Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges

Baixiang Huang
Illinois Institute of Technology
Chicago, IL
bhuang15@hawk.iit.edu

Canyu Chen
Illinois Institute of Technology
Chicago, IL
cchen151@hawk.iit.edu

Kai Shu
Emory University
Atlanta, GA
kai.shu@emory.edu

ABSTRACT

Accurate attribution of authorship is crucial for maintaining the integrity of digital content, improving forensic investigations, and mitigating the risks of misinformation and plagiarism. Addressing the imperative need for proper authorship attribution is essential to uphold the credibility and accountability of authentic authorship. The rapid advancements of Large Language Models (LLMs) have blurred the lines between human and machine authorship, posing significant challenges for traditional methods. We present a comprehensive literature review that examines the latest research on authorship attribution in the era of LLMs. This survey systematically explores the landscape of this field by categorizing four representative problems: (1) Human-written Text Attribution; (2) LLM-generated Text Detection; (3) LLM-generated Text Attribution; and (4) Human-LLM Co-authored Text Attribution. We also discuss the challenges related to ensuring the generalization and explainability of authorship attribution methods. Generalization requires the ability to generalize across various domains, while explainability emphasizes providing transparent and understandable insights into the decisions made by these models. By evaluating the strengths and limitations of existing methods and benchmarks, we identify key open problems and future research directions in this field. This literature review serves a roadmap for researchers and practitioners interested in understanding the state of the art in this rapidly evolving field. Additional resources and a curated list of papers are available and regularly updated at https://llm-authorship.github.io/.

1. INTRODUCTION

Authorship Attribution (AA) is the process of determining the author of a particular piece of writing and has significant real-world applications across various domains. In forensic investigations, authorship attribution plays a crucial role in solving murder cases disguised as suicides [Chaski, 2005; Grant, 2020], tracking terrorist threats [Winter, 2019; Cafiero and Camps, 2023], and aiding general criminal investigations [Koppel et al., 2008; Argamon, 2018; Belvisi et al., 2020]. In the digital realm, authorship attribution helps safeguard the integrity of content by preventing deceptive social media activities [Hazell, 2023], detecting account compromises [Barbon et al., 2017], and linking user profiles across various social networks [Shu et al., 2017; Sinnott and Wang, 2021]. Additionally, authorship attribution techniques are instrumental in combating misinformation [Shu et al., 2020; Chen and Shu, 2024a; Chen et al., 2022b; Hanley and Durumeric, 2024; Stiff and Johansson, 2022], protecting intellectual property rights [Meyer zu Eissen et al., 2007; Stamatatos and Koppel, 2011], and identifying fraudulent activities [Ott et al., 2011; Afroz et al., 2012].

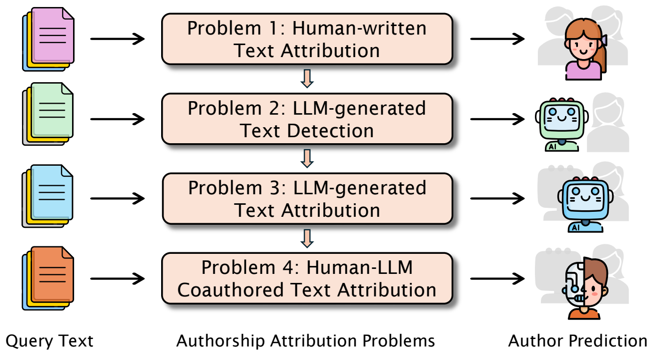

Figure 1: Representative Problems in Authorship Attribution: (1) Human-written Text Attribution, which involves attributing an unknown text to its human authors; (2) LLM-generated Text Detection, which focuses on detecting whether a text has been generated by LLMs; (3) LLM-generated Text Attribution, aimed at identifying the specific LLM or human responsible for a given text; (4) Human-LLM Co-authored Text Attribution, which classifies a text as human-written, LLM-generated, or a combination of both. *The categorization of these problems becomes increasingly complex, as indicated by the arrow, balancing complexity with practicality.*

The development of large language models (LLMs) has revolutionized text generation, offering numerous benefits but also raising significant concerns about text authenticity and originality [Brown et al., 2020; Goldstein et al., 2023]. The advent of LLMs has complicated authorship attribution, making it increasingly difficult to distinguish between LLM-generated texts and human-written texts [Clark et al., 2021; Sadasivan et al., 2023]. Identifying LLM-generated texts is challenging even for human experts, let alone traditional authorship attribution methods [Liu et al., 2023a; Gao et al., 2022]. This inability to distinguish between human and machine-generated content undermines the integrity of authorship, complicates legal and ethical responsibilities, and threatens the credibility of digital content and the safety of online space [Solaiman et al., 2023; Vidgen et al., 2024].

Over the past few decades, authorship attribution has experienced significant advancements due to the development of natural language analysis tools and innovative methods for text representation learning. Traditionally, authorship attribution relied on stylometry, which analyzes an individual's unique writing style through feature engineering to capture linguistic characteristics [Lagutina et al., 2019]. The emergence of machine learning algorithms capable of handling high-dimensional data has enabled the creation of more expressive representations. In recent years, there has been a shift towards extracting text embeddings using pre-trained language models [Fabien et al., 2020]. These approaches, while offering higher performance, often sacrifice explainability for accuracy [Rivera-Soto et al., 2021]. More recently, researchers have begun to use LLMs to extract features in conjunction with machine learning classifiers or to conduct end-to-end reasoning for authorship attribution [Patel et al., 2023; Huang et al., 2024].

Rapid advancements in LLM have significantly improved text generation, producing outputs that rival human writing in fluency and coherence. This progress underscores the imperative need to distinguish between human-written text, LLM-generated text, or a combination of both. As illustrated in Figure 1, authorship attribution can be systematically categorized into four representative problems: attributing unknown texts to human authors, detecting LLM-generated texts, identifying the specific LLM or human responsible for a text, and classifying texts as human, machine, or human-LLM co-authored. Each task presents unique challenges that necessitate corresponding solutions. Researchers continually adapt and refine attribution methods, transitioning from human-authored texts to LLM-generated content, and navigating the complex interweaving in human-LLM co-authored works. As detection methods advance, adversarial attacks also evolve to bypass these measures, creating a continuous cycle of challenge and response in the quest to distinguish and disguise authorship [Dugan et al., 2024]. Addressing these challenges will pave the way for more robust and reliable authorship attribution techniques.

Authorship attribution for both human and LLM-generated texts can be framed as either binary or multi-class classification. The LLM-generated text detection task simplifies the attribution problem by classifying each text as either originating from humans or LLMs [Jawahar et al., 2020a; Mitchell et al., 2023; Pu et al., 2023a; Sadasivan et al., 2023]. The majority of previous research on automatic detection of machine-generated text has focused on binary classification [Jawahar et al., 2020b; Mitchell et al., 2023]. In the more challenging multi-class task, the goal is not only to differentiate between human and LLM-generated text but also to classify the text according to its specific source of generative models [Uchendu et al., 2021; Li et al., 2023b]. Differences in LLM architectures, training methods, and generation techniques can influence the style of generated texts [Munir et al., 2021]. In the more complex human-LLM co-authoring problem, the goal is to distinguish texts authored by humans, LLMs, or combinations of both. Such nuanced detection provides deeper insight into the provenance of the text and is crucial for applications requiring detailed source attribution. Neural network-based detectors generally outperform metric-based methods in both human authorship attribution and LLM-generated text detection problems [He et al., 2023; Zhang et al., 2024]. However, these neural network approaches often offer less explainability compared to their metric-based counterparts.

This review serves as a valuable resource, comprehensively summarizing the existing literature and highlighting the challenges and opportunities introduced by LLMs. We provide in-depth analysis of methodologies in this evolving field. The main contributions of this paper are as follows:

- We provide a timely overview to discuss the challenges and opportunities presented by LLMs in the field of authorship attribution. By systematically categorizing authorship attribution into four representative problems and balancing problem complexity with practicality, we reveal insights into the evolving field of authorship attribution in the era of LLMs.

- We offer a comprehensive comparison of state-of-the-art methodologies, datasets, benchmarks, and commercial tools used in authorship attribution. This analysis not only improves the understanding of authorship attribution but also provides a valuable resource for researchers and practitioners to use as guidelines for approaching this direction.

- We discuss open issues and provide future directions by considering crucial aspects such as generalization, explainability, and interdisciplinary perspectives. We also discuss the broader implications of authorship attribution in real-world applications. This holistic approach ensures that authorship attribution not only yields accurate results but also provides insights that are explainable and socially relevant.

The remainder of this survey is organized as follows. Section 2 explores the attribution of human authorship, beginning with a definition of the problem, followed by an overview of various methodologies and a discussion of the associated challenges. In Section 3, we discuss LLM-generated text detection. In Section 4, we explore the attribution of LLM-generated text. Section 5 covers the attribution of human-LLM co-authored texts. Section 6 discusses resources and evaluation metrics, offering a comparison of benchmarks and datasets. Section 7 highlights opportunities and future directions. In Section 8, we discuss ethical and privacy concerns. Finally, we conclude this survey in Section 9.

2. HUMAN AUTHORSHIP ATTRIBUTION

This section explores the authorship attribution of human-written texts, discussing a range of methodologies such as stylometry, machine learning, pre-trained language models, and LLM-based approaches. It also discusses challenges such as limited data, evolving writing styles, and interpretability.

2.1 Problem Definition

Authorship attribution aims to identify the author of an unknown text from a set of known authors. This can be formulated as an open-class problem, where the true author might not be among the known authors, or a closed-class problem, where the true author is included in a finite set of authors [Stolerman et al., 2014; Andrews and Bishop, 2019]. Authorship attribution methods are typically divided into classification-based methods for a small set of candidate authors and similarity-based ranking methods for larger numbers of authors [Rivera-Soto et al., 2021; Huertas-Tato et al., 2022]. These techniques can also be adapted to related problems, such as authorship verification and profiling. Authorship verification determines whether a piece of writing was written by a specific individual [Stamatatos, 2016], while profiling infers characteristics such as age or gender from the author's writing style [Argamon et al., 2009].

2.2 Methodologies

We explore the evolution of methods used to analyze human-written text, beginning with stylometry. Over time, the focus has shifted to the use of machine learning. Recently, the integration of LLMs marks further advancements in authorship attribution.

2.2.1 *Stylometry Methods*

Stylometry, the quantitative analysis of writing style, has evolved from its initial reliance on human expertise [Mosteller and Wallace, 1963] to computational methods [Neal et al., 2017; Lagutina et al., 2019]. This discipline utilizes a variety of linguistic features to determine the authorship [Holmes, 1994; Lagutina et al., 2019], positing that each author's unique style can be captured through quantifiable characteristics [Argamon et al., 2009]. Key stylometric features include character and word frequencies [Sharma et al., 2018], parts of speech [Sundararajan and Woodard, 2018], punctuation, topics [Seroussi et al., 2014; Potha and Stamatatos, 2019; Halvani and Graner, 2021], and vocabulary richness. Important features can be categorized into the following types: lexical, syntactic, semantic, structural, and content-specific [Rudman, 1997]. Lexical features involve word choice and frequency; syntactic features pertain to sentence structure and grammar; semantic features explore the meaning and context of words; structural features relate to text organization; and content-specific features emphasize domain-specific terms [Bozkurt et al., 2007; Seroussi et al., 2014].

2.2.2 *Machine Learning Methods*

Machine learning approaches integrate stylometric features with classifiers such as logistic regression [Aborisade and Anwar, 2018; Madigan et al., 2005a], Bayesian multinomial regression [Grant, 2007; Argamon et al., 2009], and support vector machines (SVM) [Bacciu et al., 2019]. Before widespread adoption of transformer-based models, multi-headed Recurrent Neural Networks (RNNs) [Bagnall, 2015], and Long Short-Term Memory (LSTMs) were utilized at both sentence and article levels [Qian et al., 2017]. Convolutional Neural Networks (CNNs) were also applied at various levels, including characters, words, and N-grams [Ruder et al., 2016; Shrestha et al., 2017a,b]. Moreover, syntax-augmented CNN models [Zhang et al., 2018], Convolutional Siamese Networks [Saedi and Dras, 2021], and attention-based Siamese Networks [Boenninghoff et al., 2019a] were explored. Additionally, combinations of convolutions and transformers have been employed to learn embeddings for comparison tasks [Andrews and Bishop, 2019].

