Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(1), 38-49
Published Online January 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.151005

机器生成文本检测综述

孙一凯[1]，王洪俊[1,2]

[1]北京信息科技大学计算机学院，北京
[2]拓尔思信息技术股份有限公司，北京

收稿日期：2025年11月26日；录用日期：2025年12月22日；发布日期：2025年12月31日

摘　要

随着人工智能生成内容(AIGC)与人类文本之间的界限日益模糊，机器生成文本检测成为自然语言处理的重要研究方向。文章综述机器生成文本检测的技术演变，包括主动嵌入隐秘信号的水印技术、基于特征统计的传统方法、利用预训练语言模型(如RoBERTa、DeBERTa)进行判别的监督学习方法，以及结合模型预测不确定性、特征分布差异的概率检测方法。近年来，局部化检测与可解释性分析成为新的研究热点，使检测系统能够识别具体生成片段并解释判别依据。然而，跨模型泛化、多语言场景与对抗鲁棒性仍是亟待解决的难题。未来的研究将致力于构建具有更强鲁棒性和可解释性的检测框架，结合因果推理与多模态信息提升检测性能，助力推动LLM生成文本检测技术的实用化与规范化发展。

关键词

大语言模型，自然语言处理，文本检测

A Comprehensive Survey of Machine-Generated Text Detection

Yikai Sun[1], Hongjun Wang[1,2]

[1]College of Computer Science, Beijing Information Science and Technology University, Beijing
[2]TRS, Beijing

Received: November 26, 2025; accepted: December 22, 2025; published: December 31, 2025

Abstract

As the boundary between Artificial Intelligence-Generated Content (AIGC) and human-written text becomes increasingly blurred, machine-generated text detection has emerged as a critical research direction in natural language processing. This paper reviews the technological evolution of machine-generated text detection, including watermarking techniques that actively embed hidden signals,

文章引用: 孙一凯, 王洪俊. 机器生成文本检测综述[J]. 人工智能与机器人研究, 2026, 15(1): 38-49.
DOI: 10.12677/airr.2026.151005

traditional methods based on feature statistics, supervised learning approaches leveraging pre-trained language models (e.g., RoBERTa, DeBERTa) for discrimination, and probability-based detection methods that incorporate model prediction uncertainty and feature distribution differences. In recent years, localized detection and interpretability analysis have become new research hotspots, enabling detection systems to identify specific generated segments and explain the basis for discrimination. However, cross-model generalization, multilingual scenarios, and adversarial robustness remain pressing challenges to be addressed. Future research will focus on constructing detection frameworks with enhanced robustness and interpretability, integrating causal reasoning and multimodal information to improve detection performance, thereby advancing the practicalization and standardization of LLM-generated text detection technologies.

Keywords

Large Language Models, Natural Language Processing, Text Detection

1. 引言

机器生成文本检测(Machine-Generated Text Detection, MGT Detection)是近年来自然语言处理与人工智能安全领域的重要研究方向，旨在识别由大语言模型(Large Language Model, LLM)生成的文本内容。随着 ChatGPT、Claude、Gemini 等生成大语言模型的快速发展，人工智能生成内容(AIGC)与人类撰写文本之间的界限逐渐模糊，引发了关于学术诚信、内容审核和信息安全等方面的广泛关注[1]-[3]。早期方法主要基于统计特征与语言模型困惑度进行判断，而近年来，基于预训练语言模型的检测框架成为主流，如 RoBERTa、DeBERTa 和 ELECTR [4]-[6]等。这些模型通过大规模标注数据学习人类与机器文本的语义差异，在检测精度和泛化能力方面取得了显著提升。同时，局部化检测(Localization)方法[7]的提出进一步推动了该领域的发展，能够在句子或片段级别精确识别生成内容，增强了检测结果的细粒度和可解释性。此外，结合对抗学习与多模态建模的检测方法[8]，也在提高鲁棒性和跨场景适应性方面展现出潜力。尽管如此，机器生成文本检测仍面临诸多挑战，如跨模型泛化能力不足、多语言文本检测困难、对抗性文本识别不稳健，以及模型判别依据缺乏透明性。未来的研究将致力于构建具有更强鲁棒性和可解释性的检测框架，结合因果推理与多模态信息提升检测性能。

本文梳理总结了机器生成文本检测的技术演变及实现细节，讨论了现有技术的发展状况及关键挑战，以期为未来研究者探索创新方向提供参考。

2. 机器生成文本检测任务

机器生成文本检测的主要目标是判别给定的自然语言文本是由人类撰写还是由机器生成，并可进一步对文本中是否混入人机交织部分、由哪个生成模型产出、定位其具体片段等细粒度任务展开。早期检测任务通常被建模为二分类问题，待检测器输出标签{Human, Machine}，如图 1。然而随着任务与应用场景的扩展，更多子任务逐渐被提出，包括多分类[9]、局部化检测等[10]。

2.1. 二分类任务(Binary Classification)

基础的检测任务形式是二分类，即判断输入文本是否由机器生成。其典型系统输入为一句话、一段

话或整篇文章，输出为“Human”或“Machine”的布尔标签或概率分布。经典研究方法包括基于困惑度(Perplexity)分析的规则方法、监督学习模型、以及深度上下文建模的预训练检测器等。

2.2. 多分类任务(Multi-Class Classification)

随着生成模型的多样化，研究者逐渐意识到，检测模型输出的不仅应是“是否由机器生成”，还应进一步识别“由哪个模型生成”。例如，针对 GPT-2、GPT-3、LLaMA、BERT-based Generator 等不同生成架构，其生成模式、语言风格、语义连贯性等往往有所不同，因此可作为检测特征加以区分。此类任务一般在标注数据中提供多类生成模型样本，检测器训练为 K 类标签分类器(K > 2)。

