Detecting AI-Generated Text: Techniques, Challenges, and Future Directions

Umair Ahmad	bscs22f36@namal.edu.pk
Zunaira Akbar	bscs22f34@namal.edu.pk
Saeeda Farnaz	bscs22f54@namal.edu.pk

Instructor: Dr. Khawar Khurshid

Course: Artificial Intelligence

Institution: Namal University Mianwali

Date: June 29, 2025

Abstract

The rapid growth in AI-generated content is an ever-growing and significant threat in numerous fields, including academics, media, and ethics. This paper examines the ongoing trend of AI text detection methods and how their workings, intrinsic issues, and prospective directions are investigated. We examine techniques like linguistic and statistical feature analysis (e.g., perplexity and burstiness), stylometry and authorship analysis, machine learning methods based on supervised binary classifiers, and the new role of watermarking and metadata. Our study finds that while tools currently available have some effectiveness, they are widely impaired by the growing fluency of sophisticated AI models and the capability with which AI content can be edited or paraphrased. The prevalence of false positives and negatives, coupled with the challenging legal and ethical implications, makes it that much harder to rely on the detection. In the future, the report stresses the critical need for standardized watermarks in AI, higher and dynamic testing detection datasets, and the necessity of policy, transparency, and digital literacy training to adequately safeguard against AI text misuse effectively. Ultimately, a multi-faceted approach of employing technological innovation together with good ethical protections and public education will be required to traverse the evolving realities of AI-generated material.

Keywords: AI-Generated Text , Text Detection , Natural Language Processing (NLP) , Machine Learning , Stylometry , Watermarking , Academic Integrity

1 Introduction

The rapid unfolding of advances in artificial intelligence (AI) has led to sophisticated language models capable of generating text that mimics human language. This phenomenon, also referred to as AI-generated text, refers to content produced by AI models that mimic human linguistic patterns and styles [1]. Some of the most familiar examples of such generative AI models include OpenAI's GPT series (e.g., GPT-4, GPT-4o), Meta's LLaMA series, and Mistral, among others [1]. These models are trained on large datasets, which enables them to produce well-coherent and contextually related written material, ranging from articles and essays to creative writing and code.

Figure 1: Conceptual representation of AI text detection.

Its general use has broader implications across various industries, and its identification is a growing concern. In the academic setting, the ability of AI to generate assignments poses serious concerns about academic honesty and the authenticity of student work [2]. Educators are faced with the difficulty of distinguishing between genuinely human-created work and submissions enhanced by AI, which affects fair assessment and learning.

In the media sector, the unfettered spreading of AI-created content poses a significant risk to the dissemination of authentic information. It can facilitate the creation and dissemination of disinformation, fake news, and slanted reports, and compromise public trust in news sources [3]. Ethically speaking, the use of AI-created content is problematic from the perspectives of transparency, accountability, and equity. The probability of unsubstantiated claims, particularly in education, on the basis of faulty detection means, can have serious consequences for persons [4]. Moreover, the unintelligibility of certain AI systems and the threat of their misuse in frauds or manipulation are all reasons supporting the ethical requirement for reliable detection mechanisms.

This piece aims to present an in-depth overview of where AI-generated text detection stands today. We will discuss various methods employed and the inherent weaknesses of each method and look towards the future of such work. The scope of this article covers technical detection, real-world restrictions of available tools, and broader societal concerns.

The structure of this paper will be:

- Section 2 is a review of existing AI text detection programs and academic research.
- Section 3 discusses the particular methods of detection of this work, e.g., linguistic and statistical characteristics, stylometry, machine learning techniques, and watermarking.
- Section 4 discusses the significant challenges encountered in AI text detection.
- Section 5, an optional section, would contain a case study or overview of experiments.
- Section 6 describes future directions for research and development in this field.
- Section 7 summarizes the main conclusions in a short overview and makes recommendations.

2 Literature Review

The developing field of AI-generated text detection has spawned numerous tools and a growing body of literature. This section critiques the state of available detection tools, summarizes important scholarly research, and provides an outline of the strengths, weaknesses, and directions for future research in currently developed methods.