2.2.3 *Pre-trained Language Models*

Pre-trained language models (PTMs), especially BERT-based architectures [Devlin et al., 2019] such as BERT [Ippolito et al., 2020; Fabien et al., 2020; Manolache et al., 2021], Sentence-BERT [Schlicht and de Paula, 2021; Rivera-Soto et al., 2021], and RoBERTa [Huertas-Tato et al., 2022], have proven effective for learning authorship representation. These methods do not require hand-crafted features but require substantial training time and domain-specific labeled data, struggling with cross-domain generalization and explainability. Contrastive learning [Khosla et al., 2020] is often used with pre-trained language models to enhance stylistic representation by maximizing similarity between texts written by the same authors and minimizing it with texts from different authors [Huertas-Tato et al., 2022].

Barlas and Stamatatos [2020a] found that BERT performed well with large vocabularies, outperforming multi-headed RNNs. Fabien et al. [2020] fine-tuned a BERT model, showing that including additional stylometric and hybrid features in an ensemble model can improve attribution performance. Rivera-Soto et al. [2021] concluded that topic diversity and dataset size are crucial for effective cross-domain transfer. Adaptation through style transfer has not resolved cross-domain issues [Boenninghoff et al., 2019b; Wegmann et al., 2022]. Techniques like slanted triangular learning rates and gradual unfreezing can be used to avoid catastrophic forgetting during fine-tuning [Howard and Ruder, 2018].

2.2.4 *LLM-based Methods*

Despite advances in LLMs, their potential for authorship attribution remains underexplored. The natural language understanding capability allows LLMs to recognize nuances, styles, and patterns in language, which are crucial for distinguishing between authors. Authorship attribution is a complex reasoning task, and LLMs possess significant capabilities in reasoning and problem-solving, particularly in zero-shot learning within resource-limited domains [Kojima et al., 2022]. They assist in feature extraction by identifying syntactic patterns, lexical choices, and grammatical structures essential for authorship attribution. Traditionally, LLMs have been employed mainly for auxiliary tasks such as feature extraction and data annotation [Patel et al., 2023]. Notable examples include the use of GPT-3 for data annotation [Brown et al., 2020] and a T5 encoder for learning authorship signatures. Beyond feature extraction, LLMs possess the ability to identify the author of unknown text based on nuanced linguistic features, making it possible to conduct end-to-end authorship attribution [Huang et al., 2024].

The incorporation of LLMs in authorship attribution addresses several limitations of traditional methods. Unlike BERT-based models, which require computationally expensive fine-tuning and large amounts of domain-specific data for optimal performance, LLMs can generalize across various domains without fine-tuning, thereby mitigating issues related to domain specificity [Barlas and Stamatatos, 2020a]. LLMs are also effective with shorter texts, reducing the necessity for long inputs to derive meaningful representations [Eder, 2015]. Another key advantage of LLM-based approaches is their ability to provide natural language explanations for their predictions, enhancing transparency compared to hidden text embeddings [Huang et al., 2024]. This versatility marks a step forward in overcoming challenges related to data, domain specificity, text length requirements, and explainability faced by earlier methods.

2.3 Open Challenges

Human authors exhibit a diverse range of writing styles influenced by genre, topic, context, and temporal changes. This variability complicates authorship attribution, necessitating the identification of consistent and unique stylistic markers. Further complexity arises from the presence of noise due to the varying size and language of documents, requiring algorithms to identify linguistic nuances. Short or low-quality texts, such as social media posts, are often unreliable, complicating accurate attribution [Eder, 2015; Theophilo et al., 2021]. Collaborative texts, which blend multiple writing styles, further mask individual contributions and obscure distinct authorial signals [Dauber et al., 2017].

Traditional stylometric methods rely on human expertise and manually crafted features, whereas deep learning methods demand significant computational resources and extensive labeled data, with the risk of catastrophic forgetting [Ramasesh et al., 2022]. Authorship attribution using LLMs also faces several challenges: their effectiveness decreases with an increasing number of candidate authors due to context length constraints [Huang et al., 2024], and they can perpetuate biases from training data, resulting in inaccurate attributions for texts from marginalized groups and languages [Liang et al., 2023]. Additionally, LLMs can be misused to generate content that conceals true authorship by mimicking others or using LLMs to alter their work.

3. LLM-GENERATED TEXT DETECTION

LLMs excel at generating fluent and coherent text, which raises concerns about the authenticity and originality of the resulting work. Detecting LLM-generated text is crucial for several applications, including combating misinformation on social media [Gambini et al., 2022; Stiff and Johansson, 2022; Chen and Shu, 2024b], identifying spam [Jindal and Liu, 2008], preventing phishing attacks [Hazell, 2023], identifying fake reviews [Salminen et al., 2022], and detecting machine-generated scientific papers [Rodriguez et al., 2022a; Liu et al., 2023a]. As a result, the detection of LLM-generated text has garnered significant attention [Kumarage and Liu, 2023; Tang et al., 2023a; Wu et al., 2023b; Yang et al., 2023c].

3.1 Problem Definition

LLM-generated texts are included within the scope of machine-generated texts[1]. Machine-generated texts encompass any text produced by automated systems, including simpler language models or rule-based systems [Uchendu et al., 2021]. This paper focuses specifically on LLM-generated texts. The task of detecting LLM-generated text involves distinguishing text created by LLMs from that written by humans. Typically, this task is approached as a binary classification problem [Zellers et al., 2019; Solaiman et al., 2019; Jawahar et al., 2020b; Fagni et al., 2021; Mitchell et al., 2023].

3.2 Methodologies

The evaluation of the quality of machine-generated excerpts has traditionally relied on human judgement, which is considered the gold standard for open-domain generation systems [van der Lee et al., 2019; Gehrmann et al., 2019a]. However, distinguishing between LLM-generated and human-written texts poses significant challenges for humans [Dugan et al., 2023]. For example, untrained human reviewers are often unable to distinguish GPT-3-generated text from human-written text, identifying it correctly only at a rate consistent with random chance [Clark et al., 2021]. Liu et al. [2023a] found that even experienced faculty and researchers could only achieve about a 50% success rate in identifying GPT-generated academic writings. In contrast, detection algorithms frequently outperform the human in this task [Ippolito et al., 2020].

Chakraborty et al. [2023a] employed theoretical analysis to argue that detecting LLM-generated text is nearly always feasible with the collection of multiple samples, and they established precise sample complexity bounds for this detection. However, existing detectors and models for LLM-generated text are not yet fully reliable [Sadasivan et al., 2023; Wang et al., 2023; Dugan et al., 2024]. Sadasivan et al. [2023] provided theoretical insights indicating that the detection problem is becoming increasingly difficult.

LLM-generated text detectors can be categorized into metric-based and model-based methods [He et al., 2023], which are further divided into feature-based, neural network-based, zero-shot-based, and watermark-based methods. These detectors are also classified as white-box or black-box, depending on their access to the LLM weights [Tang et al., 2023b; Yang et al., 2023c]. Watermarking-based methods typically fall under the white-box detection category, while proprietary models are restricted to black-box methods.

3.2.1 *Featured-based Method*

LLM-generated texts are typically less emotional and more objective than human-written texts [Guo et al., 2023]. Human-authored texts are generally more coherent, while LLM-generated texts tend to repeat terms within a paragraph [Dugan et al., 2023]. Similarly to the problem of attribution of human authorship, linguistic characteristics such as phrasal verbs, co-reference, part-of-speech (POS) tags, and named entity (NE) tags are also useful in distinguishing LLM-generated text [Nguyen-Son et al., 2017; See et al., 2019; Fröhling and Zubiaga, 2021]. Feature-based methods are more explainable but have drawbacks, such as poor generalizability of certain features across different domains and sampling methods.

3.2.2 *Neural Network-Based Detectors*

Neural network-based detectors, particularly those utilizing BERT, have proven effective in distinguishing between human-written texts and those generated by GPT-2 [Ippolito et al., 2020; Liu et al., 2019]. Solaiman et al. [2019] fine-tuned the RoBERTa model with a dataset of GPT-2 outputs in open domain settings. Similarly, Guo et al. [2023] fine-tuned RoBERTa to detect ChatGPT-generated text. [Zhan et al., 2023] developed G3Detector by fine-tuning RoBERTa-large for the same purpose. Additionally, Chen et al. [2023a] introduced GPT-Sentinel, training both RoBERTa and T5 [Raffel et al., 2020] on their OpenGPTText dataset. In a different approach, Hu et al. [2023] created RADAR, which fine-tunes

[1]Machine-generated texts is also referred to as machine-authored, AI-generated, neural-generated, deepfake text, neural text, or synthetic text.

Vicuna 7B [Chiang et al., 2023] in a generative adversarial setting along with a paraphrase model. These efforts highlight ongoing advances in the detection of LLM-generated content using BERT-based models.

These detectors require retraining when encountering text generated by new LLMs to ensure reliable detection [Mitchell et al., 2023; Chakraborty et al., 2023b]. On the other hand, neural network-based detectors are vulnerable to adversarial and poisoning attacks [Goodfellow et al., 2015; Wang et al., 2022; Pu et al., 2023a]. These detectors also face limitations such as overfitting to training data [Uchendu et al., 2020]. The generalization ability of these detectors is critical, as they have been trained in various family models and tested in unseen models [Pu et al., 2023b; Bhattacharjee et al., 2023]. Surrogate models, which are often small language models, are also applied to train classifiers [Verma et al., 2023; Mireshghallah et al., 2023].

3.2.3 *Zero-Shot Detectors*

Zero-shot detection methods are generally statistics-based, enabling the detection of LLM-generated text without additional training [Su et al., 2023b]. Various statistical measures have been employed, including entropy [Lavergne et al., 2008; Gehrmann et al., 2019a], perplexity [Beresneva, 2016; Hans et al., 2024], average log-probability score [Solaiman et al., 2019], fluency [Holtzman et al., 2020], and Zipf's word frequency law [Zipf, 2016; Piantadosi, 2014][2]. Additional methods leverage n-grams [Yang et al., 2023b], Uniform Information Density (UID) [Venkatraman et al., 2023], log rank information [Su et al., 2023c], and various linguistic features such as part-of-speech determiners, conjunctions, auxiliary relations, vocabulary, and emotional tone [Joulin et al., 2017; Tang et al., 2023b; Gehrmann et al., 2019b].

Various zero-shot detection methods provide distinct strategies to enhance both detection accuracy and efficiency. Gallé et al. [2021] introduced an unsupervised method that identifies the over-appearance of repeated higher-order n-grams, distinguishing them from human-generated text. DetectGPT [Mitchell et al., 2023] is based on the observation that LLM-generated passages often fall into regions of negative curvature in log probability. [Bao et al., 2023] enhanced this approach and proposed Fast-DetectGPT, which increases efficiency by using conditional probability curvature. Additionally, some methods leverage LLMs themselves for text classification [Zellers et al., 2019; Solaiman et al., 2019]. Different decoding methods are often applied to generate more diverse and less repetitive text, although these can also lead to hallucinations and less verifiable content [Shakeel and Jain, 2021; Guo et al., 2023]. Fact-checking methods can mitigate these issues [Zhong et al., 2020; Schuster et al., 2020]. Additionally, Krishna et al. [2024a] developed a detector that uses information retrieval to store LLM output in a database and search for semantically similar content to identify LLM-generated text, although this method raises privacy concerns regarding the storage of user conversations.

