Computer Science Review 58 (2025) 100793

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Review article

AI-generated text detection: A comprehensive review of methods, datasets, and applications☆

Tanzila Kehkashan [a,b], Raja Adil Riaz [b,*], Ahmad Sami Al-Shamayleh [c], Adnan Akhunzada [d,*,1], Noman Ali [b], Muhammad Hamza [b], Faheem Akbar [b,2]

[a] Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
[b] Faculty of Information Technology, University of Lahore, Sargodha 40100, Pakistan
[c] Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, 19328, Jordan
[d] College of Computing & IT, Department of Data and Cybersecurity, University of Doha for Science and Technology, Doha, 24449, Qatar

ARTICLE INFO

Keywords:
AI-generated text detection
Large language models
Natural language processing
Machine learning
Deep learning
Transformer models
Text classification
Authorship attribution
Neural text generation
Detection methods
Performance evaluation
Educational integrity
Content authenticity
Watermarking

ABSTRACT

This review examines the rapidly evolving field of AI-generated text detection, which has gained critical importance following the widespread deployment of advanced large language models like ChatGPT. We analyze the technical foundations, methodological approaches, evaluation frameworks, and practical applications of detection technologies designed to distinguish between human and machine-authored content. The paper synthesizes current knowledge across key dimensions: detection techniques ranging from statistical approaches to neural architectures, datasets and their limitations, performance metrics and evaluation challenges, real-world implementations across educational, publishing, and legal domains, and emerging research directions. Our analysis reveals significant challenges, including the inherent adversarial nature of detection, cross-domain generalization difficulties, and fairness concerns regarding certain writer populations. We identify promising trends toward multi-scale analysis, human-AI collaborative frameworks, and complementary provenance-based approaches. The review concludes that effective detection remains feasible but requires combining multiple approaches, domain-specific customization, and attention to ethical implications. This comprehensive examination serves as a resource for researchers, practitioners, and policymakers navigating the complex technical and societal dimensions of AI text detection in an era of increasingly sophisticated generative AI systems.

Contents

☆ The Open Access Funding is Provided by Qatar National Library (QNL). Besides, we also appreciate the necessary support of Al-Ahliyya Amman University.
* Corresponding authors.
E-mail addresses: tanzila.kehkashan@gmail.com (T. Kehkashan), rajaadilx1@gmail.com (R.A. Riaz), a.alshamayleh@ammanu.edu.jo (A. Sami), adnan.akhunzada@udst.edu.qa (A. Akhunzada), 22nomanalinomi@gmail.com (N. Ali), muhamadhamza0301@gmail.com (M. Hamza), faheemgujjar658@gmail.com (F. Akbar).
[1] Senior Member, IEEE.
[2] Member, IEEE.

https://doi.org/10.1016/j.cosrev.2025.100793
Received 12 April 2025; Received in revised form 8 July 2025; Accepted 10 July 2025
Available online 6 August 2025

1. Introduction

Artificial Intelligence (AI) has rapidly transformed from a niche research area to a pervasive force reshaping industries, economies, and societies worldwide. The global AI market size reached approximately $136.55 billion in 2022 and is projected to expand at a compound annual growth rate (CAGR) of 37.3% from 2023 to 2030 [1]. This explosive growth has been particularly evident in natural language processing (NLP), where recent advancements have enabled machines to generate increasingly coherent, contextually appropriate, and human-like text. The emergence of Large Language Models (LLMs) like GPT-3, GPT-4, LLaMA, and PaLM has demonstrated unprecedented capabilities in tasks ranging from creative writing to complex reasoning [2].

These technological breakthroughs have introduced powerful tools with immense potential for positive applications. LLMs are being deployed to assist with content creation, summarization, translation, code generation, and numerous other productive use cases [3]. However, this same technology also presents significant challenges and risks when misused or deployed without appropriate safeguards. The ability to generate human-like text at scale raises critical concerns about misinformation, academic dishonesty, automated impersonation, and the potential erosion of trust in digital communication [4].

As LLMs become more sophisticated and widely accessible, distinguishing between human-written and machine-generated content has evolved into a critical research challenge. This distinction is particularly important in domains where content authenticity directly impacts trust, accountability, and integrity. Educational institutions are concerned about academic dishonesty as students leverage AI tools for assignments [5]. News organizations worry about the spread of synthetic misinformation at unprecedented scales [6]. Scientific publishers must safeguard research integrity against potentially fabricated submissions [7]. Even legal and medical sectors, where accuracy and authorial responsibility are paramount, face challenges from convincing AI-generated documents [8].

The release of ChatGPT in November 2022 represented a watershed moment in this domain, as it made sophisticated text generation capabilities accessible to millions of users through a conversational interface [9]. Within two months of its launch, ChatGPT had amassed over 100 million users, becoming one of the fastest-growing consumer applications in history [10]. This widespread adoption has accelerated the urgency for reliable detection methods and broader societal discussions about the implications of generative AI.

The task of detecting AI-generated text is inherently adversarial. As detection methods improve, so too do the generation models and techniques to evade detection. This creates a perpetual arms race between generation and detection technologies [11]. Moreover, the increasing sophistication of modern LLMs produces text that exhibits fewer of the statistical patterns and linguistic artifacts that earlier detection approaches relied upon. When language models are fine-tuned specifically to mimic human writing patterns or are instructed to modify their outputs to avoid detection, the challenge becomes even more formidable [12].

Despite the proliferation of AI text detection tools, both commercial and open-source, significant gaps remain in our understanding of their reliability, limitations, and appropriate applications. Current detection approaches vary widely in their methodologies, from statistical analysis of text properties to deep learning classifiers trained on labeled examples of human and AI-written content [13]. However, systematic evaluations have revealed concerning inconsistencies in their performance across different contexts, languages, and generation models [14].

Several critical knowledge gaps motivate this review:

i. *Cross-domain generalization:* Detection methods often perform well on the specific domains and models they were developed for, but struggle when applied to new contexts or against different generation systems [15]
ii. *Adversarial robustness:* Limited research exists on how detection systems perform against content specifically crafted to evade detection, such as human-edited AI text or outputs from models fine-tuned to mimic human writing patterns [16].
iii. *Theoretical foundations:* While practical detection tools abound, there remains insufficient understanding of the fundamental statistical and linguistic differences between human and AI-generated text that enable detection [17].
iv. *Ethical implications:* The deployment of detection technologies raises questions about privacy, surveillance, and potential biases that have not been thoroughly addressed in the literature [18].

This review seeks to address these gaps by synthesizing current knowledge, identifying key challenges, and highlighting promising directions for future research in this rapidly evolving field.

This review paper aims to provide a comprehensive analysis of the current state of AI-generated text detection, with a particular focus on content produced by the latest generation of large language models. Specifically, our objectives are to:

i. Systematically categorize and analyze the primary methodological approaches to AI text detection, examining their theoretical foundations, technical implementations, and reported performance metrics.
ii. Assess the strengths, limitations, and reliability of existing detection methods across different contexts, content types, and adversarial scenarios.
iii. Identify the linguistic and statistical markers that differentiate human-written from AI-generated text, and how these markers evolve as language models improve.
iv. Explore the practical challenges in deploying detection technologies in real-world settings such as education, journalism, scientific publishing, and online platforms.
v. Outline a research agenda addressing the most critical open questions and promising avenues for advancing the field. Outline a research agenda addressing the most critical open questions and promising avenues for advancing the field.

Through this analysis, we contribute a structured framework for understanding AI text detection that bridges technical, empirical, and practical perspectives. This review synthesizes insights from computer science, linguistics, and application domains to provide both researchers and practitioners with a holistic understanding of this critical technology area.

The remainder of this paper is organized as follows: Section 2 provides essential historical and technical background on the evolution of language models and text generation capabilities that necessitate detection approaches. Section 3 examines the diverse techniques and methods employed in AI text detection, from traditional statistical measures to advanced neural architectures. Section 4 surveys the datasets available for training and evaluating detection systems, analyzing their composition, limitations, and coverage across domains and languages. Section 5 investigates performance metrics for detection evaluation, highlighting the strengths and weaknesses of different assessment approaches across various application contexts. Section 6 explores real-world applications across educational, publishing, legal, and commercial domains, examining implementation challenges and successful deployment strategies. Section 7 discusses the multifaceted challenges and limitations facing detection technologies, including technical barriers, ethical concerns, and operational constraints that shape the field's development. Section 8 identifies emerging trends and promising research directions for addressing current limitations and advancing detection capabilities in an evolving technological landscape. Section 9 concludes with a synthesis of key insights and broader implications for the future of AI-generated text detection.

Fig. 1 tells the evolution of text generation models and corresponding detection approaches, highlighting the shift from traditional NLP techniques to modern deep learning architectures and the emergence of LLM-based detection methods.

This comprehensive review aims to serve as a valuable resource for researchers developing new detection methods, practitioners implementing these technologies in applied settings, and policy makers navigating the complex landscape of synthetic content in the age of generative AI.

2. Historical and technical background

The evolution of AI text detection parallels key developments in language processing, from Turing's early concepts to modern neural architectures, providing critical historical context for current approaches.

2.1. Historical background

The development of AI-generated text detection exists within the broader evolution of artificial intelligence and natural language processing (NLP). This journey spans several decades, with distinct phases marked by conceptual breakthroughs, technological limitations, and paradigm shifts.

2.1.1. Early beginnings (1950s–1980s)

The theoretical foundations for detecting machine-generated content can be traced back to Alan Turing's seminal 1950 paper, which proposed the "Imitation Game" (now known as the Turing Test) as a measure of machine intelligence [19]. This conceptual framework established the fundamental challenge that remains central to text detection today: distinguishing between human and machine communication based solely on textual output.

Early NLP systems in the 1960s and 1970s were primarily rule-based, using handcrafted grammars and pattern-matching techniques to generate and analyze text. ELIZA, developed by Joseph Weizenbaum in 1966, represented one of the first conversational agents that could produce seemingly meaningful dialogue [20]. Despite its simplicity, ELIZA successfully created the illusion of understanding in some users, highlighting how even rudimentary systems could potentially deceive human interlocutors.

The 1980s saw the first attempts at statistical language modeling with n-gram approaches, which analyzed text based on the probability distributions of word sequences. These statistical techniques would later become important in early machine-generated text detection methods, as they enabled quantification of text patterns and irregularities [21].

2.1.2. The statistical NLP era (1990s–2000s)

The 1990s marked a shift toward statistical approaches in both text generation and analysis. Hidden Markov Models (HMMs) and statistical language models became prominent, enabling more sophisticated text generation capabilities [22]. During this period, the field also witnessed early work in authorship attribution and stylometry—techniques that sought to identify the authors of texts based on statistical analysis of writing patterns. These methods, though not explicitly designed for AI text detection, established important precedents for distinguishing between different sources of text based on statistical features [23].

The early 2000s saw further refinement of statistical NLP methods and the emergence of machine learning approaches to text classification. Support Vector Machines (SVMs) and other classification algorithms were applied to various NLP tasks, including spam detection, which shared conceptual similarities with the later challenge of identifying synthetic text [24]. Researchers began using computational stylometry to distinguish between human authors, inadvertently establishing methodologies that would later be adapted for human versus machine text classification [25].

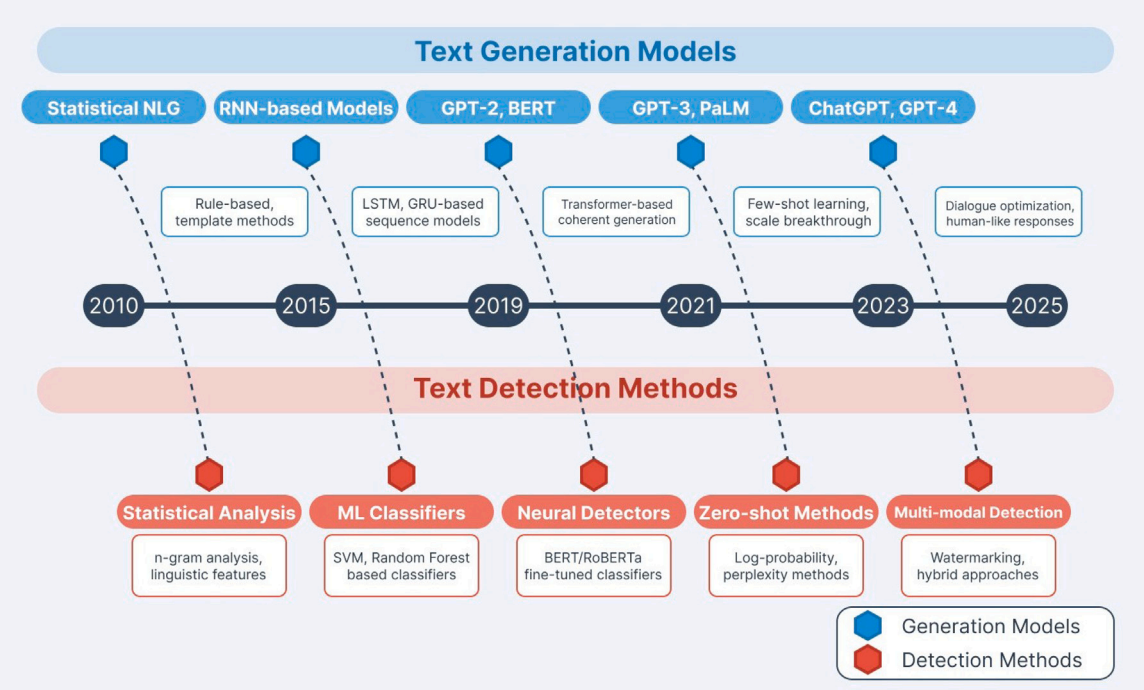

Fig. 1. How AI text generation and detection have evolved over time: A look at key breakthroughs and emerging trends.

2.1.3. The deep learning revolution (2010–2017)

The advent of deep learning transformed both text generation and detection capabilities. Word embeddings like Word2Vec (2013) and GloVe (2014) provided dense vector representations of words that captured semantic relationships, enabling more sophisticated text analysis [26]. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, significantly improved the quality of generated text by capturing long-range dependencies in language [27].

By 2015–2016, neural language models had advanced to the point where they could generate coherent paragraphs of text, prompting researchers like Yejin Choi to begin investigating methods for detecting neural-generated text [28]. This period saw the first dedicated research into neural fake text detection, as the improving quality of generated content raised concerns about potential misuse.

2.1.4. The transformer era and large language models (2017–present)

The introduction of the Transformer architecture in 2017 by Vaswani et al. revolutionized NLP and dramatically accelerated progress in text generation capabilities [29]. This architecture, which relied entirely on attention mechanisms rather than recurrence, became the foundation for a new generation of increasingly powerful language models.

The release of OpenAI's GPT (Generative Pre-trained Transformer) in 2018 marked a significant milestone, demonstrating impressive text generation abilities from a model trained on diverse internet text [30]. Shortly thereafter, researchers began developing specialized approaches for detecting GPT-generated content, recognizing the potential risks posed by increasingly convincing synthetic text [15].

GPT-2 (2019) raised further concerns with its enhanced generation capabilities, leading OpenAI to initially delay its full release due to misuse concerns [31]. This period saw the development of the first wave of dedicated GPT detection tools, including the GLTR (Giant Language model Test Room) system by Gehrmann et al. which visualized statistical patterns in text that might indicate machine generation [15].

The most recent era, characterized by models like GPT-3 (2020), ChatGPT (2022), and GPT-4 (2023), has presented unprecedented challenges for detection systems [2]. These models can generate text of sufficient quality to pass human evaluation in many contexts, necessitating increasingly sophisticated detection approaches. The emergence of specialized fine-tuned models designed to mimic human writing patterns has further complicated the detection landscape, creating an ongoing arms race between generation and detection technologies [11].

Fig. 2 illustrates the timeline of major milestones in AI text generation (blue) and detection (red) technologies, highlighting the accelerating progress since the introduction of the Transformer architecture in 2017.

Fig. 3 illustrates the comparison of computational resources required for training major language models, showing an exponential increase in FLOPs and parameter count from 2018 to 2023.

The visualization starkly illustrates the exponential growth in computational requirements underpinning modern language models, with training computation increasing by approximately eight orders of magnitude from BERT (2018) to GPT-4o (2024). This computational arms race has profound implications for text detection, as each leap in model scale has corresponded with qualitative improvements in generation capabilities that challenge existing detection methods. Particularly notable is the step change between GPT-3 and GPT-4, where both parameters and computation saw dramatic increases, coinciding with significantly more human-like text generation that required entirely new detection paradigms. The oscillating pattern in parameter counts — where some recent models like LLaMA 3 show parameter reductions while maintaining or improving capabilities — reflects the industry's growing emphasis on efficiency through architectural improvements rather than mere scale. These efficiency gains have accelerated the democratization of powerful generation models, further complicating the detection landscape by increasing the diversity of potential generation sources that detection systems must identify.

2.2. Technical background

The technical foundations of detection systems require understanding the mechanisms behind text generation, encompassing several fundamental principles that drive modern language models.

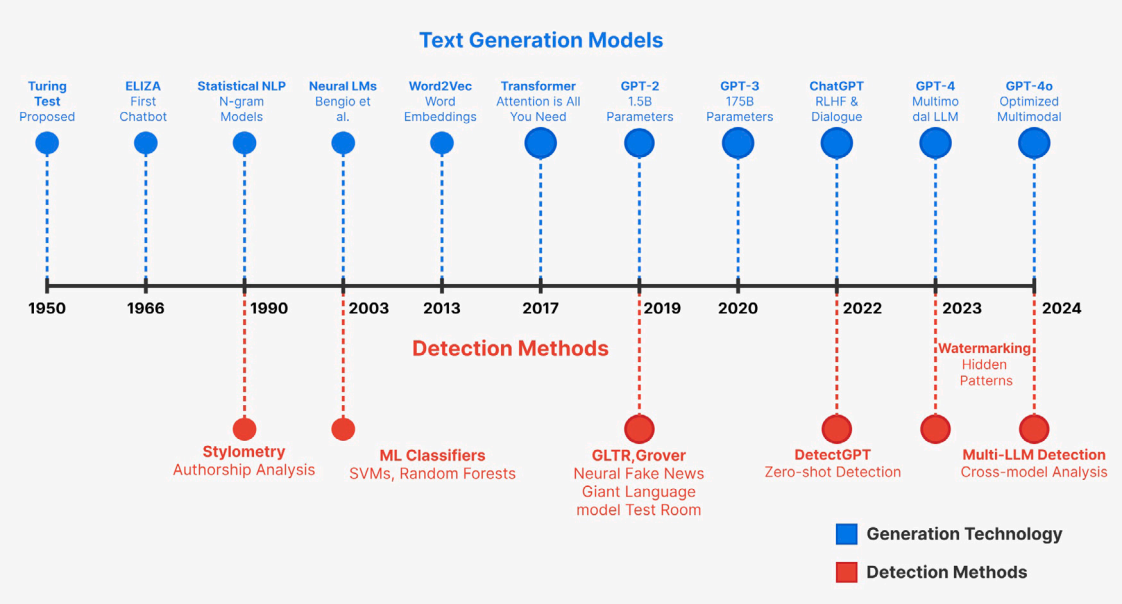

Fig. 2. AI text generation and detection: Key milestones from turing to GPT-4o.

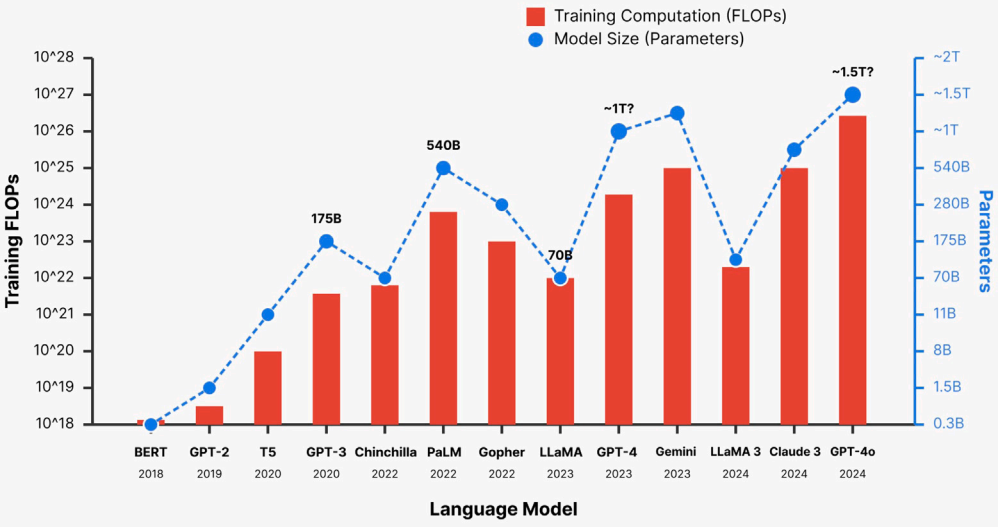

Fig. 3. Rising compute needs for AI: The growth of text generation models (2018–2025).

2.2.1. *Key concepts in AI text generation*

To understand detection systems, it is essential to first grasp the technology behind AI text generation. Modern text generation systems rely on several fundamental concepts:

- *Language Modeling* forms the foundation of text generation, involving the prediction of the probability distribution of words in a sequence. Traditional n-gram models calculated these probabilities based on fixed contexts of preceding words, while neural language models compute contextual representations dynamically [32].
- *Neural Network Architectures* for text generation have evolved significantly over time. RNNs and LSTMs process text sequentially, maintaining a hidden state that captures information from previous tokens. Transformers, the current dominant architecture, process all tokens in parallel using self-attention mechanisms to capture relationships between words regardless of their distance in the text [29].
- *Pre-training and Fine-tuning* represent the contemporary paradigm for developing language models. Models are first pre-trained on vast corpora of text using self-supervised objectives (typically next-token prediction), then fine-tuned on specific tasks or with human feedback to improve output quality and alignment with human preferences [33].
- *Decoding Strategies* determine how models generate text from probability distributions. Methods range from greedy approaches (always selecting the most probable token) to sampling-based approaches that introduce controlled randomness. The choice of decoding strategy significantly impacts the statistical properties of generated text, which in turn affects detectability [34].

