对抗微调为核心的 AI 文本对齐增强生成综述

宋安洋（24210240291）

摘　要　近年来，大语言模型（LLM）迅猛发展，在新闻创作、小说写作等诸多领域得到了广泛应用。LLM 对齐程度的加深，使得 AI 生成文本越来越难以被和人类文本区分，这导致了滥用问题如虚假新闻、网络诈骗等的加剧。本文首先概述了两类 AI 文本检测方法——基于 RoBERTa 的神经网络微调方法、基于指标的统计分析方法。本文指出，基于指标的方法尽管泛化能力较强，但本质是一种走捷径的方法，且缺少发展性，随 LLM 发展的发展效果不断变差；微调 RoBERTa 的方法上限检测能力很强，但易过拟合，泛化检测能力高度依赖数据集质量。从而本文点明持续改进数据集质量的重要性。接着本文概述了过去研究中 AI 文本检测数据集的发展脉络，指出了过去数据集仅关注于数据整体分布多样性，却存在着单一文本生成过程过于简单的致命缺陷，结合 HC-Var 的理论发现，本文指出在数据集构建过程中引入对齐增强生成的迫切性与必要性。然后本文简要介绍了包括本文期望提出的对抗微调在内的多种对齐方法。最后本文全面总结了上述目前研究所面对的问题及解决方案，并重点强调了我们计划通过“攻”——AI 文本对齐增强生成构建数据集，并在此基础上期望通过“攻防双边验证”，双角度的增强反过来证明我们“攻”的有效性，并期望由此带来实际应用贡献，增强应用到实际应用中的用户体验与友好度。

关键词　大语言模型；自然语言处理；机器文本生成；机器文本检测；强化学习；对抗学习；对齐；基准测试；数据集评估；深度学习

Abstract　In recent years, Large Language Models (LLMs) have developed rapidly and been widely used in fields like news and novel writing. Deepened LLM alignment has made AI-generated text increasingly hard to distinguish from human text, escalating abuse issues such as fake news and online fraud. This paper first outlines two AI text detection methods: RoBERTa-based neural network fine-tuning and metric-based statistical analysis. It notes that while metric-based methods have strong generalization, they are essentially shortcut approaches, lack development potential, and their effectiveness declines with LLM advancement; RoBERTa fine-tuning has high detection upper limits but is prone to overfitting, with its generalization highly dependent on dataset quality. Thus, the paper emphasizes the importance of continuously improving dataset quality. Next, it summarizes the development of AI text detection datasets in previous studies, pointing out that past datasets only focused on overall data distribution diversity but had a fatal flaw: overly simplistic single text generation processes. Combined with HC-Var theoretical findings, it highlights the urgency and necessity of introducing alignment-enhanced generation in dataset construction. Then, the paper briefly introduces various alignment methods, including the adversarial fine-tuning proposed in this study. Finally, it comprehensively summarizes the problems and solutions of current research, emphasizing that we plan to construct datasets through "attack"—AI text alignment-enhanced generation. On this basis, we aim to use "offensive-defensive bilateral verification" to prove the effectiveness of our "attack" via mutual enhancement from two perspectives, expecting to make practical application contributions and improve user experience and friendliness in real-world use.

Keywords　Large Language Model; Natural Language Processing; Machine Generated Text; Machine Text Detection; Reinforcement Learning; Adversarial Learning; Alignment; Benchmarking; Dataset Evaluation; Deep Learning

1 引言

近年来，大语言模型（Large Language Models, LLMs）发展迅猛，其语言理解能力、指令执行精准度以及生成复杂文本以完成多样任务的能力均取得了显著提升。这些技术进步使得 LLM 在日常生活中得到了广泛应用，加速了内容创作的自动化与便捷化，涵盖广告标语创作[1]、新闻报道[2]、小说写作[3]、代码生成[4,5]等在内的诸多领域。LLM 在文本生成质量上的巨大突破，同时也意味着生成的内容在语法、流畅度和表现力上都越来越难以与人类文本区分[6,7]。

而这种强大的能力带来了显著的滥用风险。Hanley 等人[8]的最新研究显示，在 2022 年 1 月 1 日到 2023 年 5 月 1 日期间，主流网站上 AI 生成新闻的占比已上升约 55.4%，令人震惊的是，在因传播不实谣言而声名狼藉的网站上，AI 生成新闻的占比飙升了约 457%。一方面，LLM 本身就存在较为严重的幻觉问题，极易受到训练数据集中虚假伪造信息、过时信息的影响[9]，导致错误的知识传播，负面影响已波及大量领域，其中对教育[10]、法律[11]、生物学[12]和医学[13]等高度依赖循证事实的领域，造成的负面影响尤为显著。另一方面，LLM 的生成高度依赖于提示词，被恶意利用时，生成虚假不实信息的操作门槛极低，且几乎无需额外成本，并且 LLM 生成的内容往往具备严密的逻辑结构与较强的说服力，极易对普通受众产生误导。这些虚假信息的危害已渗透到多个社会领域，不仅仅被用于炮制虚假新闻[14,15]、扩散网络谣言[16,17]，还成为垃圾邮件[18]、学术不端[19,20]、网络诈骗[21,22]等行为的工具，小则违背社会道德良俗，大则触及法律红线、构成违法犯罪。在学术领域，LLM 的应用引发了尤为广泛的关注与争议，导致了严重的学术信任危机[19,20]。有报告显示，仅 2023 年一年就有超过 60000 篇学术论文中包含了明显的 LLM 生成内容[23]，这一数据充分表明机器生成文本已深度渗透至学术出版体系。学生群体在作业中使用 LLM 同样普遍[24,25]，尽管在某些情境下的使用尚不构成法律或道德问题，但这一行为已引发担忧——教育过程的真正意义可能被削弱，同时也迫使教师必须面对并评判 AI 生成的作业内容。此外，当前大语言模型的训练高度依赖互联网收集的大规模语料，而 Villalobos 等人[26]指出，这类可用数据预计将在 2028 年枯竭。更关键的是，当前互联网已经充斥着大量人工智能生成内容，其中相当一部分质量低劣。这意味着，未来更高级大语言模型的训练数据中，将不可避免混入大量此类 AI 生成文本。而实际上，这一趋势已经在当前的网络语料中显现。Alemohammad 等人[27]的研究进一步证实，这种模型“自我消费”（self-consuming）的现象会导致其能力的显著退化。他们为此提出“模型自噬障碍”（Model Autophagy Disorder, MAD）这一警示性概念，明确指出当生成模型在迭代升级中逐步依赖自身生成的数据训练时，其质量与多样性将不可避免地下降。