3.2.4 *Watermarking*

Watermarking involves embedding specific patterns in text, making them imperceptible to humans but detectable through specialized methods [Topkara et al., 2005; Meral et al., 2009; Kirchenbauer et al., 2023; Zhao et al., 2023]. By imprinting distinct patterns, watermarking enables the identification of LLM-generated text. Various methods include parsed syntactic tree structures [Atallah et al., 2001; Topkara et al., 2005], synonym tables [Jalil and Mirza, 2009], adversarial watermarking [Abdelnabi and Fritz, 2021], and context-aware lexical substitution [Yang et al., 2022]. One notable approach is soft watermarking, proposed by Kirchenbauer et al. [2023], which partitions tokens into "green" and "red" lists to create patterns. A watermarked LLM samples tokens from the green list with high probability, determined by a pseudo-random generator seeded by its prefix token. The watermarking detector classifies passages with a high frequency of tokens from the green list as LLM-generated.

Other studies have improved the robustness, efficiency, and stealthiness of watermarking methods [Hou et al., 2023; Wu et al., 2023a; Zhao et al., 2023]. However, there is a trade-off between watermark effectiveness and text quality, as more reliable watermarks require more extensive text modifications [Sadasivan et al., 2023]. Additionally, watermarking presents challenges for proprietary LLMs and third-party applications due to the necessity of accessing the language model logits [Kirchenbauer et al., 2023]. Watermark-based detection methodologies are also vulnerable to paraphrasing attacks [Sadasivan et al., 2023; Krishna et al., 2024a].

3.3 Open Challenges

Detecting LLM-generated texts is challenging due to their versatile styles and contextual awareness. LLMs can emulate human writing so closely that they pose significant challenges to traditional stylometric techniques. They can incorporate complex narrative structures and varied vocabularies, making it difficult to distinguish between human and LLM-generated texts. The rapid evolution of LLMs further complicates detection, as newer versions exhibit different stylistic characteristics, making detection models quickly obsolete [Chakraborty et al., 2023b]. LLM-generated text detectors often struggle to generalize to unseen domains encountered during training [Pu et al., 2023b; Rodriguez et al., 2022b; Li et al., 2023a] and tend to perform better on LLMs they were specifically trained on [Pu et al., 2023b; Chakraborty et al., 2023b; Li et al., 2023a].

Existing detectors also lack robustness to various factors, such as alternative decoding strategies [Ippolito et al., 2020], input sequence length [Solaiman et al., 2019], different prompts [Kumarage et al., 2023; Lu et al., 2023], and repetition penalties [Fishchuk and Braun, 2023]. Additionally, detectors are vulnerable to adversarial attacks, including homoglyph attacks [Gagiano et al., 2021; Macko et al., 2024], whitespace insertion [Cai and Cui, 2023], syntactic perturbations [Bhat and Parthasarathy, 2020a], synonym replacement [Kulkarni et al., 2023], and paraphrasing [Krishna et al., 2024b; Shi et al., 2024; Becker et al., 2023].

4. LLM-GENERATED TEXT ATTRIBUTION

Identifying whether a piece of text is generated by a specific LLM or a human is crucial. This distinction helps to trace the origin of the text to ensure accountability, enhance transparency, and uphold ethical standards in information dissemination. If the content is harmful, misleading, or illegal,

[2] The frequency of a word decreases as its rank in a frequency-ordered list increases

pinpointing the exact responsible LLM is essential to address ethical concerns and fulfill legal obligations. LLM-generated text attribution builds upon techniques for LLM-generated text detection. Variations in model architecture (such as the number of layers and parameters), training methods (including pre-training and fine-tuning), and generation techniques (such as sampling parameters) influence all the characteristics of the generated texts [Munir et al., 2021].

4.1 Problem Definition

This attribution task extends beyond binary classification to handle multiple classes, increasing the complexity of LLM-generated text detection. The primary goal is to determine whether a given piece of text was created by a specific human or by one of several LLMs [Uchendu et al., 2020; Venkatraman et al., 2023; Chen et al., 2023b; He et al., 2023; Soto et al., 2024]. A sub-problem is to attribute the text solely to LLMs, also known as model sourcing [Yang et al., 2023a] or origin tracing [Li et al., 2023b].

4.2 Methodologies

Attributing texts to LLMs versus human writers involves recognizing inherent differences in their text generation capabilities. LLMs typically exhibit less diversity in word usage compared to humans [Ippolito et al., 2020; Dugan et al., 2023]. LLMs can mimic a range of styles and tones, often masking their underlying characteristics. This ability to adapt makes attribution challenging, especially as LLMs rapidly evolve and their outputs change significantly over time [Guo et al., 2023]. Additionally, LLMs may inadvertently reproduce snippets of their training data.

To simplify the classification process, it is common practice to group different human writers into a single category because humans exhibit a broader spectrum of writing styles and proficiency levels compared to machines [Uchendu et al., 2021]. For example, classifications might include comparisons such as Human vs. ChatGPT, or Human vs. LLama [Uchendu et al., 2021; He et al., 2023]. Some studies have formulated a seven-class classification that includes one human class and six LLM classes [He et al., 2023; Wang et al., 2024]. Other approaches consider multiple human classes, albeit with a limited number. For instance, a 10-class classification might include seven human classes and three LLM classes [Tripto et al., 2023]. This multiclass classification is often converted into a one-vs-rest classification for each label. Transformer-based models, such as BERT and RoBERTa, are fine-tuned on datasets containing both human-written and LLM-generated texts to conduct the attribution.

4.3 Open Challenges

Attributing texts generated by LLMs to specific humans or models presents a multi-class classification challenge. Variations in training data, model architecture, and fine-tuning processes contribute to the distinctive outputs of different LLMs, though these differences are often subtle and difficult to detect [Uchendu et al., 2021]. Effective identification requires sophisticated methods to discern unique signatures embedded in syntactic structures and lexical choices, which are influenced by specific training datasets. However, the proprietary nature of many LLMs restricts access to comparative data, posing significant hurdles. Additionally, the high degree of stylistic overlap among LLMs, especially those with similar architectures or trained on overlapping datasets, further complicates accurate classification. Continuous updates and fine-tuning of LLMs necessitate ongoing adjustments to attribution methodologies to account for evolving model characteristics [Wu et al., 2023b].

5. HUMAN-LLM CO-AUTHORED TEXT ATTRIBUTION

Besides creating text from scratch, LLMs are often used to extend sequences from human prompts. These perturbations have diminished the effectiveness of existing text detection methods [Bhat and Parthasarathy, 2020b]. Identifying text that combines input from both human authors and LLMs presents unique challenges. Hybrid texts may originate as human-written content, with LLMs employed for conditional generation, making it difficult to clearly distinguish between the stylistic features of human and machine contributions. There are fewer studies on this task due to its difficulty, and existing research often makes simplifications.

5.1 Problem Definition

A human-LLM co-authored text, also known as mixed text [Zhang et al., 2024] or collaborative human-AI writing [Richburg et al., 2024], is a piece of writing that is first created by a human and then revised or extended by LLMs, and vice versa. This task involves recognizing the nuances of multi-source authorship with fine-grained precision. Some studies categorize any text that is generated, modified, or extended by a machine as LLM-generated. This simplifies the task to either LLM-generated Text Detection or LLM-generated Text Attribution [Yang et al., 2023c; Crothers et al., 2023]. Other researchers handle LLM-revised human texts and human-revised LLM texts as a single category, alongside purely human-written and purely LLM-generated texts, approaching human-LLM co-authored text authorship attribution as a three-class classification problem [Zhang et al., 2024; Richburg et al., 2024]. One variation of this task is to detect the boundary between human-written and LLM-generated text [Cutler et al., 2021; Wang et al., 2024].

5.2 Methodologies

Human-authored texts tend to be more coherent and exhibit greater lexical diversity compared to LLM-generated texts [Guo et al., 2023; Dugan et al., 2023; Zhang et al., 2024]. Models like DNA-GPT [Yang et al., 2023a] and DetectGPT [Mitchell et al., 2023] utilized the T5 model [Raffel et al., 2020] to simulate scenarios in which humans modify LLM-generated texts. MIXSET [Zhang et al., 2024] offers a more comprehensive dataset that includes text refined by LLMs through polishing, completion, and rewriting operations.

To effectively analyze and classify human-LLM co-authored texts versus those solely authored by humans or LLMs, feature-based methods from the problem of LLM-generated text detection, such as Log-likelihood [Solaiman et al., 2019], GLTR [Gehrmann et al., 2019a], and log-rank [Mitchell et al., 2023], are adapted to this task. Additionally, neural network-based models such as BERT [Ippolito et al., 2020], Radar [Hu et al., 2023], and GPT-sentinel [Chen et al., 2023a] can also be applied. The complexity of this task increases as users may employ multiple LLMs to compose different sections of an article, further blurring the lines between human and machine-generated content. Consequently, the techniques used in the earlier detection of LLMs need to evolve continuously. This ongoing evolution in detection strategies mirrors the increasing sophistication of LLM outputs and the collaborative nature of modern text creation.

5.3 Open Challenges

Authorship attribution involving human-LLM co-authored pieces presents varying degrees of complexity, requiring different analytical approaches to accurately identify and differentiate the contributions of each author. For texts authored entirely by humans or LLMs, stylometric techniques can be utilized effectively. Human-authored texts often feature unique stylistic nuances, such as variable sentence structures and emotive language [Zhang et al., 2024]. In contrast, LLM-generated texts typically exhibit consistent syntax and a broader vocabulary [Guo et al., 2023]. Feature-based methods used in LLM-generated text detection can be adapted to classify texts by identifying these distinct patterns, thus attributing texts to their correct source.

Analyzing and classifying texts co-authored by humans and LLMs presents a significant challenge due to the blending of human and machine stylistic features. These texts often start as human drafts and are later extended or revised by LLMs, or the process might occur in reverse. This integration of styles creates a hybrid form that makes it difficult to distinguish distinct authorial markers, thereby complicating the attribution process.

Human-LLM co-authored texts pose a more intricate challenge due to the blending of stylistic and linguistic elements from both human authors and LLMs. These texts may begin as human drafts that are later extended or revised by LLMs, or vice versa, resulting in an integration of styles that obscures authorial markers [Liu et al., 2023b]. Advanced techniques are required to dissect these integrations, identifying where and how the contributions of LLM intersect with human input [Wang et al., 2024]. This involves detecting subtle shifts in style and contextual cues that indicate the extent of LLM involvement, allowing for accurate segmentation and attribution of authorship within hybrid documents.

6. RESOURCES AND EVALUATION METRICS

This section provides an in-depth examination of widely used benchmarks, datasets, and evaluation metrics in authorship attribution research, along with guidelines for selecting appropriate ones. These resources range from purely human-written texts to those generated by LLMs and human-LLM co-authored texts. This diversity is crucial for training and evaluating models across various tasks.

Traditional datasets focus exclusively on texts written by humans, while modern datasets include LLM-generated text, addressing the need to detect and attribute texts produced by LLMs. Additionally, this section covers commercial and open-source detectors commonly used to identify machine-generated text. Lastly, we summarize the common evaluation metrics employed in this field.

6.1 Benchmarks and Datasets

Authorship datasets encompass a wide range of sources, from formal literature to informal online communications, highlighting the increasing importance of user-generated content on social media. Human authorship datasets should include author identifiers and ideally contain multiple texts for each author. Manually collecting data is time-consuming and costly, motivating researchers to utilize web data sources such as Wikipedia and Reddit. In contrast, custom datasets for LLM-generated text are easier and more affordable to create, and are often built alongside human-written text to maintain a similar domain and format.

A general guideline for selecting and constructing datasets involves incorporating variations in domain, model architecture, and decoding strategies. Addressing class imbalance is crucial, as LLM-generated and human-written texts are often disproportionate. For human authorship data, it is recommended to choose content created before the widespread use of LLMs (GPT-3 [Brown et al., 2020] was released in June 2020, and ChatGPT followed in November 2022) to ensure that the texts were predominantly human-written.

Factors influencing the performance of existing authorship attribution models include the size of the training text [Hirst and Feiguina, 2007; Marton et al., 2005], the number of candidate authors [Koppel et al., 2006], and the imbalanced distribution of training texts among candidate authors [Stamatatos, 2008]. The availability of digital text in formats such as tweets, blogs, and articles has increased exponentially, providing more training data to accelerate the development of authorship attribution. However, the rapid growth of online communication has also led to shifts in writing behavior, resulting in shorter, fragmented, and less coherent social media tweets and text messages. For example, tweets are limited to 280 characters, whereas legal judgment documents contain thousands of words [Seroussi et al., 2011]. The challenge in social media stems from the brief nature of posts and a large pool of potential authors, making the attribution of short documents particularly difficult [Aborisade and Anwar, 2018; Seroussi et al., 2014; Theophilo et al., 2021].

