AI-Generated Content Detection and Lightweight Models: A Survey

www.surveyx.cn

Abstract

The rapid advancement of generative models, particularly Large Language Models (LLMs), has heightened the need for effective AI-generated content detection to safeguard the integrity of information across various domains. This survey paper explores the challenges and methodologies associated with detecting AI-generated content, emphasizing lightweight detection models, adversarial robustness, feature fusion, and edge AI. The paper highlights the significance of real-time content analysis and computational efficiency, particularly in resource-constrained environments where rapid decision-making is crucial. Key technologies such as stylometry, metadata analysis, and advanced machine learning models like BERT are examined for their efficacy in identifying AI-generated text. The survey underscores the limitations of current detection methods, including their susceptibility to adversarial attacks and high false positive rates, advocating for more robust evaluation frameworks. Lightweight models, characterized by minimal computational overhead, are pivotal for real-time applications on edge devices. Strategies to enhance adversarial robustness, such as adversarial training and feature disentanglement, are discussed to improve model resilience. Feature fusion techniques are highlighted for their role in integrating diverse data features, thereby enhancing detection accuracy and robustness. Edge AI offers significant advantages in reducing latency and energy consumption, facilitating the deployment of detection systems in dynamic environments. The paper concludes by identifying future research directions, emphasizing the need for innovative approaches to improve model robustness, transparency, and efficiency. By addressing these challenges, the development of reliable AI-generated content detection systems can be advanced, ensuring the ethical and accurate application of AI technologies.

1 Introduction

1.1 Emergence and Importance of AI-Generated Content Detection

The rapid advancements in generative models, particularly Large Language Models (LLMs), have led to a proliferation of AI-generated content that closely mimics human writing in coherence and complexity. This surge raises critical concerns regarding accountability, accuracy, and the potential amplification of biases in widely consumed information sources [1]. In academia, distinguishing between human and AI-generated content is essential to uphold the integrity and validation of scholarly work [2]. The misuse of these generative technologies could result in the dissemination of misleading or harmful information, underscoring the urgent need for effective detection mechanisms [3].

Despite the impressive performance of AI systems, particularly those employing deep learning, current detection methodologies often lack comprehensive testing and focus predominantly on isolated tasks. This limitation highlights the necessity for more robust and reliable solutions [4]. The unregulated deployment of LLMs, renowned for their capabilities in tasks like document completion and question

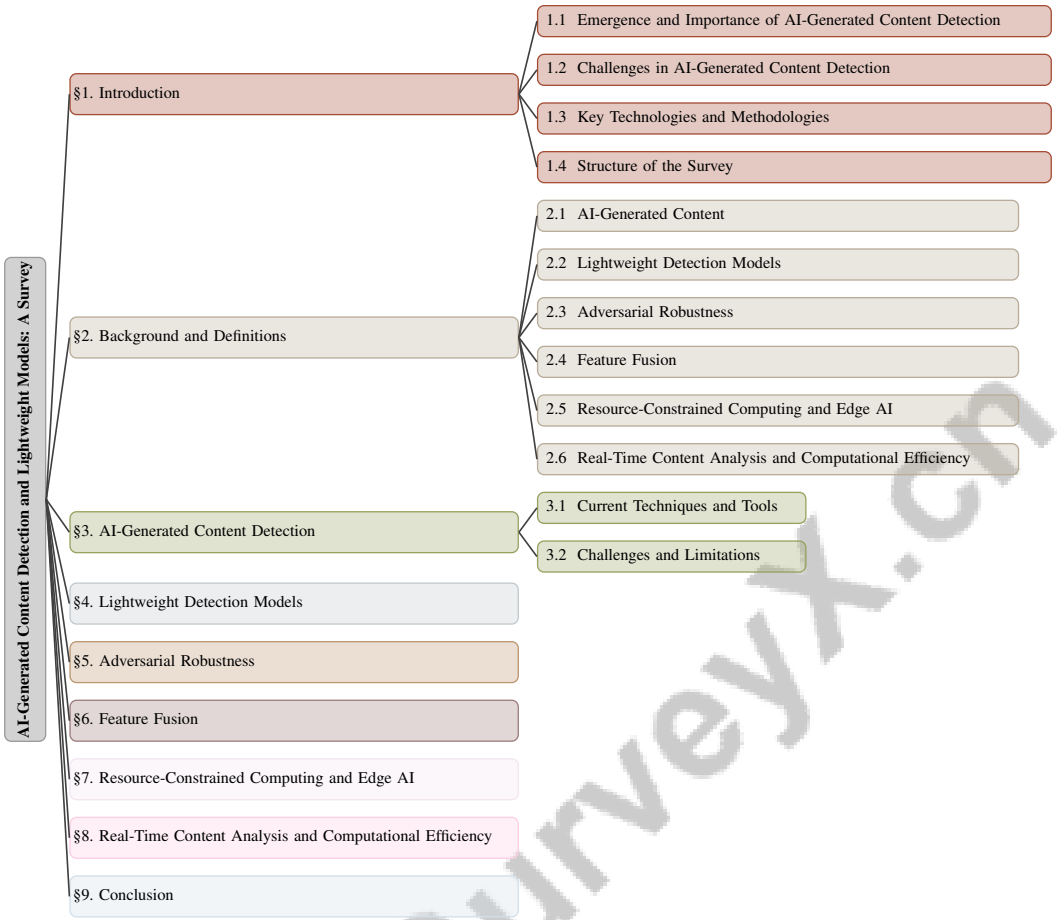

Figure 1: chapter structure

answering, poses risks such as plagiarism and the spread of misinformation [5]. Therefore, developing effective detection mechanisms is crucial for maintaining academic integrity and ensuring the ethical use of AI-generated content [6].

Additionally, the emergence of AI-generated imagery calls for advancements in detection technologies to prevent misuse in critical sectors like journalism and law [7]. The creation of such content, driven by vast datasets, necessitates benchmarks to evaluate model robustness against common corruptions and adversarial attacks [8]. This survey aims to tackle these challenges by evaluating the reliability of machine-generated text detectors and the quality of datasets used in their training, thereby contributing to the development of safe and reliable AI systems [9]. Furthermore, it seeks to provide a new perspective on AI-generated content detection by addressing knowledge gaps, particularly in computational approaches and linguistic aspects [10].

1.2 Challenges in AI-Generated Content Detection

Detecting AI-generated content presents numerous challenges due to the rapid evolution of generative models and the limitations of current detection methodologies. A primary concern is the manipulation of metadata and the sophisticated emulation of human writing styles by AI systems, complicating the differentiation between machine and human-generated content [6]. This issue is compounded by the limitations of existing detection tools, which often lack the transparency and adaptability required to address novel evasion techniques [10].

The computational demands of current detection algorithms pose another significant obstacle, especially in resource-constrained environments where deploying such models on edge devices like Raspberry Pi is impractical [11]. Balancing memory requirements with robustness against adversarial attacks remains unresolved, indicating a critical area for improvement [12].

AI-generated, for reference only.

Existing detection methods excel in specific domains, such as adversarial robustness, but struggle in areas like anomaly detection and calibration [13]. This specialization results in a lack of comprehensive performance across multiple dimensions, necessitating a holistic approach that prioritizes fairness and robustness alongside accuracy [14]. The assumption that adversarial robustness can only be evaluated post fine-tuning complicates the assessment of a model's true capabilities [15].

False positives remain a persistent challenge, as current methods often fail to distinguish between AI-generated and human-authored content, particularly when the former closely resembles human writing [9]. The reliance on handcrafted features and assumptions regarding content characteristics undermines the reliability of detection methods [7]. Furthermore, the robustness of models against real-world corruptions and adversarial attacks is frequently inadequately assessed, revealing significant gaps in understanding model performance under challenging conditions [8].