值得注意的是，M4[28]曾通过实验表明人类在区分文本由 LLM 生成还是人类书写时，表现并不理想。一方面，随着 LLM 技术发展，其类人对齐能力不断完善，生成的文本与人类创作的差异越来越小；另一方面，恶意使用者可通过调整解码、采样策略，轻易误导文本阅读者[29]。因此，在此背景下，开发并部署高效、鲁棒的 LLM 生成文本检测系统变得至

关重要，以应对其在信息传播中可能带来的潜在危害。

2 机器生成文本检测方法

目前，人工智能领域的机器文本检测方法主要可以分为两大类，第一类是基于神经网络训练的方法，第二类是基于指标的统计分析方法。本章节中，2.1、2.2 小节分别简要介绍了这两类方法的做法及其中的经典方法，2.3 小节中总结探讨了这些方法存在的问题和缺陷。

2.1 基于神经网络训练的方法

对于基于神经网络训练的方法，一般来说，用于微调的预训练分类模型默认仅选用 RoBERTa[30]。BERT[31]类模型曾引领 LLM 浪潮前自然语言处理（NLP）旧范式的最后一次技术革新浪潮。BERT 类具备强大的语义提取能力与文本分类能力，至今仍是微调神经网络方法在文本分类任务中的核心方案。在 BERT 的各类变体中，RoBERTa 在稳定性与准确率上表现突出，成为当前最主流的选择，甚至在多数场景下是唯一被采用的模型。在单一文本输入的简单文本分类任务中（如最基础的情感分类任务），不需要也尚无真正有效的方法能通过增加额外网络结构、处理 RoBERTa 提取的特征来提升分类效果。直接微调 RoBERTa 进行分类既是该类任务的 baseline 也是表现最佳的方案，此时工作重点更多集中在数据集的构建上。本文所研究的机器文本检测，正属于这类最基础的文本二分类任务。对于本文所研究的机器文本检测，事实上，M4[28]也曾基于他们的实验结果直接指出，在他们的数据集上训练，微调后的 RoBERTa 是他们所选检测器中表现最好的。

2.2 基于指标的统计分析方法

对于基于指标的统计分析方法，这类方法同时属于零样本学习(zero-shot)方法。自 GPT[32]诞生以来，基于指标的这类统计方法，本质上都是基于待检测文本在所选取的 LLM 视角下的生成概率或困惑度（Perplexity, PPL），基于一定的统计分析方法，得到这些方法认为的一般性规律，利用这个规律对人类/机器文本进行二分类。GLTR[33]是这类统计方法的开山之作（自 GPT 诞生）,对输入文本的每个 token，分析如果 GPT2[34]要生成它的话，这个 token 在 GPT2 视角下生成概率排名在词表的第几名。统计排名前十的 token 占比，再依据开发集验证设定的阈值作为最终一般性规律的阈值判断文本来源，高于阈值则认为是机器生成的，低于阈值则认为是人类文本，通过在开发集上最好的验证效果下设定的阈值作为最终一般性规律的阈值。但是，常用词在人类与 LLM 生成文本中占比均高，仅依靠 token 概率排名判断并不合理。于是，在 GLTR 之后的统计分析方法大都是基于 PPL 改造得到的指标来进行检测。

$$PPL = \exp(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i \mid w_1, \dots, w_{i-1})) \tag{1}$$

其中，N 为序列长度, w_i是 i 处的 token。-log PPL 可以理解为长度正则的序列对数概率（log probability）。PPL 是大模型预训练时的评估指标。DetectGPT[35] 首次将 PPL 用于检测，通过对文本 token 进行随机替换（即 “扰动”），发现机器文本扰动后的 PPL 会显著上升，人类文本 PPL 变化幅度小且方向不确定。据此定义扰动差$\mathbf{d}(x, p_\theta, q)$（式(2)），并依据

开发集设定阈值实现分类。

$$\mathbf{d}(x, p_\theta, q) \triangleq \log p_\theta(x) - \mathbb{E}_{\tilde{x}\sim q(\cdot|x)} \log p_\theta(\tilde{x}) \tag{2}$$