Table 1 provides a comprehensive overview of 21 widely used benchmarks and datasets. These datasets are characterized by various statistics, including domain, size, word length, language, and the LLMs used to generate text. All listed datasets support LLM-generated text detection (Problem 2). However, fewer support LLM-generated text attribution (Problem 3) and Human-LLM Co-authored Text Attribution (Problem 4). These benchmarks often originate from human-written datasets such as XSum [Narayan et al., 2018], OpenWebText [Gokaslan and Cohen, 2019], and Wikipedia. Since Problems 2, 3, and 4 frequently treat human-written text as a single category for simplicity, rather than identifying individual authors as in Problem 1, many of these datasets are unsuitable for Problem 1, which requires unique author identification. Therefore, the representative datasets for human authorship (Problem 1) are summarized as follows:

- Amazon Review [Ni et al., 2019]: Featuring reviews with ratings, text, helpfulness votes, product metadata, and related links, this dataset provides a comprehensive view of consumer opinions, ideal for commercial authorship attribution studies.

Name	Domain	Size	Length	Language	Model	P2	P3	P4
TuringBench [Uchendu et al., 2021]	News	168,612 (5.2%)	100 to 400 words	en	GPT-1,2,3, GROVER, CTRL, XLM, XLNET, FAIR, TRANSFORMER-XL, PPLM	✓	✓	
TweepFake [Fagni et al., 2021]	Social media	25,572 (50.0%)	less than 280 characters	en	GPT-2, RNN, Markov, LSTM, CharRNN	✓		
ArguGPT [Liu et al., 2023c]	Academic essays	8,153 (49.5%)	300 words on average	en	GPT2-Xl, text-babbage-001, text-curie-001, davinci-001,002,003, GPT-3.5-Turbo	✓		
AuTexTification [Sarvazyan et al., 2023]	Tweets, reviews, news, legal, and how-to articles	163,306 (42.5%)	20 to 100 tokens	en, es	BLOOM, GPT-3	✓	✓	
CHEAT [Yu et al., 2023]	Academic paper abstracts	50,699 (30.4%)	163.9 words on average	en	ChatGPT	✓		
GPABench2 [Liu et al., 2023b]	Academic paper abstracts	2.385M (6.3%)	70 to 350 words	en	ChatGPT	✓		✓
Ghostbuster [Verma et al., 2023]	News, student essays, creative writing	23,091 (87.0%)	77 to 559 (median words per document)	en	ChatGPT, Claude	✓		
HC3 [Guo et al., 2023]	Reddit, Wikipedia, medicine, finance	125,230 (64.5%)	25 to 254 words	en, zh	ChatGPT	✓		
HC3 Plus [Su et al., 2023a]	News, social media	214,498	N/A	en, zh	ChatGPT	✓		
HC-Var [Xu et al., 2023]	News, reviews, essays, QA	144k (68.8%)	50 to 200 words	en	ChatGPT	✓		
HANSEN [Tripto et al., 2023]	Transcripts of speech (spoken text), statements (written text)	535k (96.1%)	less than 1k tokens	en	ChatGPT, PaLM2, Vicuna-13B	✓	✓	
M4 [Wang et al., 2023]	Wikipedia, WikiHow, Reddit, QA, news, paper abstracts, peer reviews	147,895 (24.2%)	more than 1k characters	ar, bg, en, id, ru, ur, zh	davinci-003, ChatGPT, GPT-4, Cohere, Dolly2, BLOOMz	✓		
MGTBench [He et al., 2023]	News, student essays, creative writing	21k (14.3%)	1 to 500 words	en	ChatGPT, ChatGLM, Dolly, GPT4All, StableLM, Claude	✓	✓	
MULTITuDE [Macko et al., 2023]	News	74,081 (10.8%)	200 to 512 tokens	ar, ca, cs, de, en, es, nl, pt, ru, uk, zh	GPT-3,4, ChatGPT, Llama-65B, Alpaca-LoRa-30B, Vicuna-13B, OPT-66B, OPT-IML-Max-1.3B	✓		
OpenGPTText [Chen et al., 2023a]	OpenWebText	58,790 (50.0%)	less than 2k words	en	ChatGPT	✓		
OpenLLMText [Chen et al., 2023b]	OpenWebText	344,530 (20%)	512 tokens	en	ChatGPT, PaLM, Llama, GPT2-XL	✓	✓	
Scientic Paper [Mosca et al., 2023]	Scientific papers	29k (55.2%)	900 tokens on average	en	SCIgen, GPT-2,3, ChatGPT, Galactica	✓		
RAID [Dugan et al., 2024]	News, Wikipedia, paper abstracts, recipes, Reddit, poems, book summaries, movie reviews	523,985 (2.9%)	323 tokens on average	cs, de, en	GPT-2,3,4, ChatGPT, Mistral-7B, MPT-30B, Llama2-70B, Cohere command and chat	✓		
M4GT-Bench [Wang et al., 2024]	Wikipedia, Wikihow, Reddit, arXiv abstracts, academic paper reviews, student essays	5,368,998 (96.6%)	more than 50 characters	ar, bg, de, en, id, it, ru, ur, zh	ChatGPT, davinci-003, GPT-4, Cohere, Dolly-v2, BLOOMz	✓	✓	✓
MAGE [Li et al., 2023a]	Reddit, reviews, news, QA, story writing, Wikipedia, academic paper abstracts	448,459 (34.4%)	263 words on average	en	GPT, Llama, GLM-130B, FLAN-T5 OPT, T0, BLOOM-7B1, GPT-J-6B, GPT-NeoX-2	✓		
MIXSET [Zhang et al., 2024]	Email, news, game reviews, academic paper abstracts, speeches, blogs	3.6k (16.7%)	50 to 250 words	en	GPT-4, Llama2	✓		✓

Table 1: Summary of Authorship Attribution Datasets and Benchmarks with LLM-Generated Text. Size is shown as the sum of LLM-generated and human-written texts (with the percentage of human-written texts in parentheses). Language is displayed using the two-letter ISO 639 abbreviation. Columns P2, P3, and P4 indicate whether the dataset supports problems described in Problem 2, 3, and 4, respectively.

Detector	Price	API	Website
GPTZero	150k words at $10/month, 10k words for free per month	Yes	https://gptzero.me/
ZeroGPT	100k characters for $9.99, 15k characters for free	Yes	https://www.zerogpt.com/
Sapling	50k characters for $25, 2k characters for free	Yes	https://sapling.ai/ai-content-detector
Originality.AI	200k words at $14.95/month	Yes	https://originality.ai/
CopyLeaks	300k words at $7.99/month	Yes	https://copyleaks.com/ai-content-detector
Winston	80k words at $12/month	Yes	https://gowinston.ai/
GPT Radar	$0.02/100 tokens	N/A	https://gptradar.com/
Turnitin's AI detector	License required	N/A	https://www.turnitin.com/solutions/topics/ai-writing/ai-detector/
GPT-2 Output Detector	Free	N/A	https://github.com/openai/gpt-2-output-dataset/tree/master/detector
Crossplag	Free	N/A	https://crossplag.com/ai-content-detector/
CatchGPT	Free	N/A	https://www.catchgpt.ai/
Quil.org	Free	N/A	https://aiwritingcheck.org/
Scribbr	Free	N/A	https://www.scribbr.com/ai-detector/
Draft Goal	Free	N/A	https://detector.dng.ai/
Writefull	Free	Yes	https://x.writefull.com/gpt-detector
Phrasly	Free	Yes	https://phrasly.ai/ai-detector
Writer	Free	Yes	https://writer.com/ai-content-detector/

Table 2: Overview of LLM-Generated Text Detectors.

- Aston 100 Idiolects Corpus [Heini and Kredens, 2021]: Comprising emails, essays, text messages, and business memos from 100 individuals (ages 18–22, native English speakers), this corpus provides a broad spectrum of text types for analyzing both content and stylistic features.

- Blog Authorship Corpus [Schler et al., 2006]: Contains over 680,000 blog posts from more than 19,000 authors, with an average of 35 posts per author. The texts, averaging 79 tokens, are informal and conversational.

- Deceptive Opinion Spam [Ott et al., 2011]: Includes 400 genuine and 400 deceptive hotel reviews, with deceptive reviews generated using Amazon Mechanical Turk, useful for studying the nuances of fake versus real reviews.

- Enron Email [Klimt and Yang, 2004]: Includes around 500,000 messages from 160 employees, offering long texts and high text-per-author variance, making it ideal for studying corporate communication styles.

- Fanfiction: Collected from fanfiction.net, this dataset includes fan-written fiction [Bischoff et al., 2020; Kestemont et al., 2021], providing insights into creative writing and authorship attribution in fictional narratives.

- IMDb1M [Seroussi et al., 2014]: Features over 270,000 movie reviews by 22,000 authors, with an average of 12.3 texts per author and an average text length of 121 tokens, suitable for analyzing shorter, user-generated content.

- Pushshift Reddit [Baumgartner et al., 2020]: This dataset comprises posts and comments from various subreddits, covering diverse topics and writing styles, making it suitable for analyzing informal online discourse.

- PAN [Kestemont et al., 2021; Bevendorff et al., 2022]: Offered by PAN workshops for benchmarking authorship attribution models and are used in various authorship attribution competitions.

- VALLA [Tyo et al., 2022]: Designed for benchmarking authorship attribution models, VALLA standardizes a range of texts across various genres and writing styles.

Table 2 summarizes various commercial and open-source LLM-generated text detectors. These detectors are primarily designed for LLM-generated text detection (Problem 2). Some detectors like GPTZero [Tian and Cui, 2023] can also detect human-LLM co-authored text and identify portions likely to be LLM-generated, addressing the boundary detection sub-problem of Problem 4 [Cutler et al., 2021]. Although many of these detectors claim over 99% accuracy in detecting LLM-generated text, few are tested on shared benchmark datasets. Despite their high accuracy claims, many detectors suffer from high false positive rates, which could falsely accuse individuals of plagiarism and undermine credibility and trust in genuine authors. Additionally, these detectors often lack robustness against variations in sampling strategies, adversarial attacks, and unseen domains and language models [Dugan et al., 2024].

6.2 Evaluation Metrics

Evaluation metrics such as the F1 score and AUCROC are essential for quantifying the performance of authorship models, providing a standardized means to assess and compare the effectiveness of different authorship attribution approaches. As in other classification tasks, existing studies predominantly use the Area Under the Receiver Operating Characteristic (AUCROC) and F1 score to evaluate attribution algorithms. In human authorship attribution, where there are a large number of candidate authors, retrieval metrics such as Mean Reciprocal Rank (MRR) and recall-at-k are used [Rivera-Soto et al., 2021]. Additionally, Self-BLEU are useful metrics, with a lower score indicating higher textual diversity [Zhang et al., 2024]. Common evaluation metrics include:

- Accuracy: Measures the proportion of correctly identified authors. High accuracy indicates that the model is effective at correctly classifying authors.

- Precision, Recall, and F1-Score: Crucial in imbalanced datasets. Precision indicates the relevance of identified instances, recall measures the ability to identify all relevant instances, and the F1-Score balances both.
- Area Under the Receiver Operating Characteristic Curve (AUCROC): Represents the trade-off between true positive rates and false positive rates, where higher values indicate better performance.
- False Positive Rate (FPR) and False Negative Rate (FNR): Critical for minimizing misclassification, with FPR measuring incorrect classification of human texts as LLM-generated and FNR the reverse.
- Recall-at-k: measure the probability that the correct author appears among the top k results when ranking targets by cosine similarity to a query text.
- Mean Absolute Error (MAE): Used to evaluate the performance of human-machine text boundary detection. It measures the average absolute difference between the predicted position index and the actual change point.

