating AI-generated content from human-generated content Gallé et al. [2021], Ippolito et al. [2019], Rodriguez et al. [2022] . ML models, through classification, anomaly detection, and clustering algorithms, learn from vast datasets to identify subtle patterns indicative of synthetic creation, while DL approaches, leveraging sophisticated neural networks like Convolutional Neural Networks (CNNs) for visual content and Recurrent Neural Networks (RNNs) or Transformers for textual and sequential data, excel at uncovering nuanced features. Despite their promise, these technologies face challenges such as the need for diverse, high-quality training data and the necessity for ongoing updates to keep pace with evolving AI-generated content methods. The effectiveness of ML and DL in black box detection underscores the importance of interdisciplinary collaboration and continuous innovation to maintain and enhance detection capabilities in the dynamic landscape of digital content creation.

3.6 The Diminishing Effectiveness of Black Box Detection

As AI technologies evolve, the line distinguishing AI-generated content from human-generated content blurs, posing significant challenges to black box detection methods. The adaptive nature of AI models, especially those trained with feedback loops, means they continually learn from detection strategies and adjust to evade identification. Furthermore, improvements in AI's understanding of human nuances, statistical patterns, and factual accuracy render traditional black box methods less effective over time.

This diminishing effectiveness underscores the need for continuous innovation in detection technologies, incorporating more sophisticated statistical models, deeper linguistic analysis, and advanced fact-checking algorithms. It also highlights the importance of complementing black box approaches with white box methods, creating a more comprehensive strategy for identifying and tracking synthetic content.

4.0 White Box Detection

Recommendation

We recommend increasing investment and research in White Box Detection and watermarking techniques. Despite current vulnerabilities, these areas hold the highest potential for effective long-term solutions against synthetic content.

White box detection techniques represent a sophisticated approach to identifying synthetic content, predicated on having access to or knowledge of the internal workings of the AI models responsible for generating such content. Unlike black box methods, which analyze content from an external perspective, white box techniques delve into the model's architecture, training data, and operational algorithms to uncover distinctive markers of synthetic generation. This intrinsic analysis allows for a more targeted and often more effective identification process, leveraging the very mechanisms that produce synthetic content to detect it. Central to white box detection is the concept of watermarking—embedding unique, detectable patterns or codes within the content at the time of creation, which can later be identified to verify the content's origins. This section introduces the fundamentals of white box detection, emphasizing its reliance on deep insights into AI technology and its potential to offer robust defenses against the manipulation and misuse of synthetic content.

4.1 The Potential of Watermarking

At the heart of white box detection is watermarking, a method that embeds a unique, detectable signal or pattern into content during its creation. This signal can later be scanned to confirm the content's authenticity and origin. Watermarking strategies can be broadly categorized into two approaches: Atallah et al. [2001], Brassil et al. [1995], Jalil and Mirza [2009], Topkara et al. [2006], Abdelnabi and Fritz [2021]

- **Rule-Based Approaches:** These strategies involve embedding watermarks based on a set of predefined rules or patterns. This could include altering certain aspects of text in subtle ways or incorporating specific, detectable anomalies in digital images or audio that do not significantly affect the perceived quality. Rule-based methods are straightforward to implement but may be more susceptible to detection and removal by adversaries.
- **Neural-Based Approaches:** Leveraging the capabilities of neural networks, these approaches embed watermarks in a manner that is intrinsically linked to the content's generation process. For instance, a neural network might be trained to include specific, imperceptible patterns in the output that can be recognized by another model. This method is less obvious and more difficult for attackers to identify and erase without degrading the content's quality.

4.2 Challenges in Implementing White Box Detection

Implementing white box detection methods, particularly watermarking, presents several challenges:

- **Open-Source Models:** The open nature of these models offers an inherent challenge to the implementation of white box detection techniques. With access to the model's architecture and training data, adversaries might find ways to bypass or remove watermarks, necessitating a delicate balance between openness and security.

- **Proprietary Constraints:** In contrast, models developed and held by private entities may incorporate sophisticated watermarking techniques but face challenges in widespread adoption due to intellectual property concerns and the lack of transparency in their operation and effectiveness.

- **Technical Complexity:** The development and deployment of effective watermarking techniques require significant technical expertise and resources, potentially limiting their use to organizations with substantial R&D capabilities.

- **Adversarial Attacks:** As with any security measure, watermarking and other white box detection techniques are subject to adversarial attacks. Adversaries continually develop methods to detect, alter, or remove watermarks, requiring ongoing innovation and adaptation of watermarking technologies.

Despite these challenges, white box detection, particularly through watermarking, offers a promising avenue for authenticating and safeguarding digital content against unauthorized or malicious use. As AI technologies continue to evolve, so too will the methods for detecting and protecting against synthetic content, highlighting the importance of continued research and development in this field.