2.2.2. *Algorithm evolution in text detection*

The technical approaches to AI-generated text detection have evolved alongside generation methods:

- *Statistical Analysis* methods represent the earliest approach, examining linguistic features like sentence length distribution, vocabulary richness, and n-gram frequencies to identify statistical patterns that differ between human and machine-written text [35].
- *Supervised Classification* approaches train models on labeled examples of human and AI-generated text. Early systems used traditional machine learning algorithms with handcrafted features, while contemporary approaches often fine-tune pre-trained language models like BERT or RoBERTa to classify text sources [36].

- *Zero-shot Detection* methods require no training examples of the specific generator being detected. These approaches typically analyze the statistical properties of text using another language model as a probe. For instance, the perplexity (a measure of how well a language model predicts a text sample) can reveal whether text was likely generated by a particular model [37].
- *Model-specific Watermarking* techniques embed statistical patterns into generated text that are imperceptible to humans but detectable by specialized algorithms. Recent work has explored various watermarking strategies for large language models that minimally impact output quality while enabling reliable detection [38].

2.2.3. Frameworks and tools

Several key frameworks and platforms have emerged for AI text detection:

- OpenAI's RoBERTa Classifier(2019–2020) represented an early attempt to create a general-purpose detector for text generated by GPT-2, using a fine-tuned version of the RoBERTa language model [35].
- GLTR (Giant Language Model Test Room) provided visual analysis tools for identifying machine-generated text, highlighting tokens based on their probability according to a language model, helping humans spot patterns characteristic of synthetic text [15].
- GPTZero emerged in early 2023 as one of the first widely-used commercial tools for detecting ChatGPT-generated content, using a combination of perplexity and "burstiness" (the variation in sentence complexity) as key detection features [39].
- DetectGPT introduced a novel approach based on curvature in the log probability space, enabling zero-shot detection without requiring examples from the specific model being detected [11].
- Watermarking Systems have been developed by both academic researchers and commercial AI providers. Notable examples include Stanford's DetectLLM framework and techniques proposed by researchers at the University of Maryland [38].

2.3. Relationship between historical and technical aspects

The historical progression of text detection methods directly mirrors advances in generation technology, creating a perpetual technical "arms race". As generation systems evolved from simple rule-based approaches to sophisticated neural architectures, detection methods similarly progressed from basic statistical analysis to complex neural classifiers. The shift from identifiable patterns in early machine-generated text to the more naturalistic outputs of modern LLMs has necessitated increasingly sophisticated detection approaches. Early language models often produced text with distinctive statistical irregularities that made detection relatively straightforward. Modern models like GPT-4, however, generate text with statistical properties much closer to human writing, requiring detectors to identify increasingly subtle differences [2].

The increasing accessibility of powerful generation models has also driven the development of more user-friendly detection tools. While early detection research was primarily confined to academic settings, the public release of ChatGPT in 2022 catalyzed the development of numerous commercial and open-source detection systems designed for use in educational, professional, and content moderation contexts [39].

2.4. Challenges and limitations in historical context

Throughout the evolution of AI text detection, several persistent challenges have shaped the field:

i. The Generalizability Problem has been a consistent limitation, with detection systems often performing well on texts from models they were trained on but struggling with content from new or updated models. This challenge has become more pronounced with the rapid iteration of language models, where detection systems quickly become outdated as new generation capabilities emerge [13].
ii. The Adversarial Gap represents the fundamental asymmetry between generation and detection tasks. While generators need only produce a single convincing passage, detectors must correctly classify all possible outputs—a much more difficult challenge. This inherent disadvantage for detection systems has persisted throughout their development [11].
iii. Cross-domain Performance limitations have been observed since early detection systems, with models trained on one text domain (e.g., news articles) performing poorly when applied to different domains (e.g., creative fiction). As language models have become more versatile across domains, this challenge has only increased [17].
iv. Human Editing of machine-generated text has historically presented a significant challenge for detection systems. Even early research demonstrated that minimal human post-editing could significantly reduce the detectability of machine-generated content—a problem that remains largely unsolved despite advances in detection methods [6].
v. Ethical and Access Considerations have evolved alongside detection technology. Early concerns focused primarily on preventing malicious uses of text generation, but contemporary discussions encompass broader questions about privacy, consent, bias in detection systems, and the equitable access to detection tools across different languages and cultural contexts [4].

The historical trajectory of these challenges reveals both progress and persistent limitations. While detection methods have become more sophisticated, the fundamental challenges of generalizability, adversarial dynamics, and human-in-the-loop scenarios continue to shape the field's development and will likely define its future research priorities.

3. Techniques and methods

This section provides a systematic examination of the techniques and methodologies employed in AI-generated text detection. As the capabilities of large language models have advanced dramatically, detection methods have similarly evolved in sophistication and diversity. We categorize these approaches based on their underlying principles, technical implementation, and historical development. The section begins by exploring foundational statistical and linguistic techniques before progressing to more advanced machine learning and deep learning approaches. We then examine specialized methods designed for specific detection scenarios, followed by an analysis of emerging hybrid approaches. Finally, we present a comparative evaluation of these methods across key performance dimensions and discuss the tools and frameworks commonly used for implementation.

3.1. Statistical and linguistic approaches

These approaches fundamentally examine the inherent structural and stylistic characteristics of text by applying established computational linguistics principles to identify patterns that distinguish human from machine authorship. Rather than relying on model-specific behaviors, these methods analyze observable textual properties through

traditional statistical frameworks. N-gram Analysis represents one of the earliest and most fundamental approaches to detecting machine-generated text. This method analyzes the distribution of word or character sequences (n-grams) in text, comparing them to expected distributions in human writing. Variations include unigram (single word) frequency analysis, which can reveal unusual vocabulary distributions, and higher-order n-grams that capture phrasal patterns [35].

These methods are computationally efficient but generally less effective against sophisticated modern language models. Perplexity-based Methods leverage statistical language models to measure how "surprised" a model is by a given text. Perplexity represents the inverse probability of the test text normalized by the number of words, with higher perplexity indicating text that is less predictable to the model [15]. A key finding in this domain is that text generated by neural models often exhibits lower perplexity when evaluated by the same or similar models, as they tend to produce more predictable text patterns than humans [37]. Stylometric Analysis examines quantifiable features of writing style, such as sentence length variation, punctuation patterns, and lexical diversity. These techniques build on traditional authorship attribution methods and can be effective at identifying the more uniform stylistic patterns that often characterize machine-generated text [25]. Stylometric approaches frequently use measures like type-token ratio (lexical richness) and function word distributions, which are less dependent on specific topics and therefore more generalizable across domains.

3.2. Machine learning-based approaches

Traditional Classifiers utilize feature engineering combined with algorithms such as Support Vector Machines (SVMs), Random Forests, or logistic regression. These approaches extract handcrafted features from text — including n-gram frequencies, part-of-speech patterns, and readability metrics — and train classifiers to distinguish between human and machine-generated content [36]. While conceptually straightforward, these methods require substantial domain expertise for effective feature selection and may struggle with the increasingly human-like text from advanced LLMs. Neural Network Classifiers employ various neural architectures to learn discriminative features automatically from large datasets of human and machine-generated text. Convolutional Neural Networks (CNNs) excel at capturing local textual patterns, while Recurrent Neural Networks (RNNs) and LSTMs are effective at modeling sequential dependencies in text [40]. These approaches generally outperform traditional classifiers but require significant training data and computational resources. Transformer-Based Detectors represent the current state-of-the-art in many detection scenarios. These approaches fine-tune pre-trained language models such as BERT, RoBERTa, or DeBERTa to perform binary classification between human and AI-generated text [36]. The contextual understanding and attention mechanisms in these models enable them to capture subtle patterns that distinguish AI-written content. OpenAI's RoBERTa-based classifier and the GPTZero system are prominent examples of this approach [39].

3.3. Zero-shot and model-specific methods

This category encompasses detection techniques that exploit the probabilistic mechanisms of language models directly, operating without requiring prior training on examples from the specific target generator. These methods capitalize on the observation that language models leave distinctive statistical signatures in their outputs that can be detected through careful analysis of the generation process itself.

i. *Log-probability analysis* examines the distribution of token probabilities assigned by language models during text generation. Machine-generated text often exhibits distinctive patterns in these probability distributions, such as a higher proportion of high-probability tokens [11]. DetectGPT employs a particularly innovative approach in this category, analyzing the curvature of the log-probability function around the generated text to identify machine-generated content without requiring examples from the specific generator being detected [11].

ii. *Binoculars* Advanced Zero-Shot Detection represents one of the most sophisticated zero-shot approaches currently available. Developed by Hans et al. Binoculars leverages the observation that machine-generated text exhibits different statistical patterns when evaluated by two distinct language models compared to human-written text [41]. The method computes cross-model perplexity scores using pairs of language models, finding that human text shows more consistent perplexity patterns across different models while AI-generated text reveals systematic discrepancies. This approach has demonstrated remarkable effectiveness across diverse domains and generation models, achieving state-of-the-art performance in zero-shot scenarios without requiring any training on detection-specific datasets. The Binoculars framework particularly excels at maintaining consistent performance across different text lengths and has shown robust generalization to unseen generation models.

iii. *Watermarking techniques* embed statistical patterns into text during the generation process that are imperceptible to humans but detectable by specialized algorithms. These approaches modify the sampling distribution during generation to subtly bias the model toward certain token patterns or relationships [38]. While highly effective when applicable, watermarking requires access to and modification of the generation model, limiting its use in many real-world scenarios where the text generation system may not be under the detector's control.

iv. *Perturbation testing* measures the robustness of text to small modifications. Machine-generated text often exhibits different sensitivity patterns to perturbations compared to human writing. For instance, replacing words with synonyms or paraphrasing sentences may cause different changes in model confidence or semantic coherence between human and AI-written text [12]. These approaches can be particularly useful in adversarial scenarios where more straightforward detection methods might be deliberately evaded.

v. *Systematic evaluation through shared tasks* has become increasingly important for advancing zero-shot detection methods. The SemEval-2024 Task 8 on "Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection" represented a significant milestone in the field, providing standardized evaluation protocols and datasets for comparing different detection approaches [42]. This shared task introduced several important innovations, including the first systematic evaluation of boundary detection—identifying the specific points where human text transitions to machine-generated content within a document. The task demonstrated that while zero-shot methods like DetectGPT and Binoculars perform well in traditional binary classification scenarios, boundary detection remains significantly more challenging, with even the best systems achieving only moderate performance in identifying precise transition points between human and AI-generated segments.

3.4. Boundary detection and partial text analysis

Boundary Detection represents an emerging and particularly challenging detection modality that goes beyond traditional binary classification to identify specific locations where human-authored text transitions to machine-generated content within a single document. This approach addresses real-world scenarios where users combine human and AI-generated text, either by alternating between human writing and AI assistance or by iteratively editing and expanding content with AI tools [42].

Unlike simple hybrid text detection, which attempts to classify whether a document contains any AI-generated content, boundary detection requires precise localization of transition points. This granular analysis presents several unique technical challenges. First, the detection system must operate at multiple scales simultaneously, analyzing both local linguistic patterns that might indicate authorship changes and global coherence patterns that might reveal inconsistencies in writing style or knowledge representation [43].

Sequence Labeling Approaches adapt traditional named entity recognition and sequence labeling techniques to the boundary detection problem. These methods treat each sentence or paragraph as a token to be classified as either human-authored or machine-generated, using conditional random fields or neural sequence models to maintain consistency across predictions [44]. However, these approaches often struggle with the subtle transitions that occur when AI tools are used for editing or expansion rather than wholesale generation.

Change Point Detection methods borrowed from time series analysis have shown promise for identifying authorship transitions. These approaches model text features as time series data and apply statistical change point detection algorithms to identify locations where the underlying statistical properties shift significantly [45]. Such methods can detect both abrupt transitions, where AI-generated text begins suddenly, and gradual transitions, where human and AI contributions become increasingly intermingled.

Multi-Scale Analysis for Boundaries examines text at multiple granularities to identify inconsistencies that might indicate transitions between human and machine authorship. Rashkin et al. developed approaches that analyze coherence at the word, sentence, and paragraph levels, finding that boundary regions often exhibit characteristic patterns such as sudden changes in lexical diversity, shifts in syntactic complexity, or inconsistencies in domain-specific knowledge [45].

The SemEval-2024 shared task evaluation revealed that boundary detection remains significantly more challenging than binary classification, with the best-performing systems achieving only 23% exact match accuracy for identifying precise transition points, compared to over 85% accuracy for binary document-level classification [42]. This performance gap highlights the need for specialized approaches rather than simple adaptations of existing binary classifiers.

Temporal Sequence Considerations distinguish boundary detection from other forms of hybrid text analysis. While traditional hybrid detection focuses on identifying documents that contain both human and AI content regardless of how they were created, boundary detection specifically addresses scenarios where the temporal sequence of composition matters. For instance, a human author who writes an initial draft, then uses AI to expand certain sections, then manually edits the AI contributions creates a complex temporal pattern that influences the resulting text's linguistic characteristics [46].

These temporal considerations are crucial for developing realistic detection systems, as they reflect how generative AI is actually used in practice rather than idealized scenarios of purely human versus purely machine-generated content.

3.5. Detailed analysis of key methods

The following section examines the most influential detection approaches currently in use.

3.5.1. GLTR (giant language model test room)

GLTR represents a pioneering approach to machine text detection that combines algorithmic detection with human interpretation [15]. The system leverages GPT-2's predictive distributions to highlight tokens in text based on their likelihood.

Algorithm Description: GLTR calculates the likelihood of each word in a document using a pre-trained language model (originally GPT-2) and categorizes tokens into four bins based on their likelihood rank:

- Top 10 most likely next words (highlighted in green)
- Top 100 but not top 10 (highlighted in yellow)
- Top 1,000 but not top 100 (highlighted in red)
- Not in top 1,000 (highlighted in purple)

This color-coded visualization reveals patterns in text predictability. Machine-generated text typically contains a higher proportion of green tokens (highly predictable words) and fewer purple tokens (surprising word choices) compared to human writing.

The visual feedback system offers several practical benefits:

- Provides interpretable visual feedback that helps humans identify machine-generated text
- Requires no training data specific to the detection task
- Offers a useful framework for understanding fundamental differences between human and machine text

Despite its innovative approach, this method has notable drawbacks:

- Effectiveness diminishes against newer, more sophisticated language models
- Requires access to a language model's probability distributions
- Primarily designed as a human-in-the-loop tool rather than a fully automated detector

3.5.2. RoBERTa-based classification

Fine-tuned RoBERTa models represent a widely adopted approach for AI text detection due to their strong performance and versatility [36].

*Algorithm Description:*This approach fine-tunes the RoBERTa language model, a robustly optimized BERT variant, on a dataset containing examples of human and machine-generated text. The process involves:

- Constructing a balanced dataset of human and AI-generated texts
- Fine-tuning RoBERTa with a classification head for binary classification
- Using contextual embeddings to capture subtle linguistic patterns that distinguish between the two sources

This transformer-based approach provides compelling advantages:

- Leverages powerful pre-trained representations of language
- Capable of capturing complex contextual patterns beyond simple statistical features
- Adaptable to different domains with appropriate fine-tuning data

However, this classification method encounters several obstacles:

- Performance degrades when encountering text from generators not represented in the training data
- Requires substantial computational resources for training
- May exploit artifacts specific to certain generators rather than fundamental differences between human and machine text

3.5.3. DetectGPT

DetectGPT introduces an innovative zero-shot approach to detecting machine-generated text without requiring examples from the specific model being detected [11].

Algorithm Description: The method is based on the key insight that text generated by language models tends to occupy high-probability regions in the model's probability distribution space. DetectGPT works by:

- Computing the log probability of the original text using a language model

- Generating perturbations of the text through paraphrasing or word substitutions
- Computing log probabilities of the perturbed texts
- Measuring the average "curvature" of the log probability function around the original text
- Classifying text as machine-generated if the curvature exceeds a certain threshold

Higher curvature indicates that the original text occupies a local maximum in probability space—a characteristic more common in machine-generated text than human writing.

This probability-based method delivers significant benefits:

- Works without requiring examples from specific generators (zero-shot)
- Theoretically sound approach based on the inherent properties of language model outputs
- More robust to adversarial attacks than pure classification approaches

The computational complexity creates several implementation challenges:

- Computationally intensive, requiring multiple forward passes through a language model
- Performance varies depending on the perturbation strategy
- Less effective against text that has been extensively edited by humans

3.6. Hybrid and emerging methods

Beyond individual detection methods, researchers are increasingly exploring integrated approaches that leverage complementary strengths across multiple techniques.

3.6.1. Multi-model ensemble approaches

Combine the predictions of different detection models to achieve higher accuracy and robustness. These methods integrate diverse techniques, such as combining stylometric analysis with neural classifiers or integrating perplexity-based methods with transformer-based classifiers [40]. Ensembles typically outperform individual models, especially when the component methods capture complementary aspects of the differences between human and AI-written text.

3.6.2. Contrastive learning frameworks

Adapt self-supervised learning techniques to the detection problem by learning representations that explicitly maximize the distance between human-written and machine-generated texts in the embedding space [47]. These approaches can be more data-efficient than traditional supervised learning and potentially more robust to distribution shifts when encountering new generators.

3.6.3. Prompt-based detection

Leverages the capabilities of large language models themselves to identify machine-generated content through carefully designed prompting strategies [48]. For instance, a model might be prompted to analyze text for characteristics of machine generation, providing a meta-analysis based on its own understanding of generation patterns. This emerging approach offers a potentially adaptable method that can evolve alongside generation capabilities.

3.6.4. Multimodal detection systems

Extend beyond pure text analysis to incorporate metadata, user behavior patterns, or contextual information. These approaches recognize that detection in real-world settings often benefits from signals beyond the text itself, such as posting patterns, user history, or contextual inconsistencies [38]. While primarily deployed in social media and online content moderation contexts, these methods offer promising directions for enhancing detection robustness.

3.7. Comparison of detection methods

The effectiveness of detection methods varies significantly across different scenarios, datasets, and generation models. Table 1 provides a comparative analysis of major detection approaches along key dimensions.

Several important patterns emerge from this comparison:

- Performance Trade-offs: Methods with higher accuracy often suffer from lower generalizability across models or domains. For instance, model-specific methods like watermarking achieve near-perfect accuracy for their target models but offer no detection capability for unwatermarked models.
- *Computational Requirements:* Zero-shot methods like DetectGPT offer strong generalizability but at a significant computational cost, requiring multiple forward passes through large language models.
- *Minimal Text Length:* Most detection methods perform poorly on very short texts (fewer than 50 words), with performance improving substantially as text length increases. This represents a significant limitation for applications like comment or message moderation.
- *Adversarial Robustness:* Methods differ substantially in their resilience to adversarial modifications, with ensemble approaches and DetectGPT demonstrating superior performance against light editing of machine-generated text.

Fig. 4 presents a heatmap visualization of detection accuracy across different methods and popular benchmark datasets, revealing that performance can vary dramatically depending on the specific generation model and domain.

The heatmap reveals several critical patterns in detection method performance across different datasets and scenarios. Transformer-based approaches — particularly RoBERTa fine-tuned models — consistently outperform traditional methods on standard benchmarks, achieving impressive 92% accuracy on academic abstracts and essay datasets. However, this performance advantage narrows significantly in challenging scenarios like adversarially modified text and short-form content under 50 words. Notably, DetectGPT demonstrates the most consistent performance across all test conditions, maintaining above 78% accuracy even for short texts where other methods degrade substantially. The visualization also highlights the stark contrast between watermarking's near-perfect performance on watermarked content versus its complete inapplicability to unwatermarked text, underscoring both its power and fundamental limitation. Perhaps most instructive is the clear performance gradient visible across most methods when moving from structured content like academic abstracts to more challenging scenarios like adversarial texts—a pattern that emphasizes the importance of benchmark selection when evaluating detection systems for real-world applications.

3.8. Implementation tools and frameworks

Several tools and frameworks have emerged to facilitate the implementation of AI text detection methods:

- *Hugging Face Transformers* provides implementations of numerous transformer-based models that can be fine-tuned for detection tasks, along with pre-trained detection models shared by the community [49]. This framework has become the de facto standard for implementing and sharing transformer-based detectors.
- *OpenAI's Detector Tools* included a publicly available RoBERTa-based classifier for GPT-2 detection, though this was later deprecated due to reduced effectiveness against newer models [35]. The conceptual approach and implementation details from this work continue to influence many detection systems.

Table 1
Comparative analysis of AI text detection methods.

Characteristic	Statistical/ Linguistic	ML Classifiers	Transformer-Based	Zero-Shot Methods	Water-marking
Accuracy (GPT-3.5 text)	65%–75%	70%–85%	85%–95%	80%–90%	>98%[a]
Accuracy (GPT-4 text)	55%–65%	60%–75%	75%–85%	75%–85%	>98%[a]
Generalization to unseen models	Moderate	Low-Moderate	Moderate	High	None[b]
Minimal effective text length	200+ words	100+ words	50+ words	150+ words	25+ words
Computational requirements	Low	Moderate	Moderate-High	Very High	Low
Implementation complexity	Low	Moderate	Moderate	High	Low[c]
Adversarial robustness	Low	Low-Moderate	Moderate	High	High
Human interpretability	High	Low-Moderate	Low	Moderate	High

Note: Performance ranges are approximate and based on reported results across multiple studies. Actual performance may vary significantly across specific implementations and contexts.