其中，$x \sim p_\theta$是文本分布，$\tilde{x} \sim q(\cdot \mid x)$为扰动后文本分布。

DetectGPT 之后，DetectLLM[36]、LLMDet[37]、DNAGPT[38]、Fast-DetectGPT[39]、Ghostbuster[40]、Binoculars[41]等方法陆续出现，他们不断解决了一些问题，如模型溯源、黑盒检测等。Binoculars 将检测能力推动到了一个高峰，他尝试尽可能解决两类误判问题：一类是过于完美、正式的人类文本（例如人类学 LLM 说话，PPL 偏低导致被误分类为机器文本）；另一类是对生僻复杂问题的机器回答文本（生僻词在 LLM 中总体生成概率是很小的，直接导致 PPL 偏高被误分类为人类文本）。Binoculars 指出直接对 PPL 这个概念从形象化的角度来理解并进行分析思考，PPL 越大表示 LLM 对输入文本的困惑程度越高，即越认为是人类文本。对高频常见词，不同 LLM 眼中的概率是基本完全一样的；对低频生僻词，不同 LLM 眼中的概率分布会有相对更加明显许多的不同。于是 Binoculars 提出了由两个 LLM 构成的交叉困惑度 X-PPL，并用 X-PPL 惩罚式中$\mathcal{M}_1$视角下的 PPL，得到 Binoculars 分数（Binoculars Score）$B_{\mathcal{M}_1,\mathcal{M}_2}(s)$作为指标，用于检测。公式如下：

$$\log \text{X-PPL}_{\mathcal{M}_1,\mathcal{M}_2}(s) = -\frac{1}{L}\sum_{i=1}^{L} \mathcal{M}_1(s)_i \cdot \log(\mathcal{M}_2(s)_i) \tag{3}$$

$$B_{\mathcal{M}_1,\mathcal{M}_2}(s) = \frac{\log \text{PPL}_{\mathcal{M}_1}(s)}{\log \text{X-PPL}_{\mathcal{M}_1,\mathcal{M}_2}(s)} \tag{4}$$

其中，$\mathcal{M}_1$、$\mathcal{M}_2$分别是两个大模型，$\mathcal{M}(s)_i$表示在位置 i 处的 token i 的生成概率。Binoculars 分数在 PPL 的基础上除以了 X-PPL 进行惩罚，直观来说就是看$\mathcal{M}_1$和人类观点的分歧相比于$\mathcal{M}_1$和$\mathcal{M}_2$观点的分歧是否足够更大或者足够更小。

2.3 存在的问题

目前，上述两类检测方法内，各自存在着共性的缺陷。对于第一类方法，BERT 类模型擅长捕捉文本中的语义信息，进而尝试挖掘并利用人机文本的本质差异提升检测效果，但 BERT 类模型易过拟合、泛化能力弱，数据分布发生明显变化时，检测效果会显著下滑；第二类方法虽具备强泛化能力且支持零样本学习，但其本质均依赖 PPL，并没有对文本语义进行直接的分析与检测，是一种走捷径的方法，可靠性存疑。而正如 Ippolito 等人[29]所指出的，调整解码或采样策略就能轻易规避这类检测。Binoculars[41]也在他们的实验中进行了相关的报告，他们对于高度随机采样的机器文本，检测准确率几乎为 0。尽管过度随机的机器文本是否有被检测的必要有待商榷——一方面过度随机的文本毫无用处，另一方面人类也能轻易地直接识别出来。但 Ippolito 等人[29]也提到，合适的采样策略本就可以轻易欺骗人类——这意味着，只要能搜索出这种合适的采样策略，不仅能绕过统计分析类检测方法，还能误导人类判断，只是这样的搜索实际可执行性极差。此外，这种合适的采样策略其实与 LLM 的进化方向一致：预训练阶段虽然关注语料学习与 PPL 下降，但微调阶段更侧重指令学习、类人对齐及特定任务准确率，PPL 反而可能明显上升。随着大模型发展，基于 PPL 的统计分析方法，检测效果会持续变差，且只要依赖 PPL，就无法实现本质改进。当前的核心困境

在于，研究者尚未明确人机文本的本质差异，这种差异也可能并非单一、可简单描述的特征。**因此**，目前只能通过微调 BERT，从端到端的角度尝试捕捉并利用这种差异，而**这实质上需要持续提升用于微调的数据集的质量。**也因此，近年来，越来越多的研究聚焦于收集 LLM 生成文本，构建相应的数据集，并展开进一步的研究与分析。

3 机器生成文本检测数据集

本章节详细介绍了机器生成文本检测领域数据集/Benchmark 的发展脉络以及其中一些重要的工作，并在本章节的最后简要点明了目前对数据集的研究存在的急需改进的缺陷与不足。

数据集	原始大小	多领域	多生成器	多语言	对抗攻击	多采样参数
TuringBench[42]	200k	×	√	×	×	×
RuATD[43]	215k	√	√	×	×	×
HC3[44]	26.9k	√	×	√	×	×
MGTBench[45]	2817	√	√	×	√	×
CHEAT[46]	50k	×	×	×	√	×
MULTITuDE[47]	74.1k	×	√	√	×	×
AuText2023[48]	160k	√	×	√	×	×
M4[28]	122k	√	√	√	×	×
CCD[49]	467k	×	×	√	√	×
IMDGSP[50]	29k	×	√	×	×	×
HC-Var[51]	145k	√	×	×	×	×
MIXSET[52]	3600	√	√	×	√	×
HC3 Plus[53]	210k	√	×	√	×	×
MAGE[54]	447k	√	√	×	×	×
RAID[55]	570k	√	√	×	√	√
Beemo[56]	19.6k	√	√	×	√	×
RealDet[57]	836k	√	√	√	√	×