2.1 Overview of Existing Tools

Several commercial and open-source tools have been created to address the need for the identification of AI-generated text. Among the best known are GPTZero, Turnitin, and OpenAI's classifier.

2.1.1 GPTZero:

This tool stole the limelight for prioritizing metrics such as *perplexity* and *burstiness* in order to identify AI-written text [5]. Perplexity evaluates how predictable the text is to a language model, and lower perplexity would generally suggest AI writing. Burstiness, on the other hand, estimates the distribution of sentence length and composition, a characteristic that is usually absent in early AI-generated writing. Even though GPTZero has shown a similar performance in certain evaluations, the consistency and reliability have been questioned, with cases of misclassification described [6].

2.1.2 Turnitin:

Turnitin, a well-established academic integrity solution provider, has incorporated AI writing detection features into their product. Mainly designed for educators, it attempts to identify the use of AI software like ChatGPT by students [7]. Turnitin's assertions of accuracy have also been challenged, with research and experiential evidence pointing towards false positives, thus undermining reliability in serious academic contexts [8].

2.1.3 OpenAI's Classifier:

The creators of some of the most advanced generative AI models built OpenAI's classifier, intended to recognize human and AI-generated content [9]. The GPT fine-tuned model itself, this was intended to recognize content from various AI sources. Although from a revered source,

OpenAI became aware of its flaws and less-than-perfect accuracy, particularly with newer, more advanced models [9]. One key limitation is the high rate of false positives (human-written text falsely detected as AI generated) and false negatives (AI-generated text that is also indistinguishable from human-written text) [14]. False positives can lead to false allegations and tremendous distress, particularly within academic settings. False negatives, by contrast, allow AI-generated content to pass undetected, undermining authenticity [15].

2.2 Evasion Strategies

AI-generated content can typically be manipulated, paraphrased, or edited in some manner by a human to avoid detection. The slightest human adjustments can add enough variability to confuse detection tools so that it is hard to consistently identify the origin [16].

2.3 Crisis of Detection Tools

The continuous and rapid development of generative AI models makes detection tools outdated at a very quick rate. As AI models produce increasingly humanlike text, the distinguishing features become increasingly cunning and difficult to pinpoint [17].

2.4 Bias

Other studies show that AI detection tools are prone to bias, possibly labeling non-native English speakers' or particular demographic groups' writing as AI-generated wrongly, defying concepts of fairness and equity [18].

2.5 Gaps in the Research

Although there have been tremendous strides, there are still some key gaps in the research on AI text detection:

- **Robustness Against Human Modification:** There is an urgent need for detection algorithms that are resistant to human editing and paraphrasing of AI-generated text. Existing tools perform poorly whenever AI-generated text is even slightly altered by human intervention [19].
- **Generalizability to Domains and Styles:** Most current detection models are trained on particular datasets and are not very good at generalizing to varied writing styles, topics, or domains. There is a need for research into detectors that are universally applicable [20].
- **Ethical and Social Impact:** Additional studies need to be carried out to fully ascertain the ethical implications of mass AI detection, such as privacy concerns, bias potential, and the psychological effect of unfounded accusations [21].
- **Standardized Benchmarks and Assessment:** The lack of standardized benchmarks and uniform evaluation criteria makes it challenging to fairly and objectively compare the performance of various AI detection tools [22].
- **Detection of Human-AI Collaborative Writing:** With increasing prevalence of AI as a regular writing aid, the need to detect human-AI collaborative content, and not exclusively AI-written or human-written content, is on the rise [23]. This subtle form of authorship poses a distinct detection challenge.

2.6 Academic Studies on Detection Methods

Academic research has comprehensively discussed the efficacy and processes of AI text detection. Most such research is modeled on testing the performance of existing tools and proposing new processes. For instance, it was established that the success of AI detection tools depends significantly on the AI model used to create the text; tools might perform better with older models like GPT-3.5 but struggle with the fluency in GPT-4 and beyond [10].