[a] Only applies to text generated with watermarking enabled.

[b] Watermarking only works for specifically watermarked models.

[c] Low for detection, but requires modification of the generation process.

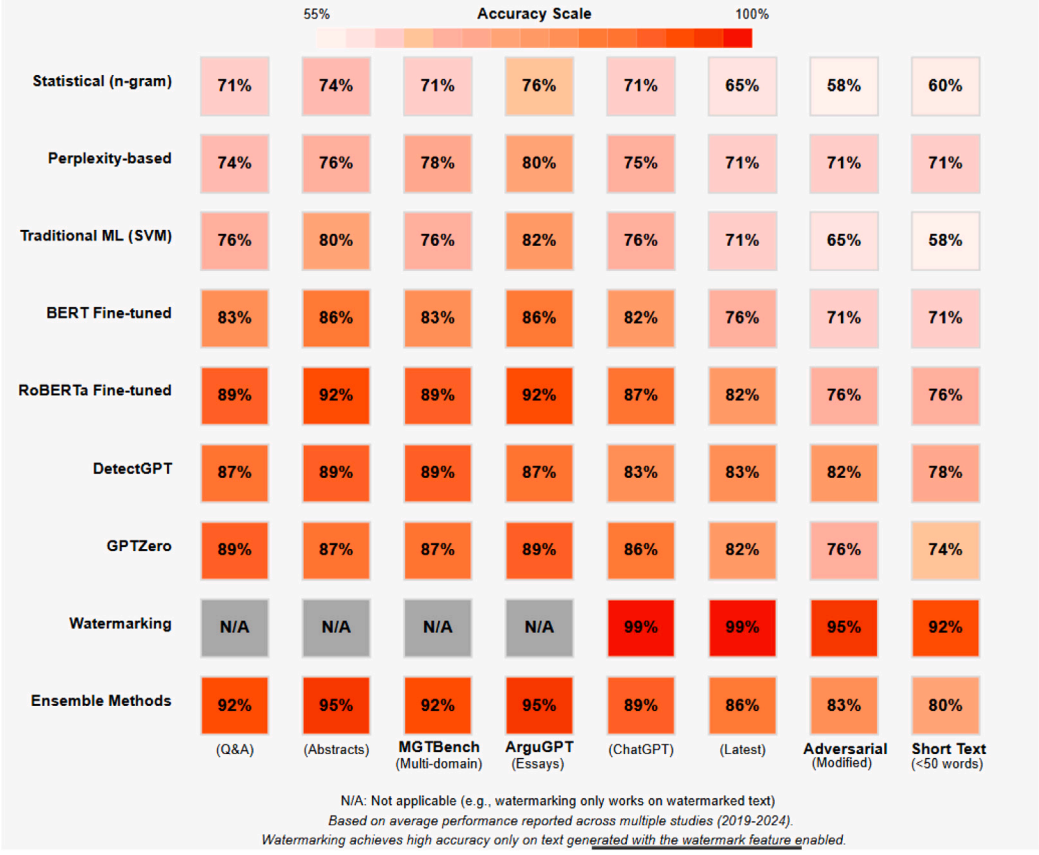

Fig. 4. AI-generated text detection: Accuracy comparison across methods and datasets.

- *GLTR Framework* offers both a web interface and programmable API for analyzing text using token likelihood visualizations [15]. While initially built around GPT-2, extensions have adapted the approach to work with newer models.
- *GPTZero* provides an accessible commercial API for text detection that integrates perplexity analysis with measures of "burstiness" (variation in perplexity across a text) [39]. This system has gained significant adoption in educational contexts.
- *DetectGPT Implementation* is available as open-source code, enabling researchers to apply and extend the zero-shot curvature-based detection approach [11]. The modular implementation facilitates experimentation with different perturbation strategies and underlying language models.
- *LLM Watermarking Libraries* such as those developed by Kirchenbauer provide tools for implementing statistical watermarking in text generation systems [38]. These libraries enable token-level biasing during the generation process and subsequent detection.

The diverse landscape of implementation tools reflects the multifaceted nature of the detection challenge. Most practitioners adopt hybrid approaches, combining multiple detection signals to achieve more robust performance across varying contexts and generation models.

4. Datasets

Datasets serve as the fundamental backbone for developing, evaluating, and benchmarking AI text detection systems. The quality, scope, and diversity of these datasets directly influence the efficacy and reliability of detection methods. As large language models continue to evolve and improve their ability to generate increasingly human-like text, the datasets used to train and evaluate detection systems must similarly evolve to remain relevant and effective. In the context of AI-generated text detection, datasets face unique challenges compared to other machine learning domains. The rapid advancement of generative models requires datasets that represent the latest generation capabilities, yet creating such datasets involves significant effort and domain expertise. Moreover, detection systems trained on text from one generation model often perform poorly when applied to text from newer or different models [16]. This moving target nature of the problem makes dataset creation and curation particularly challenging.

This section reviews the prominent datasets developed specifically for AI-generated text detection, examining their characteristics, strengths, limitations, and contributions to the field. We also discuss emerging dataset creation strategies and highlight ongoing challenges and opportunities in this critical area of research.

4.1. General characteristics of datasets

Several key characteristics define effective datasets for AI text detection, with representation balance forming a critical foundation.

4.1.1. Balanced representation

Effective AI text detection datasets require balanced representation between human-written and machine-generated text samples. This balance should extend beyond mere quantities to ensure comparable text lengths, topics, styles, and domains. Early detection datasets often exhibited significant distributional differences between human and AI-generated samples, inadvertently allowing models to exploit these artifacts rather than learning fundamental distinctions [15]. Modern datasets increasingly prioritize matching distributions across various text attributes to force models to learn more robust detection features.

4.1.2. Size and diversity

Dataset size requirements vary depending on the detection approach. While transformer-based classification methods benefit from large datasets with hundreds of thousands of examples, zero-shot methods may require smaller, more carefully curated datasets for evaluation [11]. Dataset diversity across domains (e.g., academic writing, creative fiction, technical documentation), languages, and generation models is crucial for developing generalizable detection systems. Recent research reveals that most publicly available datasets remain predominantly English-focused, with limited representation of other languages and cultural contexts [50].

Fig. 5 illustrates the distribution of dataset sizes and types for prominent AI text detection datasets, with bubble size indicating the number of text samples, while position represents the average text length and domain diversity.

The visualization reveals striking disparities in dataset composition across the detection landscape. HC3-English stands out with the largest sample size (85,449 samples) while offering moderate text length and domain diversity, contrasting with CHEAT which provides significantly longer texts (500+ words on average) with a substantial sample size. This pattern highlights a critical trade-off between text length and dataset scale that researchers must navigate. Notably, most datasets cluster either toward shorter texts with higher domain diversity (like MGTBench and HC3) or longer texts with more limited domains (like CHEAT and Writing Prompts). The adversarial dataset Human-Edited occupies a unique position with shorter texts and minimal domain diversity, potentially limiting its generalizability despite its importance for evaluating detection robustness. This visualization underscores the need for more balanced datasets that combine sufficient sample sizes with both length and domain diversity to support more generalizable detection systems.

4.1.3. Labeling quality and provenance

Unlike many machine learning tasks where ground truth can be objectively determined, AI text detection datasets rely on accurate provenance information. Human-written text must be verifiably human-authored, while machine-generated text must have clear documentation of the generation model, parameters, prompts, and any post-processing steps applied. This provenance information is increasingly recognized as essential metadata that should accompany datasets [51]. Without such documentation, it becomes difficult to contextualize a detector's performance or understand its limitations across different generation approaches.

4.1.4. Adversarial considerations

The most advanced datasets include adversarial examples designed to challenge detection systems. These may include human-edited machine text, machine-paraphrased human text, or deliberately designed examples that target known weaknesses of detection methods [12]. Such adversarial components are vital for evaluating real-world robustness, as they represent realistic scenarios where malicious actors might attempt to evade detection.

4.1.5. Ethical and privacy considerations

Dataset creation involves important ethical considerations, particularly regarding the human-authored content. Researchers must ensure proper attribution, consent, and privacy protections for human authors whose text is included in public datasets. Additionally, considerations about potential biases in dataset composition — whether related to topic, style, or demographic representation — must be addressed to avoid propagating or amplifying these biases in detection systems [4].

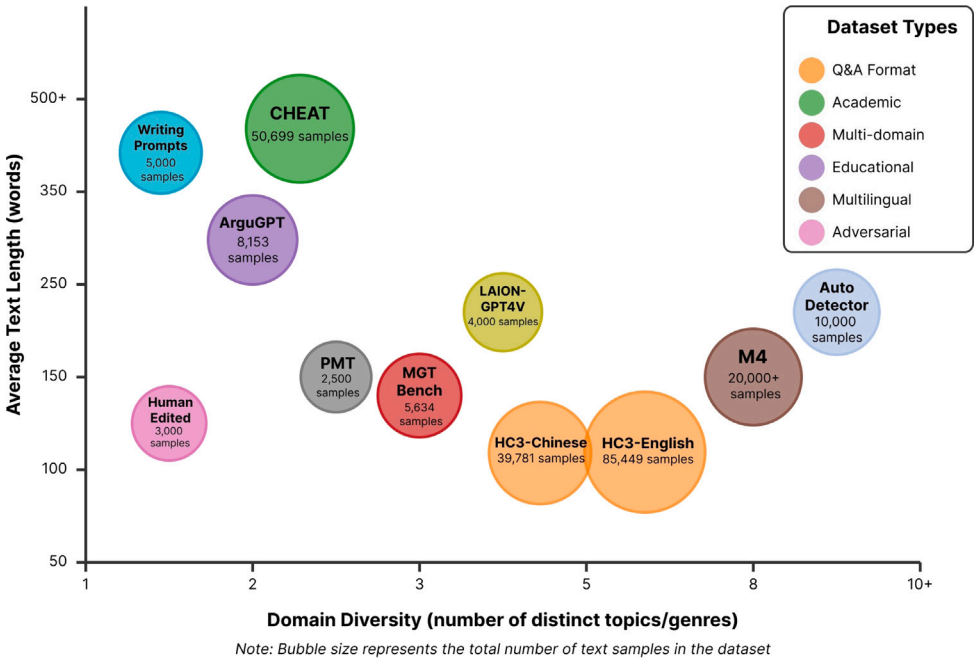

Fig. 5. Text detection datasets mapped by length, domain range, and volume.

4.2. Prominent datasets

Table 2 provides a comparative overview of major datasets for AI-generated text detection. This comparison highlights the diverse approaches to dataset construction, from HC3's broad question-answering format to CHEAT's focus on academic writing and M4's multilingual emphasis. While these datasets collectively enable evaluation across various scenarios, they exhibit significant differences in size, text length, and domain coverage that researchers must consider when selecting appropriate benchmarks. The table complements our earlier discussion by offering a structured comparison of key dataset attributes that influence detection system performance and generalizability.

4.2.1. HC3 (human ChatGPT comparison corpus)

The HC3 dataset, developed by Guo et al. stands as one of the first comprehensive datasets specifically designed for ChatGPT detection [51]. Available in both English and Chinese versions, HC3 contains question-answer pairs across multiple domains:

- *Size and Composition*: The English version includes 24,322 questions with 58,546 human answers and 26,903 ChatGPT answers. The Chinese version contains 12,853 questions with 22,259 human answers and 17,522 ChatGPT answers.
- *Data Sources*: The dataset draws human-written content from sources like Stack Exchange, Quora, and Reddit, with matched ChatGPT-generated responses created using context-sensitive prompting.
- *Strengths*: HC3's multi-domain and bilingual nature makes it particularly valuable for cross-lingual research. The inclusion of multiple human answers per question enables analysis of human writing diversity compared to machine generation.
- *Limitations*: As one of the earliest ChatGPT detection datasets, HC3 focuses only on question-answering contexts and does not include examples from more recent or advanced models.

4.2.2. CHEAT (ChatGPT-written abstract test)

Developed by Yu et al. the CHEAT dataset focuses specifically on scientific writing in the form of research paper abstracts [53]:

- *Size and Composition*: CHEAT contains 15,395 human-written abstracts and 35,304 ChatGPT-generated abstracts from computer science papers.
- *Generation Approaches*: The dataset employs three distinct generation strategies: direct generation from titles, "polishing" of human abstracts, and sentence-level mixing of human and machine content.
- *Strengths*: CHEAT's focus on scientific writing addresses a critical application domain where AI text detection has significant implications. The inclusion of adversarial examples (polished and mixed text) makes it particularly valuable for evaluating detection robustness.
- *Limitations*: The dataset is limited to computer science abstracts, potentially reducing generalizability to other scientific domains or writing styles.

4.2.3. MGTBench (machine generated text benchmark)

MGTBench represents one of the most comprehensive evaluation benchmarks for machine-generated text detection [16]:

- *Size and Composition*: The dataset includes 2,817 human-authored examples and corresponding machine-generated texts from multiple generation models, including ChatGPT, GPT-4, PaLM, and others.
- *Data Sources*: MGTBench incorporates texts from three question-answering datasets: TruthfulQA, SQuAD1, and NarrativeQA, covering factual, open-ended, and narrative questions.
- *Strengths*: By including text from multiple generation models, MGTBench facilitates comparative analysis of detection performance across different generators, addressing the crucial challenge of cross-model generalization.
- *Limitations*: Like HC3, MGTBench focuses primarily on question-answering contexts rather than encompassing other writing formats or genres.

4.2.4. ArguGPT

Focusing on argumentative writing in educational contexts, the ArguGPT dataset created by Liu et al. addresses the significant concern of AI-generated essays in academic settings [54]:

Table 2
Comparative analysis of prominent datasets for AI-generated text detection.

Dataset	Size	Average Text Length	Domain Coverage	Languages	Key Features
HC3 [52]	125,230 samples	100–150 words	Q&A (forums, social media)	English, Chinese	Multiple human answers per question; bilingual
CHEAT [53]	50,699 samples	500+ words	Academic (research abstracts)	English	Multiple generation strategies; focus on scientific writing
MGTBench [16]	5,634 samples	150 words	Question-answering (factual, narrative)	English	Multiple generation models (ChatGPT, GPT-4, PaLM)
ArguGPT [54]	8,153 samples	300 words	Educational (argumentative essays)	English	Focus on language learner writing; academic contexts
M4 [55]	20,000+ samples	200 words	News, reviews, academic	10 languages	Multilingual focus; multiple domains and models
Human-Edited [56]	3,000 samples	120 words	Web content	English	Pairs of original and human-edited AI text
PMT [12]	2,500 samples	170 words	Web content, news	English	Focus on paraphrased machine text
LAION-GPT4V [57]	4,000 samples	220 words	Image captions, descriptions	English, Japanese	Multimodal (text generated from images)
Writing Prompts [58]	5,000 samples	380 words	Creative writing	English	Focus on narrative and creative content
AutoDetector [59]	10,000 samples	240 words	Multi-domain (synthetic)	English, Spanish, French	Automatically generated dataset; domain variation

Note: Dataset statistics are approximate and compiled from multiple sources. Domain coverage indicates primary content types represented. Size figures represent the total of human and AI-generated samples combined.

- *Size and Composition*: ArguGPT contains 4,115 human-written essays and 4,038 GPT-generated essays spanning various argumentative topics and prompts.
- *Data Sources*: The dataset draws from English learning corpora, including WECCL and TOEFL11, as well as GRE preparation materials, with matched GPT-generated responses to the same prompts.
- *Strengths*: ArguGPT's focus on argumentative essays makes it particularly relevant for educational applications. The dataset includes grammatical error correction and processing to remove obvious machine-generation markers.
- *Limitations*: The essays are primarily from English language learners rather than native speakers, potentially introducing specific linguistic patterns that may not generalize to other writer populations.

4.2.5. M4 (Multilingual, Multi-domain, Multi-model Machine-generated text)

Recognizing the limitations of English-centric datasets, Feng developed M4, a multilingual dataset for AI text detection [60]:

- *Size and Composition*: M4 includes over 10,000 text pairs (human and machine-generated) across 10 languages and multiple domains, including news, reviews, and academic writing.
- *Generation Models*: The dataset incorporates text from multiple generation systems, including ChatGPT, GPT-4, BLOOM, and LLaMA.
- *Strengths*: M4's multilingual and multi-domain approach addresses critical gaps in existing datasets. By including both high-resource and low-resource languages, it enables research on cross-lingual detection transfer.
- *Limitations*: The number of examples per language varies considerably, with substantially more content in high-resource languages like English and Chinese compared to languages like Arabic and Hindi.

4.3. Emerging and synthetic datasets

Beyond standard detection datasets, researchers have developed specialized collections that focus on challenging edge cases and evasion techniques.

4.3.1. Adversarial and editing-based datasets

Several recent datasets focus specifically on challenging scenarios where machine-generated text has been deliberately modified to evade detection:

The Human-Edited GPT Dataset was introduced as a resource containing pairs of original GPT-generated text alongside corresponding human-edited versions designed to obscure their machine origins [56]. This dataset enables researchers to study how human editing affects detectability and which editing strategies are most effective at evading detection.

The Paraphrased Machine Text (PMT) dataset by Krishna focuses on the vulnerability of detectors to paraphrasing [12]. By applying various

automated paraphrasing techniques to machine-generated text, PMT enables systematic evaluation of detector robustness to semantically preserved modifications.

4.3.2. Realistic usage pattern datasets

Authentic AI Utilization Datasets represent a crucial advancement in creating more realistic evaluation scenarios. Liyanage et al. conducted groundbreaking research on the importance of developing datasets that authentically reflect how generative AI is actually used in academic writing contexts [46]. Their work demonstrates that traditional datasets, which typically present clean distinctions between purely human and purely AI-generated text, fail to capture the complex ways users interact with AI writing tools in practice.

Their study revealed that students and researchers commonly use AI tools in iterative, collaborative ways: starting with human brainstorming, using AI for initial drafts, manually revising AI outputs, requesting AI expansions of specific sections, and alternating between human editing and AI assistance throughout the writing process. This realistic usage pattern creates text with subtle linguistic boundaries and mixed authorship that significantly challenges current detection approaches.

The implications of this research extend beyond dataset construction to fundamental questions about detection system design. Liyanage et al. found that detectors trained on traditional binary datasets often fail catastrophically when applied to text created through realistic AI-human collaboration patterns, achieving accuracy rates below 60% compared to over 90% on clean datasets. This finding highlights the critical importance of developing evaluation frameworks that mirror actual AI usage rather than idealized scenarios.

4.3.3. Synthetic dataset generation

A promising approach to address the constant evolution of language models involves synthetic dataset generation. Wang proposed a framework for automatically generating detection datasets by sampling from multiple language models with varying parameters and prompts [50]. This approach potentially eliminates the need for manual dataset creation every time a new language model emerges.

Synthetic datasets also enable controlled experiments on specific linguistic or stylistic features. For instance, Guo demonstrated the creation of synthetic datasets with controlled stylistic attributes to investigate which features most contribute to detection performance [51].

4.3.4. Multimodal and cross-domain datasets

Emerging research recognizes that AI-generated content increasingly spans multiple modalities:

The LAION-GPT4V dataset contains image-text pairs where the textual descriptions were generated by multimodal models like GPT-4V in response to images [57]. This dataset enables research on detecting AI-generated content in multimodal contexts, where text generation is conditioned on visual information.

Cross-domain datasets like WritingPrompts-Human-AI pair creative writing prompts with both human and AI-generated stories, examining how detection performs in highly creative and unconstrained writing tasks [58].

4.3.5. Implications for future dataset development

The work by Liyanage et al. and others underscores the need for a fundamental shift in how detection datasets are conceptualized and constructed [46]. Future datasets should prioritize ecological validity — accurately representing the diverse ways AI tools are integrated into human writing processes — over the clean categorical distinctions that have characterized early detection research.

This shift requires more sophisticated data collection methodologies, including longitudinal studies of AI usage in natural writing contexts, process-tracing approaches that capture the temporal dynamics of human-AI collaboration, and ethnographic research into how different user communities adopt and adapt AI writing tools for their specific needs.

4.4. Future directions and challenges

Several critical challenges remain in dataset development for AI text detection:

- *Temporal Relevance*: As language models rapidly evolve, datasets quickly become outdated. Creating sustainable approaches for continuous dataset development and benchmarking remains an open challenge [61].
- *Domain Coverage*: Most existing datasets focus on specific domains like academic writing or Q&A formats. Broader coverage across diverse writing contexts (e.g., technical documentation, creative writing, journalism) is needed to develop truly robust detection systems.
- *Cross Model Generalization*: Datasets that enable systematic evaluation of cross-model generalization — detecting text from models not represented in training data — are essential but still limited [11].
- *Multilingual and Multicultural Representation*: Despite recent efforts like M4, significant disparities remain in language coverage and cultural context representation [50].

The development of more comprehensive, diverse, and challenging datasets remains a fundamental requirement for advancing AI text detection research and applications. Future efforts should prioritize addressing these challenges while maintaining rigorous documentation standards and ethical considerations.

The datasets described in this section provide the foundation for developing and evaluating AI text detection methods. As the field continues to evolve, these resources will need to expand in size, diversity, and sophistication to keep pace with advances in generation capabilities.

The language distribution chart starkly illustrates the profound English-centricity of current detection datasets, with 45% of all available samples being English-based and another 18% in Chinese. This heavy skew creates substantial challenges for developing truly multilingual detection systems, with even major world languages like Arabic, Hindi and Italian representing just 1%–2% of available data. The dataset coverage comparison below further reveals that while M4 offers the broadest language inclusion, most individual datasets contain only English or at most two languages. This limited linguistic diversity directly impacts detection fairness across non-English speaking populations, as models trained predominantly on English may exhibit degraded performance or increased bias when applied to other languages. The visualization provides compelling evidence for the urgent need to develop more balanced multilingual datasets that better represent global linguistic diversity.