5.0 Techniques for Authenticating and Labeling Synthetic Content

Authentication and labeling are pivotal in distinguishing authentic from synthetic content, tracing origins, and ensuring digital integrity. Among the various strategies employed, watermarking stands as a primary technique, complemented by digital signatures and metadata for enhanced security and traceability.

5.1 Watermarking

Previously discussed in the context of white box detection, watermarking is crucial for embedding invisible markers in content to verify authenticity. This technique, adaptable across text, images, video, and audio, is detailed in the preceding sections. It provides a robust mechanism for content provenance and integrity verification, leveraging imperceptible changes that can withstand tampering attempts. For a comprehensive understanding of watermarking's applications and advantages, refer to the dedicated section on this topic.

5.2 Digital Signatures and Metadata for Content Provenance

Recommendation

Adopt digital signatures and metadata for content provenance in general applications, recognizing their limitations against malicious use, and seek further solutions for enhanced security.

Beyond watermarking, digital signatures and metadata offer additional layers for securing and tracing digital content, each playing a unique role in the content authentication landscape.

5.2.1 Digital Signatures

Digital signatures employ cryptographic protocols to affirm the creator's identity and content's unaltered state post-signature. This method encrypts creator and content information with a private key, ensuring a secure and verifiable link between the digital artifact and its origin.

5.2.2 Utilizing Metadata

Metadata provides detailed context about the content, including creation details, authorship, and any subsequent modifications. This auxiliary data aids in mapping the lifecycle of content, enhancing transparency and aiding in origin tracing without intruding into the content itself, as watermarking does.

5.2.3 Challenges in Security and Privacy

While both digital signatures and metadata significantly contribute to the authenticity and traceability of content, they are not without challenges. The infrastructure supporting digital signatures and metadata can be susceptible to attacks, risking the compromise of content verification processes. Additionally, the collection and handling of metadata, along with the deployment of digital signatures, must navigate privacy implications, ensuring the protection of creator identities and preventing unauthorized data exploitation.

In synthesizing watermarking with digital signatures and metadata, a multifaceted approach to content authentication and labeling emerges. This strategy not only underlines the importance of technological diversity in protecting digital content but also highlights the ongoing necessity to address and mitigate associated security and privacy challenges. By advancing cryptographic and data protection methodologies, the goal of robustly authenticating and labeling synthetic content becomes increasingly attainable.

6.0 Challenges, Trade-offs, and Economic Feasibility

Recommendation

Focus on promoting the adoption of tracking solutions that are easily implementable, such as watermarks, into popular generative AI services. While incomplete, this is a pragmatic path towards providing traceability of a large quantity of generated content.

Integrating detection and authentication techniques, particularly watermarking, into synthetic content generation processes highlights significant differences in deployment complexity and costs between centralized models and decentralized or open-source environments.

6.1 Centralized Model Deployment

In a centralized model scenario, where a single model serves multiple consumers, deploying watermarking techniques can be relatively straightforward and cost-effective. This model offers several advantages:

- **Streamlined Implementation**: Watermarking can be integrated directly into the content generation pipeline, ensuring that all outputs carry the necessary markers for later authentication. This centralized approach allows for uniform application of watermarking techniques, simplifying maintenance and updates.
- **Economies of Scale**: The costs associated with developing and deploying watermarking technologies are amortized over a large volume of content and users, making it more economically viable for the entity operating the model.
- **Controlled Environment**: A centralized model affords better control over the watermarking process, including the ability to update and refine detection methods in response to emerging threats.

Even in centralized systems, challenges remain, particularly in ensuring that watermarks are robust against extraction or manipulation without compromising content quality.

6.2 Decentralized and Open-Source Models

Contrasting sharply with centralized systems, decentralized models, including small, specialized purpose models and open-source projects, present unique challenges for watermarking:

- **Increased Complexity and Costs**: The fragmented nature of decentralized systems means that watermarking must be implemented separately for each model, increasing the complexity and associated costs of deployment. This is exacerbated in open-source environments, where the model's transparency might aid adversaries in circumventing watermarking techniques.
- **Lack of Uniformity**: Each model may require a tailored approach to watermarking, considering its specific outputs and use cases. This lack of uniformity complicates the development of a one-size-fits-all solution, requiring more bespoke implementations that can drive up costs.

- **Resource Constraints**: Smaller projects and open-source models often operate with limited resources, making the investment in sophisticated watermarking and authentication technologies challenging to justify or sustain.

- **Open-Source Model Vulnerabilities**: The openness of these models, while fostering innovation and accessibility, also exposes them to a greater risk of manipulation by malign actors. The very transparency that is a hallmark of open-source projects can undermine efforts to secure content against unauthorized alterations or counterfeit reproductions.