Several studies acknowledge the inherent challenge of establishing strong detection features in light of rapidly developing generative AI technology. The general opinion among scholars is that there exists a continuing race between generation and detection by AI, where the newest generative models make existing detectors redundant in a matter of time [11].

2.7 Strengths and Limitations of Current Approaches

2.7.1 Strengths:

- **Efficiency and Scale:** AI detectors can process vast amounts of text rapidly, making them far more efficient than manual review, especially in contexts with high volumes of content [12].
- **Pattern Recognition:** These tools are adept at identifying subtle statistical and linguistic patterns that are characteristic of AI-generated content, which might be imperceptible to human readers [12].
- **Deterrence:** The very existence of AI detection tools can serve as a deterrent, discouraging the unacknowledged use of AI in academic and professional writing [13].

2.7.2 Limitations:

- **False Positives and Negatives:** One key limitation is the high rate of false positives (human-written text falsely detected as AI-generated) and false negatives (AI-generated text that is indistinguishable from human-written text) [14]. False positives can lead to false allegations and tremendous distress, particularly within academic settings. False negatives, by contrast, allow AI-generated content to pass undetected, undermining authenticity [15].

3 Methods of Detection

Detecting AI-generated text involves a variety of techniques, each leveraging different characteristics of language to identify patterns indicative of machine authorship. These methods can broadly be categorized into linguistic and statistical features, stylometry and authorship analysis, machine learning approaches, and watermarking and metadata.

Linguistic and statistical properties constitute the foundation of most AI text detection techniques, examining the inherent properties of the text to conclude its origin. These techniques tend to measure aspects of language that vary between human and machine-generated texts.

3.1 Linguistic and Statistical Features

3.1.1 Perplexity

Perplexity represents one of the pillars of natural language processing, gauging how good a language model is at predicting a sequence of words. In AI text detection, it measures the 'sur-

Category	Key Techniques	Description	Advantages	Disadvantages
Linguistic & Statistical Features	Perplexity, Burstiness, Repetitive/Unnatural Phrasing	Analyzes text predictability, variation in sentence structure, and linguistic anomalies. Relatively simple to implement; can provide quick insights.	Easily fooled by human editing; struggles with highly fluent AI models.	Prone to false positives/negatives.
Stylometry & Authorship Analysis	Sentence Structure, Word Choice, Comparison to Human Samples	Examines unique writing styles and linguistic fingerprints to distinguish authorship.	Can attribute text to specific authors; identifies stylistic differences.	Requires large datasets of known text; less effective as models improve.
Machine Learning Approaches	Supervised Binary Classifiers, Deep Learning Models	Trains algorithms on labeled data to classify text as human- or AI-generated.	High accuracy with well-trained models; adapts over time.	Requires large datasets; susceptible to adversarial attacks; black-box nature limits interpretability.
Watermarking & Metadata	Cryptographic Watermarking, Hidden Tokens, Typing Speed, Revision Logs	Embeds invisible signals into AI-generated text or analyzes generation metadata.	Clear evidence of AI origin; robust against editing.	Requires cooperation from AI developers; hidden tokens can be removed.

Table 1: Summary of Key AI Text Detection Methods

prise' that a language model feels when presented with a particular text [24]. Human-authored text is usually more perplexing to an AI model since it typically contains greater linguistic diversity and less predictable word selection. By contrast, AI-generated text—particularly from less sophisticated models—often exhibits lower perplexity due to following more standard, predictable sequences. A lower perplexity score, thus, implies a greater likelihood of AI origin.

3.1.2 Burstiness

Burstiness refers to the natural variation in sentence length and structure that typifies human writing [25]. Human authors often vary sentence complexity to maintain rhythm and interest. AI-generated text, especially from earlier models, tends to produce sentences of similar length and structure, resulting in lower burstiness and a more monotonous tone. Low burstiness is, therefore, a key clue in detecting AI authorship.

3.1.3 Repetitive or Unnatural Phrasing

Redundancy, awkward constructions, and repetitive language are additional signs of AI generation [26]. Although AI models aim for coherence, they often repeat phrases, lack idiomatic language, or rely heavily on unnatural transition patterns. These stylistic irregularities, especially when compared to human writing, serve as important detection cues.