Evasion techniques, such as recursive paraphrasing, further complicate detection efforts, as they can easily bypass current systems, including those utilizing model signatures and watermarking techniques [5]. Additionally, the incomplete evaluation of adversarial attacks and defenses hampers a comprehensive understanding of their effectiveness and limitations [16].

Addressing these challenges necessitates the development of more robust evaluation frameworks capable of assessing AI detectors across various contexts and performance dimensions. This includes enhancing the explainability and personalization of detection models in natural language processing applications and considering the environmental impact of adversarial machine learning systems, which significantly affects their carbon emissions and presents challenges for sustainable AI development [17].

1.3 Key Technologies and Methodologies

The detection of AI-generated content encompasses a diverse array of technologies and methodologies critical for accurately identifying and analyzing such content. Stylometry, which analyzes writing style to differentiate between human and machine-generated text, is a prominent technique that leverages linguistic features to enhance detection accuracy [6]. Additionally, metadata analysis and online tools like Copyleaks and Turnitin are employed to identify AI-generated content by examining document properties and comparing text against extensive databases [6].

Advanced machine learning models, such as BERT, are pivotal in detecting AI-generated content, offering improved performance over traditional models due to their contextual and semantic understanding [18]. The incorporation of neural and statistical methods further bolsters detection systems by benchmarking their robustness against adversarial attacks, providing a comprehensive evaluation of model resilience [19].

Frameworks that integrate linguistic analysis with computational metrics provide a novel perspective on AI-generated content detection. A survey by Wang et al. introduces a framework categorizing research based on linguistic analysis, enabling a deeper understanding of content characteristics beyond computational metrics [10]. This approach is complemented by the AdvCat framework, which employs model-agnostic algorithms to assess robustness across domains, enhancing the versatility and applicability of detection methodologies [20].

The innovative use of feature fusion, combining decision outputs from lightweight models like SqueezeNet and MobileNetV2, exemplifies efforts to improve detection accuracy, particularly in resource-constrained environments [7]. Furthermore, the development of robustness curves as evaluation metrics provides a nuanced understanding of model performance, facilitating the identification of strengths and weaknesses in detection systems [16].

The advancements in artificial intelligence (AI), particularly in generative models, have led to the creation of sophisticated content across various formats, including text, images, and video. This evolution presents significant challenges for distinguishing between human-generated and AI-generated content, especially in academic settings where integrity is paramount. To address these challenges, a robust framework for AI-generated content detection has been developed, employing machine learning techniques to train models on predefined datasets. This framework aims to enhance the accuracy of detecting AI-generated text in scientific research and academic publications while emphasizing the need for reliable detection systems that can adapt to the rapid advancements in generative AI technologies. The effectiveness of this framework has been benchmarked against

AI-generated, for reference only.

existing tools, highlighting the necessity for continuous improvement and adaptation in detection methodologies to mitigate the risks associated with AI misuse and maintain the integrity of scholarly work [21, 2, 22, 10, 23].

1.4 Structure of the Survey

This survey is meticulously organized to provide a comprehensive examination of AI-generated content detection and the methodologies that enhance its effectiveness. The paper begins with an **Introduction**, emphasizing the significance of detecting AI-generated content and the associated challenges, followed by an overview of the key technologies and methodologies explored. The **Background and Definitions** section offers foundational knowledge, providing definitions and explanations of core concepts such as AI-generated content, lightweight detection models, adversarial robustness, feature fusion, resource-constrained computing, edge AI, real-time content analysis, and computational efficiency.

The subsequent section, **AI-Generated Content Detection**, delves into the current state of detection technologies, examining existing techniques and tools while identifying their challenges and limitations. Following this, the paper explores **Lightweight Detection Models**, discussing their development, application, and critical role in resource-constrained environments. The discussion on **Adversarial Robustness** investigates the impact of adversarial attacks on detection models and strategies to enhance their robustness.

In **Feature Fusion**, the survey evaluates the integration of multiple data features to improve the accuracy and robustness of detection models. The section on **Resource-Constrained Computing and Edge AI** addresses the challenges and solutions for deploying detection models in environments with limited computational resources, emphasizing the advantages of edge AI. The importance of **Real-Time Content Analysis and Computational Efficiency** is examined, focusing on strategies to achieve computational efficiency without compromising accuracy and the trade-offs between speed and accuracy.

The **Conclusion** section synthesizes the primary findings of the survey on AI-generated content detection, highlighting the critical implications of the analyzed technologies and methodologies for the future landscape of detecting AI-generated content. It emphasizes the necessity of addressing the challenges posed by the increasing sophistication of AI-generated texts, the potential risks of misinformation, and the importance of integrating linguistic evaluations alongside computational approaches to enhance the robustness and explainability of detection frameworks [2, 22, 9, 10, 23]. The paper concludes with a discussion on **Future Directions and Challenges**, outlining potential areas for further research and development. This structured approach ensures a thorough understanding of the landscape of AI-generated content detection and its evolving methodologies.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 AI-Generated Content

AI-generated content encompasses various media types, including text, images, and audio, produced by AI systems using sophisticated generative models like Large Language Models (LLMs). These models generate outputs that closely mimic human-created content, complicating the differentiation between human and machine-generated materials, a challenge of particular importance in academic and scientific domains where scholarly integrity is paramount [6]. The indistinguishability of AI-generated text from human writing raises significant concerns about its potential misuse for misinformation and propaganda, jeopardizing information integrity and public trust [10, 6]. Ensuring the credibility of information sources in real-world applications necessitates effective detection mechanisms [9].

The integration of AI-generated content into platforms such as Wikipedia underscores the challenge of maintaining the accuracy of widely accessed information, especially as content may be generated by advanced language models post-release [9]. Robust detection mechanisms are essential to adapt to diverse sampling strategies and adversarial attacks, accurately identifying AI-generated content across various contexts [24]. Evaluating model resilience requires selecting metrics that address the multifaceted nature of robustness, including confidence and perturbation distances [25].

AI-generated, for reference only.

2.2 Lightweight Detection Models

Lightweight detection models are integral to AI-generated content detection, especially in resource-constrained environments. These models are designed for minimal computational overhead, making them suitable for real-time applications on edge devices [11]. Their importance lies in achieving high detection accuracy while minimizing resource consumption, crucial for on-device natural language processing (NLP) applications [26]. Techniques such as model compression and efficient architectural designs help retain the performance of larger models while reducing size and complexity [26]. Low-rank projection methods, for instance, create lightweight ensemble architectures that enhance efficiency without compromising detection capabilities [27], beneficial for real-time content analysis on mobile or edge devices [7].

Architectural innovations can enhance adversarial robustness without increasing model capacity. Exploring various configurations can improve robustness, facilitating the design of more efficient deep neural networks (DNNs) [28]. Neural Architecture Search (NAS) aids in discovering architectures with inherent adversarial robustness, reducing the need for adversarial training [29]. Integrating multiple lightweight models, such as SqueezeNet and MobileNetV2, can significantly enhance detection capabilities by leveraging the strengths of each to improve overall accuracy and robustness [7]. Feature fusion is crucial for detecting complex AI-generated content, including image forgeries, where diverse model outputs provide comprehensive analysis.

Frameworks like ADVMOE offer robust training mechanisms for mixture of expert convolutional neural networks (MoE-CNNs), optimizing router and expert parameters to enhance adversarial robustness [30]. This capability is vital for maintaining detection system integrity against evolving adversarial strategies.

2.3 Adversarial Robustness

Adversarial robustness is a key characteristic of AI-generated content detection systems, indicating their ability to maintain performance amid adversarial attacks—deliberate manipulations intended to mislead models into incorrect classifications [31]. These attacks exploit the non-convexity and non-linearity of decision boundaries in high-dimensional latent spaces, complicating the analysis of multi-modal models [32]. The vulnerability of deep neural networks (DNNs) to adversarial examples and various noise types poses significant security threats, particularly in safety-critical applications.