6.3 Navigating Economic and Operational Trade-offs

The economic and operational implications of deploying watermarking and other authentication techniques across different model architectures highlight a key trade-off between ease of implementation and the breadth of coverage. While centralized models offer a more controlled and potentially cost-effective environment for deploying these techniques, decentralized and open-source models face significant barriers in terms of complexity, cost, and vulnerability to attacks.

This divergence necessitates a strategic approach to content authentication, where the choice of technique and deployment method is closely aligned with the model's architecture and operational context. For decentralized and open-source models, collaborative efforts within the community, alongside advancements in lightweight and adaptable watermarking technologies, may provide a pathway to more effective and economically feasible solutions.

In summary, the distinction between centralized and decentralized model deployments significantly influences the practicality and economics of implementing synthetic content detection and authentication methods. Addressing these challenges requires not only technological innovation but also a nuanced understanding of the operational and economic landscapes in which these models operate.

7.0 Considerations for Different AI Applications

The application of content authentication and detection techniques must be tailored to the specific vulnerabilities and requirements of different AI applications:

- **Text and Natural Language Processing**: For AI systems specializing in text generation or natural language processing, the focus should be on detecting linguistic and statistical anomalies that indicate synthetic manipulation.

- **Image and Video Generation** AI applications involved in creating or manipulating visual content require sophisticated image analysis tools capable of identifying subtle manipulations and embedded watermarks that signal authenticity.

- **Audio and Speech Synthesis** In the realm of audio and speech synthesis, detection techniques must be able to discern alterations in sound waves and vocal patterns that may not be perceptible to the human ear but indicate synthetic origins.

By addressing these considerations and integrating content authentication and detection strategies throughout the AI development and deployment lifecycle, organizations can significantly enhance their resilience against the misuse of synthetic content. This comprehensive approach not only safeguards the integrity of digital artifacts but also reinforces public trust in AI technologies and their applications.

8.0 Preventing the Generation of Illegal Content

As AI technologies continue to evolve, there is an increasing concern over their potential to generate illegal or harmful content. This final section of the report focuses on strategies to prevent AI models, particularly language models, from producing illegal content. Key prevention measures include filtering training data, fine-tuning models with safe datasets, and implementing classifiers to scan generated content. These strategies are essential for maintaining the integrity and legality of AI-generated content.

8.1 Filtering Training Data

The first line of defense against the generation of illegal content lies in carefully curating and filtering the training data used to develop AI models. This process involves identifying and removing explicit, illegal, or harmful content from the datasets before they are fed into AI systems for training. The effectiveness of this measure largely depends on the thoroughness of the dataset curation process and the criteria used to define what constitutes illegal or harmful content. However, this approach can be resource-intensive and might not catch all potential sources of harmful content.

8.2 Fine-Tuning Models with Safe Datasets

Fine-tuning is a process where pre-trained models are further trained (or fine-tuned) on a smaller, highly curated dataset that emphasizes safe and appropriate content. This method helps the model learn context and content boundaries that are acceptable, reducing the likelihood of generating illegal content. Fine-tuning allows model developers to steer the model's outputs towards more desirable content, adhering to legal and ethical standards.

8.3 Implementing Classifiers for Content Checking

Deploying classifiers that can detect and flag illegal or harmful content generated by AI models is another crucial step. These classifiers can be trained to identify various forms of illegal content, from hate speech and threats to copyright infringement. When integrated into the AI content generation pipeline, these classifiers can automatically review and filter out inappropriate outputs, serving as a real-time safeguard against the dissemination of harmful material.

8.4 Challenges with Open Source Models

Applying these prevention measures to open-source models presents unique challenges. Once an open-source model is released, external parties can modify it, potentially reintroducing or adding capabilities for generating illegal content. This open nature makes it difficult to control or enforce content safety measures effectively. There's a risk that malicious actors might use these open-source models as a base to develop systems capable of generating harmful content, bypassing the original safety mechanisms.

8.5 Effectiveness for Centralized Providers

For centralized AI service providers, the outlined prevention measures are significantly more effective. Centralized systems allow for more controlled environments where training data can be carefully curated, models can be fine-tuned with safe datasets, and content classifiers can be consistently applied and updated. These providers have the infrastructure to enforce strict content guidelines and the capability to quickly respond to any instances of illegal content generation, ensuring compliance with legal standards and mitigating the risk of harm.

9.0 Conclusion

This report has explored the evolving landscape of synthetic content creation and the critical need for effective detection, authentication, and labeling techniques. As AI technologies advance, the capability to generate realistic digital content that blurs the line between authentic and synthetic poses significant challenges to information integrity and public trust. The key findings from our exploration underscore the complexity of detecting synthetic content, the necessity of implementing robust authentication methods, and the importance of continuous innovation to stay ahead of malign actors.