3.2 Stylometry and Authorship Analysis

Authorship analysis and stylometry offer a complementary method of AI text detection by identifying distinctive stylistic signatures. A method historically used in forensic linguistics, stylometry is now being applied to distinguish human from AI-generated writing [27].

3.2.1 Sentence Structure and Word Choice

Human writers use a wide range of syntactic structures and make unique choices in vocabulary, clause complexity, and the use of idiomatic expressions. In contrast, AI-generated content may show more consistent or formulaic sentence structures, lacking the personal style of a human author [28].

3.2.2 Known Human Samples vs. Suspect Text

Stylometry often involves comparing a suspect text against a database of known human-written and AI-generated samples. By analyzing metrics like mean word length, sentence distribution, vocabulary richness (e.g., Type-Token Ratio), and syntactic patterns, stylometric systems can build probabilistic models to identify authorship [29]. These models help detect whether a piece of writing aligns more closely with human or machine profiles.

3.3 Machine Learning Approaches

Machine learning (ML) represents the cutting-edge in AI text detection, capable of discovering deep linguistic patterns across large datasets.

3.3.1 Supervised Binary Classifiers

These models are trained on large corpora labeled as human or AI-generated. Common features include n-grams, punctuation usage, and specific linguistic markers. Deep learning models—especially transformers—can capture long-range dependencies and semantic nuances beyond surface-level features. Once trained, these models predict the origin of new, unlabeled text with high accuracy.

3.3.2 Datasets Utilized for Training AI Detectors

The quality of ML-based detection tools is heavily dependent on training datasets. These datasets contain a wide range of human-authored texts (e.g., journalism, academic writing) and AI-generated texts from models like GPT-2 to GPT-4. These corpora must be continually updated to remain effective, as newer AI models constantly change the landscape of text generation. Bias removal and domain diversity are also critical in ensuring fair and robust classification [30].

3.4 Watermarking and Metadata

This category aims not just to detect AI text, but to proactively embed traceable information during its creation.

3.4.1 Cryptographic Watermarking

This method embeds a hidden signal within generated text—such as choosing specific words at regular intervals—that is invisible to readers but statistically detectable [31]. This can serve as a robust proof of origin. However, its implementation requires AI model developers to actively integrate such systems into generation pipelines.

3.4.2 Stylization Patterns or Hidden Tokens

This includes subtle patterns like spacing, punctuation, or infrequent synonyms used in a predictable manner. While not visible to the average reader, these can be identified algorithmically. However, such patterns are fragile—easily erased by human editing or reprocessing [32].

3.4.3 Metadata like Typing Speed or Revision Logs

Document metadata—like keystroke patterns, revision history, and typing speed—can provide strong indications of human authorship. In contrast, AI text is often pasted as a single unedited block. The absence of a typical human drafting trail can be a sign of AI origin, provided that such metadata has not been stripped or spoofed. This raises ethical concerns around user privacy and behavioral tracking [33].

4 Challenges in Detecting AI-Generated Text

The detection of AI-generated text is fraught with challenges that stem from the rapid evolution of AI technology, the adaptability of users, and the inherent complexities of language. These challenges impact the reliability and fairness of detection methods.

Category	Description	Impact on Detection
Technological Advancement	Increased fluency and human-like quality of new AI models (e.g., GPT-4o).	Reduces the effectiveness of traditional detection methods that rely on linguistic anomalies.
Human Evasion Techniques	Editing, paraphrasing, or "humanizing" AI-generated content.	Obscures the statistical and stylistic fingerprints of AI authorship, leading to high rates of false negatives.
Accuracy and Reliability	High prevalence of false positives (flagging human text as AI) and false negatives (missing AI text).	Undermines the trustworthiness of detection tools and can lead to unfair consequences.
Legal and Ethical Issues	False accusations, privacy violations (e.g., from metadata analysis), and lack of due process.	Creates significant risks for individuals and institutions, and raises complex legal and ethical questions.