Current methods often inadequately evaluate adversarial robustness, relying on discrete success rates based on predefined attack budgets [33]. Thus, developing more effective methods for accurately estimating adversarial robustness is essential [33].

Strategies to enhance adversarial robustness include adversarial training, feature denoising, and architectural innovations. Adversarial training involves injecting adversarial noise into hidden layers during training to bolster resilience against diverse noise types and adversarial examples [34]. Feature disentanglement models (FDMs) aim to isolate robust features from non-robust and domain-specific features, addressing DNN vulnerabilities to adversarial examples [35].

In AI-generated content detection, enhancing adversarial robustness is crucial for ensuring system reliability and accuracy amidst adversarial threats, especially as the field shifts from cloud-centric to edge-centric AI, where quantized neural networks are increasingly utilized [36]. As adversarial strategies evolve, continuous advancements in adversarial training and robustness assessment are vital for developing resilient detection systems capable of withstanding these challenges. Maintaining classification accuracy despite adversarial perturbations is a key measure of a model's robustness and critical for the successful deployment of AI-generated content detection systems [31].

2.4 Feature Fusion

Feature fusion is a pivotal technique for enhancing the performance and robustness of AI-generated content detection models. By integrating multiple data features, feature fusion strengthens the model's ability to accurately identify and differentiate AI-generated from human-created content, effectively addressing adversarial attack challenges and improving overall resilience [37]. The process combines various feature types, such as linguistic, statistical, and contextual, creating a comprehensive representation of the analyzed content. This integration allows models to leverage the strengths of each feature type, enhancing predictive accuracy and robustness against adversarial

manipulations. Feature fusion can improve intra-class compactness and inter-class separation of feature vectors, essential for distinguishing between different content types [37].

Incorporating feature fusion into detection models also aids in marginalizing or eliminating non-robust image features often exploited in adversarial attacks. By concentrating on robust features, models can enhance their ability to withstand adversarial perturbations, maintaining high prediction accuracy [38]. This is particularly significant for neural network classifiers, where robust metrics like CLEVER offer a comprehensive understanding of model robustness independent of specific attack algorithms [39].

Feature fusion techniques bolster the robustness of deep multimodal models (DMMs), revealing vulnerabilities that traditional benchmarks might overlook. By providing a comprehensive framework for analyzing robustness, feature fusion addresses these shortcomings and enhances model performance across diverse scenarios [40].

2.5 Resource-Constrained Computing and Edge AI

Resource-constrained computing and edge AI are crucial for deploying AI-generated content detection systems, especially in environments with limited computational resources. These systems require models that operate efficiently on devices like smartphones and IoT devices [12]. Edge AI, which executes AI algorithms directly on edge devices, offers benefits such as reduced latency, improved data privacy, and decreased bandwidth usage, as data is processed locally rather than sent to a remote server [32]. This approach is particularly advantageous for applications requiring real-time decision-making, including autonomous vehicles, smart cameras, and industrial IoT applications [41].

Challenges in resource-constrained computing involve balancing model complexity and performance while minimizing computational overhead. Techniques such as model compression, quantization, and lightweight architectures address these challenges, enabling models to maintain high performance with limited resources [36]. For instance, combining quantization with Jacobian Regularization enhances model performance in resource-constrained settings by improving robustness against adversarial attacks [12].

The scalability of defenses in industrial applications remains an open question, particularly under diverse adversarial conditions [42]. Integrating domain knowledge and cognitive principles can enhance model explainability and robustness, providing a comprehensive framework for understanding adversarial threats and improving performance in resource-constrained environments [41]. Model robustness is further evaluated using benchmarks derived from real-world datasets, offering insights into performance across varied conditions [32].

2.6 Real-Time Content Analysis and Computational Efficiency

Real-time content analysis is essential for detecting AI-generated content, requiring the capability to process and analyze data as it is created or received, ensuring immediate detection and response. This capability is crucial in applications such as social media monitoring, cybersecurity, and autonomous systems, where timely decision-making is vital to prevent the spread of misleading or harmful content [42]. Achieving real-time analysis necessitates deploying models that efficiently manage large data volumes with minimal latency, maintaining content integrity and reliability in dynamic environments.

Computational efficiency is fundamental for real-time content analysis, enabling models to perform complex tasks without overwhelming available resources. Efficient algorithms and architectures are designed to optimize processing speed and reduce computational overhead, allowing effective operation even in resource-constrained environments [43]. Techniques such as model quantization, pruning, and lightweight architectures enhance computational efficiency, facilitating the deployment of robust detection systems on edge devices with limited processing power.

Furthermore, evaluating model performance in real-time content analysis involves assessing both standard accuracy and robust accuracy against adversarial attacks. Metrics like Projected Gradient Descent (PGD) and AutoAttack measure model resilience to adversarial perturbations, providing insights into robustness under various attack scenarios [42]. The ability to maintain high classification accuracy despite adversarial challenges reflects a model's robustness and computational efficiency, underscoring the importance of these metrics in developing reliable detection systems.

In recent years, the landscape of artificial intelligence (AI) has evolved dramatically, particularly in the realm of content generation. As researchers seek to understand and develop effective methods for detecting AI-generated content, it becomes essential to categorize the various techniques and tools available. Figure 2 provides a comprehensive overview of this hierarchical structure, illustrating the categorization of current detection methodologies. This figure highlights key methodologies, including pre-trained language models and CNN-based detectors, while also evaluating their performance through established benchmarks and frameworks. Furthermore, it addresses the significant challenges and limitations faced in this field, such as the rapid evolution of generative models and the vulnerabilities posed by adversarial attacks. By examining these elements, we can gain a clearer understanding of the complexities involved in AI content detection and the ongoing efforts to enhance its effectiveness.