7. OPPORTUNITIES AND FUTURE DIRECTIONS

This section explores future directions in the field of authorship attribution, focusing on leveraging the potential of LLMs while addressing associated challenges. Future efforts should aim for finer granularity in authorship attribution, leveraging LLM capabilities, improving generalization, enhancing explainability, preventing misuse, developing standardized benchmarks, and integrating interdisciplinary perspectives to enrich the field.

7.1 Finer Granularity

Current authorship attribution methods face limitations when handling a more extensive range of candidate human authors or LLMs, presenting opportunities for future research. For instance, existing approaches for LLM-generated Text Attribution typically manage only a limited number of authors or models, which restricts their applicability in real-world scenarios where the pool of potential authors or models can be vast. Previous studies have often oversimplified the problem by categorizing all human-written text into a single category [Uchendu et al., 2021; He et al., 2023]. This approach ignores the diversity among human authors and fails to leverage the rich set of characteristics that distinguish individual writing styles. Future work can build upon traditional research on human authorship to develop methods capable of attributing human-written text to individual authors even within the context of LLM-generated content. This refinement will improve the accuracy and utility of authorship attribution models, especially in mixed datasets containing both human-written and LLM-generated texts.

Similarly, for Human-LLM Co-authored Text Attribution, there is a need to attribute text more precisely to individual human authors or specific LLMs. Current work simplifies human-written text, LLM-generated content, and texts co-authored by humans and LLMs into three broad categories, without differentiating between individual human authors and specific LLMs [Zhang et al., 2024; Richburg et al., 2024]. This approach overlooks the nuanced contributions of each author or model. By improving the granularity of attribution, future models can better distinguish between various human authors and LLMs, thus increasing the practicality and reliability of authorship attribution tools. Such advancements would be particularly valuable in collaborative environments where multiple human authors and LLMs contribute to a single body of work, enabling clearer recognition of each contributor's role.

7.2 Generalization

This subsection examines the applicability of current methodologies across varying LLMs, domains, genres, and languages. Domains refer to broad areas of knowledge or topics, while genres refer to specific styles or forms of writing within any domain. Domain generalization poses significant challenges due to variability in vocabularies, syntax, and styles across different subjects, complicating accurate authorship attribution. Attribution performance tends to drop when known and query texts differ in topic or genre [Altakrori et al., 2021]. Models like BERT and RoBERTa have shown limitations in cross-domain tasks [Barlas and Stamatatos, 2020b; Huertas-Tato et al., 2022], and adapting models to new domains remains difficult due to factors like dataset size variability and writing styles. Traditional methods focused on identifying less topic-dependent features, such as function words and part-of-speech n-grams [Madigan et al., 2005b; Menon and Choi, 2011], while recent approaches highlight the importance of training more powerful transformer-based models [Rivera-Soto et al., 2021] and techniques such as adversarial training [Ganin et al., 2016; Li et al., 2017; Ben-David et al., 2010; Ganin et al., 2016].

Genre generalization involves adapting to different writing styles, such as fiction, non-fiction, and poetry, each with unique features. Authors' adaptation to various genres dilutes their identifiable stylistic traits, complicating attribution. Similarly, distinguishing between human-written and LLM-generated text in different genres requires identifying genre-specific inconsistencies. The diversity of genres demands flexible models capable of understanding various narrative structures, tones, and stylistic elements. Adapting models to handle genre variations requires more advanced and flexible approaches for effective generalization. Current authorship attribution models also struggle with out-of-distribution issues when faced with languages and LLMs not encountered during training, leading to decreased accuracy and reliability [Koppel et al., 2005; Wu et al., 2023b]. Addressing this generalization problem is crucial for developing robust models that can handle diverse and evolving linguistic and model landscapes.

Improving generalization can potentially be achieved through several strategies. First, leveraging transfer learning by pre-training on large, diverse datasets and fine-tuning on specific domains enhances adaptability and performance [Barlas and Stamatatos, 2020a; Rodriguez et al., 2022b]. Second, developing domain- and genre-invariant features would improve robustness by focusing on core stylistic elements [Argamon et al., 2003]. Third, employing hybrid models and ensemble methods that integrate domain-specific knowledge can optimize prediction accuracy by drawing on the strengths of individual models [Bacciu et al., 2019]. Additionally, incorporating contextual factors such as the writing environment or intended audience, alongside data augmentation techniques, can bolster generalization. Finally, improving attribution in multilingual contexts enables models to operate effectively across various languages [Chen et al., 2022a; Shamardina et al., 2022]. As detectors could be biased against non-native English writers [Liang et al., 2023], enhancing multilingual generalization is crucial for fairness. Collectively, these approaches foster robust and adaptable models equipped to handle diverse styles and contexts.

7.3 Explainability

Improving explainability is crucial for ensuring transparency and trust in authorship attribution models as they become more integrated into fields such as law, academia, and journalism. Developing explainable techniques for authorship attribution can lead to more transparent methodologies, where the reasoning behind attributions is clear and understandable. Explainable authorship attribution can serve as evidence in legal proceedings [Chaski, 2005; Rocha et al., 2016]. Traditional attribution methods, which rely on stylistic and linguistic features to identify an author, struggle to distinguish between human-authored texts and those generated by LLMs, which adeptly replicate these features. This challenge requires improved methodologies that not only differentiate origins but also explore how LLMs emulate specific authorial styles [Boenninghoff et al., 2019a; Danilevsky et al., 2020].

Despite attempts such as analyzing internal attention weights or employing interpretation tools and visualization techniques [Wallace et al., 2019], word-level explanations are insufficient. The challenge remains to provide higher-level explainability that aligns with human cognitive processes [Rudin, 2019]. Advances in this area may include leveraging discourse-level relations and training models with human explanations for common sense reasoning to improve the explanatory depth of model-generated attributions [Rajani et al., 2019]. For instance, Kowalczyk et al. [2022] detected GPT-2-generated fake reviews using Shapley Additive Explanations (SHAP) [Lundberg and Lee, 2017].

7.4 Misuse Prevention

Future research should focus on refining existing authorship attribution methods to detect and prevent malicious activities such as generating misinformation, plagiarism, and propaganda [Goldstein et al., 2023; Hazell, 2023; Spitale et al., 2023; Lund et al., 2023]. These methods analyze stylistic features to detect discrepancies in claimed authorship and trace the origins of content, thereby identifying suspicious texts. For plagiarism, attribution models can compare writing styles with a database of known authors. In combating misinformation and propaganda, these models could identify and flag content patterns typical of known propagandists.

To ensure effectiveness in real-world tasks, authorship attribution models should be robust against out-of-domain data and adversarial attacks. Adversarial attacks including alternative spellings, article deletion, paragraph additions, case changes, zero-width spaces, whitespace manipulation, homoglyphs, number swaps, misspellings, paraphrasing, and synonym substitution, have been shown to effectively degrade detector performance [Dugan et al., 2024]. Diversifying training data with various writing styles and topics could improve robustness. Adversarial robustness could be achieved through adversarial training and employing ensemble methods to build resilience against intentional manipulations.

7.5 Leveraging LLM Capabilities

Leveraging LLMs can enhance both traditional feature-based stylometry methods and LLM-based approaches. By integrating LLMs with existing methods, researchers can gain deeper insight into stylistic nuances, improving the robustness of authorship detection across various textual genres and lengths. The increasingly large context length of LLMs enables in-context learning (ICL) [Brown et al., 2020] by incorporating more documents, enhancing the model's ability to capture intricate writing patterns.

Another promising approach is Retrieval-Augmented Generation (RAG) [Lewis et al., 2020]. RAG can enhance authorship attribution by retrieving additional documents for each author, thereby assisting in generating more contextually accurate results. Moreover, leveraging LLMs for data augmentation and synthetic data generation can create diverse training datasets, which in turn improves the generalization of attribution models [Albalak et al., 2024].

Combining text detection with other modalities, such as images, videos, or metadata, can potentially improve the accuracy and reliability of authorship attribution. Cross-modal analysis enables the integration of various data types, providing a more comprehensive view of the content and its origins. This holistic approach not only enhances the attribution process but also paves the way for innovative methodologies that are more resilient to the evolving nature of digital content.

7.6 Developing Standardized Benchmarks

The diversity in datasets and evaluation metrics currently hinders the comparability and generalizability of different authorship attribution methods. Establishing comprehensive benchmarks that encompass a wide range of text types and sources, including human-authored, LLM-generated, and human-LLM coauthored texts, would significantly enhance the field. Unified benchmarks should incorporate diverse text corpora from various genres, lengths, and languages to reflect the breadth of real-world applications. Clear evaluation metrics are essential for providing consistent and transparent measures of attribution accuracy, robustness, and computational efficiency, enabling fair comparisons between different models. Benchmarks should also include datasets that blend human-written and machine-generated content to simulate realistic tasks and test the robustness of attribution models.

To ensure ongoing relevance and challenge for attribution methods, benchmarks must be regularly updated to include new types of LLMs and detectors. By developing and adopting standardized benchmarks, the research community can foster more rigorous, reproducible, and comparable studies. This will ultimately drive advancements in authorship attribution methodologies and applications. These standardized benchmarks would serve as a foundation for the systematic evaluation of attribution techniques, promoting innovation and progress in addressing the complexities of authorship attribution in a rapidly evolving digital landscape.

7.7 Integrating Interdisciplinary Perspectives

Authorship attribution is inherently multidisciplinary, encompassing elements of linguistics, computer science, forensic science, and psychology [Stamatatos, 2009]. Future research should continue fostering collaboration across these fields to integrate diverse perspectives. This integrative approach can lead to innovative solutions and a deeper understanding of the challenges and potential of authorship attribution. Combining insights from various disciplines can foster the creation of holistic attribution models that account for both the intricacies of human language and the complexities of LLM-generated texts. Such collaboration could also spearhead initiatives to standardize evaluation metrics for authorship attribution tools, ensuring their effectiveness across diverse contexts and compliance with ethical standards.

Linguistics can dissect textual structures and stylistic nuances, identifying unique linguistic fingerprints of authors. It explores novel features that enhance robustness across domains and improve explainability. Forensic science contributes through technological tools and methodologies, enabling a precise examination of physical and digital texts. Psychology, particularly psycholinguistics, provides insight into how the brain processes function words and grammatical markers distinctively from lexical content words, revealing correlations with socio-cultural categories such as gender, age, and native language [Chambers et al., 2013; Nerbonne, 2014; Seals and Shalin, 2023], which are pivotal in understanding identity and social affiliations [Argamon et al., 2009]. The Linguistic Inquiry and Word Count (LIWC) tool exemplifies how automated text analysis can use more than 100 psychological dimensions to analyze word use, reflecting distinct language variations among different groups in specific genres and languages [Goldstein-Stewart et al., 2009; Pennebaker et al., 2015; Dudău and Sava, 2021].

Combining these interdisciplinary perspectives enhances our ability to distinguish between human- and machine-generated texts, addressing the emerging challenges posed by sophisticated language models. By integrating linguistic theory with advanced computational techniques, forensic methodologies, and psychological insights, researchers can develop more comprehensive and nuanced authorship attribution frameworks. These frameworks will be better equipped to handle the diverse range of writing styles and contexts, ultimately leading to more accurate and reliable attribution outcomes. Furthermore, interdisciplinary collaboration can drive the development of ethical guidelines and best practices, ensuring that authorship attribution is conducted responsibly and with respect for individuals' privacy and rights.

8. ETHICAL AND PRIVACY CONCERNS

In the evolving landscape of authorship attribution, it is crucial to prioritize ethical considerations to safeguard privacy, integrity, and the rightful ownership of content. The attribution of text to specific authors or models raises significant ethical and privacy issues. Misattribution can lead to wrongful accusations or misinterpretation of an author's intent [Lund et al., 2023]. Additionally, the use of attribution technologies must balance the need for accountability with respect to individuals' privacy and the potential for misuse in surveilling or censoring content.

Authorship attribution techniques are essential in digital forensics, cybersecurity, and plagiarism detection. However, the potential to reveal the identities of anonymous authors presents significant ethical challenges. Applications such as linking user accounts across platforms and identifying compromised accounts raise privacy concerns and ethical questions about surveillance and profiling individuals based on their writing style.