9.1 Summary of Key Findings

- **Detection and Authentication Techniques**: The development and deployment of both black box and white box detection methods, alongside watermarking, digital signatures, and metadata, are essential for identifying and authenticating synthetic content. Each approach has its strengths and challenges, with the effectiveness often contingent on the type of content and the sophistication of adversarial tactics.

- **Economic and Operational Considerations:** The adoption of these techniques varies in economic feasibility and operational complexity across different organizations and applications, highlighting the need for scalable, cost-effective solutions.

- **Resilience and Adaptability:** Ensuring the resilience of detection systems against evasion and manipulation requires continuous monitoring, testing, and updates. The dynamic nature of AI-generated content necessitates adaptable and forward-looking strategies to mitigate risks.

9.2 Call to Action

The findings from this report serve as a call to action for stakeholders across industry, academia, and regulatory bodies to collaborate on the development of effective and adaptable solutions for authenticating and detecting synthetic content. This collaboration is crucial for advancing research, sharing knowledge, and establishing standards that can guide the creation and dissemination of digital content. By working together, we can create a framework that not only addresses current challenges but also anticipates future developments in AI and synthetic content generation and usage.

9.3 Balancing Innovation with Integrity and Trust

As we reflect on the ongoing challenge of balancing innovation in AI with the imperative to safeguard information integrity and public trust, it becomes clear that the path forward is not solely technological. It also requires ethical considerations, public awareness, and regulatory oversight to ensure that advancements in AI serve the greater good. The potential of AI to enrich and enhance our lives is immense, but so is the responsibility to use it wisely, ethically, and with a commitment to transparency and accountability.

In conclusion, the journey toward securing digital content authenticity is ongoing, with new challenges and opportunities emerging as technology evolves. By fostering a collaborative ecosystem that embraces both innovation and ethical stewardship, we can navigate these complexities and build a digital future that upholds the highest standards of integrity and trust.

10.0 About Ozni AI

Ozni AI is a U.S.-based applied R&D organization that specializes in the application of AI/ML into mission-relevant areas. We have established a strong working relationship with Northrop Grumman, particularly in the topics of Trust in AI and AI Safety, demonstrating our commitment to developing and implementing AI technologies with a keen focus on safety and ethical considerations.

Our technical expertise was recently showcased as a winner in the U.S. General Services Administration (GSA) Applied AI competition in 2023 with our development of "Jibber Jabber." Jibber Jabber enables users with limited Radio Frequency (RF) domain knowledge to make natural language requests to collect, process, and interpret RF data. This innovation opens the RF landscape to a diverse array of use cases, including disaster response management, emergency services coordination, missing person investigations, law enforcement, and scientific research. Currently, Jibber Jabber and our other innovative solutions are under evaluation by Northrop Grumman for applications in Airborne SIGINT and space-based commercial RF.

References

Sahar Abdelnabi and Mario Fritz. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 121–140. IEEE, 2021.

Mikhail J Atallah, Victor Raskin, Michael Crogan, Christian Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In *Information Hiding: 4th International Workshop, IH 2001 Pittsburgh, PA, USA, April 25–27, 2001 Proceedings 4*, pages 185–200. Springer, 2001.

Paras Bhatt and Anthony Rios. Detecting bot-generated text by characterizing linguistic accommodation in human-bot interactions. *arXiv preprint arXiv:2106.01170*, 2021.

Jack T Brassil, Steven Low, Nicholas F. Maxemchuk, and Lawrence O'Gorman. Electronic marking and identification techniques to discourage document copying. *IEEE Journal on Selected Areas in Communications*, 13(8):1495–1504, 1995.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Jennifer C Lai, and Robert L Mercer. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40, 1992.

Leon Fröhling and Arkaitz Zubiaga. Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. *PeerJ Computer Science*, 7:e443, 2021.

Matthias Gallé, Jos Rozen, Germán Kruszewski, and Hady Elsahar. Unsupervised and distributional detection of machine-generated text. *arXiv preprint arXiv:2111.02878*, 2021.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*, 2019.

Zunera Jalil and Anwar M Mirza. A review of digital watermarking techniques for text documents. In *2009 International Conference on Information and Multimedia Technology*, pages 230–234. IEEE, 2009.

Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21:1112–1130, 2014.

Juan Diego Rodriguez, Todd Hay, David Gros, Zain Shamsi, and Ravi Srinivasan. Cross-domain detection of gpt-2-generated technical text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1213–1233, 2022.

Umut Topkara, Mercan Topkara, and Mikhail J Atallah. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions. In *Proceedings of the 8th workshop on Multimedia and security*, pages 164–174, 2006.

Wanjun Zhong, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. Neural deepfake detection with factual structure of text. *arXiv preprint arXiv:2010.07475*, 2020.