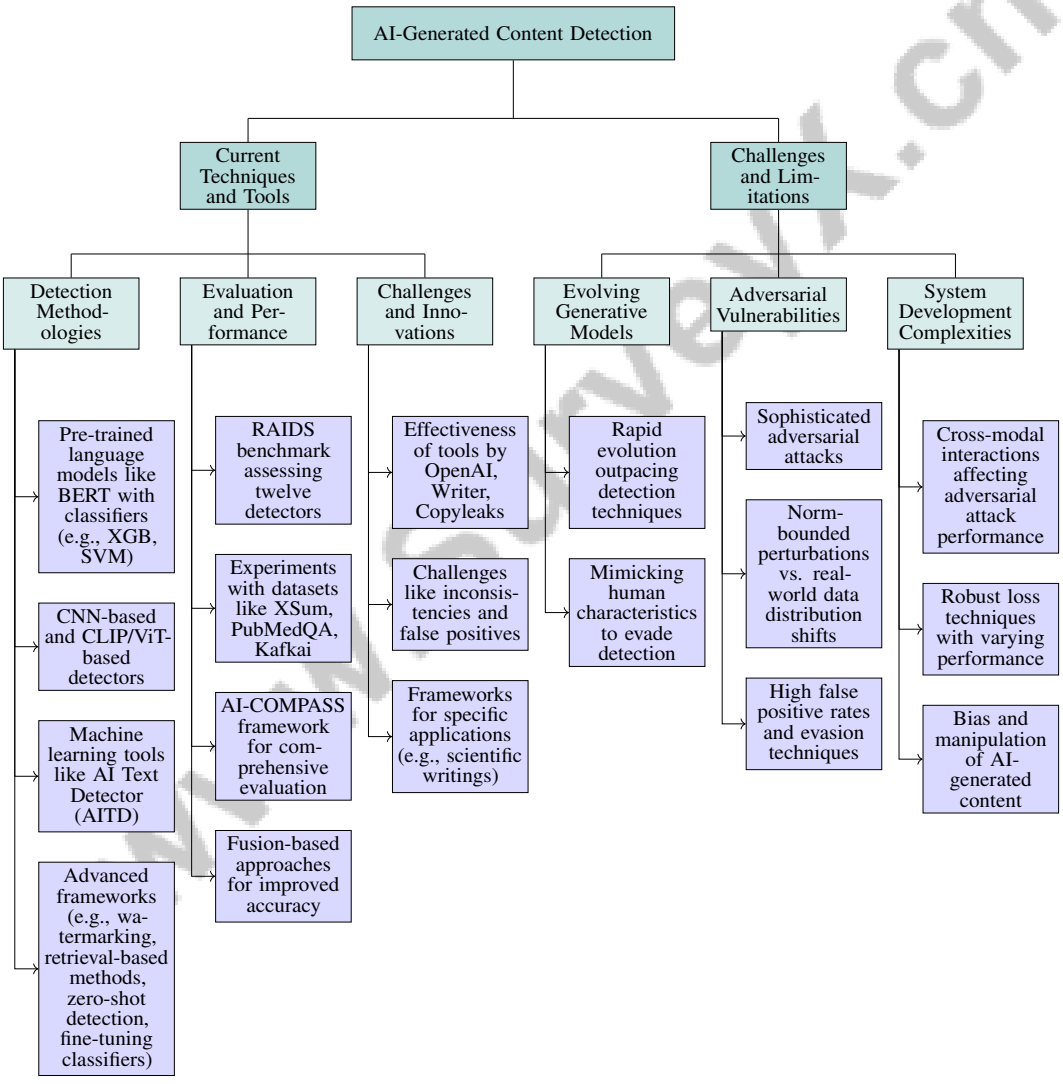

Figure 2: This figure depicts the hierarchical structure of AI-generated content detection, categorizing current techniques and tools, evaluation and performance, and challenges and limitations. It highlights methodologies like pre-trained language models and CNN-based detectors, evaluates performance through benchmarks and frameworks, and addresses challenges such as evolving generative models and adversarial vulnerabilities.

AI-generated, for reference only.

3 AI-Generated Content Detection

3.1 Current Techniques and Tools

The field of AI-generated content detection has progressed, employing various methodologies to tackle challenges posed by advanced generative models. Pre-trained language models like BERT leverage contextual and semantic analysis to detect AI-generated text, often enhanced by classifiers such as XGB and SVM for improved accuracy [18]. CNN-based and CLIP/ViT-based detectors are extensively evaluated for adversarial robustness [44], as demonstrated by the RAIDS benchmark, which assessed twelve detectors, showcasing the diversity of available tools [45].

The AI Text Detector (AITD) exemplifies machine learning tools that classify text by analyzing features through multilayer neural networks [2]. Surveys indicate that supervised detectors using advanced language models outperform traditional methods [3]. Advanced frameworks categorize research into methodologies like watermarking, retrieval-based methods, zero-shot detection, and fine-tuning classifiers, identifying gaps and areas for improvement [9]. Versatile tools like AdvCat apply across domains, including fake news and intrusion detection, using datasets like FakeNewsNet [20].

Experiments with datasets such as XSum, PubMedQA, and Kafkai, featuring AI-generated texts, evaluate various detectors, including watermarking-based and neural network-based approaches [5]. These techniques are categorized to highlight their effectiveness and limitations in identifying AI-generated content [6]. The AI-COMPASS framework provides a comprehensive evaluation of detection performance across multiple dimensions, highlighting model vulnerabilities and strengths [4]. Fusion-based approaches significantly improve accuracy, showcasing the effectiveness of lightweight deep learning models for image forgery detection [7].

To visualize the complexity and diversity of methodologies within this evolving landscape, Figure 3 illustrates the hierarchical categorization of current techniques, tools, and challenges in AI-generated content detection. This figure underscores the intricate challenges faced by researchers and practitioners in the field.

Advanced detection techniques and tools represent innovation at the forefront of this field. Studies underscore the effectiveness of tools developed by OpenAI, Writer, and Copyleaks in distinguishing AI-generated texts, while also revealing challenges like inconsistencies and false positives. Frameworks for specific applications, such as assessing scientific writings, highlight the potential for enhancing detection capabilities. As AI-generated content evolves, ongoing refinement and collaboration between human reviewers and detection tools are essential for maintaining academic integrity and addressing technological challenges [23, 22, 10, 2].

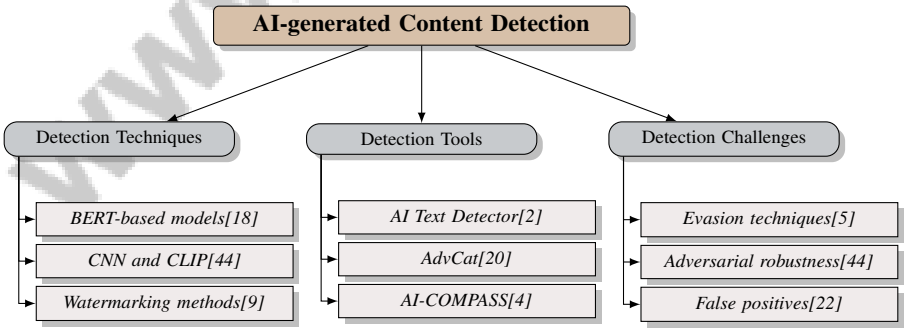

Figure 3: This figure illustrates the hierarchical categorization of current techniques, tools, and challenges in AI-generated content detection, showcasing the diversity of methodologies and the complexity of challenges faced in the field.

3.2 Challenges and Limitations

Detecting AI-generated content faces challenges and limitations that hinder effectiveness across diverse domains. A key issue is the rapid evolution of generative models, which often outpace detection techniques, causing inconsistent performance [46]. This inconsistency is compounded by

biases, as AI-generated content can mimic human characteristics to evade detection [6]. Detection methods are vulnerable to sophisticated adversarial attacks exploiting non-robust features, leading to incorrect classifications [35]. Despite various defenses, many remain susceptible to evolving adversarial strategies.

Current methodologies focus on norm-bounded perturbations, neglecting unbounded real-world data distribution shifts that impact model robustness [47]. This limitation underscores the need for comprehensive evaluation frameworks to assess model performance under broader adversarial conditions. Techniques like CIFS enhance adversarial robustness but may reduce natural accuracy on certain datasets, indicating an area for further exploration [48]. High false positive rates in existing methods pose significant limitations, easily circumvented by advanced evasion techniques [9], particularly in security-sensitive applications.

The lack of understanding of cross-modal interactions affecting adversarial attack performance complicates robust system development [49]. Although robust loss techniques show promise, performance varies across perturbation thresholds, necessitating more adaptable detection methods [46]. Potential bias and manipulation of AI-generated content to appear human-like further complicate effective detection system development [6].

4 Lightweight Detection Models

4.1 Development and Application of Lightweight Models

The development of lightweight detection models is pivotal for identifying AI-generated content, especially in environments with limited resources. These models are crucial for addressing misinformation and preserving academic integrity, aligning with ethical research standards [2, 9, 22, 10]. Designed to minimize computational demands, lightweight models are ideal for real-time applications on devices like smartphones and IoT systems, where high detection accuracy and low resource consumption are imperative.

Key techniques such as pruning and quantization are instrumental in creating lightweight models by reducing model size and complexity without compromising performance [50]. These methods streamline model architectures, removing unnecessary parameters to facilitate deployment on devices with constrained computational capabilities [11]. Additionally, adversarial robustness is a significant consideration in lightweight model design. Techniques like Neural Architecture Search (NAS) help explore topologies that enhance resilience without additional training data [29], adjusting parameters such as model width and depth to improve efficiency and robustness [50].