2.3. 局部化检测任务(Localization)

局部化检测(Localizing Machine Text)相较于整体判别任务更具挑战性，其目标是确定输入文本中哪些句子或片段由机器生成。该任务一般被建模为序列标注或句子级分类问题，输出格式可为连续位置标注或句子粒度判断(图 1)。

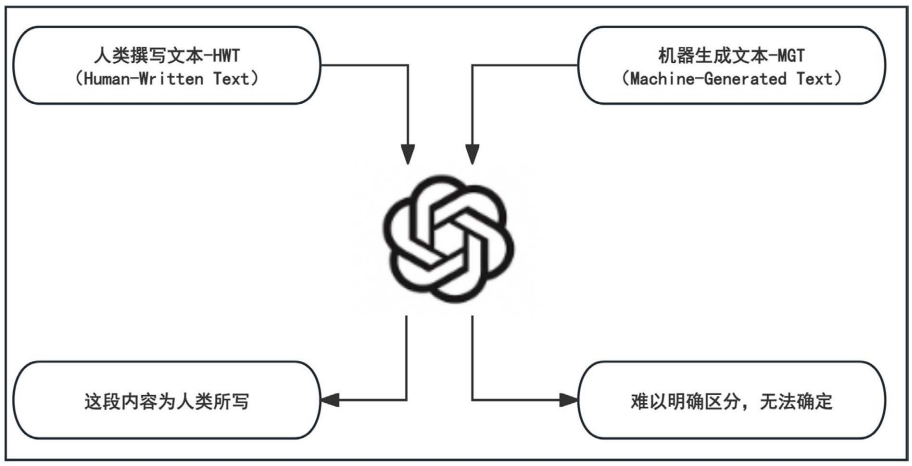

Figure 1. Using a ChatGPT-based detector to distinguish text
图 1. 使用 ChatGPT 检测器区分文本

3. 机器文本检测数据集

在机器生成文本检测中，高质量、丰富多样的数据集成为检测模型训练、验证与评估的关键资源。在早期研究中，学者们多依赖有限领域或单一模型生成的数据集，主要关注新闻文本或维基文本的二分类任务。然而，随着生成模型能力的提升及应用场景多样化，研究者开始构建大规模、多语言、多模型以及对抗场景的综合数据集，以推动检测方法的泛化与鲁棒性研究。目前公开且广泛使用的数据集包括：MAGE、MULTITUDE、RAID、MAGRET、MGTBench、M4GT-Bench、M4、MultiSocial 等。

3.1. MAGE

MAGE 数据集[11]旨在为真实场景下的机器文本检测提供基准，该数据集涵盖 27 种主流大语言模型生成的文本，覆盖十余种写作任务，包括新闻、评论、问答、故事等领域。其一个重要特点是“野外场景”，用于评估检测器在未见模型与未见领域下的泛化能力。数据集中每条样本标注为“人类撰写”或“机器生成”，并提供生成模型来源信息，使研究者可进行细粒度分析。该数据集规模庞大，训练与测试样本总计超过 200 万条，能够支持深度学习检测模型的大规模训练。

3.2. MULTITUDE

MULTITuDE 数据集[12]是一个多语言机器文本检测基准数据集，覆盖 11 种语言，包括英语、阿拉伯语、中文、俄语等。该数据集收录了由 8 种多语言大语言模型生成的文本，同时包含相应的人类写作样本。MULTITuDE 不仅提供文本分类标签，还提供生成模型类型、语言和文本任务类别信息。该数据集的多语言、多任务特点使其成为跨语言检测方法研究的重要资源。其总规模约 500 万条文本样本，是目前规模较大且覆盖面广的数据集之一。

3.3. RAID

RAID 数据集[13]在评估机器文本检测方法在对抗和鲁棒性场景下的性能。该数据集包含超过 600 万条生成文本，覆盖 11 种生成模型、8 个文本领域和 11 种对抗攻击策略，包括同义词替换、句法重排以及混合人机文本。RAID 的设计特别强调检测器在应对改写与攻击生成文本时的稳健性，对研究跨模型和跨领域的检测技术具有重要意义。数据集提供了详细的攻击类型标注，可用于分析检测模型在不同攻击场景下的性能差异，是目前研究对抗机器文本检测的重要资源。

3.4. MAGRET

MAGRET 数据集[14]用于识别机器生成文本并进行“改写追踪”任务。该数据集采用同一提示由模型生成原文，再由模型或人类进行改写，以构建“原生成 → 改写”序列。研究发现：即便是闭源大模型生成的文本，在改写后仍可通过统计或语义关系检测。MAGRET 支持检测与追溯两种任务，适用于文本来源识别研究。

3.5. MGTBench

MGTBench 数据集[15]是首个专门针对于强大大语言模型(如 ChatGPT-turbo、Claude)生成文本的综合检测基准。数据集中包含多个子数据集、由多种 LLM 生成文本、人类文本对照、以及完整的对抗改写攻击实验(如 paraphrasing、随机空格、扰动)用于评估检测器的鲁棒性。研究发现，文本长度较大通常带来更好检测性能，同时许多检测方法在更少训练样本下即可达到类似效果。MGTBench 支持二分类检测和来源归属(attribution)任务，是当前检测研究中不可或缺的标准基准。

3.6. M4GT-Bench

M4GT-Bench 数据集[16]面向“黑盒”机器生成文本检测场景，设计了三项任务：单语言与多语言的二分类检测(人类 vs 机器)；多分类检测任务，要求识别文本是由哪一款生成模型生成；混合人类－机器文本检测任务，要求定位人写与机器写之间的边界。数据集涵盖多语言(包括英语、阿拉伯语、中文、俄语、德语、意大利语、乌尔都语、保加利亚语、印尼语)、多领域(news、Wikipedia、问答论坛、论文摘要等)及多生成器，从而提高检测任务的复杂度与现实适用性。