表 1 机器生成文本检测领域已公开发表数据集的对比

表 1 展示的是机器生成文本检测领域已公开发表数据集的对比，从宏观角度展示了过去工作中对数据集/Benchmark 的研究的发展脉络。其中，原始数据集大小（original size of raw text）的对比方式是 RealDet[57]提出的，表示人类来源文本加机器生成的文本的总量，不包含机器生成文本被规则化对抗攻击简单改写的文本，如简单地加了几个随机空格就不算

作一条新的原始数据，这主要影响到 RAID[55]数据集的大小对比，相比 RAID[55]宣称的 6M 条文本，RealDet[57]明确指出 RAID[55]的原始大小是 570k 条。

TuringBench[42]是自 GPT 诞生以来已公开发表的最早用于检测大语言模型生成文本的 Benchmark，该工作主要面向虚假新闻泛滥的问题，收集了 10k 条人工撰写的美国 CNN 英文新闻文章，使用包含 GPT-1、GPT-2、GPT-3 在内的 19 个早期 LLM 生成了共计约 190k 条机器生成新闻，构建了旨在探索图灵测试（即人类/机器文本二分类问题）的数据集。显而易见地是，TuringBench 数据集中的文本十分单一，于是后来的工作不断尝试扩充数据多样性，这个发展过程在 M4[28]达到了第一个高峰。

M4[28]首次系统提出了多生成器（Multi-Generator）、多领域（Multi-Domain）、多语言的概念（Multi-Lingual），并将其应用于 M4 数据集的构建，以扩充数据集的多样性，并论证了其意义。M4 的实验初步表明了微调神经网络类检测器，泛化到它们在训练集上没见过的生成器或领域，是表现不佳的、具有挑战性的。M4 的设置用于微调神经网络类检测器有助于提高其泛化表现。受启发于 M4，MAGE[54]对 M4 的多样性进行了进一步地扩充，并系统提出了一个测试平台，专门面向于泛化检测问题。MAGE 系统地提出了 8 种泛化检测设置，分别为：1.固定领域&固定生成器 2.任意领域&固定生成器 3.固定领域&任意生成器 4.任意领域&任意生成器 5.未见过的生成器 6.未见过的领域 7.未见过的领域&未见过的生成器 8.释义攻击。

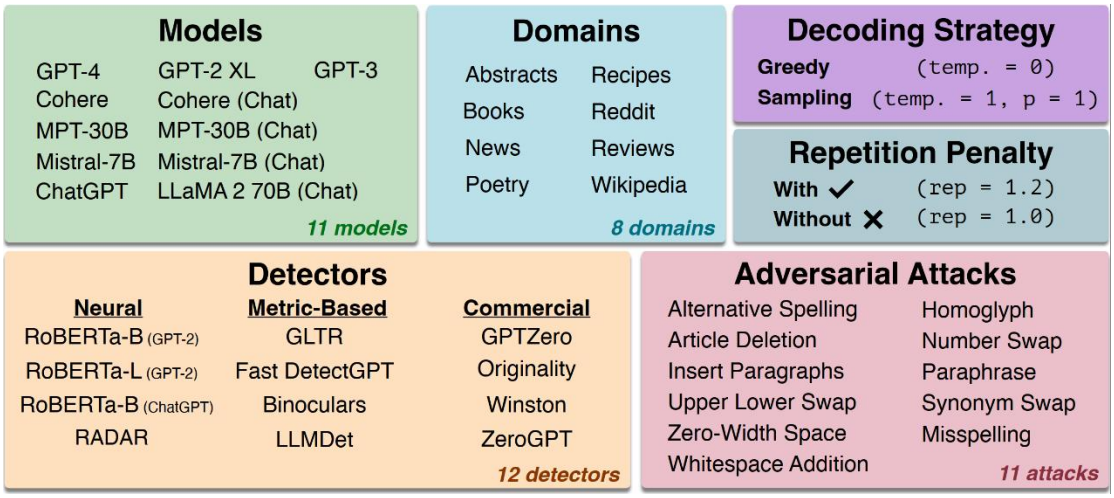

图 1 RAID 数据集与 Benchmark 结构图概览

在这之后不久，RAID[55]将数据集的质量推动到了一个新的高峰。如图 1 所示，在 M4 的基础上，RAID 引入了对生成解码采样参数多样性的考量，并扩充了大量相对常见且易于执行的基于规则的攻击手法，以测试检测器的鲁棒性。RealDet[57]对 RAID 进一步扩充，通过补充了中文文本，来弥补了 RAID 在多语言上的不足，但是 RealDet 没有对采样参数多样性进行探讨和研究。

在此期间的其他工作也提出、分析并尝试解决一些其他有意思的问题。RuATD[43]和

AuText2023[48]是竞赛数据集，RuATD 专门面向俄语检测问题，AuText2023 专门面向西班牙语检测问题。HC3[44]首次提出从词法结构、语法依存分析、情感分析等角度尝试分析探索人类/机器文本的本质差异究竟是什么，HC3 Plus[53]是其扩充版本。MGTBench[45]首次引入对抗性攻击，包括释义文本攻击、随机加入空格等少量最基本的方法。CHEAT[46]首次系统提出人机混合文本的概念，后来 MIXSET[52]和 Beemo[56]对人机混合过程进行了改进，并陆续提出了三分类和四分类任务（人类文本、机器文本、人类改写的机器文本、机器改写的人类文本）。CCD[49]首次提出机器生成代码的检测任务。IMDGSP[50]首次提出判断一篇完整的科学论文是否机器生成的检测任务。