5. Performance metrics

Evaluation metrics play a pivotal role in quantifying the effectiveness of AI-generated text detection systems. Unlike many standard classification tasks, the detection of machine-generated content presents unique challenges that require careful consideration when selecting appropriate metrics. The rapidly evolving capabilities of large language models, combined with the high stakes in domains like education, journalism, and scientific publishing, make rigorous and multi-faceted evaluation particularly crucial.

At its core, AI text detection resembles a binary classification problem—distinguishing between human-written and machine-generated text. However, this seemingly straightforward task encompasses multiple dimensions of performance that must be captured by different metrics [13]. A detector might excel at identifying obvious machine-generated passages while struggling with more subtle cases, or it might perform well on certain text domains while failing on others. Moreover, the consequences of false positives (incorrectly flagging human-written

Table 3
Comparison of key evaluation metrics in AI-generated text detection.

Metric	Definition	Strengths	Limitations	Typical Range
Accuracy	Proportion of correctly classified samples (both human and AI)	Intuitive interpretation; easily communicated to stakeholders	Misleading for imbalanced datasets; masks performance disparities	0.65–0.95
F1 Score	Harmonic mean of precision and recall	Balances false positives and false negatives; robust to class imbalance	Assumes equal importance of precision and recall	0.70–0.92
AUC-ROC	Area under the Receiver Operating Characteristic curve; measures discrimination across thresholds	Threshold-invariant; robust performance indicator across decision boundaries	Overly optimistic on imbalanced datasets; abstract interpretation	0.75–0.98
AUC-PR	Area under the Precision-Recall curve	Superior for imbalanced datasets; focuses on positive class performance	Less intuitive than basic metrics; requires computational visualization	0.72–0.95
CMGS	Ratio of detection performance on unseen models to performance on known models	Measures critical generalization capability; identifies overfitting to specific generators	Requires multiple model evaluations; relatively new with limited standardization	0.65–0.88

Note: Performance ranges are approximate and vary significantly based on dataset properties, model architecture, and evaluation methodology. Higher values indicate better performance for all metrics.

text as machine-generated) and false negatives (failing to identify machine-generated text) may carry different weights depending on the application context [62].

The evaluation of AI text detectors must also account for their performance under realistic conditions. This includes robustness against adversarial modifications, generalization to unseen models, and performance on text samples from diverse domains, languages, and cultural contexts [63]. Given these multifaceted requirements, no single metric can fully capture a detector's performance, necessitating a comprehensive evaluation framework.

5.1. Common metrics

Table 3 provides a comprehensive comparison of principal evaluation metrics employed in AI text detection research. While accuracy remains the most widely reported measure due to its interpretational simplicity, the field increasingly recognizes its limitations, particularly when assessing detector performance across diverse application contexts. Metrics like F1 score offer more balanced assessment by considering both false positives and negatives—a critical consideration in high-stakes deployment scenarios where error types carry different consequences. The relatively newer Cross-Model Generalization Score addresses a fundamental challenge by explicitly quantifying a detector's ability to identify content from generation models not represented in training data. This comparative framework illustrates why reliance on any single metric provides incomplete performance assessment, with each measure capturing distinct aspects of detection capability while exhibiting specific weaknesses. Researchers increasingly report multiple complementary metrics to provide more transparent and comprehensive performance evaluation.

5.1.1. Binary classification metrics

Accuracy represents the most straightforward metric for AI text detection, measuring the proportion of correctly classified samples across both classes. While widely reported due to its intuitive interpretation, accuracy can be misleading when classes are imbalanced, which often occurs in real-world scenarios where the prevalence of AI-generated text may vary significantly [11].

Precision and Recall provide more nuanced insights into detector performance. Precision (the proportion of samples identified as machine-generated that are actually machine-generated) reflects the reliability of positive predictions, while recall (the proportion of actual machine-generated samples that are correctly identified) measures completeness. These metrics are particularly relevant in contexts with asymmetric costs for different types of errors [14]. For instance, educational applications might prioritize high precision to avoid falsely accusing students of using AI tools, while content moderation systems might emphasize recall to maximize the detection of potentially problematic AI-generated content.

F1 Score balances precision and recall through their harmonic mean, offering a single metric that considers both false positives and false negatives. This metric is widely reported in AI text detection literature as it provides a more balanced view than accuracy alone [16]. However, standard F1 scores still assume equal importance of precision and recall, which may not always align with application requirements.

Area Under the ROC Curve (AUC-ROC) measures a detector's ability to distinguish between classes across different threshold settings. This threshold-invariant metric is particularly valuable for comparing different detection approaches and assessing their discriminative power independent of specific decision thresholds [15]. AUC-ROC values approaching 1.0 indicate superior class separation, while values near 0.5 suggest performance equivalent to random guessing.

Area Under the Precision-Recall Curve (AUC-PR) complements AUC-ROC by focusing on performance at various precision–recall tradeoffs. This metric is especially relevant for imbalanced datasets where maintaining high precision at acceptable recall levels is crucial [36].

5.1.2. Domain-specific and advanced metrics

Detection Error Tradeoff (DET) Curves visualize the relationship between false negative and false positive rates on a log–log scale, providing a more detailed view of detector performance, particularly for high-security applications where very low error rates are required [12].

Cross-Model Generalization Score (CMGS), introduced by Guo et al. measures a detector's ability to identify text from generation models not represented in its training data [51]. This metric is calculated as the ratio of detection performance on unseen models to performance

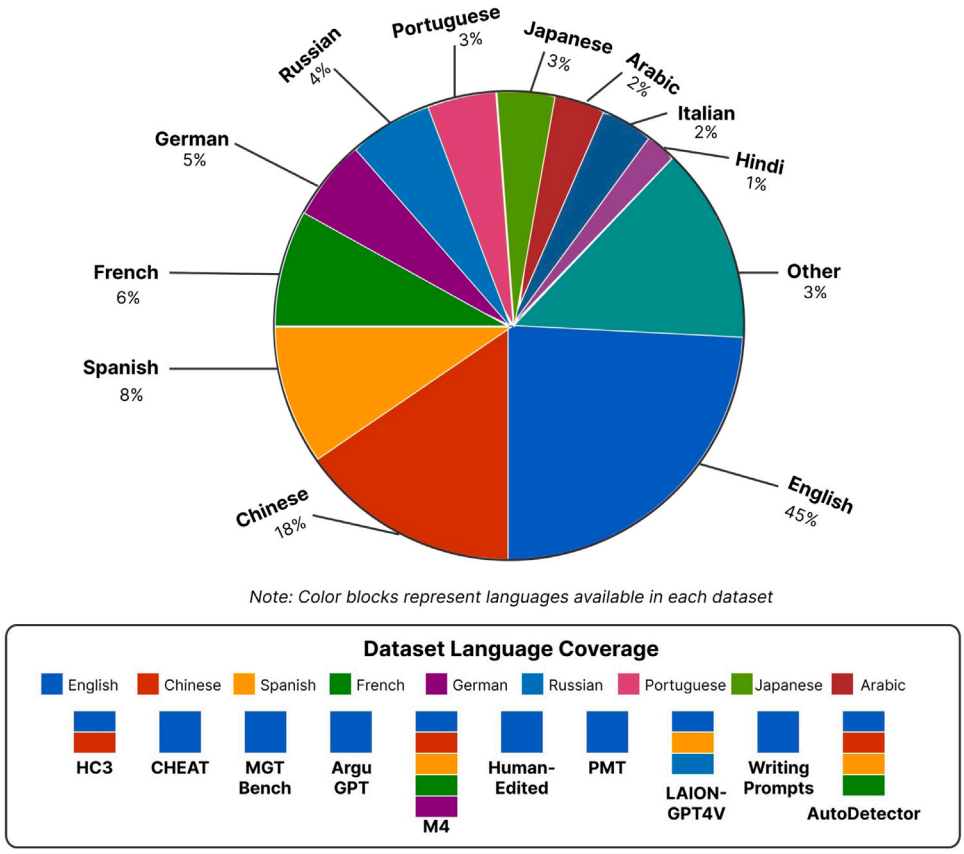

Fig. 6. Language diversity in AI detection benchmarks with English predominance.

on known models, with higher values indicating superior generalization capabilities.

Adversarial Robustness Index (ARI) quantifies a detector's resilience against various adversarial modifications, such as paraphrasing, human editing, or other evasion techniques [56]. ARI measures the relative drop in performance when evaluated on adversarially modified texts compared to unmodified samples.

Language Diversity Performance Gap (LDPG) assesses disparities in detector performance across different languages, highlighting potential biases in multilingual settings [14]. LDPG is typically measured as the standard deviation or maximum difference in performance metrics across languages, with lower values indicating more consistent cross-lingual performance.

Minimum Effective Text Length (METL) identifies the shortest text length at which a detector can achieve reliable performance, typically defined as exceeding a specified accuracy or F1 score threshold [64]. This metric is particularly relevant for applications involving short-form content, such as social media posts or short answers in educational assessments.

Fig. 6 visualizes the language distribution in multilingual AI text detection datasets, highlighting the significant imbalance toward English content across the field.

The visualization reveals critical patterns in how different evaluation metrics respond across dataset types. While all metrics show strong agreement on structured datasets like CHEAT (academic abstracts), they diverge significantly on more challenging scenarios. Notably, AUC-ROC consistently provides the most optimistic assessment across all datasets, potentially creating misleading impressions of detector robustness if used in isolation. The Cross-Model Generalization Score (CMGS) demonstrates unique sensitivity to dataset variation, performing distinctively worse than other metrics on HC3 but showing greater stability on adversarially modified texts. Most concerning is the universal performance collapse across all metrics when evaluating short text samples under 50 words—dropping to the 0.68-0.77 range regardless of metric choice. This visualization empirically demonstrates why relying on any single evaluation metric, particularly accuracy alone, provides an incomplete and potentially misleading assessment of detection capabilities across diverse real-world scenarios.

5.2. Limitations and opportunities

Despite the array of available metrics, several significant limitations remain in current evaluation approaches for AI text detection.

Overemphasis on Aggregate Performance represents a primary concern, as most studies report only average performance across test datasets without examining variations across different text types, domains, or generation parameters [63]. This approach can mask critical weaknesses that might emerge in specific scenarios, providing an incomplete picture of detector capabilities.

Limited Evaluation of Real-World Robustness constitutes another major gap. Many evaluations utilize clean, artificially constructed datasets where human and machine texts are clearly distinguished and cover similar topics and styles. Such idealized conditions rarely match real-world scenarios where machine-generated text may be edited, combined with human writing, or intentionally disguised [12].

Inadequate Representation of Diverse Text Types represents a persistent issue in current evaluation frameworks. Most studies focus predominantly on formal, academic, or otherwise structured writing, with insufficient attention to creative, conversational, or specialized domain texts [62]. This limitation is particularly problematic given that detection difficulty varies considerably across different writing styles and purposes.

Lack of Interpretability Metrics presents another critical gap. While current metrics quantify how well a detector performs, they provide little insight into the factors driving its decisions or failure cases. Metrics that assess a detector's ability to provide meaningful explanations for its classifications would substantially enhance their utility in practical applications [65].

These limitations create numerous opportunities for developing more comprehensive evaluation frameworks:

Context-Sensitive Evaluation that weighs errors based on their potential impact in specific application domains could provide more relevant performance assessments [14]. For instance, false positives in educational contexts might be weighted more heavily than in content moderation scenarios.

Progressive Difficulty Benchmarks could systematically evaluate detectors across increasingly challenging scenarios, from straightforward cases to adversarial examples with human editing or domain transfer [16]. Such graduated evaluation would provide a more nuanced understanding of detector capabilities and limitations.

Human-AI Collaborative Assessment metrics could evaluate how effectively detectors can assist human reviewers in identifying machine-generated content, rather than focusing solely on fully automated performance [15]. This approach acknowledges that many practical applications involve human-in-the-loop systems rather than fully automated detection.

5.3. Challenges and recommendations

Despite significant advances in detection methods, several persistent issues continue to limit their effectiveness.

5.3.1. Key challenges

Rapid Evolution of Generation Models represents perhaps the most significant evaluation challenge. Detection methods typically demonstrate degraded performance when applied to text from newer or different generation models than those they were developed for [11]. This moving target makes consistent longitudinal evaluation extremely difficult.

Ethical Constraints in Dataset Creation introduce additional complications. Collecting and annotating diverse, balanced datasets while respecting copyright, privacy, and informed consent requirements presents significant practical challenges [4]. These constraints can limit the diversity and representativeness of evaluation datasets.

Cross-Cultural and Linguistic Biases permeate current evaluation approaches. Studies have demonstrated that detectors often perform significantly worse on non-native English writing or texts from certain cultural contexts, raising serious concerns about fairness and potential discrimination [66].

Inconsistent Evaluation Protocols across research teams make direct comparison of different detection methods problematic. Variations in dataset composition, preprocessing steps, evaluation metrics, and reporting practices complicate meta-analysis and reliable progress tracking in the field [13].

5.3.2. Recommendations

Standardized Evaluation Frameworks with clearly defined protocols, metrics, and benchmark datasets would significantly enhance comparability across studies. Initiatives similar to GLUE or SuperGLUE for natural language understanding could provide the necessary infrastructure for consistent evaluation [16].

Multidimensional Performance Reporting should become standard practice, including not only aggregate metrics but also performance breakdowns by text length, domain, generation model, and adversarial conditions [63]. Such comprehensive reporting would provide a more complete picture of detector capabilities and limitations.

Regular Benchmark Updates are essential to keep pace with rapidly evolving generation capabilities. Evaluation frameworks should incorporate mechanisms for continuous addition of samples from new generation models and emerging adversarial techniques [61].

Inclusive Dataset Development practices should actively address linguistic and cultural biases by including diverse text sources and engaging researchers from varied backgrounds [4]. This approach would help ensure that detectors can serve diverse user populations fairly.

Human Factors Integration in evaluation would acknowledge the sociotechnical nature of AI text detection, considering how metrics relate to human judgment and real-world usage patterns [15]. Such human-centered evaluation would better align with practical applications of detection technology.

The implementation of these recommendations would substantially strengthen the evaluation of AI text detectors, providing more reliable, comprehensive, and meaningful assessments of their capabilities and limitations. As the field continues to evolve, our evaluation methodologies must similarly advance to ensure that detection technologies can effectively and fairly meet the growing challenges posed by increasingly sophisticated text generation models.

Fig. 7 illustrates the performance comparison of different evaluation metrics across primary AI text detection datasets. The graph illustrates how various metrics yield different assessments of detector performance, highlighting the importance of multi-metric evaluation.

6. Applications

The application landscape for AI text detection has expanded dramatically since the public release of powerful large language models like ChatGPT, Claude, and GPT-4. What once existed primarily as a niche research domain has rapidly transformed into a critical technology with implementations across numerous sectors. This transition reflects both the unprecedented capabilities of modern text generation systems and growing societal concerns about their potential misuse.

AI text detection now serves multiple crucial functions: maintaining integrity in educational and academic settings, preventing misinformation spread, enabling content moderation at scale, supporting forensic analysis, protecting intellectual property, and fostering healthy human-AI collaboration [13]. The practical importance of these applications has driven substantial investment in both commercial and open-source detection technologies, despite the significant technical challenges detailed in previous sections.

The deployment contexts for these technologies vary considerably in their requirements, priorities, and constraints. Each application domain presents unique considerations regarding acceptable error rates, interpretability needs, and the balance between false positives and false negatives [62]. For instance, educational applications might prioritize low false positive rates to avoid unfairly penalizing students, while content moderation systems might emphasize high recall to minimize the proliferation of synthetic misinformation.

As we examine these diverse applications, several cross-cutting themes emerge: the need for human-in-the-loop approaches, the importance of transparency in deployment, the challenge of maintaining effectiveness as generation technologies evolve, and the critical nature of fairness and equity considerations in system design and implementation [4].

Fig. 8 illustrates the visual representation of domain-specific AI text detection applications, showing key implementation features, integration points, and user interfaces across educational, publishing, and legal contexts.

The visualization captures the striking differences in how detection interfaces are implemented across sectors to meet domain-specific needs. Educational implementations like Turnitin's AI Writing Score emphasize interpretability through granular confidence levels and explanation options, reflecting the pedagogical imperative to provide teachable moments rather than binary judgments. In contrast, Reuters'

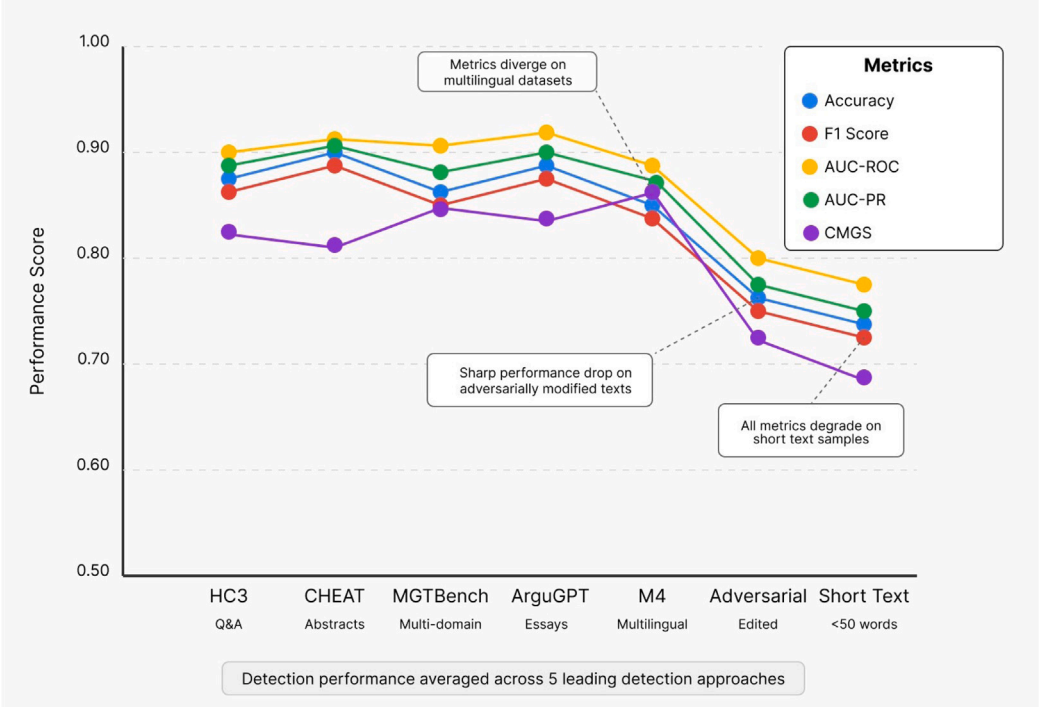

Fig. 7. AI detection struggles with adversarial and short texts.

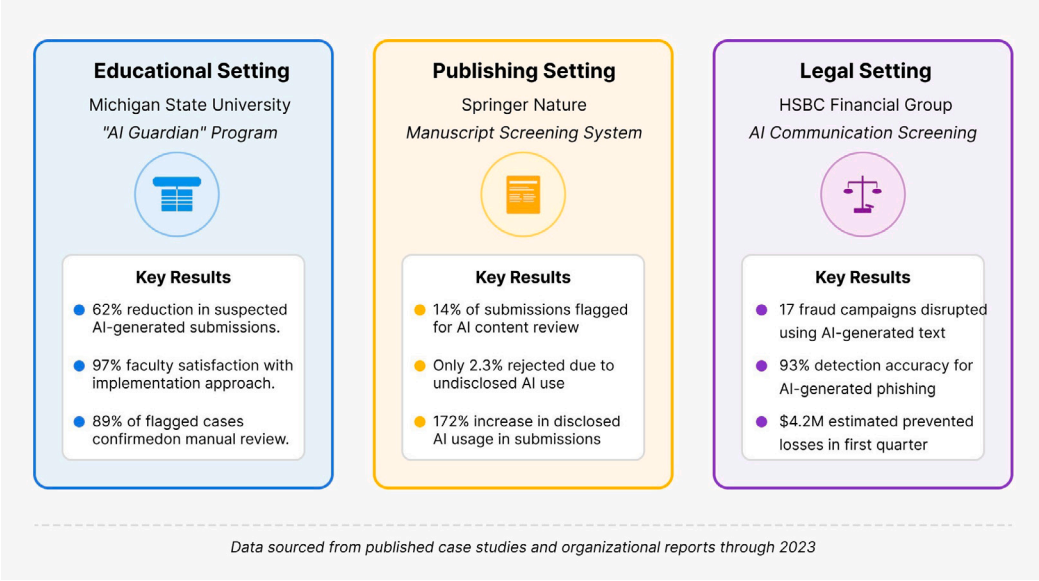

Fig. 8. Real-world case studies of AI text detection systems in educational, publishing, and legal contexts.

publishing interface highlights workflow integration with clear action prompts ("Flag for Review") and prominent probability indicators designed for quick editorial decision-making. Most distinctive is Everlaw's legal implementation, which embeds detection within a comprehensive document analysis framework — including tampering checks and chain of custody verification — illustrating how detection in legal contexts must satisfy rigorous evidence standards. These interfaces reveal more than mere design preferences; they embody fundamental differences in priorities across domains: educational systems emphasizing learning opportunities, publishing platforms prioritizing efficient content screening, and legal applications focusing on verification chains that could withstand courtroom scrutiny. The consistency of these patterns suggests detection technologies are increasingly evolving toward domain-optimized implementations rather than one-size-fits-all solutions.