3.7. M4

M4 数据集[17]的目标是在“黑盒”场景中检测机器生成文本，即模型来源未知，仅通过文本决策。研究发现，当检测器面对来自未见域或未见生成器时，误判率显著上升，表明真实场景下的泛化能力仍是重大挑战。因此，M4 已成为当前 MGT 研究中的标准基准(benchmark)与压力测试(stress test)，非常适合用来评估模型的泛化能力、鲁棒性以及跨域或跨语言适应性。

3.8. MultiSocial

MultiSocial 数据集[18]是第一个聚焦于社交媒体文本场景(短文本、口语化、多语言平台)的大规模检

测数据集。该数据集覆盖 22 种语言、5 个社交媒体平台，共约 472,097 条文本，其中约 58 k 为人类撰写，其余由 7 种多语言大模型生成。MultiSocial 针对社交媒体文本特点提供了更细粒度的标注，包括是否带有情绪极性、是否存在平台特有结构(如 hashtag、emoji、URL)、以及是否属于对话链条的一部分。MultiSocial 已逐渐成为评估模型在真实用户场景下鲁棒性的重要 benchmark，特别适合用于测试跨平台泛化能力与短文本检测性能。

表 1 为机器生成文本检测数据集的总结。该表提供了数据集特征的简要概述，便于研究人员比较与选择符合需求的数据集。

Table 1. Datasets for machine-generated text detection
表 1. 机器文本检测数据集

数据集	语言	规模	生成模型数量	主要任务	用途
MAGE	多语言 (主要为英语)	200 万	27 种主流 LLM	二分类 (人类 vs 机器)	覆盖新闻、评论、问答；支持细粒度分析与泛化评估。
MULTITUDE	11 种语言 (英、汉、俄等)	500 万	8 种多语言 LLM	多语言文本检测	提供语言、任务、模型类型标签。
RAID	多语言	600 万	11 种	对抗与鲁邦检测	包含 11 种对抗攻击策略(同义替换、句法重排等)。
MAGRET	英文为主	数十万级	多种开源与闭源模型	检测 + 改写追踪	构建“原生成→改写”序列，支持检测与追溯任务。
MGTBench	英文	数十万级	多种(如 ChatGPT-turbo、Claude)	二分类 + 来源归属	含多种对抗改写攻击；评估鲁棒性。
M4GT-Bench	多语言	百万级	多生成器	二分类 + 多分类 + 混合检测	面向“黑盒”检测场景；支持模型识别与人机边界定位。
M4	多语言	百万级	多种(ChatGPT、GPT-4、BLOOMz 等)	黑盒检测	聚焦未见域与未见生成器下的检测泛化；分析误判问题。
MultiSocial	22 种语言	数十万级	多种 LLM	社交媒体文本检测	面向短文本、多语言平台、提高实际场景适用性。

4. MGT 方法

本节系统梳理机器生成文本检测技术的发展历程，并按其核心技术机制进行分类总结如图 2，进而揭示其优势、局限与关键挑战。

4.1. 基于风格和统计特征的检测

统计特征方法通过提取词频、词汇多样性、句长、词性分布、突发性及困惑度等显式或隐式统计指标，捕捉人机文本的分布差异。其优势在于实现简单、计算高效且可解释性强，但随着大语言模型生成

质量提升，其在面对改写、润色或跨域场景时鲁棒性显著下降。

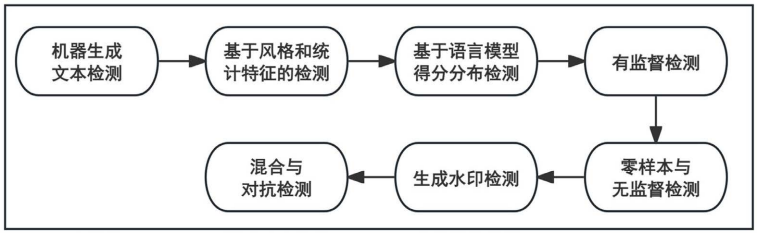

Figure 2. Machine-generated text detection methods
图 2. 机器生成文本检测方法

Gehrmann 等[19]提出 GLTR，GLTR 利用 GPT-2 计算每个词的预测排名，将其分为 top-10、top-100、top-1000 等区间，并通过可视化与统计直方图展示文本的可预测性分布。人类文本的分布更分散，而机器文本的 top-10 占比更高，因此可据此区分两者。成为早期“可解释化”检测工具的重要基线与教学范例。Zellers 等[20]在 Grover 工作中提出用大规模新闻语料训练可控生成器，基于生成器或相似架构计算候选文章的生成概率轨迹与词级置信度分布，并将这些分布性特征输入到判别器中以判断真假新闻，形成“自适应”的统计检测机制。Mitchell 等[21]提出从概率曲面结构角度对文本施加小幅扰动并观测扰动前后对数概率的变化(即概率曲率)，计算对数概率差异并据此得分，最终基于曲率统计决定“机器或人”归属。Bao 等[22]提出以条件概率曲率为核心，用采样替代扰动步骤，融合似然与熵的特性，在白盒和黑盒设置下检测准确率相对提升约 75%，速度提升 340 倍，且鲁棒性强。Li 等[11]使用 20+种 LLM 在同一提示下生成对照生成文本；针对每条文本计算一组统计特征(如 perplexity、token-rank 分布、局部熵、词性统计等)，并将这些特征用于构建综合检测器或作为基线可视化工具。Venkatraman 等[23]提出的 GPT-who 检测器，以均匀信息密度(UID)原则为核心，通过计算文本 token 惊讶度的均值、方差、相邻差异及最大或最小均匀片段等特征，结合逻辑回归实现机器生成文本检测。