在这之中，HC-Var[51]获得了非常有趣的重要发现，并从数学理论分析的角度进行了证明。

HC-Var 在研究跨领域泛化问题时，发现越类人对齐的文本越有利于微调得到的 RoBERTa 的泛化检测能力。例如对 P1:泛领域开放问答 P2:Reddit 帖子生成 P3:Reddit ELI5 帖子生成，从 P1 到 P3 任务越来越人性化，他们发现 P3 生成的文本上微调得到的 RoBERTa 可以取得非常好的检测效果，P2 到 P1 亦是如此，而反过来从 P1 泛化到 P2、P3，P2 泛化到 P3，效果就非常不好。

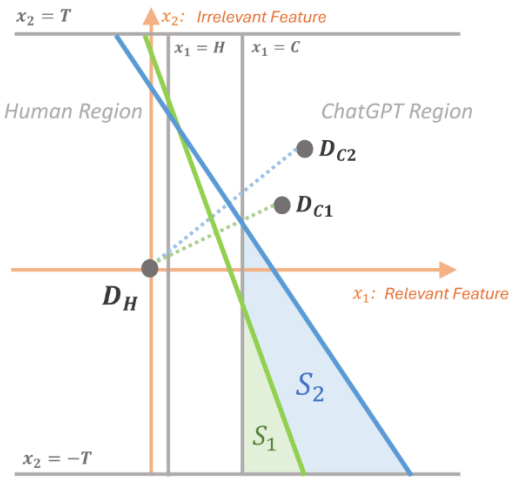

图 2 HC-Var 所提出定理的对应图解

HC-Var 对此进行了进一步的研究与证明。他们提出了两个概念定义和一条定理。定义 1：相关特征与无关特征。其中，对于无关特征，例如长度，尽管 ChatGPT 可能偏好于生成长文本，但本质上长度有多长是完全自由、无规律且高度可控的，这不是人类与机器文本间的真正差异，不应当被检测器聚焦并用于检测，HC-Var 将其定义为无关特征。对于相关特征，即人机文本本质的差异，目前的研究仍然不清楚这个差异究竟是什么，但是 HC-Var 的实验表明，更加类人对齐的机器文本在相关特征上更接近人类文本。定义 2：假阴性区域（False Negative Area, FNA），即机器文本被错误检测为人类文本的区域，如图 2 所示，S1 和 S2 即 FNA。

于是，如图 2 所示，记相关特征为x_1，无关特征为x_2，设$x_1 = C(C > 0)$的右侧区域为机器生成，$x_1 = H(H > 0)$的左侧区域为人类撰写，记人类文本符合高斯分布$\mathcal{D}_H = \mathcal{N}(0, \sigma^2 I)$，假设机器文本在以$x_1 \geq C$的区域某个中心$x_1 = \theta$同样符合高斯分布，于是对相关特征距离人类文本较近的机器文本分布$\mathcal{D}_{C1}$和距离较远的$\mathcal{D}_{C2}$可以不妨记作如下分布：

$$\begin{cases} \mathcal{D}_{C1} = \mathcal{N}(\theta_1, \sigma^2 I), ||\theta_1||_2 = d, \\ \mathcal{D}_{C2} = \mathcal{N}(\theta_2, \sigma^2 I), ||\theta_2||_2 = K \cdot d, \end{cases} \tag{5}$$

其中，$\mathrm{d} \geq C, \mathrm{K} > 1$。

记给定检测器模型 f，FNA 即被 $\mathrm{f}, \mathrm{x}_1 = C, \mathrm{x}_2 = \pm T$包围区域，其中$\mathrm{T} > 0$，T 是控制$\mathrm{x}_2$限制范围的一个阈值。记模型 f 的 FNA 为 $\Gamma(f)$。

通过一定的证明，我们可以得到定理：在数据规模控制变量的前提下，模型 f_2 的 FNA 会以明显较高概率大于模型 f_1 的 FNA 的上限，且 K 越高，成立的概率越高。具体来说，以至少$\left(1 - \left(\frac{\pi}{2} - \frac{C}{d} + \Omega(\frac{C}{d})^3\right) / \left(\frac{\pi}{2} - \frac{C}{Kd}\right)\right)$的概率，如下关系式（6）成立。

$$(\frac{\Gamma(f_2)}{sup\Gamma(f_1)})^2 \geq (1 + (K-1) \cdot \frac{1}{1+2T \cdot \Omega(1/d)}) > 1 \tag{6}$$

过去机器文本检测领域数据集构建的发展过程，核心旨在从数据集整体分布的角度进行拓宽、加大分类难度，生成过程往往是很简单的，反复套用几条或十几条相对固定的提示词模板，**缺少一定的增强。**尽管以 RAID[55]为首的许多工作高度重视了检测器鲁棒性问题，但基于规则的对抗性攻击终究难以穷尽，对泛用性的检测十分有限。且其更多的意义在于测试，而非训练。随机加入空格、随机加入换行、随机删除 a/an/the、随机删除段落文本等做法，破坏了语义结构，提高了文本的困惑度 PPL，类似于随机文本，尽管正如 Binoculars[41]所指出的那样，根据随机程度的不同，可能对基于指标的统计分析类检测方法产生极其有效的攻击。但视程度不同，人类也可能轻易地直接识别出来，是否有检测的意义有待商榷，且破坏语义结果不利于微调 RoBERTa 学习、捕捉人类与机器文本的本质差异。