AI text detection has found practical applications across various sectors, each addressing unique challenges in maintaining content authenticity:

i. *Educational Applications:* Educational institutions have emerged as early adopters of AI text detection technologies, driven by concerns about academic integrity in the wake of increasingly accessible and capable text generation tools. These applications range from standalone plagiarism detection systems to integrated learning management solutions.

ii. *Publishing and Media Applications:* Publishers, news organizations, and media platforms utilize AI text detection to maintain content quality, verify authorship, and combat misinformation. These applications span from pre-publication screening to post-publication analysis and monitoring.
iii. *Content Moderation Applications:* Social media platforms, online forums, and user-generated content sites employ AI text detection as part of broader content moderation strategies, helping to identify potentially problematic synthetic content at scale.
iv. *Legal and Forensic Applications:* Law enforcement, legal professionals, and digital forensics experts apply text detection technologies to identify potential fraud, investigate digital evidence, and authenticate document provenance.
v. *Research and Academic Integrity Applications:* Scientific publishers, research institutions, and funding agencies leverage detection technologies to maintain research integrity and ensure proper attribution in scholarly communications.
vi. *Commercial and Enterprise Applications:* Businesses across various sectors implement text detection to protect brand integrity, verify customer communications, authenticate reviews, and secure internal communications.

6.1. Detailed discussion of key applications

Fig. 9 illustrates the case studies highlighting successful implementations of AI text detection across different domains, including measured impact on user behavior and system performance.

These case studies reveal striking contrasts in how detection systems function across different sectors. In educational settings, Michigan State's "AI Guardian" program demonstrates that policy-focused implementation — combining detection technology with revised teaching approaches — can dramatically reduce AI misuse while maintaining high faculty satisfaction. The publishing domain shows a more nuanced picture, with Springer Nature's system identifying significant undisclosed AI usage (14% of submissions flagged) while simultaneously documenting a substantial increase in transparent AI assistance as authors adapt to new disclosure norms. Perhaps most compelling is HSBC's financial application, where detection technology translated directly to quantifiable business impact through fraud prevention ($4.2M in prevented losses) and exceptional accuracy rates for identifying AI-generated phishing attempts. These implementations illustrate a key pattern: detection technology delivers optimal value when tailored to domain-specific priorities — academic integrity in education, disclosure compliance in publishing, and fraud prevention in financial contexts — rather than pursuing generic identification goals

6.1.1. Publishing and media applications

Student assignment authentication. Educational institutions worldwide have rapidly adopted AI text detection to address concerns about academic integrity in the era of accessible generative AI. Turnitin, a leading provider of academic integrity solutions used by over 16,000 institutions globally, integrated AI text detection capabilities into its platform in April 2023. Their system analyzes linguistic patterns, consistency, and other markers to provide an AI Writing Score that helps educators identify potentially AI-generated content in student submissions [67].

A study conducted at Stanford University examined the implementation of AI text detection in undergraduate courses, finding that simply informing students about detection capabilities reduced AI-assisted cheating attempts by approximately 23% [68]. However, the study also highlighted the challenge of distinguishing between legitimate AI assistance and unauthorized use, suggesting the need for more nuanced policies and detection approaches.

Michigan State University has pioneered an alternative approach through their AI Guardian program, which combines detection technology with educational workshops and revised assignment structures. This program reported a 62% reduction in suspected AI-generated submissions while maintaining student satisfaction with course experiences [69]. Their multi-faceted approach demonstrates how detection technologies can be effectively integrated within broader educational frameworks.

Assessment design and adaptation. Beyond simply detecting AI-generated text, educational institutions are using these technologies to redesign assessment strategies. The University of California system has developed an "AI-Resistant Assignment Framework" that combines detection tools with assignments specifically designed to minimize the value of AI-generated responses [70].

This framework incorporates elements that current AI systems struggle with, such as specific personal experiences, cross-disciplinary integration, and in-class components. When combined with targeted detection, this approach addresses the root causes of academic dishonesty while accommodating legitimate AI use for learning support.

Feedback and learning support. Innovative applications are emerging that use AI text detection not merely as an enforcement tool but as a teaching opportunity. WriteLab, an educational technology platform, implements detection alongside instructional feedback that helps students understand how to properly cite AI assistance and develop their authentic voice [71].

Their approach targets identification of AI-generated passages within predominantly student-authored work, providing specific guidance on sections that may need revision or proper attribution. During a pilot study at five universities, this supportive approach resulted in a 78% reduction in unattributed AI use and improved overall writing quality scores [71].

News verification systems. Major news organizations have implemented AI text detection as part of their verification workflows to combat synthetic misinformation. The Associated Press deployed a system called "AP Verify" in 2023 that combines detection technology with traditional fact-checking approaches to identify potentially AI-generated news submissions [72]. This system has successfully flagged numerous instances of synthetic content designed to mimic legitimate news reports.

Similarly, Reuters has integrated detection capabilities into their "Reuters Fact Check" operation, using a combination of statistical analysis and journalistic expertise to investigate viral content suspected of being machine-generated [73]. Their approach emphasizes human judgment in conjunction with technological tools, recognizing the limitations of fully automated detection.

Scientific publishing screening. Springer Nature has integrated AI-driven tools to strengthen publication integrity. Their June 2024 implementation includes Geppetto and SnappShot, systems specifically developed to identify AI-generated content in submitted manuscripts. The Geppetto tool examines manuscript segments, generates probability scores for AI authorship, and highlights concerning sections for human evaluation. This forward-thinking strategy has successfully blocked multiple fraudulent papers from publication [74].

The Lancet Group has taken a similar approach but employs a combination of three different detection systems to minimize false positives, given the high stakes of wrongly accusing researchers of misconduct [75]. This multi-detector approach illustrates how high-consequence applications often require redundancy and conservative thresholds to maintain fairness.

Media content authentication. The New York Times joined Adobe and Twitter as founding partners of the Content Authenticity Initiative in November 2019. This collaboration developed Content Credentials, an open standard for provenance metadata designed to combat misinformation and verify digital content origins. While initially centered on image and video authentication, the framework has expanded

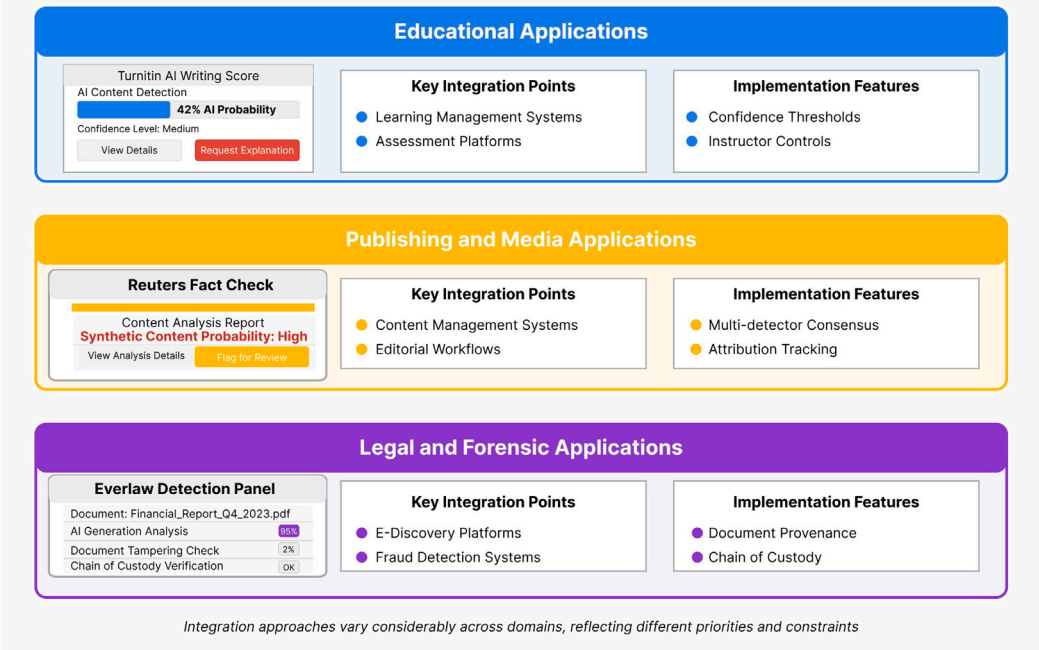

Fig. 9. Implementation examples of AI text detection across educational, publishing, and legal domains.

to explore text verification approaches that could help identify AI-generated content and maintain publishing integrity across multiple media formats [76].

This system has proven particularly valuable for user-generated content and eyewitness accounts, where the risk of synthetic manipulation is heightened. During its first six months of operation, the system identified over 300 instances of potentially AI-generated content submitted as authentic human accounts [76].

6.1.2. Content moderation applications

Online platforms increasingly rely on detection systems to manage the growing volume of synthetic content across their ecosystems:

i. *Social Media Screening:* Social media platforms have integrated AI text detection into their content moderation systems to identify potentially problematic synthetic content. Meta (formerly Facebook) implemented "Synthetic Content Detection" across its platforms in 2023, focusing particularly on political content, health information, and financial advice [77]. Their approach combines text detection with metadata analysis and user reporting to prioritize content for human review. Reddit has taken a different approach by leveraging its community structure, deploying detection tools that subreddit moderators can activate based on their community's specific needs and concerns [78]. This decentralized implementation recognizes the varying importance of AI content detection across different community contexts.

ii. *Comment and Review Filtering:* E-commerce platforms and review sites have adopted detection technologies to maintain the authenticity of user reviews. Amazon implemented an enhanced review verification system in 2023 that includes AI text detection as one component of a multi-faceted approach to combat fake reviews [79]. Their system targets both human-written and AI-generated fake reviews, focusing on linguistic inconsistencies and patterns common in synthetic text. Yelp has similarly integrated detection into their "Recommended" review filter, combining traditional signals like user history and engagement patterns with linguistic analysis to identify potentially machine-generated reviews [80]. Their approach emphasizes minimizing false positives to avoid unfairly excluding legitimate customer feedback.

6.1.3. Legal and forensic applications

Legal professionals increasingly turn to detection technologies to verify document authenticity and investigate potential fraud in digital communications:

i. *Legal Document Authentication:* Law firms and legal departments have adopted text detection to verify document authenticity and prevent fraud. The American Bar Association has published guidelines for implementing detection technologies in legal practice, emphasizing their role in due diligence and evidence authentication [81]. These guidelines acknowledge that while detection technology alone cannot prove or disprove authenticity, it provides valuable signals that complement traditional verification approaches. Notable implementations include the system developed by Everlaw, a legal technology provider, which incorporates AI text detection into its e-discovery platform to flag potentially synthetic documents for further review [82]. Their approach integrates detection results into the broader context of document metadata, provenance information, and legal relevance.

ii. *Fraud Investigation:* Financial institutions have integrated text detection capabilities into their fraud detection systems to identify potentially synthetic communications used in scams and social engineering attacks. HSBC implemented an "AI Communication Screening" system in 2023 to analyze suspicious emails and messages reported by customers [83]. This system has proven particularly effective at identifying sophisticated phishing attempts that use AI-generated text to mimic legitimate communications, detecting linguistic patterns that might escape human notice. During its first three months of operation, the system helped identify and disrupt 17 distinct fraud campaigns using AI-generated communications [83].

6.1.4. Research and academic integrity applications

Research institutions and funding bodies are implementing detection systems to ensure transparency and maintain integrity in scholarly communications:

i. *Grant Proposal Screening:* Research funding organizations have begun implementing AI text detection as part of their proposal

review process. The National Science Foundation piloted a detection system for grant proposals in late 2023, focusing on identifying potential instances of undisclosed AI use [84]. Importantly, their approach emphasizes transparency rather than prohibition, requiring disclosure of AI assistance while allowing its appropriate use. This pilot program has informed an emerging framework for balancing innovation with integrity, recognizing that AI tools can legitimately enhance research productivity when used ethically and transparently. The detection system serves primarily to ensure compliance with disclosure requirements rather than penalizing AI use itself.

ii. *Conference Submission Verification:* Academic conferences have increasingly adopted detection technologies to maintain submission integrity. The Neural Information Processing Systems (NeurIPS) conference, one of the premier venues for machine learning research, implemented a detection system for its 2023 submissions [85]. Their approach includes a tiered review system where papers flagged by automated detection undergo additional human review before any action is taken. This implementation demonstrates the importance of human oversight in high-stakes applications, with conference organizers emphasizing that detection results are treated as signals rather than definitive verdicts. The conference reported that approximately 2% of submissions were flagged for additional review, with the vast majority ultimately being accepted after human assessment [86].

6.1.5. Commercial and enterprise applications

Businesses across industries are adopting detection technologies to maintain authentic communications and comply with transparency requirements:

i. *Customer Service Authentication:* Businesses have implemented detection technologies to verify the authenticity of customer communications and support interactions. Zendesk, a customer service platform, integrated detection capabilities in 2023 to help businesses identify potentially AI-generated support tickets and inquiries [87]. Their system highlights unusual linguistic patterns or inconsistencies that might indicate synthetic content. This application addresses the emerging challenge of "AI-powered customer harassment", where individuals use generative AI to create large volumes of seemingly unique complaints or inquiries. The detection system helps customer service teams prioritize genuine customer needs while identifying potentially automated campaigns.

ii. *Marketing and Advertisement Compliance:* Major advertising platforms have integrated detection systems alongside updated policies to ensure transparency around AI-generated content. Google Ads took a significant regulatory step in September 2023, requiring election advertisers to explicitly disclose any synthetic elements in their campaign materials. This mandate aims to preserve authenticity in political messaging by making AI use immediately apparent to voters viewing such advertisements [88]. This implementation reflects growing regulatory attention to transparency in AI-generated commercial content. The system focuses particularly on industries with heightened disclosure requirements, such as financial services, healthcare, and political advertising. Google reported that in the first two months after implementation, approximately 4% of new advertisements were flagged for additional review due to potential undisclosed AI use [89].

6.2. Emerging applications

Several innovative applications move beyond conventional detection toward more nuanced integration of AI and human capabilities.

6.2.1. AI-human collaborative writing

A promising emerging application moves beyond binary classification (human vs. AI) toward more nuanced collaboration between human writers and AI systems. Microsoft has pioneered this approach with their Co-pilot Attribution system integrated into Microsoft 365 applications [90]. This system helps track which portions of a document received significant AI input, enabling proper attribution without stigmatizing AI assistance.

This collaborative approach marks a fundamental pivot from detection as an enforcement tool to detection as a mechanism for attribution and transparency. Microsoft 365 Copilot exemplifies this evolution by weaving AI assistance capabilities throughout its applications while clearly identifying AI contributions. Rather than simply policing content, the system emphasizes understanding AI's role in document creation. Initial user studies suggest that explicit attribution functionality builds greater trust in AI writing assistance while simultaneously enhancing the perceived credibility of documents containing AI elements. [91].

6.2.2. Source attribution and tracing

Beyond binary detection, emerging applications focus on tracing AI-generated content to specific models and training data. Stanford University's AI Provenance Project has developed techniques to identify not just whether text is AI-generated, but which specific model likely generated it [92]. This approach has significant implications for intellectual property disputes and attribution of harmful content.

The technology leverages subtle model fingerprints that persist in generated text, similar to how forensic linguistics can identify individual human authors. Early results demonstrate the ability to distinguish between text generated by GPT-4, Claude, LLaMA, and other large language models with approximately 87% accuracy [92].

6.2.3. Multimodal detection systems

As generative AI expands beyond text to encompass images, audio, and video, emerging applications are exploring holistic approaches to synthetic content detection. The Coalition for Content Provenance and Authenticity (C2PA) has developed standards and tools for detecting AI-generated content across modalities, including combined text-image compositions [93].

Their approach emphasizes cryptographic provenance records that travel with content, creating an auditable trail from creation through publication. This initiative, supported by major technology companies including Adobe, Microsoft, and Intel, represents a shift toward comprehensive trust infrastructure rather than post-hoc detection alone.

6.2.4. Continuous learning detection frameworks

To address the challenge of detection systems becoming obsolete as generation models evolve, emerging applications focus on continuous learning frameworks. OpenAI has released research on Adversarial Detection Frameworks that continuously update based on new generation capabilities [94]. This approach implements a red team/blue team methodology where detection systems and generation systems co-evolve through continuous competition.

Early implementations of this approach have demonstrated more robust performance against new generation models compared to static detection systems. However, significant challenges remain in maintaining this continuous learning without extensive computational resources, potentially exacerbating disparities between well-resourced and resource-constrained organizations.

6.2.5. Specialized domain detectors

Rather than pursuing general-purpose detection, an emerging trend focuses on highly specialized detection systems for specific domains with unique requirements. The MITRE Corporation has developed "Domain-Adaptive Detection" frameworks that leverage domain-spec-

ific linguistic patterns for more accurate detection in specialized contexts like medical documentation, legal contracts, or technical specifications [95].

These specialized systems achieve significantly higher accuracy within their target domains compared to general-purpose detectors, with some medical text detectors reaching 97% accuracy on clinical documents while maintaining false positive rates below 1% [95]. This domain specialization represents a promising direction for applications where general-purpose detection proves insufficient.

The applications of AI text detection technology continue to evolve rapidly, shaped by technical capabilities, ethical considerations, regulatory requirements, and user needs. While detection technology alone cannot address all challenges posed by increasingly powerful generative AI, thoughtfully designed and contextually appropriate implementations can support human decision-making, enhance transparency, and promote responsible use of these powerful new tools.

7. Challenges and limitations

Understanding the challenges and limitations of AI text detection is essential not only for advancing the field technically but also for setting realistic expectations about its capabilities and appropriate deployment contexts. While previous sections have documented promising approaches and applications, a critical examination of current limitations is necessary for researchers to focus their efforts, for practitioners to make informed implementation decisions, and for policymakers to develop appropriate guidelines and regulations [13].

The challenges facing AI text detection span multiple dimensions—technical, operational, ethical, and societal. These challenges are evolving rapidly as both generative and detection technologies advance, creating a dynamic landscape where today's solutions may become tomorrow's vulnerabilities [14]. This section aims to provide a comprehensive overview of these challenges, examining their implications for different stakeholders and identifying potential pathways toward their resolution.

What makes these challenges particularly significant is their deeply interconnected nature. Technical limitations influence ethical concerns, data issues impact operational deployment, and societal considerations shape research priorities. Understanding these interconnections is crucial for developing holistic approaches to advancing the field while minimizing potential harms [4].

Fig. 10 illustrates the concept map of key challenges facing AI text detection, highlighting interconnections between technical, ethical, data-related, and operational issues.. The concept map illustrates how challenges in AI text detection form an interconnected ecosystem rather than isolated problems. Technical challenges like the moving target problem directly influence data challenges through continual shifts in training data requirements.

Similarly, false positive issues—a technical limitation—feed directly into ethical concerns about bias against non-native speakers and impacts on creative expression. These cross-domain relationships explain why addressing challenges in isolation often proves ineffective. Particularly notable are the bidirectional connections between ethical and operational domains. Privacy concerns shape integration possibilities within existing systems, while resource disparities determine which organizations can implement bias mitigation techniques. The map reveals subtle relationships that might otherwise go unnoticed, such as how prompt contamination (a data challenge) amplifies the moving target problem (a technical challenge) by blurring distinctions between human and machine writing patterns over time.

These visualized interconnections underscore a critical insight: progress in AI text detection requires simultaneous advancement across multiple challenge domains. When researchers or implementers focus exclusively on technical improvements without addressing corresponding ethical or operational considerations, solutions often create new problems elsewhere in the ecosystem. This holistic perspective helps explain why technically impressive detection methods sometimes fail in practical deployment contexts.

7.1. General challenges in AI

AI text detection faces fundamental challenges inherent to adversarial machine learning contexts that complicate sustainable solutions.

7.1.1. The moving target problem

AI text detection faces a fundamental challenge common to adversarial AI domains: detection systems must continually evolve to keep pace with rapidly advancing generation technologies. This creates an asymmetric burden where detection methods must adapt to each new generation technique, while generators need only find a single weakness to evade detection [11]. This asymmetry is exacerbated by the fact that many organizations developing the most powerful generation models have substantially greater resources than those focused on detection.

The practical impact of this moving target problem became evident in 2023 when several leading detection systems, including OpenAI's classifier and Turnitin's initial solution, required significant revisions after their performance degraded with the introduction of newer language models [96]. Each new generation model introduces subtle changes in text patterns, necessitating retraining or redesigning detection approaches.

7.1.2. Computational demands

High-quality detection systems often require substantial computational resources, particularly for methods involving multiple model inferences or ensemble approaches. Zero-shot methods like DetectGPT typically demand more computation than simpler classifier-based approaches, creating a tradeoff between detection performance and computational efficiency [11].

This computational burden has significant implications for real-time applications, where low-latency detection is essential. Resource constraints also exacerbate disparities in access to detection technologies, as smaller organizations or educational institutions with limited infrastructure may be unable to deploy the most effective detection solutions [97].

7.1.3. Fundamental theoretical limitations

Some challenges stem from more fundamental theoretical limitations in distinguishing human from machine-generated text. As language models increasingly approximate human linguistic patterns through pre-training on human text, the decision boundary between human and machine-generated text becomes inherently blurrier [63]. This creates an upper bound on detection performance that may be insufficient for high-stakes applications requiring extremely high accuracy.

Recent theoretical work suggests that for certain generation models and sufficiently long text samples, perfect detection may be theoretically possible. However, for shorter texts or more advanced models with sophisticated sampling strategies, reliable detection faces fundamental mathematical constraints [38].

7.2. Domain-specific challenges

Fig. 11 illustrates the comparative impact of various challenges on detection performance across different domains and text types, showing relative degradation in detection accuracy.