The use of authorship attribution methods must be carefully managed to protect individual privacy and adhere to ethical standards, particularly in sensitive areas such as journalism, political dissent, and corporate whistle-blowing [Sison et al., 2023]. Ensuring that these methods are not used to undermine privacy rights or expose individuals to risks without their consent is essential. Despite existing measures to prohibit the unethical use of LLMs, these restrictions could be evaded through prompt engineering and jail-breaking, posing risks of phishing and fraud scams.

Furthermore, the increasing difficulty in distinguishing between human and LLM-generated content raises concerns about intellectual property, plagiarism, and accountability. Accurate attribution is crucial for maintaining academic and creative integrity, yet tools and methods for achieving this must evolve rapidly to keep up with technological advancements. The deployment of LLMs in generating content across various domains—from journalism to literature—necessitates a rethinking of authorship norms and the legal frameworks governing creative works.

9. CONCLUSION

The field of authorship attribution is experiencing both unprecedented challenges and remarkable opportunities with the advent of LLMs. Whether the objective is to identify human authors, differentiate between human- and machine-generated texts, attribute texts to specific LLMs, or manage the complexities of human-LLM co-authored texts, ongoing innovation is imperative. Effectively addressing these multifaceted issues requires interdisciplinary approaches and collaborative efforts among researchers. This survey explores various problems within authorship attribution, offering a comprehensive comparison of methodologies and datasets. By integrating robustness, explainability, and interdisciplinary perspectives, we highlight the importance of developing methods that are not only accurate but also socially relevant and trustworthy. We highlight the strengths and limitations of current approaches, identify key open problems, and outline future research directions. This holistic analysis equips researchers and practitioners with the knowledge necessary to navigate the evolving landscape of authorship attribution, emphasizing critical areas for future research and development.

Acknowledgments

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-07-04, NSF awards (SaTC-2241068, IIS-2339198, and POSE-2346158), a Cisco Research Award, and a Microsoft Accelerate Foundation Models Research Award. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security and the National Science Foundation.

References

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

Opeyemi Aborisade and Mohd Anwar. 2018. Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 269–276.

Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *2012 IEEE Symposium on Security and Privacy*. IEEE, 461–475.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. 2024. A survey on data selection for language models. *ArXiv preprint* abs/2402.16827 (2024). https://arxiv.org/abs/2402.16827

Malik Altakrori, Jackie Chi Kit Cheung, and Benjamin C. M. Fung. 2021. The Topic Confusion Task: A Novel Evaluation Scenario for Authorship Attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 4242–4256. https://doi.org/10.18653/v1/2021.findings-emnlp.359

Nicholas Andrews and Marcus Bishop. 2019. Learning Invariant Representations of Social Media Users. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 1684–1695. https://doi.org/10.18653/v1/D19-1178

Shlomo Argamon. 2018. Computational forensic authorship analysis: Promises and pitfalls. *Language and Law/Linguagem e Direito* 5, 2 (2018), 7–37.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text & talk* 23, 3 (2003), 321–346.

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM* 52, 2 (2009), 119–123.

Mikhail J Atallah, Victor Raskin, Michael Crogan, Christian Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. 2001. 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*. Springer, 185–200.

Andrea Bacciu, Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, Valerio Neri, and Julinda Stefa. 2019. Cross-Domain Authorship Attribution Combining Instance Based and Profile-Based Features.. In *CLEF (Working Notes)*.

Douglas Bagnall. 2015. Author identification using multi-headed recurrent neural networks. *ArXiv preprint* abs/1506.04891 (2015). https://arxiv.org/abs/1506.04891

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *ArXiv preprint* abs/2310.05130 (2023). https://arxiv.org/abs/2310.05130

Sylvio Barbon, Rodrigo Augusto Igawa, and Bruno Bogaz Zarpelão. 2017. Authorship verification applied to detection of compromised accounts on online social networks: A continuous approach. *Multimedia Tools and Applications* 76 (2017), 3213–3233.

Georgios Barlas and Efstathios Stamatatos. 2020a. Cross-domain authorship attribution using pre-trained language models. In *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I 16*. Springer, 255–266.

Georgios Barlas and Efstathios Stamatatos. 2020b. Cross-domain authorship attribution using pre-trained language models. In *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I 16*. Springer, 255–266.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 830–839.

Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2023. Paraphrase detection: Human vs. machine content. *ArXiv preprint* abs/2303.13989 (2023). https://arxiv.org/abs/2303.13989

Nicole Mariah Sharon Belvisi, Naveed Muhammad, and Fernando Alonso-Fernandez. 2020. Forensic authorship analysis of microblogging texts using n-grams and stylometric features. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 1–6.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* 79 (2010), 151–175.

Daria Beresneva. 2016. Computer-generated text detection using machine learning: A systematic review. In *Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings 21*. Springer, 421–426.

Janek Bevendorff, Berta Chulvi, Elisabetta Fersini, Annina Heini, Mike Kestemont, Krzysztof Kredens, Maximilian Mayerl, Reynier Ortega-Bueno, Piotr Pęzik, Martin Potthast, et al. 2022. Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 382–394.

Meghana Moorthy Bhat and Srinivasan Parthasarathy. 2020a. How Effectively Can Machines Defend Against Machine-Generated Fake News? An Empirical Study. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*. Association for Computational Linguistics, Online, 48–53. https://doi.org/10.18653/v1/2020.insights-1.7

Meghana Moorthy Bhat and Srinivasan Parthasarathy. 2020b. How Effectively Can Machines Defend Against Machine-Generated Fake News? An Empirical Study. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*. Association for Computational Linguistics, Online, 48–53. https://doi.org/10.18653/v1/2020.insights-1.7

Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. 2023. Conda: Contrastive domain adaptation for ai-generated text detection. *ArXiv preprint* abs/2309.03992 (2023). https://arxiv.org/abs/2309.03992

Sebastian Bischoff, Niklas Deckers, Marcel Schliebs, Ben Thies, Matthias Hagen, Efstathios Stamatatos, Benno Stein, and Martin Potthast. 2020. The importance of suppressing domain style in authorship analysis. *ArXiv preprint* abs/2005.14714 (2020). https://arxiv.org/abs/2005.14714

Benedikt T. Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M. Nickel. 2019a. Explainable Authorship Verification in Social Media via Attention-based Similarity Learning. In *2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, December 9-12, 2019*. IEEE, 36–45. https://doi.org/10.1109/BigData47090.2019.9005650

Benedikt T. Boenninghoff, Robert M. Nickel, Steffen Zeiler, and Dorothea Kolossa. 2019b. Similarity Learning for Authorship Verification in Social Media. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, 2457–2461. https://doi.org/10.1109/ICASSP.2019.8683405

Ilker Nadi Bozkurt, Ozgur Baghoglu, and Erkan Uyar. 2007. Authorship attribution. In *2007 22nd international symposium on computer and information sciences*. IEEE, 1–5.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Florian Cafiero and Jean-Baptiste Camps. 2023. Who could be behind QAnon? Authorship attribution with supervised machine-learning. *ArXiv preprint* abs/2303.02078 (2023). https://arxiv.org/abs/2303.02078

Shuyang Cai and Wanyun Cui. 2023. Evade ChatGPT detectors via a single space. *ArXiv preprint* abs/2307.02599 (2023). https://arxiv.org/abs/2307.02599

Megha Chakraborty, SM Tonmoy, SM Zaman, Krish Sharma, Niyar R Barman, Chandan Gupta, Shreya Gautam, Tanay Kumar, Vinija Jain, Aman Chadha, et al. 2023b. Counter Turing Test CTˆ 2: AI-Generated Text Detection is Not as Easy as You May Think–Introducing AI Detectability Index. *ArXiv preprint* abs/2310.05030 (2023). https://arxiv.org/abs/2310.05030

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023a. On the possibilities of ai-generated text detection. *ArXiv preprint* abs/2304.04736 (2023). https://arxiv.org/abs/2304.04736

Jack K Chambers, Peter Trudgill, and Natalie Schilling-Estes. 2013. *The handbook of language variation and change*. Wiley Online Library.

Carole E Chaski. 2005. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International journal of digital evidence* 4, 1 (2005), 1–13.

Canyu Chen and Kai Shu. 2024a. Can LLM-Generated Misinformation Be Detected?. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=ccxD4mtkTU

Canyu Chen and Kai Shu. 2024b. Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine* (2024). https://doi.org/10.1002/aaai.12188

Canyu Chen, Haoran Wang, Matthew Shapiro, Yunyu Xiao, Fei Wang, and Kai Shu. 2022b. Combating health misinformation in social media: Characterization, detection, intervention, and open issues. *ArXiv preprint* abs/2211.05289 (2022). https://arxiv.org/abs/2211.05289

Xingyuan Chen, Peng Jin, Siyuan Jing, and Chunming Xie. 2022a. Automatic detection of Chinese generated essayss based on pre-trained BERT. In *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, Vol. 10. IEEE, 2257–2260.

Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. 2023a. Gpt-sentinel: Distinguishing human and chatgpt generated content. *ArXiv preprint* abs/2305.07969 (2023). https://arxiv.org/abs/2305.07969

Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. 2023b. Token prediction as implicit classification to identify LLM-generated text. *ArXiv preprint* abs/2311.08723 (2023). https://arxiv.org/abs/2311.08723

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565

Evan Crothers, Nathalie Japkowicz, and Herna L Viktor. 2023. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access* (2023).

Joseph Cutler, Liam Dugan, Shreya Havaldar, and Adam Stein. 2021. Automatic Detection of Hybrid Human-Machine Text Boundaries. (2021).

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Suzhou, China, 447–459. https://aclanthology.org/2020.aacl-main.46

Edwin Dauber, Rebekah Overdorf, and Rachel Greenstadt. 2017. Stylometric authorship attribution of collaborative documents. In *Cyber Security Cryptography and Machine Learning: First International Conference, CSCML 2017, Beer-Sheva, Israel, June 29-30, 2017, Proceedings 1*. Springer, 115–135.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Diana Paula Dudău and Florin Alin Sava. 2021. Performing multilingual analysis with Linguistic Inquiry and Word Count 2015 (LIWC2015). An equivalence study of four languages. *Frontiers in Psychology* 12 (2021), 570568.

Liam Dugan, Alyssa Hwang, Filip Trhlik, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors. *ArXiv preprint* abs/2405.07940 (2024). https://arxiv.org/abs/2405.07940

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 12763–12771.

Maciej Eder. 2015. Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities* 30, 2 (2015), 167–182.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for Authorship Attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*. NLP Association of India (NLPAI), Indian Institute of Technology Patna, Patna, India, 127–137. https://aclanthology.org/2020.icon-main.16

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. TweepFake: About detecting deepfake tweets. *Plos one* 16, 5 (2021), e0251415.

Vitalii Fishchuk and Daniel Braun. 2023. Efficient Black-Box Adversarial Attacks on Neural Text Detectors. *ArXiv preprint* abs/2311.01873 (2023). https://arxiv.org/abs/2311.01873

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

Rinaldo Gagiano, Maria Myung-Hee Kim, Xiuzhen Zhang, and Jennifer Biggs. 2021. Robustness Analysis of Grover for Machine-Generated News Detection. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*. Australasian Language Technology Association, Online, 119–127. https://aclanthology.org/2021.alta-1.12

Matthias Gallé, Jos Rozen, Germán Kruszewski, and Hady Elsahar. 2021. Unsupervised and distributional detection of machine-generated text. *ArXiv preprint* abs/2111.02878 (2021). https://arxiv.org/abs/2111.02878

Margherita Gambini, Tiziano Fagni, Fabrizio Falchi, and Maurizio Tesconi. 2022. On pushing DeepFake Tweet Detection capabilities to the limits. In *14th ACM Web Science Conference 2022*. 154–163.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* 17, 1 (2016), 2096–2030.

Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. 2022. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *BioRxiv* (2022), 2022–12.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019a. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Florence, Italy, 111–116. https://doi.org/10.18653/v1/P19-3019

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019b. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Florence, Italy, 111–116. https://doi.org/10.18653/v1/P19-3019

Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText Corpus. http://Skylion007.github.io/OpenWebTextCorpus

Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv:2301.04246 [cs.CY]

Jade Goldstein-Stewart, Ransom Winder, and Roberta Sabin. 2009. Person Identification from Text and Speech Genre Samples. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Association for Computational Linguistics, Athens, Greece, 336–344. https://aclanthology.org/E09-1039

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6572

Tim Grant. 2007. Quantifying evidence in forensic authorship analysis. *International Journal of Speech, Language & the Law* 14, 1 (2007).

Tim Grant. 2020. Text messaging forensics: Txt 4n6: idiolect-free authorship analysis? In *The Routledge handbook of forensic linguistics*. Routledge, 558–575.

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

Oren Halvani and Lukas Graner. 2021. Posnoise: An effective countermeasure against topic biases in authorship analysis. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*. 1–12.

Hans WA Hanley and Zakir Durumeric. 2024. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 542–556.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. *ArXiv preprint* abs/2401.12070 (2024). https://arxiv.org/abs/2401.12070

Julian Hazell. 2023. Large language models can be used to effectively scale spear phishing campaigns. *ArXiv preprint* abs/2305.06972 (2023). https://arxiv.org/abs/2305.06972

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. Mgtbench: Benchmarking machine-generated text detection. *ArXiv preprint* abs/2303.14822 (2023). https://arxiv.org/abs/2303.14822

Pezik-P. Heini, A. and K. Kredens. 2021. GPTZero: Towards detection of AI-generated text using zero-shot and supervised methods. http://fold.aston.ac.uk/handle/123456789/17

Graeme Hirst and Ol'ga Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing* 22, 4 (2007), 405–417.

David I Holmes. 1994. Authorship attribution. *Computers and the Humanities* 28 (1994), 87–106.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=rygGQyrFvH

Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. 2023. Semstamp: A semantic watermark with paraphrastic robustness for text generation. *ArXiv preprint* abs/2310.03991 (2023). https://arxiv.org/abs/2310.03991

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 328–339. https://doi.org/10.18653/v1/P18-1031

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in Neural Information Processing Systems* 36 (2023), 15077–15095.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can Large Language Models Identify Authorship? *ArXiv preprint* abs/2403.08213 (2024). https://arxiv.org/abs/2403.08213

Javier Huertas-Tato, Alvaro Huertas-Garcia, Alejandro Martin, and David Camacho. 2022. PART: Pre-trained Authorship Representation Transformer. *ArXiv preprint* abs/2209.15373 (2022). https://arxiv.org/abs/2209.15373

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic Detection of Generated Text is Easiest when Humans are Fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1808–1822. https://doi.org/10.18653/v1/2020.acl-main.164

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

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020a. Automatic Detection of Machine Generated Text: A Critical Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 2296–2309. https://doi.org/10.18653/v1/2020.coling-main.208

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020b. Automatic Detection of Machine Generated Text: A Critical Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 2296–2309. https://doi.org/10.18653/v1/2020.coling-main.208

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008*, Marc Najork, Andrei Z. Broder, and Soumen Chakrabarti (Eds.). ACM, 219–230. https://doi.org/10.1145/1341531.1341560

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 427–431. https://aclanthology.org/E17-2068

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2021. Overview of the Cross-Domain Authorship Verification Task at PAN 2021.. In *CLEF (Working Notes)*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*. PMLR, 17061–17084.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings 15*. Springer, 217–226.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. 2006. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 659–660.

Moshe Koppel, Jonathan Schler, and Eran Messeri. 2008. Authorship attribution in law enforcement scenarios. *NATO Security Through Science Series D-Information and Communication Security* 15 (2008), 111.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. 624–628.

Peter Kowalczyk, Marco Röder, Alexander Dürr, and Frédéric Thiesse. 2022. Detecting and understanding textual deepfakes in online reviews. (2022).

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2024a. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems* 36 (2024).

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2024b. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems* 36 (2024).

Pranav Kulkarni, Ziqing Ji, Yan Xu, Marko Neskovic, and Kevin Nolan. 2023. Exploring Semantic Perturbations on Grover. *ArXiv preprint* abs/2302.00509 (2023). https://arxiv.org/abs/2302.00509

Tharindu Kumarage and Huan Liu. 2023. Neural Authorship Attribution: Stylometric Analysis on Large Language Models. *ArXiv preprint* abs/2308.07305 (2023). https://arxiv.org/abs/2308.07305

Tharindu Kumarage, Paras Sheth, Raha Moraffah, Joshua Garland, and Huan Liu. 2023. How reliable are ai-generated-text detectors? an assessment framework using evasive soft prompts. *ArXiv preprint* abs/2310.05095 (2023). https://arxiv.org/abs/2310.05095

Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and PG Demidov. 2019. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*. IEEE, 184–195.

Thomas Lavergne, Tanguy Urvoy, and François Yvon. 2008. Detecting Fake Content with Relative Entropy Scoring. *Pan* 8, 27-31 (2008), 4.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Linyang Li, Pengyu Wang, Ke Ren, Tianxiang Sun, and Xipeng Qiu. 2023b. Origin tracing and detecting of llms. *ArXiv preprint* abs/2304.14072 (2023). https://arxiv.org/abs/2304.14072

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023a. MAGE: Machine-generated Text Detection in the Wild. https://arxiv.org/abs/2305.13242

Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-End Adversarial Memory Network for Cross-domain Sentiment Classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, Carles Sierra (Ed.). ijcai.org, 2237–2243. https://doi.org/10.24963/ijcai.2017/311

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns* 4, 7 (2023).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint* abs/1907.11692 (2019). https://arxiv.org/abs/1907.11692

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023c. ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models. *ArXiv preprint* abs/2304.07666 (2023). https://arxiv.org/abs/2304.07666

Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2023a. On the Detectability of ChatGPT Content: Benchmarking, Methodology, and Evaluation through the Lens of Academic Writing. https://arxiv.org/abs/2306.05524

Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2023b. On the detectability of chatgpt content: benchmarking, methodology, and evaluation through the lens of academic writing. *arXiv e-prints* (2023), arXiv–2306.

Ning Lu, Shengcai Liu, Rui He, Qi Wang, Yew-Soon Ong, and Ke Tang. 2023. Large language models can be guided to evade ai-generated text detection. *ArXiv preprint* abs/2305.10847 (2023). https://arxiv.org/abs/2305.10847

Brady D Lund, Ting Wang, Nishith Reddy Mannuru, Bing Nie, Somipam Shimray, and Ziang Wang. 2023. ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology* 74, 5 (2023), 570–581.

Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Samuel Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, et al. 2023. MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark. *ArXiv preprint* abs/2310.13606 (2023). https://arxiv.org/abs/2310.13606

Dominik Macko, Robert Moro, Adaku Uchendu, Ivan Srba, Jason Samuel Lucas, Michiharu Yamashita, Nafis Irtiza Tripto, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2024. Authorship obfuscation in multilingual machine-generated text detection. *ArXiv preprint* abs/2401.07867 (2024). https://arxiv.org/abs/2401.07867

David Madigan, Alexander Genkin, David D Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye. 2005b. Author identification on the large scale. In *Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA)*.

David Madigan, Alexander Genkin, David D Lewis, and Dmitriy Fradkin. 2005a. Bayesian multinomial logistic regression for author identification. In *AIP conference proceedings*, Vol. 803. American Institute of Physics, 509–516.

Andrei Manolache, Florin Brad, Elena Burceanu, Antonio Barbalau, Radu Ionescu, and Marius Popescu. 2021. Transferring bert-like transformers' knowledge for authorship verification. *ArXiv preprint* abs/2112.05125 (2021). https://arxiv.org/abs/2112.05125

Yuval Marton, Ning Wu, and Lisa Hellerstein. 2005. On compression-based text classification. In *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27*. Springer, 300–314.

Rohith Menon and Yejin Choi. 2011. Domain Independent Authorship Attribution without Domain Adaptation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Association for Computational Linguistics, Hissar, Bulgaria, 309–315. https://aclanthology.org/R11-1043

Hasan Mesut Meral, Bülent Sankur, A Sumru Özsoy, Tunga Güngör, and Emre Sevinç. 2009. Natural language watermarking via morphosyntactic alterations. *Computer Speech & Language* 23, 1 (2009), 107–125.

Sven Meyer zu Eissen, Benno Stein, and Marion Kulig. 2007. Plagiarism detection without reference collections. In *Advances in Data Analysis: Proceedings of the 30 th Annual Conference of the Gesellschaft für Klassifikation eV, Freie Universität Berlin, March 8–10, 2006*. Springer, 359–366.

Niloofar Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick. 2023. Smaller language models are better black-box machine-generated text detectors. *ArXiv preprint* abs/2305.09859 (2023). https://arxiv.org/abs/2305.09859

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *ArXiv preprint* abs/2301.11305 (2023). https://arxiv.org/abs/2301.11305

Edoardo Mosca, Mohamed Hesham Ibrahim Abdalla, Paolo Basso, Margherita Musumeci, and Georg Groh. 2023. Distinguishing Fact from Fiction: A Benchmark Dataset for Identifying Machine-Generated Scientific Papers in the LLM Era.. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. 190–207.

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *J. Amer. Statist. Assoc.* 58, 302 (1963), 275–309.

Shaoor Munir, Brishna Batool, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2021. Through the Looking Glass: Learning to Attribute Synthetic Text Generated by Language Models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 1811–1822. https://doi.org/10.18653/v1/2021.eacl-main.155

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 1797–1807. https://doi.org/10.18653/v1/D18-1206

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)* 50, 6 (2017), 1–36.

John Nerbonne. 2014. The secret life of pronouns. what our words say about us. *Literary and Linguistic Computing* 29, 1 (2014), 139–142.

Hoang-Quoc Nguyen-Son, Ngoc-Dung T Tieu, Huy H Nguyen, Junichi Yamagishi, and Isao Echi Zen. 2017. Identifying computer-generated text using statistical analysis. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1504–1511.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 188–197. https://doi.org/10.18653/v1/D19-1018

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 309–319. https://aclanthology.org/P11-1032

Ajay Patel, Delip Rao, and Chris Callison-Burch. 2023. Learning Interpretable Style Embeddings via Prompting LLMs. *ArXiv preprint* abs/2305.12696 (2023). https://arxiv.org/abs/2305.12696

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.

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

Nektaria Potha and Efstathios Stamatatos. 2019. Improving author verification based on topic modeling. *Journal of the Association for Information Science and Technology* 70, 10 (2019), 1074–1088.

Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, and Bimal Viswanath. 2023a. Deepfake text detection: Limitations and opportunities. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1613–1630.

Xiao Pu, Jingyu Zhang, Xiaochuang Han, Yulia Tsvetkov, and Tianxing He. 2023b. On the zero-shot generalization of machine-generated text detectors. *ArXiv preprint* abs/2310.05165 (2023). https://arxiv.org/abs/2310.05165

Chen Qian, Tianchang He, and Rao Zhang. 2017. Deep learning based authorship identification. *Report, Stanford University* (2017), 1–9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. http://jmlr.org/papers/v21/20-074.html

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, 4932–4942. https://doi.org/10.18653/v1/P19-1487

Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. 2022. Effect of scale on catastrophic forgetting in neural networks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=GhVS8_yPeEa

Aquia Richburg, Calvin Bao, and Marine Carpuat. 2024. Automatic Authorship Analysis in Human-AI Collaborative Writing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 1845–1855.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning Universal Authorship Representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 913–919. https://doi.org/10.18653/v1/2021.emnlp-main.70

Anderson Rocha, Walter J Scheirer, Christopher W Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne RB Carvalho, and Efstathios Stamatatos. 2016. Authorship attribution for social media forensics. *IEEE transactions on information forensics and security* 12, 1 (2016), 5–33.