Explainability is another crucial aspect, with models like DistilBERT, LSTM, CNN, and Random Forest Classifiers serving as baselines to improve transparency in detection systems [51]. By clarifying decision-making processes, these models enhance interpretability, which is vital for sensitive applications. Advanced training techniques, such as the RobustDet framework, further strengthen the robustness of lightweight models against adversarial attacks through diverse imperceptible adversarial examples [52]. Incorporating semantic information, like targeted universal perturbations, can also enhance robustness by focusing on critical decision boundary regions [53].

Lightweight models are extensively applied in detecting AI-generated content, classifying text as human or machine-generated [9]. By integrating advanced language models within lightweight architectures, detection systems can significantly improve accuracy and robustness, even in resource-constrained conditions. This capability is crucial for real-time content analysis, ensuring rapid and precise identification of AI-generated content across various platforms.

4.2 Importance in Resource-Constrained Environments

Deploying AI-generated content detection systems in resource-limited settings, like edge devices and IoT systems, necessitates lightweight models. These models are engineered for efficiency, allowing real-time analysis without sacrificing accuracy or robustness against adversarial attacks. Their importance lies in delivering high-performance detection capabilities while minimizing resource consumption, essential for applications requiring low latency and immediate response [41].

As illustrated in Figure 4, the hierarchical structure of AI-generated content detection systems emphasizes the role of lightweight models and advanced training techniques, while also highlighting the challenges and solutions pertinent to their deployment in resource-constrained environments.

Pruning and quantization techniques enhance the efficiency and robustness of lightweight models, making them apt for edge device deployment. By reducing model complexity and size without significantly impacting performance, these methods enable robust detection systems in power-constrained environments [50]. This is especially relevant for applications like autonomous vehicles and smart cameras, where real-time decision-making is critical.

Feature fusion techniques in lightweight models further enhance detection accuracy and robustness, as demonstrated by the superior performance of fusion models over individual models in reducing false positives and improving reliability [54]. These approaches allow models to leverage diverse feature types, strengthening resilience against adversarial attacks.

Advanced training methodologies are essential for achieving computational efficiency in resource-constrained environments. Techniques such as the CNC method enable robust performance without extensive adversarial training, making them suitable for environments with limited computational resources. Similarly, the FDN approach addresses the need for effective adversarial robustness while being implementable in resource-constrained contexts [34].

Despite progress in natural language processing, challenges persist in balancing model complexity and performance, particularly in minimizing computational overhead. This balance is crucial as researchers work to enhance the robustness of detection frameworks for AI-generated text while mitigating vulnerabilities such as misinformation and adversarial attacks [9, 55]. Exploring architectural configurations and integrating domain knowledge are vital for overcoming these challenges, ensuring the effective deployment of efficient and robust AI-generated content detection systems in dynamic, resource-limited settings.

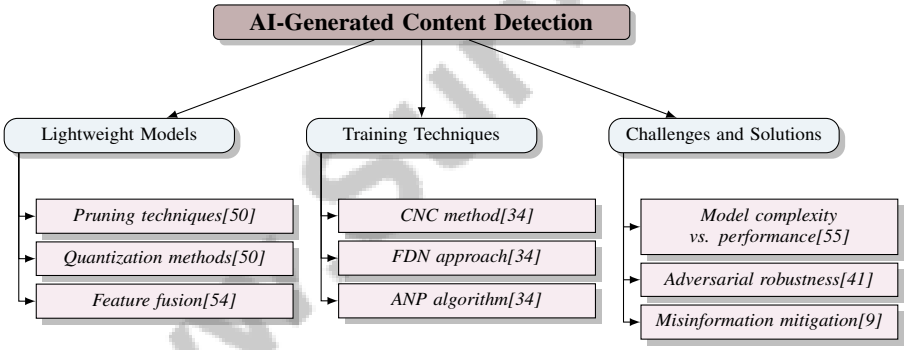

Figure 4: This figure illustrates the hierarchical structure of AI-generated content detection systems, focusing on lightweight models, advanced training techniques, and the challenges and solutions associated with deploying these systems in resource-constrained environments.

5 Adversarial Robustness

5.1 Understanding Adversarial Attacks

Adversarial attacks pose a significant threat to AI-generated content detection systems by exploiting deep neural networks' (DNNs) vulnerabilities. These attacks introduce subtle perturbations to input data, leading to misclassifications and degraded model performance, exacerbated by the high non-convexity and non-linearity of DNN decision boundaries in high-dimensional latent spaces [32]. This threat is critical in domains like computer vision and natural language processing, where precision is essential [10]. The fragility of DNNs is linked to their reliance on non-robust features, which are easily perturbed, resulting in incorrect predictions [35]. Traditional adversarial example generation methods, often based on ascending the gradient of loss, are limited, underscoring the need for more advanced techniques [54].

Innovative methods like the Collaborative Multimodal Adversarial Attack (Co-Attack) have emerged, targeting both image and text modalities to enhance adversarial effectiveness [56]. Additionally,

data-driven subsampling reduces computational complexity while improving adversarial robustness, addressing significant limitations of existing strategies [41]. Balancing model complexity with resilience is crucial, as simply increasing parameters does not guarantee enhanced robustness. Optimizing architectural configurations for resilience within constrained parameter budgets is essential, especially in resource-limited environments [28, 29].

5.2 Strategies for Enhancing Robustness

Enhancing the robustness of AI-generated content detection systems against adversarial attacks necessitates multifaceted strategies and innovative defense mechanisms. Adversarial training techniques, such as the A3T method, involve training a discriminator alongside the classifier to differentiate between real and adversarial examples, thereby enhancing feature robustness and enforcing feature invariance [31]. Feature disentanglement techniques are also vital, with a proposed three-branch architecture that separates feature representations into robust, non-robust, and domain-specific components, improving performance under adversarial conditions [35].

Quantized neural networks offer another pathway for enhancing adversarial robustness. By modeling these networks' adversarial robustness, researchers can identify critical attack strengths where quantization minimally affects accuracy, maintaining robust performance against adversarial threats [36]. Furthermore, estimation ensembles represent a significant advancement in estimating adversarial distance, combining various methods to provide more accurate estimations than existing techniques and offering valuable insights into classifier robustness [33]. Future research should focus on refining upper bound estimations through iterative adversarial attacks and exploring alternative certification methods for more reliable lower bounds [33].

6 Feature Fusion

6.1 Techniques for Feature Fusion

Feature fusion plays a pivotal role in enhancing the adversarial robustness and performance of AI-generated content detection models by integrating diverse data features such as linguistic, statistical, and contextual elements. This comprehensive representation improves the model's accuracy in distinguishing between AI-generated and human-created content. A notable approach involves manipulating feature vectors to enhance robustness against adversarial attacks, as discussed in Smith et al.'s taxonomy [37].

By combining multiple feature types, models can leverage the unique advantages of each category, enhancing predictive accuracy and resilience against adversarial threats. Studies have shown that neural features excel in performance, while statistical features significantly contribute to adversarial resilience, making them essential in ensemble detection models. This synergy allows models to effectively navigate challenges posed by evolving generative technologies, ensuring high standards of text quality and reliability [19, 2]. This approach is particularly effective in enhancing intra-class compactness and inter-class separation of feature vectors, crucial for distinguishing content types and mitigating adversarial perturbations.

Innovative frameworks for feature fusion, such as the one proposed by Zhang et al., highlight the importance of prompt-based mask prediction over traditional label prediction methods. This novel perspective facilitates more effective integration of diverse data features within AI-generated content detection [57].