The visualization quantifies how various challenges degrade detection performance across three critical application domains, revealing notable domain-specific vulnerabilities. Publishing and media applications suffer most severely from short text limitations, with accuracy declining by nearly 45% when analyzing brief content snippets—a particular concern for news headlines and social media monitoring. Educational contexts show more consistent degradation patterns but remain especially vulnerable to hybrid content scenarios where students combine AI-generated passages with their own writing. Perhaps most

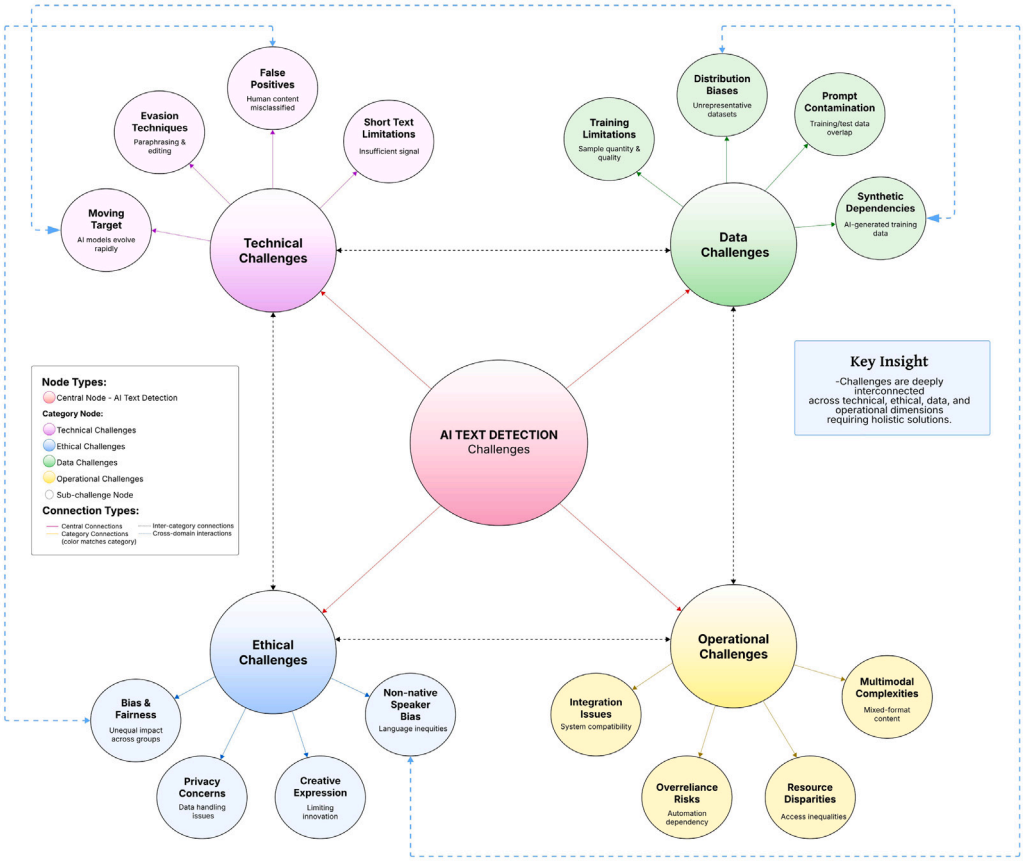

Fig. 10. Concept map showing key challenges in AI text detection systems and their interconnections.

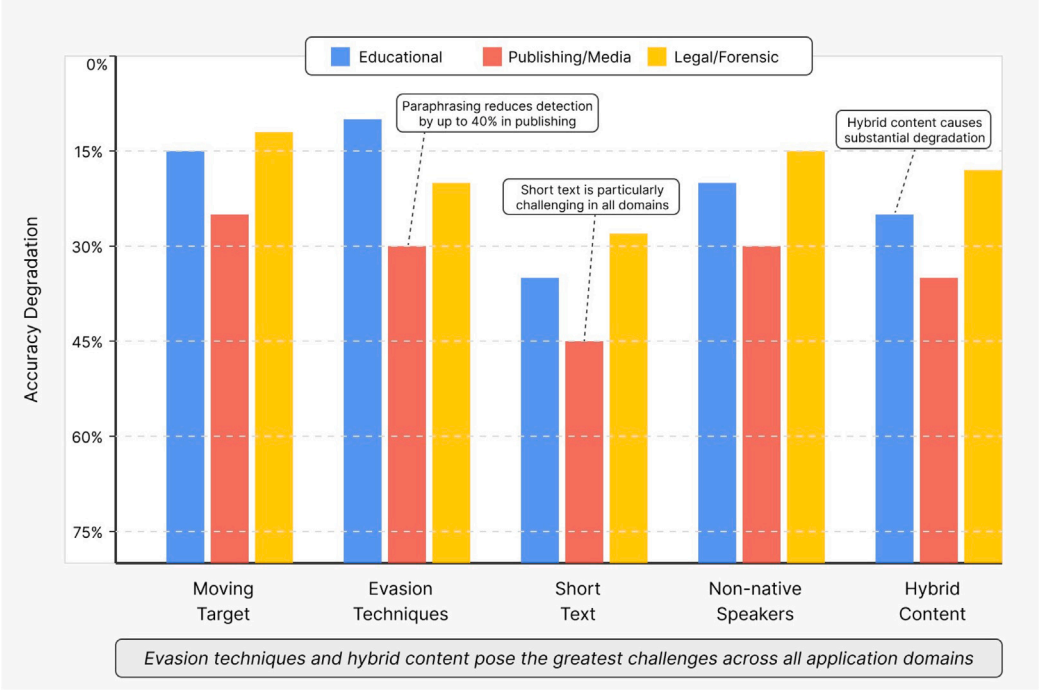

Fig. 11. Impact of detection challenges across educational, publishing, and legal domains.

striking is how evasion techniques, particularly paraphrasing, cause substantial accuracy drops of 30%–40% across all domains. Legal and forensic applications, while generally more robust to individual challenges, show concerning vulnerability to non-native speaker content, which may raise significant justice and equity concerns in international legal contexts. These quantified performance impacts underscore why technical advances in isolation often fail to address real-world implementation needs, as each domain faces a unique profile of critical vulnerabilities that must be specifically addressed.

7.2.1. The hybridization challenge

A significant real-world challenge for AI text detection is the prevalence of "hybrid" text—content that combines both human and machine-generated elements. As AI writing assistants become more integrated into composition workflows, purely binary classification (human or AI) becomes insufficient [15]. Users may generate initial text with AI, then extensively edit it, or they might use AI tools to expand or refine human-written content, creating blended texts that defy simple categorization.

This hybridization challenge is particularly acute in educational settings, where students might legitimately use AI tools for brainstorming or editing while producing primarily original work. Multiple studies have shown that current detection systems struggle with accurately classifying such hybrid content, often producing inconsistent or misleading results [68].

7.2.2. Cross-domain generalization

Detection systems trained on texts from specific domains (e.g., academic writing) often perform poorly when applied to different domains (e.g., creative writing or social media posts). This domain shift problem stems from the substantial variations in language patterns, vocabulary, and structure across different text types [36].

Educational implementations of detection technologies face this challenge acutely, as student assignments span diverse disciplines and genres. A detector trained primarily on argumentative essays may perform inadequately when applied to scientific reports, creative writing, or technical documentation. This necessitates either domain-specific detectors or more sophisticated approaches capable of cross-domain generalization [95].

7.2.3. Short text limitation

Most detection approaches perform significantly worse on short text samples (typically under 100 words), creating a substantial blind spot for applications involving brief communications [64]. This limitation is particularly problematic for social media monitoring, comment moderation, and assessment of short-answer questions in educational settings.

The fundamental cause of this limitation is that shorter texts provide fewer statistical patterns and linguistic features for detectors to analyze. While some specialized approaches for short text detection have been proposed, they typically achieve lower accuracy than methods applied to longer texts [40].

7.3. Technical limitations

Beyond general AI challenges, several specific technical limitations create persistent obstacles for detection systems.

7.3.1. False positive concerns

One of the most significant technical limitations across detection systems is the persistent issue of false positives—human-written text incorrectly flagged as machine-generated. False positives can have serious consequences, particularly in high-stakes settings like education or content moderation, where they may lead to unjust accusations or inappropriate content removal [63].

Recent studies have documented systematic patterns in false positive rates, with certain writing styles and non-native English writers being disproportionately affected. For instance, research by Du et al. found that texts written by non-native English speakers were up to 2.5 times more likely to be incorrectly classified as machine-generated by widely used detection systems [98].

7.3.2. Evasion techniques

The development of increasingly sophisticated AI text detection systems has catalyzed a parallel evolution in methods designed to circumvent these technologies, creating substantial challenges for reliable detection implementation. Understanding these evasion strategies proves crucial for developing robust detection frameworks and anticipating potential vulnerabilities in deployed systems.

Numerous techniques have been developed to deliberately evade AI text detection, creating significant challenges for reliable implementation. These include text transformation methods like paraphrasing, word replacement, intentional grammatical errors, or character substitutions [12]. More sophisticated techniques leverage the specific detection mechanisms employed, essentially exploiting knowledge of how detectors work to generate text specifically designed to bypass them.

The proliferation of detection-evasion methods has grown dramatically, with various tools and detailed guides now readily accessible online. Krishna revealed how even basic paraphrasing strategies can substantially undermine detection capabilities across several prominent systems, allowing machine-generated content to bypass identification mechanisms that would otherwise flag it [12].

Automated Transformation Tools represent one of the most accessible evasion vectors currently available. Commercial services such as QuillBot and similar paraphrasing platforms can systematically rephrase AI-generated content while preserving semantic meaning, often reducing detection accuracy by 40%–60% across multiple detection systems [63]. Lexical substitution attacks involve replacing words with synonyms or contextually appropriate alternatives, with research demonstrating that strategic replacement of just 10%–15% of words can reduce detection accuracy from 85% to below 50% for several prominent systems [99].

Human-AI Collaboration Techniques present sophisticated evasion approaches where authors systematically modify AI-generated content through iterative editing cycles. This method exploits the observation that even minimal human intervention can significantly alter the statistical signatures that detection systems rely upon [11]. Mixed authorship strategies involve strategic interleaving of human-written and AI-generated segments, creating hybrid texts that challenge binary classification approaches and capitalize on the difficulty many detection systems face when analyzing documents containing both human and machine contributions [63].

Advanced Linguistic Manipulation encompasses techniques that modify syntactic structures, stylistic patterns, and grammatical constructions while maintaining semantic equivalence. These approaches target detection systems that analyze specific linguistic patterns characteristic of different generation models. Experimental studies have shown that systematic modifications such as voice conversion, clause reordering, and stylistic variation injection can significantly degrade detector performance by obscuring the uniform signatures often exhibited by machine-generated text [100].

Countermeasures and Defense Strategies have evolved in response to these evasion challenges. Adversarial training approaches incorporate known evasion methods during detector development, while ensemble defense strategies combine multiple detection approaches with different vulnerability profiles [11]. Continuous learning frameworks represent an emerging response that automatically adapts to new evasion techniques, though significant challenges remain in balancing adaptability with computational efficiency [63].

The evasion landscape continues to evolve rapidly, driven by the inherent tension between increasingly capable detection systems and equally sophisticated circumvention techniques. This dynamic creates substantial challenges for practical deployment of detection technologies, particularly in high-stakes applications where evasion attempts are likely to be frequent and technically sophisticated.

7.3.3. Interpretability deficits

Many high-performing detection approaches, particularly those based on deep learning, lack transparency in their decision-making processes. This "black box" nature creates challenges for understanding detection errors, improving system performance, and providing meaningful explanations to users [65].

The interpretability deficit is particularly problematic in educational and legal contexts where stakeholders need to understand the basis for detection claims. While some approaches like GLTR offer visual explanations, they often sacrifice detection accuracy for interpretability, creating a challenging tradeoff between performance and explainability [15].

7.4. Data-related issues

Data quality and availability create unique challenges that directly impact detection effectiveness across applications.

7.4.1. Training data limitations

Detection systems require diverse, representative training data encompassing both human-written and machine-generated texts. Obtaining high-quality, properly labeled data remains challenging, particularly for newer generation models where limited examples may be available [16]. This creates a bootstrapping problem where effective detection requires data that may not be readily accessible.

The rapid evolution of generation models exacerbates this challenge, as training data representing older models may not adequately capture the characteristics of text from newer models. This necessitates continuous data collection and annotation efforts, which are resource-intensive and difficult to scale [60], with additional complexity arising in multilingual scenarios that demand robust preprocessing pipelines and effective stemming algorithms [101,102].

7.4.2. Data distribution biases

Training datasets for detection systems often contain subtle biases in content, style, or topic distribution that can lead to biased detection outcomes. For instance, if training examples of human-written text come predominantly from professional writers while machine-generated examples come from generic prompts, the detector may learn to discriminate based on writing skill rather than genuine human versus machine differences [51].

These distribution biases can manifest in systematic detection errors for certain content types, potentially disadvantaging specific user groups or content domains. Studies have shown that current detection systems often perform worse on texts addressing non-Western topics or using culturally specific references, raising serious equity concerns [98].

7.4.3. Prompt contamination

A growing challenge for detection systems is "prompt contamination"
—the inclusion of human-prompted machine-generated text in the pretraining data of newer language models. As models are increasingly trained on internet data that includes outputs from earlier models, the statistical boundary between human and machine text becomes blurrier [97].

This recursive contamination may fundamentally alter the characteristics that detection systems rely on, potentially diminishing their effectiveness over time. This challenge is particularly concerning as it suggests a potential future where reliable detection becomes increasingly difficult regardless of methodological advances.

7.5. Ethical and societal concerns

Detection technologies raise significant ethical questions that extend beyond technical performance metrics.

7.5.1. Bias and fairness issues

Significant evidence indicates that current detection systems exhibit biases against non-native English writers, certain cultural expressions, and particular writing styles. These biases create serious fairness concerns, particularly in educational and professional contexts where detection results may impact opportunities or evaluations [98].

Research by Weber et al. documented substantial disparity in false positive rates across languages, with texts in languages other than English being misclassified at rates up to four times higher than English texts [14]. These disparities raise serious questions about the equitable application of detection technologies in multilingual or multicultural contexts.

7.5.2. Surveillance and privacy concerns

The widespread deployment of text detection systems raises concerns about surveillance and privacy, particularly in educational settings where student work is routinely analyzed without explicit consent. These concerns are amplified when detection systems retain or utilize submitted text for model improvement or other purposes beyond the immediate detection task [4].

Privacy frameworks and regulations like GDPR and FERPA create complex compliance challenges for detection system deployment, particularly for global implementations spanning multiple jurisdictional contexts. Finding the appropriate balance between detection effectiveness and privacy protection remains an ongoing challenge for the field.

7.5.3. Impact on creative expression

Concerns have emerged about how widespread detection deployment might impact creative expression and writing processes. Writers aware that their text will be subjected to algorithmic analysis may consciously or unconsciously alter their writing to avoid false positive detection, potentially stifling creativity or reinforcing conventional writing patterns [103].

This "chilling effect" on creative expression is particularly concerning in educational contexts, where developing authentic voice and experimental writing approaches are important learning objectives. Balancing the integrity benefits of detection with the potential negative impacts on creative expression presents a significant challenge for institutional implementation.

7.6. Operational and deployment issues

Practical implementation considerations often determine detection success regardless of underlying algorithm quality.

7.6.1. Integration challenges

Organizations implementing detection systems face substantial integration challenges with existing workflows and technologies. Educational institutions must navigate complex integrations with learning management systems, plagiarism detection tools, and assessment platforms. These technical challenges often require significant resources and expertise that may be unavailable to smaller organizations [69].

Effective integration also demands careful consideration of user experience design to avoid creating friction or confusion in existing workflows. Poor integration can lead to low adoption rates or improper use of detection capabilities, undermining their potential benefits.

7.6.2. Overreliance risks

A significant operational risk in detection deployment is overreliance
—treating detection results as definitive rather than probabilistic signals requiring human judgment. This risk is exacerbated by the tendency to view technological solutions as more objective or reliable than human assessment, despite their documented limitations [86].

Educational implementations face particular challenges in this regard, as time-constrained educators may default to accepting detection outcomes rather than exercising critical judgment. Establishing appropriate guidelines and training for integrating detection results into human decision processes remains an ongoing challenge.

7.6.3. Resource disparities

The resources required for effective detection — including technical infrastructure, expertise, and ongoing maintenance — are not equally distributed across organizations. This creates significant disparities in detection capabilities between well-resourced institutions and those with more limited means, potentially exacerbating existing inequities [84].

These resource disparities are particularly concerning in educational contexts, where wealthy institutions may implement sophisticated detection approaches while less-resourced schools rely on less effective or more error-prone methods. Addressing these disparities requires consideration of sustainable, accessible detection approaches that do not widen the digital divide.

7.7. Emerging challenges

Several emerging trends in AI generation create new detection challenges that current approaches are unprepared to address.

7.7.1. Multimodal generation

As generation technologies expand beyond text to encompass multimodal outputs combining text with images, audio, or video, detection systems face increasing complexity. Current text-specific approaches are ill-equipped to handle these multimodal artifacts, creating potential blind spots as generated content evolves [93].

The challenge becomes even more significant as multimodal models like GPT-4V can generate text based on visual inputs or create text-image combinations where the text directly references the image content. These capabilities create new avenues for evading text-only detection approaches.

7.7.2. Smaller, more accessible models

While much detection research focuses on content from large, commercial models like ChatGPT or Claude, the proliferation of smaller, locally-runnable open-source models presents distinct challenges. These models, which can run on consumer hardware without API access, may generate text with different characteristics than their larger counterparts [92].

The diversity of these models, combined with their rapid development and customization capabilities, creates a fragmented detection landscape where approaches optimized for major commercial models may perform poorly. Developing detection approaches that generalize across this varied ecosystem represents a significant emerging challenge.

7.7.3. Synthetic data dependencies

An ironic emerging challenge is the growing dependency on synthetic data for training both generation and detection systems. As human-labeled datasets become insufficient for the scale of modern AI systems, synthetic data generation offers an alternative—but one that potentially creates recursive loops between generation and detection [94].

This synthetic data dependency raises fundamental questions about whether detection systems trained partly on synthetic data can reliably distinguish between human and machine-generated content, particularly as the boundaries between these categories become increasingly blurred.

The challenges and limitations outlined in this section reveal the complex, multifaceted nature of AI text detection. While significant technical progress continues, these systems exist within broader sociotechnical contexts that introduce additional constraints and considerations. Addressing these challenges requires not only technical innovation but also careful attention to ethical implications, operational realities, and societal impacts.

8. Emerging trends and future directions

The field of AI-generated text detection stands at an inflection point, where rapid advancements in generation capabilities are matched by innovative detection approaches emerging across technical, practical, and ethical dimensions. This section examines these current developments alongside promising future research directions, mapping the trajectory of a discipline responding to increasingly sophisticated challenges.

8.1. Emerging trends

Fig. 12 illustrates the Evolution of research focus in AI text detection, showing the shifting emphasis from binary classification approaches to more nuanced detection paradigms across three distinct phases of development (2019–2024).

The visualization reveals a clear evolutionary trajectory in AI text detection research across three distinct phases. During Phase 1 (2019–2020), research predominantly focused on binary classification approaches
—distinguishing human from machine-generated text through straightforward classification methods. This early period also saw modest attention to linguistic feature analysis and initial exploration of zero-shot methods, reflecting the field's nascent understanding of detection challenges.

Phase 2 (2021–2022) marks an important transition as researchers responded to more sophisticated generation capabilities in the GPT-3 era. While binary classification remained significant, we observe a substantial increase in adversarial robustness research—a direct response to emerging evasion techniques. This middle period also witnessed growing interest in zero-shot methods and the emergence of multimodal detection approaches, signaling the field's recognition that purely text-based analysis might prove insufficient as generation technologies evolved.

Most revealing is the dramatic shift in Phase 3 (2023–2024), where ethical considerations and human-in-loop approaches have surged to prominence, collectively surpassing binary classification research volume for the first time. This recent emphasis on fairness, transparency, and human collaboration reflects growing awareness of bias issues and false positive concerns documented by researchers like Du et al. and Weber et al. Simultaneously, provenance and watermarking research has emerged as a significant new direction, offering complementary approaches to purely statistical detection methods.

The landscape of AI-generated text detection is evolving rapidly, shaped by both advancements in generation technologies and innovations in detection methodologies. This section examines key emerging trends that are transforming the field, with significant implications for future research, implementation, and policy considerations.

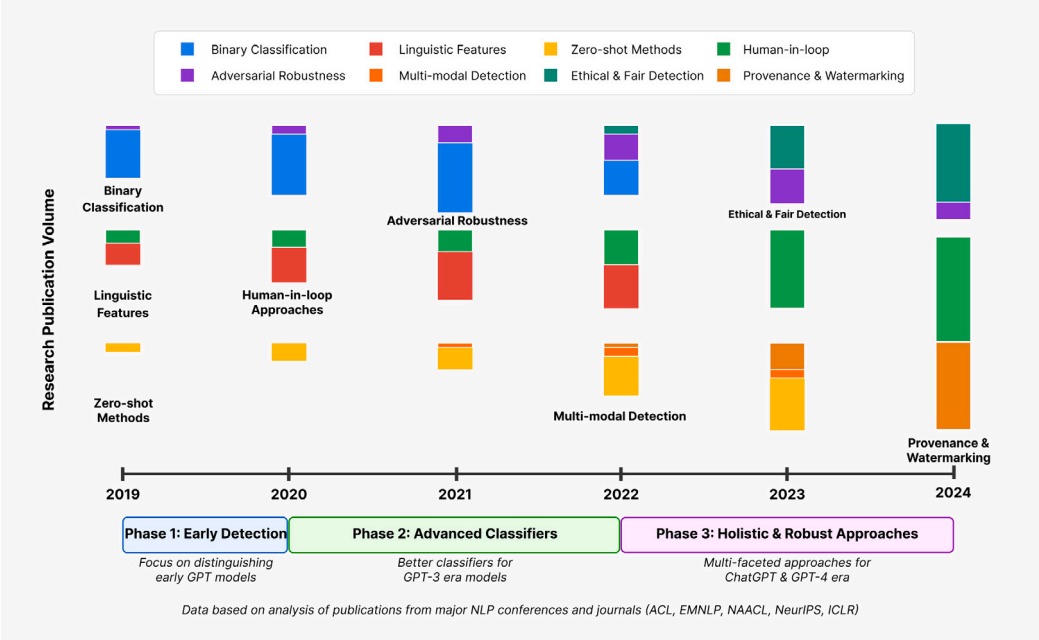

Fig. 12. Timeline of research trends in AI text detection from 2019 to 2024.