因此，从类人对齐的角度对机器文本的生成过程进行增强，存在着显著的必要性与迫切性。

4 对齐生成

4.1 强化学习

提及对齐，最常见最直接最根本的手段是强化学习（Reinforce Learning，RL），因此可以将 RoBERTa 检测器作为 RL 微调 LLM 的奖励模型，通过 LLM 和 RoBERTa 的对抗微调不断提升 LLM 生成文本类人对齐程度和 RoBERTa 检测能力。但是 RoBERTa 存在着 in-distribution 过拟合的问题，直接对抗微调是不可行的，一方面我们无法筛选出无法被 RoBERTa 检测出的那不足 2%甚至近似 0%的机器文本采样，另一方面这种情况下的筛选可靠性实际并无法保证，因为根据 HC-Var 的理论，RoBERTa 会过拟合到无关特征。RADAR[58]是机器文本检测领域对抗微调的唯一已发表工作，RADAR 面向于解决得到一个高质量的释义文本检测器的问题，这也是为了使对抗微调能够实际可行。测试层面而言，释义文本被视

作一种攻击，但并不适合直接视作机器文本用于训练，其微调得到的检测器，检测机器文本其实充满着不确定性，近期工作如 MIXSET[52]、Beemo[56]已陆续提出包含人机混合文本的三分类、四分类问题。此外，RADAR 还将 RoBERTa 微调的 step 集成到了 LLM 的 PPO step 中，尽管这更加自动化地通过快速大量多轮对抗取得持续提升，但是却可能带来训练不稳定、前期 RoBERTa 经验不足导致收敛方向不正确、无法结合多生成器经验等问题，导致效果可能几乎只能与常规的 M4 扩充数据集多样性分布的构建手法微调得到的检测器相提并论。但是若将对抗微调解耦出一般形式，并**将 HC-Var 的 RoBERTa 在训练集分布外（Out-of-Distribution, OOD）主要借助本质特征进行检测的特性与对抗微调进行结合，进行跨域/跨生成器交叉奖励**，便可应用到常规的机器文本中，并解决 RADAR 所面对的所有问题。此外，还需要注意的是，之所以通过对抗微调，而不是知识蒸馏商业模型或者不妨直接仅使用商业模型，是因为需要保持模型生成风格并参与数据集构建。CogLM[59]曾通过一定的对比分析指出，不同参数量大小的 LLM 可以被视作不同年龄的人类。以 M4[28]、RAID[55]为首的诸多工作都表明了引入多种不同参数量的 LLM 对效果提升有着至关重要的必要性。

4.2 其他方法

除对抗微调外，还可通过提示词对齐、推理过程增强来对机器生成文本进行类人对齐增强。可以借助于 LLM system 角色的强力设定能力让 LLM 进行**粗粒度角色扮演**来对机器生成文本进行类人对齐增强[60,61]。可以通过专门的 **BPO 模型**[62]为提示词附加信息，使得生成的机器文本能够更好地类人对齐，这也真正意义上可以实现每条提示词各不同样，不再是完全千篇一律地简单套用提示词模板。可以通过**反思精炼**[63,64]增强推理过程来对机器生成文本进行类人对齐增强。这些对齐增强手法之间，以及**与对抗微调之间，并不冲突，可以相互结合**。并且因为增加了提示词的多样性，与对抗微调结合还**具备着相辅相成的作用**。

5 总结

本文围绕 AI 文本对齐增强生成展开综述，梳理分析机器文本检测方法、数据集发展脉络及对齐技术，厘清当前领域核心矛盾与未来突破方向。从检测方法看，基于指标的统计分析方法虽泛化性强，但本质是依赖 PPL 的 “捷径方案”，易被采样策略规避且随 LLM 对齐能力提升效果衰减；微调 RoBERTa 能捕捉文本深层语义，却受限于数据集质量，存在严重过拟合与泛化不足问题，持续改进数据集质量便十分重要。数据集层面，现有研究均聚焦于多领域、多生成器等宏观分布多样性，却忽视单文本生成过程增强，且规则化对抗攻击（如随机加空格、删词）破坏语义结构，难以支撑检测器学习人机文本本质差异；而 HC-Var 在跨域泛化检测实验中发现，类人对齐程度更高的领域的文本，可帮助 RoBERTa 聚焦相关特征（潜在人机本质差异），显著提升其泛化检测能力，并且也通过了数学理论分析进行了证明。

当前研究仍面临一些其他核心挑战。例如人机文本本质差异（相关特征）未明确，使检测模型缺乏清晰优化目标，易过拟合于长度、句式等无关特征；直接的对抗微调由于 RoBERTa 的过拟合问题因而不具备实际的可执行性。通过本文期望的 AI 文本对齐增强生成

构建得到的数据集，既可填补 AI 文本检测领域数据集构建中生成过程缺少增强的核心缺陷，又可推动对人机文本本质差异探索与研究的进一步深化。通过结合 HC-Var 理论，借助 RoBERTa 面对 OOD 数据无法依赖过拟合无关特征检测、主要借助本质特征进行检测的特性，将选取的生成器和领域分成两组，RoBERTa 检测器交叉奖励对抗可以让对抗微调具备实际可执行性。并且与多种其他对齐方法的结合，可以相辅相成进一步提升 AI 文本对齐增强生成的效果。