6.2 Feature Fusion and Model Performance

Feature fusion is crucial for enhancing the performance and robustness of AI-generated content detection models against adversarial attacks, addressing concerns about the misuse of generative models for misinformation and other malicious activities. By integrating neural and statistical features, detection frameworks achieve greater resilience against sophisticated adversarial strategies, ensuring reliable identification of AI-generated text across various contexts [19, 9, 22, 21]. This integration enables models to capitalize on the strengths of various data types, enhancing their ability to differentiate between AI-generated and human-authored content.

Benchmark	Size	Domain	Task Format	Metric
AIDB[22]	4	Academic Integrity	Text Classification	Confidence Percentage, Accuracy
AIDT[18]	3,000	Text Detection	Text Classification	Accuracy, F1-score
CAA[58]	1,220,092	Credit Scoring	Binary Classification	Adversarial Accuracy
ARES-Bench[59]	1,000,000	Image Classification	Robustness Evaluation	Robustness Curves
AR-CLIP[44]	2,000	Image Forensics	Image Detection	Success Attack Rate
RAID[45]	6,287,820	Text Generation	Text Classification	Accuracy, F1-score
E-P-R[60]	1,000,000	Natural Language Inference	Classification	Accuracy, ADP
Attack-SAM[57]	1,000	Image Segmentation	Mask Prediction	mIoU

Table 1: The table presents a comprehensive overview of various benchmarks used in evaluating AI-generated content detection models. It details the size, domain, task format, and performance metrics of each benchmark, providing insight into their diverse applications and evaluation criteria.

Benchmark	Size	Domain	Task Format	Metric
AIDB[22]	4	Academic Integrity	Text Classification	Confidence Percentage, Accuracy
AIDT[18]	3,000	Text Detection	Text Classification	Accuracy, F1-score
CAA[58]	1,220,092	Credit Scoring	Binary Classification	Adversarial Accuracy
ARES-Bench[59]	1,000,000	Image Classification	Robustness Evaluation	Robustness Curves
AR-CLIP[44]	2,000	Image Forensics	Image Detection	Success Attack Rate
RAID[45]	6,287,820	Text Generation	Text Classification	Accuracy, F1-score
E-P-R[60]	1,000,000	Natural Language Inference	Classification	Accuracy, ADP
Attack-SAM[57]	1,000	Image Segmentation	Mask Prediction	mIoU

Table 2: The table presents a comprehensive overview of various benchmarks used in evaluating AI-generated content detection models. It details the size, domain, task format, and performance metrics of each benchmark, providing insight into their diverse applications and evaluation criteria.

Benchmark	Size	Domain	Task Format	Metric
AIDB[22]	4	Academic Integrity	Text Classification	Confidence Percentage, Accuracy
AIDT[18]	3,000	Text Detection	Text Classification	Accuracy, F1-score
CAA[58]	1,220,092	Credit Scoring	Binary Classification	Adversarial Accuracy
ARES-Bench[59]	1,000,000	Image Classification	Robustness Evaluation	Robustness Curves
AR-CLIP[44]	2,000	Image Forensics	Image Detection	Success Attack Rate
RAID[45]	6,287,820	Text Generation	Text Classification	Accuracy, F1-score
E-P-R[60]	1,000,000	Natural Language Inference	Classification	Accuracy, ADP
Attack-SAM[57]	1,000	Image Segmentation	Mask Prediction	mIoU

Table 3: The table presents a comprehensive overview of various benchmarks used in evaluating AI-generated content detection models. It details the size, domain, task format, and performance metrics of each benchmark, providing insight into their diverse applications and evaluation criteria.

Smith et al.'s taxonomy illustrates the effectiveness of diverse ensemble defenses, suggesting that combining strengths from multiple categories provides a more robust defense against adversarial attacks than any single method [37]. Feature fusion improves intra-class compactness and inter-class separation, critical for accurate content classification and robust performance.

Moreover, feature fusion enhances the model's ability to mitigate adversarial manipulations by focusing on robust feature representations, maintaining high accuracy levels across various contexts, and ensuring detection system reliability in dynamic environments. As adversarial strategies in AI-generated content evolve, feature fusion becomes crucial for developing robust detection systems. This approach fortifies detection methodologies against sophisticated adversarial attacks and underscores the necessity of combining neural and statistical features to enhance accuracy in distinguishing between human and AI-generated text, addressing potential generative model misuse challenges in academia, journalism, and online communication [2, 22, 9, 44, 19]. Table 3 provides an overview of key benchmarks utilized in the evaluation of AI-generated content detection models, highlighting their diverse characteristics and metrics.

7 Resource-Constrained Computing and Edge AI

7.1 Edge AI: Concepts and Advantages

Edge AI facilitates the deployment of artificial intelligence algorithms on edge devices, such as smartphones and IoT sensors, rather than relying on centralized cloud systems. This localized processing is particularly advantageous in resource-constrained environments where real-time data processing is crucial, as it reduces latency, enhances data privacy, and decreases bandwidth usage [61]. A significant benefit of edge AI is its energy efficiency, exemplified by the RobustEdge framework, which achieves substantial reductions in energy consumption while enhancing adversarial detection performance, making it suitable for devices with limited power resources [61].

Edge AI also bolsters adversarial robustness through secure subsampling strategies, optimizing detection systems' resilience against adversarial attacks in dynamic environments [41]. This capability is critical for applications requiring immediate decision-making, such as autonomous vehicles and smart surveillance systems. Implementing edge AI necessitates balancing efficiency and robustness in deep neural network (DNN) applications to maintain high performance under limited computational resources [62]. By leveraging efficient computational strategies and architectural innovations, edge AI enables robust AI-generated content detection models to function effectively in resource-constrained settings.

7.2 Applications of Edge AI in Content Detection

Edge AI is revolutionizing content detection by embedding advanced AI algorithms into edge devices, including smartphones, cameras, and IoT sensors, enabling real-time analysis to differentiate between human-generated and AI-generated content. This capability is increasingly vital in fields such as academic integrity and media verification, as it enhances the ability to swiftly address misinformation and academic misconduct, promoting ethical content creation and consumption [2, 22, 9, 23].

In smart surveillance systems, edge AI processes video feeds locally, allowing real-time anomaly detection without continuous data transmission to central servers, thereby reducing latency and enhancing data privacy [61]. In autonomous vehicles, edge AI enables rapid sensor data processing for obstacle detection and real-time decision-making, crucial for safety and performance in dynamic driving conditions [41].

Furthermore, edge AI is critical for deploying AI-generated content detection systems on mobile devices with limited computational resources. Utilizing lightweight models and efficient architectures, edge AI facilitates the detection of AI-generated text and images directly on devices, providing immediate feedback and improving information reliability [6]. The energy efficiency of edge AI, as demonstrated by the RobustEdge framework, is especially beneficial in scenarios with constrained power resources, such as remote environmental monitoring, where devices must operate autonomously for extended periods [61].

8 Real-Time Content Analysis and Computational Efficiency

The evolution of AI-generated content underscores the pivotal role of real-time content analysis in enhancing decision-making and ensuring information integrity. Timely evaluations are fundamental in mitigating misinformation risks, highlighting the need for efficient content processing mechanisms.

8.1 Importance of Real-Time Content Analysis

Real-time content analysis is crucial for detecting AI-generated content, enabling immediate evaluation of data as it is produced or received. This capability is essential in contexts demanding quick decision-making to prevent the dissemination of misleading information. For instance, in social media, real-time analysis allows for rapid identification and correction of false information, preserving content integrity [42]. In cybersecurity, it is vital for preventing security breaches by swiftly identifying threats and protecting sensitive information [42]. Autonomous systems, like self-driving cars and drones, rely on real-time content analysis to ensure safety and operational efficiency, as they need immediate sensor data interpretation to make informed decisions in dynamic environments [43]. Achieving the computational efficiency required for real-time analysis involves deploying optimized

algorithms and architectures that minimize latency and resource use, with techniques such as model quantization and pruning enhancing processing speed [43].