8.1.1. Algorithmic innovations

Multi-scale Analysis approaches are gaining prominence in text detection, analyzing content at multiple granularities simultaneously—character, word, sentence, and document levels. This technique, pioneered by researchers at Carnegie Mellon University, has demonstrated improved ability to detect sophisticated AI-generated text by identifying inconsistencies that might only be apparent at specific linguistic levels [104]. Their multi-scale transformer architecture achieved a 7% improvement in detection accuracy compared to single-scale approaches when tested against the latest language models.

Self-supervised Contrastive Learning represents another promising algorithmic direction, enabling detectors to learn distinctions between human and machine-generated text without requiring extensive labeled datasets. Fu et al. developed a contrastive framework that creates positive and negative pairs from unlabeled texts by applying controlled perturbations, achieving strong generalization to unseen models [47]. This approach particularly excels at generalizing to new generation models, with only a 9% performance drop when evaluating text from models not represented in training data, compared to the 22%–35% drops observed in supervised approaches.

Graph-based Methods for analyzing text structure and coherence patterns are emerging as powerful tools for detection. These approaches model text as graphs where nodes represent semantic concepts and edges represent relationships between them, capturing deeper structural patterns that differentiate human from machine writing [105]. Graph neural networks trained on these representations have shown particular promise in identifying AI-generated scientific and technical content, where logical structure is especially important.

Neural Architecture Search (NAS) for detection models is gaining traction as researchers seek optimal architectures specifically tailored to the detection task. Wang et al. employed automated NAS to discover specialized architectures that outperform manual designs by adapting to the specific statistical properties of machine-generated text [103]. Their approach demonstrated a 5%–8% improvement over manually designed architectures across multiple datasets.

8.1.2. Applications expansion

Cross-lingual Detection frameworks are addressing the critical need for detection capabilities beyond English. wang developed a multilingual detection approach that employs language-agnostic features and cross-lingual transfer learning to achieve consistent performance across 10 languages with minimal language-specific training data [55]. This work represents an important step toward more inclusive detection technologies that serve global user populations.

Domain-specialized Detection systems are emerging for high-stakes fields with unique linguistic patterns. In the medical domain, researchers have developed detection models specifically trained on clinical narratives and medical literature, achieving substantially higher accuracy compared to general-purpose detectors when evaluating AI-generated medical content. Similar specialized approaches are appearing in legal and financial sectors, where field-specific terminology and writing conventions present unique detection challenges.

Multimodal Content Analysis is expanding detection capabilities to encompass text generated in conjunction with other modalities such as images or audio. The LAION research group has pioneered detection methods for identifying AI-generated image captions and image-prompted text, addressing the increasingly multimodal nature of generated content [57]. Their approach combines textual features with image-text alignment signals to identify AI-generated content in multimodal contexts with 89% accuracy.

Real-time Detection Systems are being developed for streaming contexts like live chat, social media, and collaborative workspaces. These approaches emphasize computational efficiency and incremental analysis to provide detection capabilities with minimal latency. Microsoft Research has demonstrated a streaming text detection system capable of analyzing text with only a 200 ms delay while maintaining 85% of the accuracy of non-streaming approaches [90].

8.1.3. Sustainable AI detection

Lightweight Detection Models are addressing the computational demands of detection through knowledge distillation and model compression techniques. Wei et al. developed a distilled detector model with only 7% of the parameters of its teacher model while retaining 95% of its detection performance [106]. Such approaches are particularly important for making detection technologies accessible on resource-constrained devices and in settings with limited computational infrastructure.

Parameter-efficient Fine-tuning methods like LoRA (Low-Rank Adaptation) and prompt tuning are enabling the adaptation of large pre-trained models for detection with minimal additional parameters. These

approaches reduce the computational and storage costs of detection while maintaining high performance. Zhang et al. demonstrated that LoRA fine-tuning for detection requires only 0.1% of the parameters needed for full fine-tuning while achieving comparable performance [107].

Energy-efficient Inference techniques are gaining traction as deployment scales increase. Approaches like dynamic computation, early exiting for confident predictions, and mixed-precision inference are being applied to detection models to reduce their energy footprint without compromising effectiveness. Tang et al. demonstrated a 72% reduction in energy consumption through dynamic sparsity and mixed precision techniques while maintaining detection accuracy within 2% of the full-precision model [108].

8.1.4. Ethical AI detection

Fairness-aware Detection approaches are emerging to address documented biases in current systems. Researchers at MIT have developed detection methods that explicitly optimize for consistent performance across different writer demographics and language varieties [98]. Their approach incorporates fairness constraints during training and evaluation, reducing false positive disparities between native and non-native English writers from 2.4x to 1.2x.

Explainable Detection Systems are gaining prominence as transparency becomes increasingly important for high-stakes applications. Mitrović et al. developed a detection approach that not only classifies text but also highlights specific features and patterns that influenced the classification decision [65]. This transparency enables more informed human judgment about detection results and promotes trust in automated assessments.

Privacy-preserving Detection methods are addressing concerns about surveillance and data privacy. Federated learning approaches enable distributed training of detection models without sharing sensitive text data, while differential privacy techniques add noise to training processes to protect individual privacy. Princeton researchers demonstrated a federated detection approach that achieved 92% of the performance of centralized training while providing stronger privacy guarantees [109].

8.1.5. Integration with other technologies

Blockchain-based Content Authentication systems are complementing detection approaches by creating verifiable records of content provenance. The Coalition for Content Provenance and Authenticity (C2PA) has developed standards for cryptographically signing and verifying content origins, allowing recipients to confirm whether content was human-created or AI-generated [93]. These approaches provide an important additional layer beyond statistical detection.

LLM-powered Detection represents an intriguing circular development, where large language models themselves are being employed to identify AI-generated content. OpenAI's research on using GPT-4 to evaluate outputs from other models has shown promising results, with GPT-4 achieving detection performance comparable to specialized detectors across diverse content types [94]. This approach leverages the sophisticated pattern recognition capabilities of advanced LLMs for meta-analysis.

API Integration Frameworks are standardizing how detection capabilities can be incorporated into diverse applications and workflows. The LangChain project has developed interoperable components for connecting detection services with content creation platforms, educational systems, and content moderation pipelines [110]. These standardized interfaces reduce implementation barriers and promote wider adoption of detection capabilities.

8.2. Future directions

Fig. 13 illustrates the Impact assessment of future research directions in AI text detection, comparing projected importance, implementation difficulty, and expected timeline across key development areas.

This matrix visualization organizes research directions into four strategic quadrants based on potential impact and implementation difficulty. The "Quick Wins" quadrant highlights approaches like cross-lingual detection and human-AI collaboration that offer substantial benefits with relatively manageable implementation challenges. These areas present immediate opportunities for meaningful advancement within 1-3 year timeframes.

In contrast, the "Major Projects" quadrant identifies high-impact directions that face significant implementation hurdles, including watermarking and multimodal detection. While these approaches require greater resource investment and technical sophistication, they potentially offer transformative capabilities for detection systems over 2–3 year development cycles. The varying bubble sizes across the visualization reflect the current research activity intensity within each area, with larger bubbles indicating more active investigation.

Notably, several approaches occupy intermediate positions that defy simple categorization. For instance, domain-specific detection presents moderate implementation difficulty but varied impact potential depending on application context. The diagram also illustrates resource requirement trends, with higher-impact technologies generally demanding more substantial computational resources – a pattern that reinforces concerns about accessibility disparities discussed earlier in this review.

8.2.1. Scalability and efficiency

Detection-as-a-Service (DaaS) frameworks represent a promising direction for making detection capabilities more accessible and scalable. Cloud-based detection services with standardized APIs could dramatically reduce the technical barriers to implementing detection, enabling broader adoption across application domains. Future research should focus on developing flexible, efficient service architectures that support customization while maintaining performance at scale [109].

Hardware-optimized Detection approaches designed specifically for edge devices and specialized AI accelerators will be crucial for expanding the reach of detection technologies. Research is needed on model architectures and inference strategies that exploit specialized hardware features like neural processing units and domain-specific accelerators. The development of hardware-specific optimizations could enable real-time detection even on resource-constrained devices [106].

Continual Learning Systems capable of adapting to new generation models and techniques without complete retraining will be essential for maintaining effectiveness in a rapidly evolving landscape. Initial work by Pu et al. on meta-learning approaches for detection shows promise, but substantial research is still needed on efficient adaptation mechanisms, knowledge preservation, and update strategies [111].

8.2.2. Data-centric AI detection

Synthetic Data Generation for training detection systems represents an important future direction as human-labeled examples become insufficient for the scale and diversity required. Approaches that systematically generate challenging test cases by manipulating generation parameters, combining models, or applying structured transformations could substantially improve detector robustness. Early work by Wang et al. demonstrates the potential of this approach, but considerable research is needed to ensure that synthetic training data adequately represents real-world generation patterns [103].

Cross-domain Annotation Standards would enable more effective dataset sharing and comparison across research groups and application domains. The development of standardized formats for encoding provenance information, model parameters, and generation contexts would

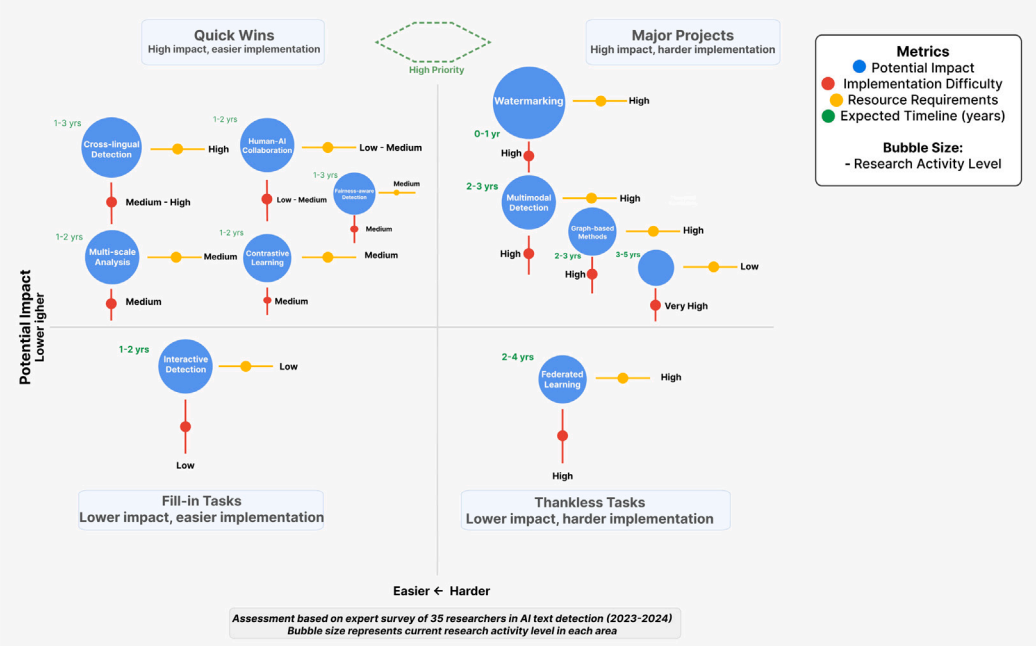

Fig. 13. Impact assessment matrix of future research directions in AI text detection.

facilitate more rapid progress in detection research. Future efforts should focus on creating flexible, extensible annotation schemas that capture the diverse attributes relevant to detection [16].

Active Learning Frameworks that strategically select the most informative examples for human annotation could dramatically improve data efficiency for detection model training. Research in this direction should explore uncertainty-based sampling, diversity-maximizing selection, and adversarial example mining to reduce annotation requirements while maintaining detection performance [60].

8.2.3. Human-AI collaboration

Interactive Detection Systems that combine algorithmic analysis with human judgment represent a promising direction for high-stakes applications. Future research should explore effective interfaces for presenting detection evidence, eliciting human input, and integrating multiple sources of information. Work by Gehrmann et al. on visual analytics for text evaluation provides an initial framework, but substantial research is needed on workflow integration and decision support features [15].

Detection-aware Content Creation tools that provide real-time feedback during the composition process could help authors maintain transparency about AI contributions. Research is needed on non-disruptive interfaces for attribution, stylistic analysis, and content verification that support ethical AI use rather than simply policing it. Microsoft's work on attribution in collaborative editing environments offers initial insights, but more comprehensive approaches are needed [90].

Cultural and Contextual Adaptation of detection systems to accommodate diverse writing styles, linguistic expressions, and creative practices represents an important future challenge. Research is needed on methods that can distinguish between machine-generated content and culturally distinct human writing patterns without reinforcing biases against non-dominant styles. Work by Du et al. highlights these challenges but only begins to address the necessary adaptations [98].

8.2.4. Generalization and robustness

Transfer Learning Techniques specifically designed for cross-model and cross-domain generalization will be critical for developing detection systems that remain effective as generation technologies evolve. Research should explore representation learning approaches that capture fundamental distinguishing features rather than model-specific artifacts. Initial work by Sadasivan et al. highlights the challenge, but substantial research is needed on transferable feature extraction and domain-adaptive architectures [63].

Adversarial Training Frameworks that systematically incorporate evasion attempts during the development process could substantially improve detector robustness. Future research should explore automated red-teaming approaches, adversarial example generation, and defensive distillation techniques to create more resilient detection systems. Work by Krishna et al. demonstrates the vulnerability of current approaches to adversarial modifications, but comprehensive defense strategies remain an open research area [12].

Theoretical Foundations for detection represent an important long-term research direction. Deeper understanding of the fundamental statistical and information-theoretic properties that distinguish human from machine-generated text could guide the development of more principled detection approaches. Mitchell et al. have begun exploring these theoretical aspects through probability curvature analysis, but much remains to be discovered about the mathematical foundations of detection [11].

8.2.5. Regulatory and policy frameworks

Detection Standards and Benchmarks would enable more systematic evaluation and comparison of detection technologies. Future efforts should focus on developing comprehensive benchmark suites that reflect diverse generation models, content types, and evasion scenarios. The MGTBench initiative by He et al. represents an important step in this direction, but more extensive and regularly updated benchmarks are needed [16].

Cross-border Detection Governance frameworks will become increasingly important as detection technologies are deployed globally. Research is needed on approaches that balance cultural and legal differences while maintaining consistent protection against harmful applications of text generation. Early discussions of these issues appear in AI ethics literature, but concrete governance mechanisms for detection technologies remain underdeveloped [4].

Educational Policy Development for text detection in academic contexts represents a critical need as educational institutions grapple with the implications of generative AI. Research should explore effective policies that maintain academic integrity while accommodating legitimate AI use as a learning tool. Work by Yang et al. examines detection

impact on student behavior, but comprehensive policy frameworks are still emerging [68].

8.3. Barriers to adoption of trends

Several significant barriers could hinder the adoption and effectiveness of these emerging trends and future directions:

The Moving Target Problem remains perhaps the most fundamental barrier to effective detection. As generation models continue to advance, detection systems face a perpetual adaptation challenge. This asymmetric advantage for generation creates substantial uncertainty about the long-term viability of purely statistical detection approaches [11].

Resource Disparities between well-funded technology companies developing generation models and the broader ecosystem of detection researchers and practitioners create an uneven playing field. Many promising approaches require substantial computational resources or specialized expertise that may be inaccessible to important stakeholders like educational institutions or small content platforms [84].

Cross-disciplinary Collaboration Gaps between NLP researchers, human–computer interaction experts, ethicists, and domain specialists impede the development of holistic detection approaches. Future progress will require more effective integration of insights across these disciplines, but structural and communication barriers often hinder such collaboration [4].

Lack of Standardized Evaluation methodologies makes it difficult to assess and compare different detection approaches, potentially slowing adoption of the most effective techniques. Consistent, comprehensive evaluation frameworks that reflect real-world usage scenarios are needed to guide technology development and implementation decisions.

Privacy and Ethics Concerns surrounding detection technologies could limit their deployment, particularly in sensitive contexts or regions with strong data protection frameworks. Addressing these concerns requires both technical innovations in privacy-preserving detection and careful consideration of the ethical implications of widespread text analysis [109].

9. Conclusion

The field of AI-generated text detection stands at a critical juncture, with both significant challenges and promising innovations on the horizon. The trends and directions outlined in this section suggest a future where detection technologies become more sophisticated, accessible, and integrated into diverse workflows and contexts. However, realizing this potential will require sustained research efforts, thoughtful policy development, and careful attention to ethical implications. The integration of multiple analytical approaches — combining statistical, linguistic, and contextual signals — appears essential for robust detection as generation capabilities advance. No single method is likely to provide a complete solution, but complementary approaches can collectively address different aspects of the detection challenge. The shift toward collaborative human-AI detection frameworks, rather than fully automated classification, aligns with the inherently sociotechnical nature of the challenge. Detection technologies are most valuable when they support human judgment rather than attempting to replace it entirely. The development of more inclusive, adaptable detection approaches that serve diverse user populations and content domains represents both an ethical imperative and a practical necessity. Detection systems that exhibit biases or perform inconsistently across languages and contexts will ultimately undermine trust and limit adoption. The balance between detection and provenance-based approaches suggests that the long-term solution may involve both retrospective analysis of potentially AI-generated content and proactive mechanisms for establishing and verifying content origins. As generative AI technologies continue to evolve and proliferate, detection capabilities will remain essential for maintaining text authenticity, supporting appropriate use policies, and preserving trust in digital communication. The research directions outlined here offer promising pathways toward meeting these challenges, though substantial work remains to transform emerging trends into robust, accessible, and ethically sound detection solutions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] Grand View Research, Artificial intelligence market size report, 2023–2030, 2023.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.

[3] R. Bommasani, D.A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M.S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, 2021, arXiv preprint arXiv:2108.07258.

[4] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, et al., Taxonomy of risks posed by language models, FAccT ' 22: Proc. 2022 ACM Conf. Fairness, Account. Transpar. (2022) 214–229.

[5] D.R. Cotton, P.A. Cotton, J.R. Shipway, Chatting and cheating: Ensuring academic integrity in the era of ChatGPT, Innov. Educ. Teach. Int. (2023) 1–12, http://dx.doi.org/10.1080/14703297.2023.2190148.

[6] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi, Defending against neural fake news, 2019, arXiv preprint arXiv:1905.12616.

[7] C.A. Gao, F.M. Howard, N.S. Markov, E.C. Dyer, S. Ramesh, Y. Luo, A.T. Pearson, Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers, 2022, http://dx.doi.org/10.1101/2022.12.23.521610, bioRxiv preprint.

[8] N. Anderson, D.L. Belavy, S.M. Perle, S. Hendricks, L. Hespanhol, E. Verhagen, A.R. Memon, AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in sports & exercise medicine manuscript generation, BMJ Open Sport. Exerc. Med. 9 (1) (2023) e001568.

[9] OpenAI, ChatGPT: Optimizing language models for dialogue, 2022, OpenAI Blog. URL https://openai.com/blog/chatgpt.

[10] K. Hu, ChatGPT sets record for fastest-growing user base - analyst note, Reuters Technol. News (2023) URL https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[11] E. Mitchell, Y. Lee, A. Khazatsky, C.D. Manning, C. Finn, DetectGPT: Zero-shot machine-generated text detection using probability curvature, 2023, arXiv preprint arXiv:2301.11305.

[12] S. Krishna, E. Shen, J.Z. Kolter, T. Berg-Kirkpatrick, A. Berg, Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense, 2023, arXiv preprint arXiv:2303.13408.

[13] E. Crothers, N. Japkowicz, H. Viktor, Machine generated text: A comprehensive survey of threat models and detection methods, 2023, arXiv preprint arXiv:2210.07321.

[14] D. Weber-Wulff, A. Anohina-Naumeca, S. Bjelobaba, T. Foltýnek, J. Guerrero-Dib, O. Popoola, P. Šigut, L. Waddington, Testing of detection tools for AI-generated text, Int. J. Educ. Integr. 19 (1) (2023) 1–39.

[15] S. Gehrmann, H. Strobelt, A. Rush, 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, 2019, pp. 111–116.

[16] X. He, X. Shen, Z. Chen, M. Backes, Y. Zhang, MGTBench: Benchmarking machine-generated text detection, 2023, arXiv preprint arXiv:2303.14822.

[17] G. Jawahar, M. Abdul-Mageed, V. Lakshmanan, Automatic detection of machine generated text: A critical survey, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, 2020, pp. 2296–2309.

[18] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al., Holistic evaluation of language models, 2022, arXiv preprint arXiv:2211.09110.

[19] A.M. Turing, Computing machinery and intelligence, Mind 59 (236) (1950) 433–460.

[20] J. Weizenbaum, ELIZA—A computer program for the study of natural language communication between man and machine, Commun. ACM 9 (1) (1966) 36–45.

[21] F. Jelinek, Language modeling for speech recognition, Proc. IEEE Work. Speech Recognit. 1 (1980) 1–8.

[22] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.

[23] D.I. Holmes, The evolution of stylometry in humanities scholarship, Lit. Linguist. Comput. 13 (3) (1998) 111–117.