4.2. 基于语言模型得分分布的检测

困惑度与语言模型通过计算文本的困惑度或平均对数概率并设定阈值进行判别，虽实现简单、数据依赖低，但易受模型迭代、采样策略变化及对抗改写影响，鲁棒性有限，成为后续研究亟需突破的方向。

Ippolito 等[24]在数据预处理阶段标准化文本长度、剔除格式噪声，并在特征工程层面组合多维统计量以提升判别稳定性，从而将统计特征的单一信号扩展为多元判别视角。Welleck 等[25]提出在训练目标中加入“消极似然”项，显式惩罚模型给出不当高概率(如重复 token)的行为，进而改变模型的概率分布特性。Megías 等[26]探讨了困惑度(perplexity)作为分类信号在检测中的作用，但其效果受生成模型类型和语言域差异限制。Holtzman 等[27]提出了通过计算困惑度以及基于生成模型的概率分布来区分机器生成文本和人类文本的方法。通过对比多个生成模型在相同文本上的表现，设计了一个基于困惑度的判别框架，通过优化模型的生成概率进行训练。实验表明，在一定条件下，基于困惑度的方法能有效区分机器生成文本。

4.3. 有监督检测方法

有监督检测通过标注数据训练分类器区分人类与机器生成文本，甚至溯源至具体生成模型，依托 Transformer 等深度架构，有效捕捉语义与结构层面的细微差异，显著优于依赖表层统计或困惑度的传统手段，展现出更强的判别能力。

Uchendu 等[28]提出将 RoBERTa 应用于二分类检测任务的方法，将文本输入预训练的 RoBERTa 模型，通过 Transformer 编码器提取上下文特征。随后在输出层添加全连接分类器，将编码后的向量映射为二分类标签，形成端到端的有监督检测流程。He 等[29]提出 DeBERTa，其在多项 NLP 任务上显著超越 BERT 与 RoBERTa，首次实现单模型在 SuperGLUE 基准上超越人类表现，展现了强大的语言理解与生成能力。Clark 等[6]将输入文本转换为 token 序列，然后利用 ELECTRA 的生成判别器(discriminator)进行特征提取。输出特征向量经过线性层映射为二分类标签，并在有监督训练下优化交叉熵损失。Kuznetsov 等[30]提出基于受限嵌入的鲁棒 AI 生成文本检测方法，通过移除嵌入空间中有害线性子空间，剔除领域特异性伪特征。实验显示，其对 RoBERTa 和 BERT 嵌入的优化，使分布外分类分数最高提升 14%，鲁棒性显著优于现有方法。Zhi 等[31]提出基于对比增强混合特征与支持向量机(SVM)的高效 AI 生成文本检测模型。该模型在 SAID_quora 数据集上准确率与 F1 值均达 0.89，加权准确率 0.93，推理时间仅 16 秒，较 RoBERTa 效率提升 30%，兼顾检测性能与计算效率。Wei 等[32]通过微调 LLM，结合校准损失优化检测边界，在 21 个领域和 4 种 LLM 数据上，其 ID 场景 AUROC 达 0.90，OOD 场景达 0.66，对抗攻击下仍保持 0.87 以上，较 Fast-DetectGPT 等基线最高提升 48.66%。

4.4. 零样本与无监督检测方法

零样本检测其核心假设是 AI 生成文本在概率结构或统计分布上与人类文本存在系统性差异，因而具备更强的跨模型与跨域泛化能力，尤其适用于未知生成器或未见领域的检测场景。

Jiao 等[33]提出 M-RangeDetector，以人机写作策略差异为领域无关特征，融合多范围注意力模块，实现机器生成文本检测的泛化提升。该模块通过全局、带状、扩张、随机四种注意力掩码，捕获不同范围的上下文表征，伪分类器增强特征多样性。Guo 等[34]提出 AuthentiGPT 模型，通过融合黑盒 LLM 去噪、语义相似度对比与高斯混合模型聚类技术，在生物医学领域数据集上 AUROC 达 0.918，优于主流商业检测器，适配性强。Pu 等[35]提出融合中型模型数据的集成检测方案，剔除大型模型数据后性能损失极小，为资源受限场景提供高效解决方案。Sadiq 等[36]提出融合 FastText 词嵌入于 CNN 的社交媒体深度伪造推文检测方法，用于识别机器生成的虚假内容，在 Tweepfake 数据集上准确率达 93%，F1 值 0.93，显著优于 BERT、RoBERTa 等模型(准确率均为 89%)。Yan 等[37]提出轻量级零样本机器生成文本检测器，针对中文场景优化，解决现有方法资源消耗大的问题。Hans 等[38]提出 Binoculars 零样本机器生成文本检测器，在 0.01%误报率下，对 ChatGPT 生成文本的检测率超 90%，能识别多种 LLMs 输出，泛化性优于 GPTZero 等方法。Guo 等[39]提出 DeTeCtive 框架，通过多任务辅助的多层对比学习解决 AI 生成文本检测泛化性不足的问题。在 Deepfake 等数据集上，其 ID 场景 AvgRec 达 96.15%，OOD 场景对未见过的模型和领域分别超现有方法 5.58%和 14.20%。