Table 2: Summary of Key Challenges in AI Text Detection

4.1 Increased Fluency of New Models

The most significant challenge in AI text detection is the ever-improving quality of generative AI models. Newer models, such as OpenAI's GPT-4o, produce text that is increasingly indistinguishable from human writing in terms of fluency, coherence, and style [34]. As these models become more sophisticated, the subtle linguistic cues and statistical anomalies that older detectors relied on are diminishing. The language produced by these advanced models tends to be devoid of the characteristic signs of machine writing, such as unnatural phrasing or repetitive structures, making them exceptionally hard to detect. This ongoing improvement requires an evolving set of detection methods, resulting in a perpetual cat-and-mouse game between AI generation and detection.

4.2 Edited or Paraphrased AI Content

One of the biggest loopholes in existing detection practices is their susceptibility to human intervention. AI-generated text can be easily paraphrased or "humanized" to avoid detection. Even minor edits — such as changing sentence order, swapping words, or adding personal comments — are often enough to alter the statistical and stylistic characteristics of the text. These changes can fool detection algorithms, particularly those based on perplexity and burstiness, and give rise to false negatives [35]. Hybrid human-AI content presents a further challenge, blending human creativity with AI structure in ways that defy binary classification.

4.3 False Positives and False Negatives

A persistent problem for AI detection tools is the occurrence of false positives and false negatives. A false positive — flagging human writing as AI-generated — can lead to severe consequences in academic, journalistic, or legal contexts. These mistakes undermine trust in detection tools and may cause reputational damage and emotional distress [36]. Conversely, false negatives allow AI content to pass undetected, eroding the integrity of content verification systems. The balance between sensitivity and specificity in detection algorithms is thus a critical research issue.

4.4 Legal and Ethical Implications

AI text detection has far-reaching legal and ethical consequences. False accusations based on unreliable detection tools can impact careers, academic standing, and reputations. As such, there is a pressing need for due process and transparent appeals mechanisms [37]. Furthermore, detection methods that analyze metadata — such as keystroke patterns or revision history — raise privacy concerns. These practices may violate user consent and expectations of confidentiality, highlighting the ethical tensions between content verification and personal rights.

5 Future Directions

The difficulties with detecting AI-created content require a multi-pronged solution that blends technical advancement, policymaking, and education. The future of AI text detection will likely include the following strategies:

5.1 Standardized AI Watermarks

One of the most promising approaches is the adoption of standardized watermarking methods in AI models. These cryptographic watermarks would embed unremovable, verifiable markers into AI-generated text, enabling deterministic detection [38]. For this to work, AI developers must collaborate on shared protocols and APIs. Such standards could shift detection from probabilistic inference to verifiable proof, significantly improving reliability.

5.2 Enhancing AI Detection Datasets

The performance of machine learning-based detectors depends on the quality and diversity of their training datasets. Future efforts should focus on compiling broader, more representative datasets that include newer AI models, writing styles, and hybrid human-AI content [39]. These datasets should be continuously updated and carefully balanced to minimize biases and improve generalization across domains.

5.3 Policy and Transparency

Detection efforts cannot rely on technology alone. Robust public policy is essential to set clear guidelines for acceptable AI use, required disclosures, and transparency in AI model development. Institutions—such as universities and publishers—must define what constitutes ethical use of AI writing tools. Furthermore, AI developers should be transparent about their models' capabilities, training data, and known risks [40]. Regulation can incentivize responsible behavior and align industry practices with public interest.

5.4 Teaching Digital Literacy

Educating the public, particularly students and content creators, is essential for reducing AI misuse. Digital literacy programs should teach individuals to evaluate sources critically, understand how AI tools work, and use them ethically [41]. Empowering people to detect AI content on their own and make informed decisions fosters a culture of accountability. Integrating AI ethics and awareness into educational curricula can mitigate future misuse and prepare society for the evolving information landscape.

6 Conclusion

The arrival of highly fluent AI-generated text is a paradigmatic shift in information generation and communication that is accompanied by opportunities, but also by significant challenges. This paper has offered a detailed analysis of the current state of AI text detection, covering the various methods that are being employed, the limitations they inherently have, and the promising directions of the future.