Juan Rodriguez, Todd Hay, David Gros, Zain Shamsi, and Ravi Srinivasan. 2022a. 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*. Association for Computational Linguistics, Seattle, United States, 1213–1233. https://doi.org/10.18653/v1/2022.naacl-main.88

Juan Rodriguez, Todd Hay, David Gros, Zain Shamsi, and Ravi Srinivasan. 2022b. 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*. Association for Computational Linguistics, Seattle, United States, 1213–1233. https://doi.org/10.18653/v1/2022.naacl-main.88

Sebastian Ruder, Parsa Ghaffari, and John G Breslin. 2016. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *ArXiv preprint* abs/1609.06686 (2016). https://arxiv.org/abs/1609.06686

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

Joseph Rudman. 1997. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities* 31 (1997), 351–365.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *ArXiv preprint* abs/2303.11156 (2023). https://arxiv.org/abs/2303.11156

Chakaveh Saedi and Mark Dras. 2021. Siamese networks for large-scale author identification. *Computer Speech & Language* 70 (2021), 101241.

Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services* 64 (2022), 102771.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *ArXiv preprint* abs/2309.11285 (2023). https://arxiv.org/abs/2309.11285

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging.. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, Vol. 6. 199–205.

Ipek Baris Schlicht and Angel Felipe Magnossão de Paula. 2021. Unified and multilingual author profiling for detecting haters. *ArXiv preprint* abs/2109.09233 (2021). https://arxiv.org/abs/2109.09233

Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The Limitations of Stylometry for Detecting Machine-Generated Fake News. *Computational Linguistics* 46, 2 (2020), 499–510. https://doi.org/10.1162/coli_a_00380

SM Seals and Valerie L Shalin. 2023. Long-form analogies generated by chatGPT lack human-like psycholinguistic properties. *ArXiv preprint* abs/2306.04537 (2023). https://arxiv.org/abs/2306.04537

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. Do Massively Pretrained Language Models Make Better Storytellers?. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, Hong Kong, China, 843–861. https://doi.org/10.18653/v1/K19-1079

Yanir Seroussi, Russell Smyth, and Ingrid Zukerman. 2011. Ghosts from the high court's past: Evidence from computational linguistics for Dixon ghosting for Mctiernan and rich. *University of New South Wales Law Journal, The* 34, 3 (2011), 984–1005.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship Attribution with Topic Models. *Computational Linguistics* 40, 2 (2014), 269–310. https://doi.org/10.1162/COLI_a_00173

Danish Shakeel and Nitin Jain. 2021. Fake news detection and fact verification using knowledge graphs and machine learning. *ResearchGate preprint* 10 (2021).

Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the the ruatd shared task 2022 on artificial text detection in russian. *ArXiv preprint* abs/2206.01583 (2022). https://arxiv.org/abs/2206.01583

Abhay Sharma, Ananya Nandan, and Reetika Ralhan. 2018. An investigation of supervised learning methods for authorship attribution in short hinglish texts using char & word n-grams. *ArXiv preprint* abs/1812.10281 (2018). https://arxiv.org/abs/1812.10281

Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. 2024. Red teaming language model detectors with language models. *Transactions of the Association for Computational Linguistics* 12 (2024), 174–189.

Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017a. Convolutional Neural Networks for Authorship Attribution of Short Texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 669–674. https://aclanthology.org/E17-2106

Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017b. Convolutional Neural Networks for Authorship Attribution of Short Texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 669–674. https://aclanthology.org/E17-2106

Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Mining disinformation and fake news: Concepts, methods, and recent advancements. *Disinformation, misinformation, and fake news in social media: Emerging research challenges and opportunities* (2020), 1–19.

Kai Shu, Suhang Wang, Jiliang Tang, Reza Zafarani, and Huan Liu. 2017. User identity linkage across online social networks: A review. *Acm Sigkdd Explorations Newsletter* 18, 2 (2017), 5–17.

Richard Sinnott and Zijian Wang. 2021. Linking user accounts across social media platforms. In *2021 IEEE/ACM 8th International Conference on Big Data Computing, Applications and Technologies (BDCAT'21)*. 18–27.

Alejo Jose G Sison, Marco Tulio Daza, Roberto Gozalo-Brizuela, and Eduardo C Garrido-Merchán. 2023. ChatGPT: more than a "weapon of mass deception" ethical challenges and responses from the human-centered artificial intelligence (HCAI) perspective. *International Journal of Human–Computer Interaction* (2023), 1–20.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *ArXiv preprint* abs/1908.09203 (2019). https://arxiv.org/abs/1908.09203

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé III, Jesse Dodge, Isabella Duan, et al. 2023. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949* (2023).

Rafael Rivera Soto, Kailin Koch, Aleem Khan, Barry Chen, Marcus Bishop, and Nicholas Andrews. 2024. Few-Shot Detection of Machine-Generated Text using Style Representations. *ArXiv preprint* abs/2401.06712 (2024). https://arxiv.org/abs/2401.06712

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. AI model GPT-3 (dis) informs us better than humans. *Science Advances* 9, 26 (2023), eadh1850.

Efstathios Stamatatos. 2008. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management* 44, 2 (2008), 790–799.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60, 3 (2009), 538–556.

Efstathios Stamatatos. 2016. Authorship verification: a review of recent advances. *Research in Computing Science* 123 (2016), 9–25.

Efstathios Stamatatos and Moshe Koppel. 2011. Plagiarism and authorship analysis: introduction to the special issue. *Language Resources and Evaluation* 45 (2011), 1–4.

Harald Stiff and Fredrik Johansson. 2022. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics* 13, 4 (2022), 363–383.

Ariel Stolerman, Rebekah Overdorf, Sadia Afroz, and Rachel Greenstadt. 2014. Breaking the closed-world assumption in stylometric authorship attribution. In *Advances in Digital Forensics X: 10th IFIP WG 11.9 International Conference, Vienna, Austria, January 8-10, 2014, Revised Selected Papers 10*. Springer, 185–205.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023b. DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. *ArXiv preprint* abs/2306.05540 (2023). https://arxiv.org/abs/2306.05540

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023c. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *ArXiv preprint* abs/2306.05540 (2023). https://arxiv.org/abs/2306.05540

Zhenpeng Su, Xing Wu, Wei Zhou, Guangyuan Ma, and Songlin Hu. 2023a. Hc3 plus: A semantic-invariant human chatgpt comparison corpus. *ArXiv preprint* abs/2309.02731 (2023). https://arxiv.org/abs/2309.02731

Kalaivani Sundararajan and Damon Woodard. 2018. What represents "style" in authorship attribution?. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2814–2822. https://aclanthology.org/C18-1238

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023a. The science of detecting llm-generated texts. *ArXiv preprint* abs/2303.07205 (2023). https://arxiv.org/abs/2303.07205

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023b. The science of detecting llm-generated texts. *ArXiv preprint* abs/2303.07205 (2023). https://arxiv.org/abs/2303.07205

Antonio Theophilo, Romain Giot, and Anderson Rocha. 2021. Authorship attribution of social media messages. *IEEE Transactions on Computational Social Systems* (2021).

Edward Tian and Alexander Cui. 2023. GPTZero: Towards detection of AI-generated text using zero-shot and supervised methods. https://gptzero.me

Mercan Topkara, Cuneyt M Taskiran, and Edward J Delp III. 2005. Natural language watermarking. In *Security, Steganography, and Watermarking of Multimedia Contents VII*, Vol. 5681. SPIE, 441–452.

Nafis Irtiza Tripto, Adaku Uchendu, Thai Le, Mattia Setzu, Fosca Giannotti, and Dongwon Lee. 2023. HANSEN: human and AI spoken text benchmark for authorship analysis. *ArXiv preprint* abs/2310.16746 (2023). https://arxiv.org/abs/2310.16746

Jacob Tyo, Bhuwan Dhingra, and Zachary C Lipton. 2022. On the state of the art in authorship attribution and authorship verification. *ArXiv preprint* abs/2209.06869 (2022). https://arxiv.org/abs/2209.06869

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship Attribution for Neural Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 8384–8395. https://doi.org/10.18653/v1/2020.emnlp-main.673

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 2001–2016. https://doi.org/10.18653/v1/2021.findings-emnlp.172

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*. Association for Computational Linguistics, Tokyo, Japan, 355–368. https://doi.org/10.18653/v1/W19-8643

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2023. Gpt-who: An information density-based machine-generated text detector. *ArXiv preprint* abs/2310.06202 (2023). https://arxiv.org/abs/2310.06202

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2023. Ghostbuster: Detecting text ghostwritten by large language models. *ArXiv preprint* abs/2305.15047 (2023). https://arxiv.org/abs/2305.15047

Bertie Vidgen, Adarsh Agrawal, Ahmed M Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Borhane Blili-Hamelin, et al. 2024. Introducing v0. 5 of the ai safety benchmark from mlcommons. *arXiv preprint arXiv:2404.12241* (2024).

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Association for Computational Linguistics, Hong Kong, China, 7–12. https://doi.org/10.18653/v1/D19-3002

Wenxiao Wang, Alexander Levine, and Soheil Feizi. 2022. Improved Certified Defenses against Data Poisoning with (Deterministic) Finite Aggregation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 22769–22783. https://proceedings.mlr.press/v162/wang22m.html

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohanned Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, et al. 2024. M4GT-Bench: Evaluation Benchmark for Black-Box Machine-Generated Text Detection. *to appear in ACL 2024* (2024).

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, et al. 2023. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *ArXiv preprint* abs/2305.14902 (2023). https://arxiv.org/abs/2305.14902

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same Author or Just Same Topic? Towards Content-Independent Style Representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Dublin, Ireland, 249–268. https://doi.org/10.18653/v1/2022.repl4nlp-1.26

Jana Winter. 2019. Exclusive: FBI document warns conspiracy theories are a new domestic terrorism threat. *Yahoo News* 1 (2019).

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F Wong, and Lidia S Chao. 2023b. A survey on llm-gernerated text detection: Necessity, methods, and future directions. *ArXiv preprint* abs/2310.14724 (2023). https://arxiv.org/abs/2310.14724

Yihan Wu, Zhengmian Hu, Hongyang Zhang, and Heng Huang. 2023a. Dipmark: A stealthy, efficient and resilient watermark for large language models. *ArXiv preprint* abs/2310.07710 (2023). https://arxiv.org/abs/2310.07710

Han Xu, Jie Ren, Pengfei He, Shenglai Zeng, Yingqian Cui, Amy Liu, Hui Liu, and Jiliang Tang. 2023. On the Generalization of Training-based ChatGPT Detection Methods. arXiv:2310.01307 [cs.CL]

Xianjun Yang, Wei Cheng, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023a. DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text. *ArXiv preprint* abs/2305.17359 (2023). https://arxiv.org/abs/2305.17359

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023b. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *ArXiv preprint* abs/2305.17359 (2023). https://arxiv.org/abs/2305.17359

Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Petzold, William Yang Wang, and Wei Cheng. 2023c. A survey on detection of llms-generated content. *ArXiv preprint* abs/2310.15654 (2023). https://arxiv.org/abs/2310.15654

Xi Yang, Jie Zhang, Kejiang Chen, Weiming Zhang, Zehua Ma, Feng Wang, and Nenghai Yu. 2022. Tracing Text Provenance via Context-Aware Lexical Substitution. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 11613–11621. https://ojs.aaai.org/index.php/AAAI/article/view/21415

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. 2023. Cheat: A large-scale dataset for detecting chatgpt-written abstracts. *ArXiv preprint* abs/2304.12008 (2023). https://arxiv.org/abs/2304.12008

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 9051–9062. https://proceedings.neurips.cc/paper/2019/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html

Haolan Zhan, Xuanli He, Qiongkai Xu, Yuxiang Wu, and Pontus Stenetorp. 2023. G3Detector: General GPT-generated text detector. *ArXiv preprint* abs/2305.12680 (2023). https://arxiv.org/abs/2305.12680

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, et al. 2024. LLM-as-a-Coauthor: Can Mixed Human-Written and Machine-Generated Text Be Detected?. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 409–436.

Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. 2018. Syntax Encoding with Application in Authorship Attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2742–2753. https://doi.org/10.18653/v1/D18-1294

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023. Provable robust watermarking for ai-generated text. *ArXiv preprint* abs/2306.17439 (2023). https://arxiv.org/abs/2306.17439

Wanjun Zhong, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Neural Deepfake Detection with Factual Structure of Text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 2461–2470. https://doi.org/10.18653/v1/2020.emnlp-main.193

George Kingsley Zipf. 2016. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books.