4.5. 生成水印检测技术

生成水印技术通过在生成阶段向模型输出中嵌入特定分布或信号，使得生成文本可在后续用轻量计算有效识别。如 Fu 等[40]提出语义感知水印算法，结合输入上下文，通过词向量相似度筛选语义相关 token 纳入“绿色列表”，平衡水印随机性与语义关联性。Yang 等[41]提出适用于黑盒语言模型的文本水印框架，解决传统水印无法适配 API 调用场景的问题。在检测时采用统计检验，提供快速和精确两种模式，在中英文本上均保持语义保真度，且能抵抗重译、润色等攻击，鲁棒性优异。Hou 等[42]提出 SEMSTAMP 语义水印算法。面对 Pegasus 双词复述攻击，KGW 算法 AUC 降 7.9%，而该算法仅降 3.5%，且生成文本困惑度 10.20，接近无水印模型的 10.02，对复述攻击抵抗力与文本质量均更优。Piet 等[43]提出 MARKMYWORDS 基准，用于系统评估 LLM 输出水印方案。Huang 等[44]基于贝叶斯规则提出水印检测

器(BRWD)，在数学任务中，BRWD 将 1% FPR 下的 TPR 从低于 60%提升至 91%，代码任务中相对准确率最高提升 50%，且在无原始提示场景下仍保持优势。Xu 等[45]针对开源 LLM 滥用问题，定义 IP 侵权和生成文本违规两类场景，提出后门水印与推理式水印蒸馏两种方案，可靠性强。Wang 等[46]借助代理语言模型实现词汇概率均衡划分，在生成时嵌入多比特定制信息，该方法在 OPT、LLaMA 系列模型上，水印成功率达 95%，且生成文本困惑度接近无水印文本，质量损失小。Zhao 等[47]提出 BranchWM 黑盒模型水印协议，解决现有方法导致模型主任务性能下降的问题。其通过添加并行水印分支解耦原任务与取证任务，基于 EUF-CMA 安全的 MAC 构造触发机制。

4.6. 混合与对抗检测

现有混合与对抗检测技术通过细粒度建模、对抗训练引入扰动样本、软标签估计、多视角特征融合以及可解释性注意力引导等策略，提升模型对人机混合文本和对抗改写的鲁棒判别能力，推动检测从整体判断迈向精准定位。

Macko 等[48]提出的多语言机器生成文本(MGT)检测基准与规避研究框架，通过整合 10 类作者混淆(AO)方法、37 种 MGT 检测模型及 11 种语言数据，构建了首个多语言对抗性评估体系。Koike 等[49]提出 OUTFOX 框架，通过上下文学习让检测器与攻击者互学，攻击者生成难检测文本，检测器据此强化识别能力。Przybyła 等[50]提出 TREPAT 框架，在假新闻等 4 类任务的 BODEG 得分达 0.33，人工评估语义保留率超基线 30%以上。Teja 等[51]提出基于句子级分割的 AI 生成文本细粒度检测模型，在 TriBERT 和 M4GT 数据集上表现优异，有效识别混合文本中的 AI 生成片段。Jiang 等[52]提出 SenDetEX 框架，通过融合风格与上下文信息，实现人机混合文本的句子级 AI 生成文本检测。Corizzo 等[53]提出 One-GPT 单类深度融合模型，融合 Doc2Vec 文档嵌入的上下文特征与文本、重复性等五类语言特征，通过自编码器架构训练，展现跨语言检测能力。Li 等[54]提出基于语法树的抗扰动 LLM 生成文本检测器(PRDetect)，解决现有方法易受文本扰动影响的问题。Bethany 等[55]基于 T5 编码器提取文本嵌入，结合多分类器与子聚类技术，适配不同生成器的独特特征。在 9 个领域和 9 个生成器上，对未见过的生成器和领域的 F1 分数平均提升 11.9%，生成器归因准确率达 93.6%，对抗扰动后 F1 仍保持 67.2%，泛化性与鲁棒性优异。Wang 等[56]提出 SeqXGPT 方法提取白盒 LLM 的词级对数概率作为波浪状特征，在 SeqXGPT-Bench 数据集上，其句子级检测 Macro-F1 达 95.7%，文档级达 94.2%，OOD 场景仍保持 92.8%，显著超越 DetectGPT 等基线，泛化性优异。

5. 关键挑战及未来研究

机器生成文本检测技术近年来取得了显著进展，但仍面临以下关键挑战：① 模型升级与泛化能力不足：生成模型迭代速度极快，新模型往往显著提升了生成文本的自然性、逻辑一致性与语义流畅性，从而加剧了检测难度。例如，检测器基于 GPT-3 样本训练，但面对 GPT-4 或 GPT-5 的文本时检测性能显著下降。② 人机混合文本带来的模糊边界：在现实场景下，单纯由人或机器生成的“干净文本”较为罕见，常见的是机器先生成草稿，人类再对其进行修改润色(或反之)。例如，一名学生使用 GPT-4 生成文章框架后再人工扩充，高质量的人机混合文本极大干扰检测器识别。③ 对抗性文本与改写攻击：简单的 GPT 检测器往往会被少量改写所迷惑，例如通过同义重写、强调被动语态、插入错别字或语气词、或增加低频词等方式。④ 多语言与跨领域检测困难：大多数现有检测方法主要针对英文生成文本开展，在多语言模型兴起后，支持不同语言的文本生成检测需求愈发明显。此外，对于不同领域(如医学、法律、社交媒体)内容，其语体规范、术语频率等因素的变化也会影响检测器的鲁棒性和可迁移性。

未来研究可聚焦于几个方向：一是突破表层统计特征，挖掘推理链一致性、知识调用方式、错误模

式等更深层的人类写作特征；二是构建跨模型、跨任务的通用检测框架，提高在未知模型上的泛化能力；三是发展具备对抗鲁棒性的检测方法，通过不变特征、多视角融合或对抗训练应对 paraphrasing 攻击；四是探索水印、生成轨迹等模型内部信号，实现更可信的源追踪；五是构建覆盖多语言、多体裁、真实改写场景的大规模基准，以推动该领域向更稳健、更可解释、更规范化的方向发展。