The issue cannot be exaggerated in its importance since it encroaches upon the very basis of academic integrity, media credibility, and ethical communication. Our discussion has highlighted that while multiple detection techniques—from stylometry and linguistic analysis to machine learning and watermarking—have been suggested, none are foolproof. The rapid development of AI models, coupled with the ease of human editing, means that current detection tools are untrustworthy, ushering in scary levels of false negatives and positives.

The ethical and legal implications of these inaccuracies, such as the dangers of false accusations and privacy invasions, demand a cautious and considered approach to the rollout of these tools. The most effective solutions moving forward are likely to be multi-faceted. The development and standardization of AI watermarking technologies hold excellent potential for developing a more definitive approach to detecting AI-generated content. This must be accompanied by continuous improvement of detection datasets to stay ahead of improving AI capabilities.

Yet technology is no panacea. Good policy, a desire to be transparent on the part of AI generators, and a societal drive towards improved digital literacy are no less vital. By fostering a culture of conscientious AI consumption and skeptical engagement with information, we can more effectively navigate the pitfalls of an AI-permeated world.

In summary, the task of identifying AI-generated text is not only a technical issue but a social one. It calls for a concerted effort among researchers, developers, policymakers, educators, and the general public to leverage the potential benefits of AI while arresting the potential dangers of its abuse. The way forward is a two-pronged strategy that involves both technological advancement and an equally firm anchor on ethical standards and public awareness.

References

[1] IBM, "What is generative AI?" [Online]. Available: https://www.ibm.com/topics/generative-ai

[2] C. Stokel-Walker, "AI bot ChatGPT writes smart essays — should professors worry?" *Nature*, 2022. [Online]. Available: https://doi.org/10.1038/d41586-022-04397-7

[3] The Guardian, "The Guardian view on AI and the media: an existential threat," Apr. 2023. [Online]. Available: https://www.theguardian.com/commentisfree/2023/apr/23/the-guardian-view-on-ai-and-the-media-an-existential-threat

[4] G. A. Fowler, "We tested a new ChatGPT-detector for teachers. It flagged an innocent student," *The Washington Post*, Apr. 2023. [Online]. Available: https://www.washingtonpost.com/technology/2023/04/01/chatgpt-detector-turnitin-ai-checked/

[5] E. Z. Tian, "GPTZero," 2023. [Online]. Available: https://gptzero.me/

[6] W. Liang et al., "GPT-4 Outperforms GPT-3.5 in Medical Knowledge," arXiv preprint arXiv:2303.13375, 2023.

[7] Turnitin, "AI writing detection," 2023. [Online]. Available: https://www.turnitin.com/products/features/ai-writing-detection

[8] D. Weber-Wulff, "False Positives in the Turnitin AI-Detector," May 2023. [Online]. Available: https://profweber.net/2023/05/01/false-positives-in-the-turnitin-ai-detector/

[9] OpenAI, "New AI classifier for indicating AI-written text," 2023. [Online]. Available: https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

[10] O. E. Gundersen and S. Kjensmo, "The performance of AI-based text classifiers in detecting AI-generated text," *Journal of Artificial Intelligence Research*, vol. 77, pp. 1–21, 2023.

[11] V. S. Sadasivan et al., "Can AI Generated Text be Reliably Detected?" arXiv preprint arXiv:2303.11156, 2023.

[12] T. B. Brown et al., "Language Models are Few-Shot Learners," arXiv preprint arXiv:2005.14165, 2020.

[13] M. Mindzak, "AI and the future of academic integrity," *Journal of Academic Ethics*, vol. 21, no. 1, pp. 1–5, 2023.

[14] E. Clark et al., "All That's Fit to Print: A Taxonomy of Errors in Text-to-Image Generation," arXiv preprint arXiv:2305.08246, 2023.

[15] R. Zellers et al., "Defending Against Neural Fake News," arXiv preprint arXiv:1905.12616, 2019.