通过对抗微调为核心的 AI 文本对齐增强生成来构建数据集，测试现有检测器，期望现有检测器在该数据集上的检测表现明显下滑，并期望在该数据集上微调得到的 RoBERTa 检测器的泛化能力提升。单纯地难以被现有检测器检测出来，也可能是走捷径地以某种方式 hack 掉现有检测器，而通过上述“攻防双边验证”则可以反过来很好地证明本文期望的 AI 生成文本的对齐增强程度的提升，证明我们的“攻”的有效性。通过对抗微调为核心的 AI 文本对齐增强生成提供一个更高质量的数据集，并且基于该数据集微调的 RoBERTa 泛化检测性能更优，为未来学术研究提供了可复用的实践参考。最终又落脚于 AI 文本对齐增强生成，让大语言模型输出文本的整体口吻更贴近人类表达，在实际场景应用中可增强用户友好度，有效提升用户体验，带来实际应用价值。

参 考 文 献

[1] Murakami S, Hoshino S, Zhang P. Natural language generation for advertising: A survey[J]. arXiv preprint arXiv:2306.12719, 2023.

[2] Yanagi Y, Orihara R, Sei Y, et al. Fake news detection with generated comments for news articles[C]//2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES). IEEE, 2020: 85-90.

[3] Yuan A, Coenen A, Reif E, et al. Wordcraft: story writing with large language models[C]//Proceedings of the 27th International Conference on Intelligent User Interfaces. 2022: 841-852.

[4] Becker B A, Denny P, Finnie-Ansley J, et al. Programming is hard-or at least it used to be: Educational opportunities and challenges of ai code generation[C]//Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1. 2023: 500-506.

[5] Zheng Q, Xia X, Zou X, et al. Codegeex: A pre-trained model for code generation with multilingual benchmarking on humaneval-x[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023: 5673-5684.

[6] Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models[J]. ACM transactions on intelligent systems and technology, 2024, 15(3): 1-45.

[7] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.

[8] Hanley H W A, Durumeric Z. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites[C]//Proceedings of the international AAAI conference on web and social media. 2024, 18: 542-556.

[9] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM computing surveys, 2023, 55(12): 1-38.

[10] Susnjak T, McIntosh T R. ChatGPT: The end of online exam integrity?[J]. Education Sciences, 2024, 14(6): 656.

[11] Cui J, Li Z, Yan Y, et al. Chatlaw: Open-source legal large language model with integrated external knowledge bases[J]. CoRR, 2023.

[12] Piccolo S R, Denny P, Luxton-Reilly A, et al. Many bioinformatics programming tasks can be automated with ChatGPT[J]. arXiv preprint arXiv:2303.13528, 2023.

[13] Thirunavukarasu A J, Ting D S J, Elangovan K, et al. Large language models in medicine[J]. Nature medicine, 2023, 29(8): 1930-1940.

[14] Zellers R, Holtzman A, Rashkin H, et al. Defending against neural fake news[J]. Advances in neural information processing systems, 2019, 32.

[15] Zhou J, Zhang Y, Luo Q, et al. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions[C]//Proceedings of the 2023 CHI conference on human factors in computing systems. 2023: 1-20.

[16] Pagnoni A, Graciarena M, Tsvetkov Y. Threat scenarios and best practices to detect neural fake news[C]//Proceedings of the 29th International Conference on Computational Linguistics. 2022: 1233-1249.

[17] Lin S, Hilton J, Evans O. TruthfulQA: Measuring How Models Mimic Human Falsehoods[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 3214-3252.

[18] Mirsky Y, Demontis A, Kotak J, et al. The threat of offensive ai to organizations[J]. Computers & Security, 2023, 124: 103006.

[19] Stokel-Walker C. AI bot ChatGPT writes smart essays-should professors worry?[J]. Nature, 2022.
[20] Kasneci E, Seßler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education[J]. Learning and individual differences, 2023, 103: 102274.
[21] Weidinger L, Mellor J, Rauh M, et al. Ethical and social risks of harm from language models[J]. arXiv preprint arXiv:2112.04359, 2021.
[22] Ayoobi N, Shahriar S, Mukherjee A. The looming threat of fake and llm-generated linkedin profiles: Challenges and opportunities for detection and prevention[C]//Proceedings of the 34th ACM conference on hypertext and social media. 2023: 1-10.
[23] Gray A. ChatGPT" contamination": estimating the prevalence of LLMs in the scholarly literature[J]. arXiv preprint arXiv:2403.16887, 2024.
[24] Koike R, Kaneko M, Okazaki N. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(19): 21258-21266.
[25] Ma Y, Liu J, Yi F, et al. AI vs. Human--differentiation analysis of scientific content generation[J]. arXiv preprint arXiv:2301.10416, 2023.
[26] Villalobos P, Sevilla J, Heim L, et al. Will we run out of data? an analysis of the limits of scaling datasets in machine learning[J]. arXiv preprint arXiv:2211.04325, 2022, 1: 1.
[27] Alemohammad S, Casco-Rodriguez J, Luzi L, et al. Self-consuming generative models go mad[C]//The Twelfth International Conference on Learning Representations. 2023.
[28] Wang Y, Mansurov J, Ivanov P, et al. M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection[C]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 1369-1407.
[29] Ippolito D, Duckworth D, Eck D. Automatic Detection of Generated Text is Easiest when Humans are Fooled[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 1808-1822.
[30] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
[31] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.
[32] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.
[33] Gehrmann S, Strobelt H, Rush A M. GLTR: Statistical Detection and Visualization of Generated Text[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2019: 111-116.
[34] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
[35] Mitchell E, Lee Y, Khazatsky A, et al. Detectgpt: Zero-shot machine-generated text detection using probability curvature[C]//International conference on machine learning. PMLR, 2023: 24950-24962.
[36] Su J, Zhuo T, Wang D, et al. DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. 2023: 12395-12412.
[37] Wu K, Pang L, Shen H, et al. LLMDet: A Third Party Large Language Models Generated Text