6. 结语

本文系统梳理了机器生成文本检测领域的技术框架、核心任务、关键数据集与发展挑战。从早期的统计特征分析到当前基于预训练语言模型的深度检测器，该领域的研究范式已实现了显著演进。然而，随着生成式人工智能技术的爆发式发展，检测技术正面临着前所未有的严峻考验。

展望未来，机器生成文本检测绝不会止步于一个简单的二分类问题。它必将迈向一个更复杂、更需鲁棒性与可解释性的综合体系。未来的研究需要将因果推理、对抗训练、多模态信息乃至伦理考量更深层次地融入检测框架，致力于构建不仅“检测得准”，更能“解释得清”且“适应得快”的新一代系统。最终，这项技术的发展，不仅关乎技术本身，更是在塑造人机协作新范式下，一个可信、安全、负责任的数字生态。

参考文献

[1] Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., *et al.* (2023) Testing of Detection Tools for AI-Generated Text. *International Journal for Educational Integrity*, **19**, 1-39. https://doi.org/10.1007/s40979-023-00146-z

[2] Najjar, A.A., Ashqar, H.I., Darwish, O.A., *et al.* (2025) Detecting AI-Generated Text in Educational Content: Leveraging Machine Learning and Explainable AI for Academic Integrity. arXiv:2501.03203.

[3] Zhou, Y., He, B. and Sun, L. (2024) Humanizing Machine-Generated Content: Evading AI-Text Detection through Adversarial Attack. arXiv:2404.01907.

[4] Yadagiri, A., Shree, L., Parween, S., *et al.* (2024) Detecting AI-Generated Text with Pre-Trained Models Using Linguistic Features. *Proceedings of the* 21*st International Conference on Natural Language Processing* (*ICON*), Chennai, 15-18 December 2024, 188-196.

[5] He, P., Liu, X., Gao, J., *et al.* (2021) DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. *Proceedings of the International Conference on Learning Representations*, Vienna, 3-7 May 2021, 1-17.

[6] Clark, K., Luong, M.T., Le, Q.V., *et al.* (2020) Electra: Pre-Training Text Encoders as Discriminators Rather Than Generators. arXiv:2003.10555.

[7] Zhang, Z., Qin, W. and Plummer, B. (2024) Machine-Generated Text Localization. Findings of the Association for Computational Linguistics ACL 2024, Bangkok, 11-16 August 2024, 8357-8371. https://doi.org/10.18653/v1/2024.findings-acl.495

[8] Kadhim, A.K., Jiao, L., Shafik, R., *et al.* (2025) Adversarial Attacks on AI-Generated Text Detection Models: A Token Probability-Based Approach Using Embeddings. arXiv:2501.18998.

[9] Zeng, C., Tang, S., Chen, Y., *et al.* (2025) Human Texts Are Outliers: Detecting LLM-Generated Texts via Out-of-Distribution Detection. arXiv:2510.08602.

[10] Tao, Z., Li, Z., Chen, R., *et al.* (2024) Unveiling Large Language Models Generated Texts: A Multi-Level Fine-Grained Detection Framework. arXiv:2410.14231.

[11] Li, Y., Li, Q., Cui, L., Bi, W., Wang, Z., Wang, L., *et al.* (2024) MAGE: Machine-Generated Text Detection in the Wild. *Proceedings of the* 62*nd Annual Meeting of the Association for Computational Linguistics*, **1**, 36-53. https://doi.org/10.18653/v1/2024.acl-long.3

[12] Macko, D., Moro, R., Uchendu, A., Lucas, J., Yamashita, M., Pikuliak, M., *et al.* (2023) Multitude: Large-Scale Multilingual Machine-Generated Text Detection Benchmark. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 9960-9987. https://doi.org/10.18653/v1/2023.emnlp-main.616

[13] Dugan, L., Hwang, A., Trhlík, F., Zhu, A., Ludan, J.M., Xu, H., *et al.* (2024) RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors. *Proceedings of the* 62*nd Annual Meeting of the Association for*

Computational Linguistics, **1**, 12463-12492. https://doi.org/10.18653/v1/2024.acl-long.674

[14] Huang, Y., Cao, J., Luo, H., Guan, X., Liu, B. (2025) MAGRET: Machine-Generated Text Detection with Rewritten Texts. *Proceedings of the* 31*st International Conference on Computational Linguistics*, Abu Dhabi, 19-24 January 2025, 8336-8346.

[15] He, X., Shen, X., Chen, Z., Backes, M. and Zhang, Y. (2024) MGTbench: Benchmarking Machine-Generated Text Detection. *Proceedings of the* 2024 *on ACM SIGSAC Conference on Computer and Communications Security*, Salt Lake City, 14-18 October 2024, 2251-2265. https://doi.org/10.1145/3658644.3670344

[16] Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., *et al.* (2024) M4GT-Bench: Evaluation Benchmark for Black-Box Machine-Generated Text Detection. *Proceedings of the* 62*nd Annual Meeting of the Association for Computational Linguistics*, **1**, 3964-3992. https://doi.org/10.18653/v1/2024.acl-long.218

[17] Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., *et al.* (2024) M4: Multi-Generator, Multi-Domain, and Multi-Lingual Black-Box Machine-Generated Text Detection. *Proceedings of the* 18*th Conference of the European Chapter of the Association for Computational Linguistics* (Volume 1: Long Papers), St. Julian's, 17-22 March 2024, 1369-1407. https://doi.org/10.18653/v1/2024.eacl-long.83

[18] Macko, D., Kopál, J., Moro, R. and Srba, I. (2025) Multisocial: Multilingual Benchmark of Machine-Generated Text Detection of Social-Media Texts. *Proceedings of the* 63*rd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Vienna, 27 July-1 August 2025, 727-752. https://doi.org/10.18653/v1/2025.acl-long.36

[19] Gehrmann, S., Strobelt, H. and Rush, A. (2019) GLTR: Statistical Detection and Visualization of Generated Text. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*: *System Demonstrations*, Florence, 28 July-2 August 2019, 111-116. https://doi.org/10.18653/v1/p19-3019

[20] Zellers, R., Holtzman, A., Rashkin, H., *et al.* (2019) Defending against Neural Fake News. *Proceedings of the* 33*rd International Conference on Neural Information*, Vancouver, 8-14 December 2019, 9051-9062.