[24] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: European Conference on Machine Learning, Springer, 1998, pp. 137–142.

[25] E. Stamatatos, A survey of modern authorship attribution methods, J. Am. Soc. Inf. Sci. Technol. 60 (3) (2009) 538–556.

[26] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: 1st International Conference on Learning Representations, ICLR 2013, 2013.

[27] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[28] Z. Yao, D. Cai, Y. Hu, L. Liu, Z. Huang, Automated essay scoring by capturing relative writing quality, Computer Journal 60 (9) (2017) 1318–1331.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[30] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, OpenAI Blog (2018).

[31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.

[32] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, J. Mach. Learn. Res. 3 (2003) 1137–1155.

[33] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2019, arXiv preprint arXiv:1810.04805.

[34] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, The curious case of neural text degeneration, Int. Conf. Learn. Represent. (2020).

[35] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J.W. Kim, S. Kreps, et al., Release strategies and the social impacts of language models, 2019, arXiv preprint arXiv:1908.09203.

[36] A. Uchendu, T. Le, K. Shu, D. Lee, Authorship attribution for neural text generation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020, pp. 8384–8395.

[37] D. Ippolito, D. Duckworth, C. Callison-Burch, D. Eck, 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, 2020, pp. 1808–1822.

[38] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, T. Goldstein, A watermark for large language models, 2023, arXiv preprint arXiv:2301.10226.

[39] E. Tian, GPTZero: A Revolutionary AI Classifier, Princeton University, 2023.

[40] T. Fagni, F. Falchi, M. Gambini, A. Martella, M. Tesconi, TweepFake: About detecting deepfake tweets, PLoS One 16 (5) (2021) e0251415.

[41] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, T. Goldstein, Spotting llms with binoculars: Zero-shot detection of machine-generated text, 2024, arXiv preprint arXiv:2401.12070.

[42] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, C. Whitehouse, A.F. Aji, N. Habash, I. Gurevych, et al., SemEval-2024 task 8: Multigenerator, multidomain, and multilingual black-box machine-generated text detection, in: Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Association for Computational Linguistics, 2024, pp. 4262–4275.

[43] Z. Zeng, S. Liu, L. Sha, Z. Li, K. Yang, S. Liu, D. Gašević, G. Chen, Towards detecting AI-generated text within human-AI collaborative hybrid texts, 2024, arXiv, arXiv:2403.03506.

[44] L. Kushnareva, T. Gaintseva, G. Magai, S. Barannikov, D. Abulkhanov, K. Kuznetsov, E. Tulchinskii, I. Piontkovskaya, S. Nikolenko, AI-generated text boundary detection with RoFT, 2023, arXiv preprint arXiv:2311.08349.

[45] Z. Zeng, L. Sha, Y. Li, K. Yang, D. Gašević, G. Chen, Towards automatic boundary detection for human-AI collaborative hybrid essay in education, 2023, arXiv:2307.12267.

[46] V. Liyanage, D. Buscaldi, Detecting artificially generated academic text: The importance of mimicking human utilization of large language models, in: International Conference on Applications of Natural Language To Information Systems, Springer, 2023, pp. 558–565.

[47] Z. Fu, J. Xuan, W.X. Zhao, Z. Niu, J. Wang, L. Hou, Y. Yang, H. Li, X. Jin, F. Zhu, Detection of machine-generated text using contrastive learning, 2023, arXiv preprint arXiv:2307.04303.

[48] A. Wang, Y. Kordi, S. Mishra, L. Castricato, T. Laino, Y. Choi, K. Narasimhan, Prompting GPT-3 to be reliable, in: International Conference on Learning Representations, 2023.

[49] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, Proc. 2020 Conf. Empir. Methods Nat. Lang. Process.: Syst. Demonstr. (2020) 38–45.

[50] X. Wang, W. Zhang, S. Rajtmajer, Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey, 2024, arXiv preprint arXiv:2410.18390.

[51] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, Y. Wu, How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection, 2023, arXiv preprint arXiv:2301.07597.

[52] Z. Su, X. Wu, W. Zhou, G. Ma, S. Hu, Hc3 plus: A semantic-invariant human chatgpt comparison corpus, 2023, arXiv preprint arXiv:2309.02731.

[53] P. Yu, J. Chen, X. Feng, Z. Xia, CHEAT: A large-scale dataset for detecting ChatGPT-written AbsTracts, 2023, arXiv preprint arXiv:2304.12008.

[54] Y. Liu, Z. Zhang, W. Zhang, S. Yue, X. Zhao, X. Cheng, Y. Zhang, H. Hu, ArguGPT: Evaluating, understanding and identifying argumentative essays generated by GPT models, 2023, arXiv preprint arXiv:2304.07666.

[55] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, C. Whitehouse, O.M. Afzal, T. Mahmoud, T. Sasaki, et al., M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection, 2023, arXiv preprint arXiv:2305.14902.

[56] Human-ChatGPT texts dataset, 2023, https://paperswithcode.com/dataset/human-chatgpt-texts. (Accessed 25 March 2025).

[57] L. Research, LAION-GPT4v: A multimodal dataset of image-text pairs generated by GPT-4V, 2024, Available online. Multimodal image-text dataset generated using GPT-4V. URL https://huggingface.co/datasets/qnguyen3/laion-gpt4v.

[58] A.I. and Literature Research Group, WritingPrompts-human-AI: A cross-domain dataset for creative writing detection, 2024, Available online. Cross-domain dataset pairing creative writing prompts with human and AI-generated stories. URL https://paperswithcode.com/dataset/writingprompts.

[59] Kaggle Community, LLM - detect AI generated text, 2024, https://www.kaggle.com/competitions/llm-detect-ai-generated-text. (Accessed 25 March 2025).

[60] J. Feng, J. Chen, C. Tao, S. Chai, Y. Zhu, B. Li, T. Wang, M. Kuang, Y. Yao, R. Yan, et al., M4: Multi-generator, multilingual, and multi-domain benchmark for machine-generated text detection, 2023, arXiv preprint arXiv:2311.09573.

[61] D. Tang, W. Cheng, S. Aldana, M. Ploenzke, J. Rao, Z. Lu, C. An, H. Zhao, T. Yarkoni, H. Lee, et al., Science in the age of large language models, Nat. Rev. Phys. 5 (12) (2023) 782–794.

[62] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al., Holistic evaluation of language models, 2023, arXiv preprint arXiv:2211.09110.

[63] V. Sadasivan, A. Sriram, D. Yang, P. Liang, T. Hashimoto, Can AI-generated text be reliably detected?, 2023, arXiv preprint arXiv:2303.11156.

[64] J. Yang, H. Du, T. Li, S. Wu, T. Liu, H. Zhao, J. Pan, C. Chen, Z. Wang, Q. Xu, et al., Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond, 2023, arXiv preprint arXiv:2304.13712.

[65] S. Mitrović, D. Andreoletti, O. Ayoub, ChatGPT or human? Detect and explain. Explaining decisions of machine learning model for detecting short ChatGPT-generated text, 2023, arXiv preprint arXiv:2301.13852.

[66] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, J. Zou, GPT detectors are biased against non-native english writers, 2023, arXiv:2304.02819.

[67] Turnitin, Launch of Turnitin AI Writing Detection Capabilities: Technical Overview and Validation, Tech. Rep., Turnitin, LLC, 2023, URL https://www.turnitin.com/papers/ai-writing.

[68] S. Yang, J. Prather, C. Piech, Deterring AI-assisted academic dishonesty: The impact of detection systems and educational interventions, Proc. Learn. Scale Conf. (2023) 78–89.

[69] J. Robinson, J. Grabill, A. Whyte, Integrating AI Text Detection in Higher Education: The Michigan State University AI Guardian Program, Tech. Rep., Michigan State University, 2023, URL https://tech.msu.edu/technology/ai/.

[70] U. of California Office of the President, AI-Resistant Assignment Framework: Guidelines and Best Practices, Tech. Rep., University of California, 2023, URL https://www.ucop.edu/academic-affairs/index.html.

[71] K. Aluthgama-Baduge, J. Moore, J. Grabill, From detection to instruction: Using AI text detection as a teaching tool, Comput. Compos. 69 (2023) 102711.

[72] D. Klepper, Learning to lie: AI Tools Adept at Creating Disinformation, Associated Press, 2023, URL https://apnews.com/article/technology-science-business-artificial-intelligence-afb4618ff593db9e3e51ecbd91dc3eef.

[73] Reuters News Agency, Reuters Fact Check Annual Report, Tech. Rep., Reuters, 2023, URL https://reutersinstitute.politics.ox.ac.uk/.

[74] Springer Nature, Springer nature unveils two new AI tools to protect research integrity: Geppetto and SnappShot playing important role in stopping fake research from being published, 2024, https://group.springernature.com/in/group/media/press-releases/new-research-integrity-tools-using-ai/27200740?utm_source=chatgpt.com. (Accessed 26 March 2025).

[75] The Lancet Digital Health, AI text detection and scientific publishing: navigating new challenges in research integrity, Lancet Digit. Heal. 5 (7) (2023) e368–e369.

[76] Wikipedia contributors, Content Authenticity Initiative, 2024, URL https://en.wikipedia.org/wiki/Content_Authenticity_Initiative. ((Accessed 26 March 2025)).

[77] K. Paul, Meta to start labeling AI-generated images from companies like openai, google, Reuters (2024) URL https://www.reuters.com/technology/meta-start-labeling-ai-generated-images-companies-like-openai-google-2024-02-06/. (Accessed 26 March 2025).
[78] Reuters, Reddit unveils new content moderation and analytics tools to boost user engagement, Reuters (2025) URL https://www.reuters.com/technology/reddit-unveils-new-content-moderation-analytics-tools-boost-user-engagement-2025-03-06/?utm_source=chatgpt.com. (Accessed 25 March 2025).
[79] Amazon, Robust proactive controls: AI-powered fake review detection, 2023, https://trustworthyshopping.aboutamazon.com/approach/robust-proactive-controls. (Accessed 26 March 2025).
[80] I. Yelp, Yelp releases 2024 trust & safety report, BusinessWire (2024) URL https://www.businesswire.com/news/home/20250205544346/en/Yelp-Releases-2024-Trust-Safety-Report. (Accessed 26 March 2025).
[81] A.B. Association, ABA Guidelines on the Use of AI Detection Technologies in Legal Practice, Tech. Rep., American Bar Association, 2023, URL https://www.americanbar.org/groups/law_practice/resources/law-technology-today/2025/responsible-ai-use-in-attorney-well-being/.
[82] Everlaw, AI Text Detection in E-Discovery: Capabilities and Limitations, Tech. Rep., Everlaw, Inc., 2023, URL https://www.everlaw.com/blog/ai-and-law/unlocking-justice-ai-evidence-analysis-forensics/.
[83] HSBC Holdings plc, HSBC Cybersecurity Annual Report: AI-Enabled Threat Detection, Tech. Rep., HSBC, 2023, URL https://www.hsbc.com/-/files/hsbc/investors/hsbc-results/2024/annual/pdfs/hsbc-holdings-plc/250219-annual-report-and-accounts-2024.pdf.
[84] National Science Foundation, NSF Pilot Program on AI Detection for Grant Proposals: Implementation Report, Tech. Rep., National Science Foundation, 2023, URL https://www.nsf.gov/focus-areas/artificial-intelligence/nairr.
[85] Neural Information Processing Systems, NeurIPS 2023 Guidelines on LLM Usage and Detection, Tech. Rep., NeurIPS Foundation, 2023, URL https://neurips.cc/virtual/2024/poster/96876.
[86] Neural Information Processing Systems, NeurIPS 2023 Conference Report: LLM Usage and Detection Statistics, Tech. Rep., NeurIPS Foundation, 2023, URL https://blog.neurips.cc/2023/12/12/neurips-2023-conference-report.
[87] Zendesk, Current and upcoming zendesk betas and early access programs (EAPs), 2023, URL https://support.zendesk.com/hc/en-us/articles/4408829663642-Current-and-upcoming-Zendesk-betas-and-early-access-programs-EAPs. (Accessed 26 March 2025).
[88] Reuters, Google to make disclosure of AI-generated content mandatory for election advertisers, 2023, URL https://www.reuters.com/technology/google-make-disclosure-ai-generated-content-mandatory-election-advertisers-2023-09-06/. (Accessed 26 March 2025).
[89] Google, Transparency report: Advertising content policies update - AI disclosure requirements, Google Transpar. Rep. (2023) Retrieved March 26, 2025. xURL https://transparencyreport.google.com/advertising-policies/updates.
[90] Microsoft Corporation, Introducing microsoft 365 copilot: Your copilot for work, 2023, URL https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/. (Accessed 26 March 2025).
[91] Microsoft Corporation, Microsoft 365 copilot release notes, 2023, URL https://learn.microsoft.com/en-us/copilot/microsoft-365/release-notes. (Accessed 26 March 2025).
[92] Stanford University Human-Centered Artificial Intelligence, Human writer or ai? Scholars build a detection tool, 2023, URL https://hai.stanford.edu/news/human-writer-or-ai-scholars-build-detection-tool. (Accessed 26 March 2025).
[93] Coalition for Content Provenance and Authenticity, C2PA Technical Specification: Content Provenance and Authentication, Tech. Rep., C2PA, 2023, URL https://c2pa.org/specifications/specifications/1.3/specs/C2PA_Specification.html.
[94] OpenAI, Adversarial Detection Frameworks: Technical Overview, Tech. Rep., OpenAI, 2023, URL https://openai.com/research/adversarial-detection.
[95] The Mitre Corporation, Domain-Adaptive AI Text Detection: Technical Approach and Evaluation, Tech. Rep., MITRE, 2023, URL https://www.mitre.org/sites/default/files/2024-07/PR-24-01019-6-Repeatable-Process-ssuring-AI-enabled-Systems.pdf.
[96] OpenAI, AI-Generated Content: Policy Updates, Tech. Rep., OpenAI, 2023, URL https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.
[97] S. Levy, M.S. Saxon, W.Y. Wang, The truth is out there: Investigating conspiracy theories in text generation, 2021, arXiv. arXiv:2101.00379.
[98] J. Du, Y. Han, A. Dafoe, J. Graber, T.B. Roettger, ChatGPT detectors are biased against non-native english writers, 2023, arXiv preprint arXiv:2307.14525.
[99] M. Wolff, S. Wolff, Attacking neural text detectors, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 2712–2726.
[100] X. Hu, P.-Y. Chen, T.-Y. Ho, RADAR: Robust AI-text detection via adversarial learning, 2023, arXiv:2307.03838.
[101] A. Jabbar, S. Iqbal, A. Akhunzada, Q. Abbas, An improved urdu stemming algorithm for text mining based on multi-step hybrid approach, J. Exp. Theor. Artif. Intell. 30 (5) (2018) 703–723.
[102] A. Jabbar, S. Iqbal, M.I. Tamimy, S. Hussain, A. Akhunzada, Empirical evaluation and study of text stemming algorithms, Artif. Intell. Rev. 53 (2020) 5559–5588.
[103] K. Wang, J. Zhu, M. Ren, Z. Liu, S. Li, Z. Zhang, C. Zhang, X. Wu, Q. Zhan, Q. Liu, et al., A survey on data synthesis and augmentation for large language models, 2024, arXiv preprint arXiv:2410.12896.
[104] Y. Tian, H. Chen, X. Wang, Z. Bai, Q. Zhang, R. Li, C. Xu, Y. Wang, Multiscale positive-unlabeled detection of AI-generated texts, 2023, arXiv:2305.18149.
[105] J. Yu, Y. Ren, C. Gong, J. Tan, X. Li, X. Zhang, Leveraging large language models for node generation in few-shot learning on text-attributed graphs, 2023, arXiv.
[106] M. Ballout, U. Krumnack, G. Heidemann, K.-U. Kühnberger, Efficient knowledge distillation: Empowering small language models with teacher model insights, Appl Nat Lang Data Base (2024) 32–46.
[107] D. Zhang, T. Feng, L. Xue, Y. Wang, Y. Dong, J. Tang, Parameter-efficient fine-tuning for foundation models, 2025, arXiv preprint arXiv:2501.13787.
[108] H. Chen, F. Tan, A. Kouris, R. Lee, H. Fan, S.I. Venieris, Progressive Mixed-Precision decoding for efficient LLM inference, 2024, arXiv. arXiv:2410.13461.
[109] N. Truong, K. Sun, S. Wang, F. Guitton, Y. Guo, Privacy preservation in federated learning: An insightful survey from the GDPR perspective, Comput. Secur. 110 (102402) (2021) 102402.
[110] H. Chase, LangChain: Building applications with LLMs through composability, 2023, https://github.com/hwchase17/langchain.
[111] Y. Xianjun, Z. Kexun, C. Haifeng, P. Linda, W.Y. Wang, C. Wei, Zero-shot detection of machine-generated codes, 2023, arXiv:2310.05103.

Dr. Tanzila Kehkashan is an Assistant Professor at the Faculty of Information Technology, University of Lahore, Pakistan. She earned her Ph.D. from Universiti Teknologi Malaysia (UTM), Malaysia, and holds a master's degree from the University of Central Punjab (UCP), Pakistan. In addition to her teaching role, Dr. Kehkashan serves as the Principal Investigator of the Vision & Language Computing Matrix Lab (VLCMatrixLab) at the University of Lahore, Pakistan, where she leads research projects in the areas of Artificial Intelligence, Computer Vision, and Natural Language Processing. Her work has been published in prestigious journals and conference proceedings, showcasing her commitment to advancing research in these fields. Her research interests include artificial intelligence, machine learning, deep learning, computer vision, natural language processing, and the development of innovative solutions for real-world problems in healthcare, agriculture, education, and cybersecurity. Her dedication to both research and academic practice reflects her passion for the field of computer science.

Raja Adil Riaz is a final-year student in the Bachelor's program in Computer Science at the University of Lahore. He has a well-established foundation in core computer science principles and has specialized in programming, software development, and their practical implementations, as a research member of the Vision & Language Computing Matrix Lab (VLC Matrix Lab). He is actively engaged in collaborative academic projects and ongoing research activities. Throughout his academic journey, he has taken a keen interest in conducting research and practical projects on significant topics, such as predictive modeling and interpretability in health-related applications. His dedication to the field is evident through his involvement in academic publishing and continuous contributions to impactful research. Always eager to explore new knowledge and experiences, he remains open to challenges that inspire innovation and growth.

Dr. Ahmad Sami Al-Shamayleh received the master's degree in information systems from The University of Jordan, Jordan, in 2014, and the Ph.D. degree in artificial intelligence from the University of Malaya, Malaysia, in 2020. He is currently an Assistant Professor at the Faculty of Information Technology, Al-Ahliyya Amman University, Jordan. His research interests include artificial intelligence, human–computer interaction, the IoT, Arabic NLP, Arabic sign language recognition, language resources production, the design and evaluation of interactive applications for handicapped people, multimodality, and software engineering.

Dr. Adnan Akhunzada (Senior Member, IEEE) Distinguished Senior Member of IEEE and Professional Member of ACM, brings 15 years of expertise in Research and Development (R&D) at the nexus of the ICT industry and academia. Renowned for his high-impact publications, US patents, and commercial products, Dr. Akhunzada's patented innovations in cybersecurity and AI have secured multi-million-dollar projects with global entities such as Vinnova and EU Horizon. In 2023, Stanford University recognized him as one of the top 2% scientists globally for his outstanding scholarly contributions. Leveraging his robust cybersecurity skills and advanced technological knowledge, Prof. Akhunzada excels in solving industrial challenges and developing state-of-the-art security tools and frameworks. His expertise spans Cybersecurity & AI, Secure Future Internet, Secure & Dependable Software Defined Networks, and Large-Scale Distributed Systems (including Cloud, Fog, Edge, IoT, IoE, IIoT, CPS). Additionally, his work on Lightweight Cryptographic Communication Protocols, QoS/QoE, and Adversarial Machine Learning is shaping the future of secure and dependable systems.

Noman Ali is a diligent student pursuing a degree in Computer Science from the University of Lahore. His research interests and studies focus on cutting-edge sectors, including Machine Learning, Natural Language Processing, Agricultural Technologies, and Medical Imaging. Throughout his academic journey, Noman has actively sought knowledge in applying Artificial Intelligence across various domains and has contributed innovative solutions to real-world challenges. His dedication to leveraging technology for relevant research is evident through his involvement in agricultural systems and healthcare diagnostics projects. Noman remains open to opportunities that challenge and inspire him as he continuously expands his knowledge and technical skills.

Muhammad Hamza is a diligent student pursuing a degree in Computer Science from the University of Lahore. His research interests and studies revolve around some of the latest sectors, including Machine Learning, Natural Language Processing, Agricultural Technologies, and Medical Imaging. During his academic career, Hamza has actively pursued knowledge in the application of Artificial Intelligence across different domains and was able to contribute innovative solutions to practical challenges. His commitment to harnessing technology for relevant research is demonstrated through his participation in agricultural systems and healthcare diagnostics. Hamza keeps himself open to opportunities that challenge and inspire him and continually expands his knowledge and skills.

Faheem Akbar is a dedicated Bachelor of Computer Science student at the University of Lahore (UOL), Pakistan, currently in his final year. He is a dedicated research member of the Vision and Language Computing Matrix Lab (UOL VLCMatrixLab, Pakistan), where he actively contributes to research projects focused on artificial intelligence and healthcare innovation. His research focuses on integrating machine learning, natural language processing, and medical imaging, with a specialization in applying advanced AI techniques for biomedical applications. With a strong background in Python, TensorFlow, and data analysis, he has gained substantial experience in model training, validation, and data visualization. Faheem has a keen interest in using technology to improve clinical decision-making, healthcare systems, and predictive analytics. Upon graduation, he plans to further his career in AI research and contribute to transformative advancements in the field.