Detection Tool[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. 2023: 2113-2133.

[38] Yang X, Cheng W, Wu Y, et al. DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text[C]//The Twelfth International Conference on Learning Representations.

[39] Bao G, Zhao Y, Teng Z, et al. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature[C]//The Twelfth International Conference on Learning Representations.

[40] Chaman A, Wang J, Sun J, et al. Ghostbuster: Detecting the presence of hidden eavesdroppers[C]//Proceedings of the 24th annual international conference on mobile computing and networking. 2018: 337-351.

[41] Hans A, Schwarzschild A, Cherepanova V, et al. Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text[C]//International Conference on Machine Learning. PMLR, 2024: 17519-17537.

[42] Uchendu A, Ma Z, Le T, et al. TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation[C]//2021 Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021. Association for Computational Linguistics (ACL), 2021: 2001-2016.

[43] Shamardina T, Mikhailov V, Chernianskii D, et al. Findings of the the ruatd shared task 2022 on artificial text detection in russian[J]. arXiv preprint arXiv:2206.01583, 2022.

[44] Guo B, Zhang X, Wang Z, et al. How close is chatgpt to human experts? comparison corpus, evaluation, and detection[J]. arXiv preprint arXiv:2301.07597, 2023.

[45] He X, Shen X, Chen Z, et al. Mgtbench: Benchmarking machine-generated text detection[C]//Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. 2024: 2251-2265.

[46] Yu P, Chen J, Feng X, et al. Cheat: A large-scale dataset for detecting chatgpt-written abstracts[J]. IEEE Transactions on Big Data, 2025.

[47] Macko D, Moro R, Uchendu A, et al. MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 9960-9987.

[48] Sarvazyan A M, González J Á, Franco-Salvador M, et al. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains[J]. arXiv preprint arXiv:2309.11285, 2023.

[49] Wang J, Liu S, Xie X, et al. Evaluating AIGC detectors on code content[J]. arXiv preprint arXiv:2304.05193, 2023.

[50] Mosca E, Abdalla M H I, Basso P, et al. Distinguishing Fact from Fiction: A Benchmark Dataset for Identifying Machine-Generated Scientific Papers in the LLM Era[C]//Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023). 2023: 190-207.

[51] Xu H, Ren J, He P, et al. On the Generalization of Training-based ChatGPT Detection Methods[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 7223-7243.

[52] Zhang Q, Gao C, Chen D, et al. LLM-as-a-Coauthor: Can Mixed Human-Written and Machine-Generated Text Be Detected?[C]//Findings of the Association for Computational Linguistics: NAACL 2024. 2024: 409-436.

[53] Su Z, Wu X, Zhou W, et al. Hc3 plus: A semantic-invariant human chatgpt comparison corpus[J]. arXiv preprint arXiv:2309.02731, 2023.

[54] Li Y, Li Q, Cui L, et al. MAGE: Machine-generated Text Detection in the Wild[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 36-53.

[55] Dugan L, Hwang A, Trhlík F, et al. RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 12463-12492.

[56] Artemova E, Lucas J S, Venkatraman S, et al. Beemo: Benchmark of Expert-edited Machine-generated Outputs[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025: 6992-7018.

[57] Zhu X W, Ren Y B, Cao Y N, et al. Reliably Bounding False Positives: A Zero-Shot Machine-Generated Text Detection Framework via Multiscaled Conformal Prediction[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 12298-12319.

[58] Hu X, Chen P Y, Ho T Y. Radar: Robust ai-text detection via adversarial learning[J]. Advances in neural information processing systems, 2023, 36: 15077-15095.

[59] Wang X, Yuan P, Feng S, et al. CogLM: Tracking Cognitive Development of Large Language Models[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025: 73-87.

[60] Chen N, Wang Y, Deng Y, et al. The oscars of ai theater: A survey on role-playing with language models[J]. arXiv preprint arXiv:2407.11484, 2024.

[61] Shivagunde N, Lialin V, Rumshisky A. Larger Probes Tell a Different Story: Extending Psycholinguistic Datasets Via In-Context Learning[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 2094-2107.

[62] Cheng J, Liu X, Zheng K, et al. Black-Box Prompt Optimization: Aligning Large Language Models without Model Training[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 3201-3219.

[63] Madaan A, Tandon N, Gupta P, et al. Self-refine: Iterative refinement with self-feedback[J]. Advances in Neural Information Processing Systems, 2023, 36: 46534-46594.

[64] Welleck S, Lu X, West P, et al. Generating Sequences by Learning to Self-Correct[C]//The Eleventh International Conference on Learning Representations.