[21] Mitchell, E., Lee, Y., Khazatsky, A., *et al.* (2023) DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature. *Proceedings of the* 40*th International Conference on Machine Learning*, Honolulu, 23-29 July 2023, 24950-24962.

[22] Bao, G., Zhao, Y., Teng, Z., *et al.* (2024) Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. *Proceedings of the Twelfth International Conference on Learning Representations*, Vienna, May 2024, 1-9.

[23] Venkatraman, S., Uchendu, A. and Lee, D. (2024) GPT-Who: An Information Density-Based Machine-Generated Text Detector. *Findings of the Association for Computational Linguistics*: *NAACL* 2024, Mexico, 16-21 June 2024, 103-115. https://doi.org/10.18653/v1/2024.findings-naacl.8

[24] Ippolito, D., Duckworth, D., Callison-Burch, C. and Eck, D. (2020) Automatic Detection of Generated Text Is Easiest When Humans Are Fooled. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, 5-10 July 2020, 1808-1822. https://doi.org/10.18653/v1/2020.acl-main.164

[25] Welleck, S., Kulikov, I., Roller, S., *et al.* (2019) Neural Text Generation with Unlikelihood Training. arXiv:1908.04319.

[26] Megías, A.J.G., Ureña-López, L.A. and Martínez-Cámara, E. (2024) The Influence of the Perplexity Score in the Detection of Machine-Generated Texts. *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, Lancaster, 29-30 July 2024, 80-85.

[27] Holtzman, A., Buys, J., Du, L., *et al.* (2020) The Curious Case of Neural Text Degeneration. *Proceedings of the International Conference on Learning Representations* (*ICLR*), April 2020.

[28] Uchendu, A., Le, T., Shu, K. and Lee, D. (2020) Authorship Attribution for Neural Text Generation. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Online, 16-20 November 2020, 8384-8395. https://doi.org/10.18653/v1/2020.emnlp-main.673

[29] He, P., Liu, X., Gao, J., *et al.* (2020) DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. arXiv:2006.03654.

[30] Kuznetsov, K., Tulchinskii, E., Kushnareva, L., Magai, G., Barannikov, S., Nikolenko, S., *et al.* (2024) Robust AI-Generated Text Detection by Restricted Embeddings. *Findings of the Association for Computational Linguistics*: *EMNLP* 2024, Miami, 12-16 November 2024, 17036-17055. https://doi.org/10.18653/v1/2024.findings-emnlp.992

[31] Zhi, L., Fang, L. and Cai, M. (2025) Efficient AI-Generated Text Detection Based on Contrastively Enhanced Hybrid Features and Support Vector Machine. 2025 2*nd International Conference on Intelligent Perception and Pattern Recognition* (*IPPR*), Chongqing, 15-17 August 2025, 386-391. https://doi.org/10.1109/ippr66507.2025.11198205

[32] Hao, W., Li, R., Zhao, W., Yang, J. and Mao, C. (2025) Learning to Rewrite: Generalized LLM-Generated Text Detection. *Proceedings of the* 63*rd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Vienna, Vienna, 27 July-1 August 2025, 6421-6434. https://doi.org/10.18653/v1/2025.acl-long.322

[33] Jiao, K., Wang, Q., Zhang, L., Guo, Z. and Mao, Z. (2025) M-Rangedetector: Enhancing Generalization in Machine-Generated Text Detection through Multi-Range Attention Masks. *Findings of the Association for Computational Linguistics*: *ACL* 2025, Vienna, 27 July-1 August 2025, 8971-8983. https://doi.org/10.18653/v1/2025.findings-acl.469

[34] Guo, Z. and Yu, S. (2023) AuthentiGPT: Detecting Machine-Generated Text via Black-Box Language Models Denoising. arXiv:2311.07700.

[35] Pu, X., Zhang, J., Han, X., Tsvetkov, Y. and He, T. (2023) On the Zero-Shot Generalization of Machine-Generated Text Detectors. *Findings of the Association for Computational Linguistics*: *EMNLP* 2023, Singapore, 6-10 December 2023, 4799-4808. https://doi.org/10.18653/v1/2023.findings-emnlp.318

[36] Sadiq, S., Aljrees, T. and Ullah, S. (2023) Deepfake Detection on Social Media: Leveraging Deep Learning and Fasttext Embeddings for Identifying Machine-Generated Tweets. *IEEE Access*, **11**, 95008-95021. https://doi.org/10.1109/access.2023.3308515

[37] Yan, J., Zhao, W. and Guo, H. (2025) A Lightweight Detector: Zero-Shot Detection of Machine-Generated Text with Once Call. 2025 5*th International Conference on Artificial Intelligence, Big Data and Algorithms* (*CAIBDA*), Beijing, 20-22 June 2025, 1331-1334. https://doi.org/10.1109/caibda65784.2025.11183347

[38] Hans, A., Schwarzschild, A., Cherepanova, V., *et al.* (2024) Spotting LLMs with Binoculars: Zero-Shot Detection of Machine-Generated Text. arXiv:2401.12070.