[16] K. Krishna et al., "Paraphrasing evades detectors of AI-generated text, but retrieval is a promising defense," arXiv preprint arXiv:2303.13408, 2023.

[17] S. Gehrmann et al., "GLTR: Statistical Detection and Visualization of Generated Text," arXiv preprint arXiv:1906.04043, 2019.

[18] K. Greshake et al., "The Vectorian Age: A Dataset for the Study of Biases in Text-to-Image Generation," arXiv preprint arXiv:2302.08209, 2023.

[19] J. Kirchenbauer et al., "A Watermark for Large Language Models," arXiv preprint arXiv:2301.10226, 2023.

[20] I. Solaiman et al., "Release Strategies and the Social Impacts of Language Models," arXiv preprint arXiv:1908.09203, 2019.

[21] L. Weidinger et al., "Ethical and social risks of harm from Language Models," arXiv preprint arXiv:2112.04359, 2021.

[22] R. Bommasani et al., "On the Opportunities and Risks of Foundation Models," arXiv preprint arXiv:2108.07258, 2021.

[23] H. Lee et al., "CoAuthor: A Human-AI Collaborative Writing Dataset," arXiv preprint arXiv:2305.09808, 2023.

[24] F. Jelinek et al., "Perplexity—a measure of the difficulty of speech recognition," *The Journal of the Acoustical Society of America*, vol. 62, no. S1, pp. S63–S63, 1977.

[25] K. W. Church and W. A. Gale, "Poisson mixtures," *Natural Language Engineering*, vol. 1, no. 2, pp. 163–190, 1995.

[26] C. van der Lee et al., "Best-Worst Scaling for the Annotation of Sentence Level Redundancy," *NAACL HLT*, vol. 1, pp. 3496–3501, 2019.

[27] M. Koppel et al., "Computational methods in authorship attribution," *JASIST*, vol. 60, no. 1, pp. 9–26, 2009.

[28] A. C. Frery and R. H. Caldeira, "On the stylometry of literary translations," *Digital Scholarship in the Humanities*, vol. 36, Suppl. 2, pp. ii65–ii80, 2021.

[29] M. Eder, "Does size matter? Authorship attribution, small samples, big problem," *Digital Scholarship in the Humanities*, vol. 30, no. 2, pp. 167–182, 2015.

[30] D. Hovy and S. L. Spruit, "The social impact of natural language processing," *ACL*, vol. 2, pp. 591–598, 2016.

[31] M. Christ et al., "Undetectable Watermarks for Language Models," arXiv preprint arXiv:2306.09194, 2023.

[32] J. He et al., "CATER: A diagnostic dataset for controllable text generation," arXiv preprint arXiv:2205.14235, 2022.

[33] S. Carter et al., "What's in a Log? A Study of Human-Computer Interaction in a Collaborative Writing Setting," *CHI*, pp. 1–15, 2023.

[34] OpenAI, "Hello GPT-4o," 2024. [Online]. Available: `https://openai.com/index/hello-gpt-4o/`

[35] S. Chakraborty et al., "On the possibilities of AI-assisted plagiarism: A case study of ChatGPT," arXiv preprint arXiv:2303.04168, 2023.

[36] T. Lancaster, "The Challenge of AI-Generated Text in Higher Education," 2023. [Online]. Available: https://www.hepi.ac.uk/2023/01/26/the-challenge-of-ai-generated-text-in-highereducation/

[37] The European Parliament, "EU AI Act: first regulation on artificial intelligence," 2023. [Online]. Available: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/euai-act-first-regulation-on-artificial-intelligence

[38] S. Aaronson, "My AI safety lecture for UT Austin's new online Master's in AI," 2023. [Online]. Available: https://scottaaronson.blog/?p=7288

[39] E. M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *ACM FAccT*, pp. 610–623, 2021.

[40] J. Anderson and L. Rainie, "The Future of Digital Life and Well-Being," *Pew Research Center*, 2023. [Online]. Available: https://www.pewresearch.org/internet/2023/06/28/the-future-ofdigital-life-and-well-being/

[41] D. Buckingham, *The Media Education Manifesto*, John Wiley & Sons, 2019.