8.2 Strategies for Achieving Computational Efficiency

Achieving computational efficiency in AI-generated content detection systems is critical for real-time applications and resource-constrained environments. Strategies have been developed to optimize deployment and operational components, addressing the challenges of accurately identifying AI-generated content and mitigating misuse in academic contexts [22, 3, 9]. Cloud-based solutions, such as AWS EC2, enhance scalability and reliability, facilitating the efficient handling of varying loads and ensuring consistent performance [2]. Optimizing model architectures for edge devices is also crucial; for example, Lightweight Convolutional Neural Networks (L-CNNs) effectively process video frames for object detection with minimal computational overhead [11]. Introducing noise into hidden layers during training, as demonstrated in Adversarial Noise Propagation (ANP), enhances robustness without significant computational costs, increasing model resilience against adversarial attacks [34]. These strategies highlight the necessity of balancing computational efficiency and accuracy in AI-generated content detection systems, particularly as these tools face challenges in distinguishing between human and AI-generated text [23, 22].

8.3 Trade-offs Between Speed and Accuracy

Balancing speed and accuracy in AI-generated content detection systems involves navigating trade-offs that affect model performance and operational efficiency. Rapid processing capabilities often require compromises in detection accuracy, especially in resource-constrained environments [11]. Lightweight models, while reducing computational overhead, may not achieve the same accuracy as more complex models, particularly against sophisticated adversarial attacks [12]. Techniques like model compression and pruning can enhance processing speed but may degrade accuracy if not carefully managed [50]. Integrating feature fusion techniques can improve model robustness and accuracy but may introduce additional computational complexity, impacting processing speed [37]. In real-time applications, such as autonomous vehicles and smart surveillance systems, the trade-off between speed and accuracy is critical as processing delays could have significant consequences [43]. Managing these trade-offs is essential to ensure reliable performance across diverse applications. Advanced optimization techniques and architectural innovations can enhance both speed and accuracy, enabling the deployment of robust and efficient detection systems in dynamic and resource-constrained environments [42].

9 Conclusion

9.1 Future Directions and Challenges

The advancement of AI-generated content detection systems relies on overcoming several critical challenges while exploring innovative research directions to enhance their robustness and reliability. Key to these efforts is the development of models that prioritize linguistic analysis and transparency, which are essential for improving detection accuracy and fostering trust in such systems [10].

A promising research avenue involves enhancing adversarial robustness strategies, particularly through the optimization of Collaborative Multimodal Adversarial Attack (Co-Attack). Investigating its applicability across various multimodal tasks and devising effective defenses against adversarial attacks can significantly improve detection system reliability [56]. Furthermore, integrating physiological components to validate findings and examining visual perception aspects related to adversarial robustness presents an exciting opportunity for field advancement [63].

Refining techniques like FADER to bolster robustness against attacks while minimizing computational overhead remains a crucial research focus. Exploring architectural variants and incorporating domain knowledge can enhance the explainability and resilience of detection models, especially in resource-constrained environments [41].

To tackle the challenges posed by sophisticated adversarial attacks, future research should prioritize the development of adaptable and resilient detection methods capable of withstanding evolving adversarial strategies. This entails integrating advanced adversarial training methods and verification

techniques to bolster the robustness of few-shot classifiers [64]. Additionally, creating comprehensive evaluation frameworks that accurately assess model performance under diverse adversarial conditions is vital for the field's progression [47].

In resource-constrained contexts, ongoing research should investigate the potential of edge AI and resource-efficient computational strategies, such as quantization and model compression, to enhance detection systems' performance and robustness [12]. Moreover, integrating domain knowledge and cognitive principles into detection models is essential for improving their explainability and adaptability to evolving adversarial strategies [41].

References

[1] Creston Brooks, Samuel Eggert, and Denis Peskoff. The rise of ai-generated content in wikipedia. *arXiv preprint arXiv:2410.08044*, 2024.

[2] Paria Sarzaeim, Arya Doshi, and Qusay Mahmoud. A framework for detecting ai-generated text in research publications. In *Proceedings of the International Conference on Advanced Technologies*, volume 11, pages 121–127, 2023.

[3] Tharindu Kumarage, Garima Agrawal, Paras Sheth, Raha Moraffah, Aman Chadha, Joshua Garland, and Huan Liu. A survey of ai-generated text forensic systems: Detection, attribution, and characterization, 2024.

[4] Zhiyu Zhu, Zhibo Jin, Hongsheng Hu, Minhui Xue, Ruoxi Sun, Seyit Camtepe, Praveen Gauravaram, and Huaming Chen. Ai-compass: A comprehensive and effective multi-module testing tool for ai systems, 2024.

[5] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

[6] Levent Uzun. Chatgpt and academic integrity concerns: Detecting artificial intelligence generated content. *Language Education and Technology*, 3(1), 2023.

[7] Amit Doegar, Srinidhi Hiriyannaiah, Siddesh Gaddadevara Matt, Srinivasa Krishnarajanagar Gopaliyengar, and Maitreyee Dutta. Image forgery detection based on fusion of lightweight deep learning models. *Turkish Journal of Electrical Engineering and Computer Sciences*, 29(4):1978–1993, 2021.

[8] A. Emin Orhan. Robustness properties of facebook's resnext wsl models, 2019.

[9] Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit Bedi. A survey on the possibilities & impossibilities of ai-generated text detection. *Transactions on Machine Learning Research*, 2023.

[10] Yu Wang. Survey for detecting ai-generated content. *Advances in Engineering Technology Research*, 11(1):643–643, 2024.

[11] Seyed Yahya Nikouei, Yu Chen, Sejun Song, Ronghua Xu, Baek-Young Choi, and Timothy R Faughnan. Real-time human detection as an edge service enabled by a lightweight cnn. In *2018 IEEE International Conference on Edge Computing (EDGE)*, pages 125–129. IEEE, 2018.

[12] Ferheen Ayaz, Idris Zakariyya, José Cano, Sye Loong Keoh, Jeremy Singer, Danilo Pau, and Mounia Kharbouche-Harrari. Improving robustness against adversarial attacks with deeply quantized neural networks, 2023.

[13] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures, 2022.

[14] Moninder Singh, Gevorg Ghalachyan, Kush R. Varshney, and Reginald E. Bryant. An empirical study of accuracy, fairness, explainability, distributional robustness, and adversarial robustness, 2021.

[15] Cuong Dang, Dung D. Le, and Thai Le. A curious case of searching for the correlation between training data and adversarial robustness of transformer textual models, 2024.

[16] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 321–331, 2020.

[17] Syed Mhamudul Hasan, Abdur R. Shahid, and Ahmed Imteaj. Towards sustainable secureml: Quantifying carbon footprint of adversarial machine learning, 2024.

[18] Nuzhat Prova. Detecting ai generated text based on nlp and machine learning approaches. *arXiv preprint arXiv:2404.10032*, 2024.

[19] Evan Crothers, Nathalie Japkowicz, Herna Viktor, and Paula Branco. Adversarial robustness of neural-statistical features in detection of generative transformers, 2022.

[20] Helene Orsini, Hongyan Bao, Yujun Zhou, Xiangrui Xu, Yufei Han, Longyang Yi, Wei Wang, Xin Gao, and Xiangliang Zhang. Advcat: Domain-agnostic robustness assessment for cybersecurity-critical applications with categorical inputs, 2022.

[21] Kathleen C Fraser, Hillary Dawkins, and Svetlana Kiritchenko. Detecting ai-generated text: Factors influencing detectability with current methods. *arXiv preprint arXiv:2406.15583*, 2024.