[39] Feng, W., Guo, X., He, Y., Huang, H., Ma, C., Zhang, S., *et al.* (2024) Detective: Detecting AI-Generated Text via Multi-Level Contrastive Learning. *Advances in Neural Information Processing Systems*, **37**, 88320-88347. https://doi.org/10.52202/079017-2802

[40] Fu, Y., Xiong, D. and Dong, Y. (2024) Watermarking Conditional Text Generation for AI Detection: Unveiling Challenges and a Semantic-Aware Watermark Remedy. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 18003-18011. https://doi.org/10.1609/aaai.v38i16.29756

[41] Yang, X., Chen, K., Zhang, W., *et al.* (2023) Watermarking Text Generated by Black-Box Language Models. arXiv:2305.08883.

[42] Hou, A., Zhang, J., He, T., Wang, Y., Chuang, Y., Wang, H., *et al.* (2024) Semstamp: A Semantic Watermark with Paraphrastic Robustness for Text Generation. *Proceedings of the* 2024 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies* (Volume 1: Long Papers), Mexico, 16-21 June 2024, 4067-4082. https://doi.org/10.18653/v1/2024.naacl-long.226

[43] Piet, J., Sitawarin, C., Fang, V., Mu, N. and Wagner, D. (2025) Markmywords: Analyzing and Evaluating Language Model Watermarks. 2025 *IEEE Conference on Secure and Trustworthy Machine Learning* (*SaTML*), Copenhagen, 9-11 April 2025, 68-91. https://doi.org/10.1109/satml64287.2025.00012

[44] Huang, B., Su, D., Sun, F., Cao, Q., Shen, H. and Cheng, X. (2025) Low-Entropy Watermark Detection via Bayes' Rule Derived Detector. *Findings of the Association for Computational Linguistics*: *ACL* 2025, Vienna, 27 July-1 August 2025, 14330-14344. https://doi.org/10.18653/v1/2025.findings-acl.739

[45] Xu, Y., Liu, A., Hu, X., *et al.* (2025) Mark Your LLM: Detecting the Misuse of Open-Source Large Language Models via Watermarking. arXiv:2503.04636.

[46] Wang, L., Yang, W., Chen, D., *et al.* (2023) Towards Codable Watermarking for Injecting Multi-Bits Information to LLMs. arXiv:2307.15992.

[47] Zhao, N., Chen, K., Zhang, W. and Yu, N. (2025) Performance-Lossless Black-Box Model Watermarking. *IEEE Transactions on Dependable and Secure Computing*, 1-17. https://doi.org/10.1109/tdsc.2025.3622253

[48] Macko, D., Moro, R., Uchendu, A., Srba, I., Lucas, J.S., Yamashita, M., *et al.* (2024) Authorship Obfuscation in Multilingual Machine-Generated Text Detection. *Findings of the Association for Computational Linguistics*: *EMNLP* 2024, Miami, 12-16 November 2024, 6348-6368. https://doi.org/10.18653/v1/2024.findings-emnlp.369

[49] Koike, R., Kaneko, M. and Okazaki, N. (2024) Outfox: LLM-Generated Essay Detection through In-Context Learning with Adversarially Generated Examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 21258-21266. https://doi.org/10.1609/aaai.v38i19.30120

[50] Przybyła, P., McGill, E. and Saggion, H. (2025) Attacking Misinformation Detection Using Adversarial Examples Generated by Language Models. *Proceedings of the* 2025 *Conference on Empirical Methods in Natural Language Processing*, Suzhou, 4-9 November 2025, 27614-27630. https://doi.org/10.18653/v1/2025.emnlp-main.1405

[51] Teja, L.S., Yadagiri, A., Chunka, C., *et al.* (2025) Fine-Grained Detection of AI-Generated Text Using Sentence-Level Seg-Mentation. arXiv:2509.17830.

[52] Jiang, L., Wu, D. and Zheng, X. (2025) Sendetex: Sentence-Level AI-Generated Text Detection for Human-AI Hybrid Content via Style and Context Fusion. *Proceedings of the* 2025 *Conference on Empirical Methods in Natural Language Processing*, Suzhou, 4-9 November 2025, 5287-5302. https://doi.org/10.18653/v1/2025.emnlp-main.268

[53] Corizzo, R. and Leal-Arenas, S. (2023) One-GPT: A One-Class Deep Fusion Model for Machine-Generated Text

Detection. 2023 *IEEE International Conference on Big Data* (*BigData*), Sorrento, 15-18 December 2023, 5743-5752. https://doi.org/10.1109/bigdata59044.2023.10386674

[54] Li, X., Yin, Z., Tan, H., Jing, S., Su, D., Cheng, Y., *et al.* (2025) PRDetect: Perturbation-Robust LLM-Generated Text Detection Based on Syntax Tree. *Findings of the Association for Computational Linguistics*: *NAACL* 2025, Albuquerque, 29 April-4 May 2025, 8290-8301. https://doi.org/10.18653/v1/2025.findings-naacl.464

[55] Bethany, M., Wherry, B., Bethany, E., *et al.* (2024) Deciphering Textual Authenticity: A Generalized Strategy through the Lens of Large Language Semantics for Detecting Human vs. Machine-Generated Text. 33*rd USENIX Security Symposium* (USENIX Security 24), Philadelphia, 14-16 August 2024, 5805-5822.

[56] Wang, P., Li, L., Ren, K., Jiang, B., Zhang, D. and Qiu, X. (2023) SeqxGPT: Sentence-Level AI-Generated Text Detection. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 1144-1156. https://doi.org/10.18653/v1/2023.emnlp-main.73