[22] N Ladha, K Yadav, and P Rathore. Ai-generated content detectors: Boon or bane for scientific writing. *Indian Journal of Science and Technology*, 16(39):3435–3439, 2023.

[23] Ahmed M Elkhatat, Khaled Elsaid, and Saeed Almeer. Evaluating the efficacy of ai content detection tools in differentiating between human and ai-generated text. *International Journal for Educational Integrity*, 19(1):17, 2023.

[24] Cenyuan Zhang, Xiang Zhou, Yixin Wan, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. Improving the adversarial robustness of nlp models by information bottleneck, 2022.

[25] Ping Guo, Cheng Gong, Xi Lin, Zhiyuan Yang, and Qingfu Zhang. Exploring the adversarial frontier: Quantifying robustness via adversarial hypervolume, 2024.

[26] Yao Qiang, Supriya Tumkur Suresh Kumar, Marco Brocanelli, and Dongxiao Zhu. Adversarially robust and explainable model compression with on-device personalization for text classification, 2021.

[27] Ruoxi Qin, Linyuan Wang, Xuehui Du, Xingyuan Chen, and Bin Yan. Dynamic ensemble selection based on deep neural network uncertainty estimation for adversarial robustness, 2023.

[28] Hanxun Huang, Yisen Wang, Sarah Monazam Erfani, Quanquan Gu, James Bailey, and Xingjun Ma. Exploring architectural ingredients of adversarially robust deep neural networks, 2022.

[29] Chaitanya Devaguptapu, Devansh Agarwal, Gaurav Mittal, Pulkit Gopalani, and Vineeth N Balasubramanian. On adversarial robustness: A neural architecture search perspective, 2021.

[30] Yihua Zhang, Ruisi Cai, Tianlong Chen, Guanhua Zhang, Huan Zhang, Pin-Yu Chen, Shiyu Chang, Zhangyang Wang, and Sijia Liu. Robust mixture-of-expert training for convolutional neural networks, 2023.

[31] Akram Erraqabi, Aristide Baratin, Yoshua Bengio, and Simon Lacoste-Julien. A3t: Adversarially augmented adversarial training, 2018.

[32] Juncheng B Li, Shuhui Qu, Xinjian Li, Po-Yao Huang, and Florian Metze. On adversarial robustness of large-scale audio visual learning, 2022.

[33] Georg Siedel, Ekagra Gupta, and Andrey Morozov. A practical approach to evaluating the adversarial distance for machine learning classifiers, 2024.

[34] Aishan Liu, Xianglong Liu, Chongzhi Zhang, Hang Yu, Qiang Liu, and Dacheng Tao. Training robust deep neural networks via adversarial noise propagation, 2020.

[35] Hong Wang, Yuefan Deng, Shinjae Yoo, and Yuewei Lin. Exploring robust features for improving adversarial robustness, 2023.

[36] Micah Gorsline, James Smith, and Cory Merkel. On the adversarial robustness of quantized neural networks, 2021.

[37] Leslie N. Smith. A useful taxonomy for adversarial robustness of neural networks, 2019.

[38] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*, 2018.

[39] Gavin Weiguang Ding, Luyu Wang, and Xiaomeng Jin. advertorch v0.1: An adversarial robustness toolbox based on pytorch, 2019.

[40] Nishant Vishwamitra, Hongxin Hu, Ziming Zhao, Long Cheng, and Feng Luo. Understanding and measuring robustness of multimodal learning, 2021.

[41] Abu Shafin Mohammad Mahdee Jameel, Ahmed P. Mohamed, Jinho Yi, Aly El Gamal, and Akshay Malhotra. Data-driven subsampling in the presence of an adversarial actor, 2024.

[42] Yunjuan Wang, Hussein Hazimeh, Natalia Ponomareva, Alexey Kurakin, Ibrahim Hammoud, and Raman Arora. Dart: A principled approach to adversarially robust unsupervised domain adaptation, 2024.

[43] Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. Adversarial robustness comparison of vision transformer and mlp-mixer to cnns, 2021.

[44] Vincenzo De Rosa, Fabrizio Guillaro, Giovanni Poggi, Davide Cozzolino, and Luisa Verdoliva. Exploring the adversarial robustness of clip for ai-generated image detection, 2024.

[45] Liam Dugan, Alyssa Hwang, Filip Trhlik, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. Raid: A shared benchmark for robust evaluation of machine-generated text detectors, 2024.

[46] Niklas Risse, Christina Göpfert, and Jan Philip Göpfert. How to compare adversarial robustness of classifiers from a global perspective, 2020.

[47] Alexander Robey, Hamed Hassani, and George J. Pappas. Model-based robust deep learning: Generalizing to natural, out-of-distribution data, 2020.

[48] Hanshu Yan, Jingfeng Zhang, Gang Niu, Jiashi Feng, Vincent Y. F. Tan, and Masashi Sugiyama. Cifs: Improving adversarial robustness of cnns via channel-wise importance-based feature selection, 2021.

[49] Jiwei Guan, Tianyu Ding, Longbing Cao, Lei Pan, Chen Wang, and Xi Zheng. Probing the robustness of vision-language pretrained models: A multimodal adversarial attack approach, 2024.

[50] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan Yuille, and Kaiming He. Feature denoising for improving adversarial robustness, 2019.

[51] Jie Wang, Jun Ai, Minyan Lu, Haoran Su, Dan Yu, Yutao Zhang, Junda Zhu, and Jingyu Liu. A survey of neural network robustness assessment in image recognition, 2024.

[52] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models, 2021.

[53] Yang Wang, Bo Dong, Ke Xu, Haiyin Piao, Yufei Ding, Baocai Yin, and Xin Yang. A geometrical approach to evaluate the adversarial robustness of deep neural networks, 2023.

[54] Andras Rozsa, Manuel Günther, and Terrance E. Boult. Adversarial robustness: Softmax versus openmax, 2017.

[55] Seong-Il Park and Jay-Yoon Lee. Toward robust ralms: Revealing the impact of imperfect retrieval on retrieval-augmented language models, 2024.

[56] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models, 2022.

[57] Chenshuang Zhang, Chaoning Zhang, Taegoo Kang, Donghun Kim, Sung-Ho Bae, and In So Kweon. Attack-sam: Towards attacking segment anything model with adversarial examples, 2023.

[58] Thibault Simonetto, Salah Ghamizi, Antoine Desjardins, Maxime Cordy, and Yves Le Traon. Constrained adaptive attacks: Realistic evaluation of adversarial examples and robust training of deep neural networks for tabular data, 2023.

[59] Chang Liu, Yinpeng Dong, Wenzhao Xiang, Xiao Yang, Hang Su, Jun Zhu, Yuefeng Chen, Yuan He, Hui Xue, and Shibao Zheng. A comprehensive study on robustness of image classification models: Benchmarking and rethinking, 2023.

[60] Xiaojing Fan and Chunliang Tao. Towards resilient and efficient llms: A comparative study of efficiency, performance, and adversarial robustness, 2024.

[61] Abhishek Moitra, Abhiroop Bhattacharjee, Youngeun Kim, and Priyadarshini Panda. Robust-edge: Low power adversarial detection for cloud-edge systems, 2023.

[62] Yiwen Guo, Chao Zhang, Changshui Zhang, and Yurong Chen. Sparse dnns with improved adversarial robustness. *Advances in neural information processing systems*, 31, 2018.

[63] Anne Harrington and Arturo Deza. Finding biological plausibility for adversarially robust features via metameric tasks, 2022.

[64] Mathias Lundteigen Mohus and Jinyue Li. Adversarial robustness in unsupervised machine learning: A systematic review, 2023.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.