arXiv:2304.01487v2 [cs.LG] 5 Apr 2023

要 ChatGPT，还是不要 ChatGPT：这就是问题！

Alessandro Pegoraro*、Kavita Kumari*、Hossein Fereidooni*、Ahmad-Reza Sadeghi*

*系统安全实验室，达姆施塔特工业大学，德国

摘要——ChatGPT 已风靡全球。随着 ChatGPT 和其他大型语言模型（LLM）的出现，人们越来越担心它们被以各种方式滥用，例如传播虚假新闻、抄袭、操纵公众舆论、作弊和欺诈。因此，区分人工智能生成的文本与人工生成的文本变得日益重要。研究人员提出了多种检测方法，从基础的二分类器到更复杂的深度学习模型。一些检测技术依赖统计特征或句法模式，而另一些则结合语义或上下文信息来提高准确率。

本研究的主要目标是全面评估 ChatGPT 检测领域最新的技术。此外，我们还评估了其他未专门声称能够检测 ChatGPT 生成内容的人工智能生成文本检测工具，以考察它们检测 ChatGPT 生成内容的表现。为开展评估，我们整理了一个基准数据集，其中包含来自 ChatGPT 和人类的提示词，涵盖医学、开放式问答和金融等领域的多样化问题，以及来自热门社交网络平台的用户生成回复。该数据集可作为评估各种技术检测 ChatGPT 生成内容表现的参考。评估结果表明，现有方法均无法有效检测 ChatGPT 生成的内容。

I. 引言

由 OpenAI 开发的 ChatGPT 引起了广泛关注，并在自然语言处理（NLP）社区及其他多个领域引发了热烈讨论。ChatGPT 是 OpenAI 于 2022年11月 推出的人工智能聊天机器人。它利用了 OpenAI GPT-3.5 和 GPT-4 系列大语言模型的强大能力。然而，ChatGPT 并非这些模型的简单延伸，而是经过了基于人类反馈的监督学习和强化学习微调过程 [7], [23]。这种迁移学习方法使 ChatGPT 能够从现有数据中学习，并针对对话应用优化其性能。它还促成了 ChatGPT 在各种具有挑战性的自然语言处理任务中取得出色表现 [5]–[11], [27], [34], [38]。媒体对 ChatGPT 的推介引发了一连串反应：新闻和媒体公司利用它优化内容创作，教师和学术界用它制定课程目标和计划，个人则用它进行跨语言内容翻译。不幸的是，和这类技术常见的情况一样，滥用也随之出现。学生用它生成项目和编程作业 [6], [8], ，而学者则用它撰写论文 [15]。恶意行为者利用它在社交媒体平台传播虚假新闻 [19], [9], ，教育机构则未经学生同意，利用它为学生提供心理健康教育 [2]，此外还有其他用途。此外，ChatGPT 还可能生成看似真实的故事，欺骗毫无防备的读者 [4], [32]。因此，开发一种高效的检测算法，以区分人工智能生成文本（尤其是 ChatGPT 生成的文本）与人工撰写文本，已吸引了众多研究人员的关注。

一般而言，使用机器学习检测人工智能生成文本主要有两种类型：黑盒检测和白盒检测。黑盒检测依赖于对语言模型的API级访问，因此限制了其检测合成文本的能力 [35]。这种方法涉及数据收集、特征提取以及构建用于检测的分类器，其中包括二元逻辑回归等简单分类器 [33]。相比之下，白盒检测可以完全访问语言模型，从而能够控制模型的行为并追踪结果 [35]。其中包括零样本检测 [25], [33], [39] ，该方法利用 GPT-*2* [33] 或 Grover [39] 等预训练生成模型，以及针对该任务进行微调的预训练语言模型。

大量研究致力于为 AI机器人生成的文本构建检测器 [8], [15], [16], [17], [20], [21], [22], [25], [29], [33], [39], [40]。此外，有些研究声称其 AI文本检测器能够区分 ChatGPT 生成的文本与人工撰写文本 [1], [3], [6], [10], [12], [13], [14], [18], [26], [28], [36], [37]。因此，我们旨在使用一个基准数据集（第 III-A 节）测试所有工具（通用 AI文本检测器以及针对 ChatGPT 生成文本的检测器）。该数据集包含 ChatGPT提示词和人类回答，涵盖不同领域。下一节将详细介绍各项工具及其功能。

本文的目标如下：

- 我们探讨了人工智能生成文本检测领域的研究，重点关注 ChatGPT 检测。我们概述了文献中提出的各种白盒和黑盒检测方案。我们还探讨了教育、学术/科学写作中的检测方案，以及在线提供的检测工具。

- 我们评估了旨在区分 ChatGPT 生成的回答与人工生成的回答的各种工具的效果，并比较了它们的准确率和可靠性。我们的

评估涵盖了声称能够检测 ChatGPT 提示词的工具，以及其他并非专门针对 ChatGPT 生成内容的人工智能生成文本检测工具。本次评估的主要目标是衡量这些工具检测 ChatGPT 生成内容的效果。我们的分析显示，检测生成文本的效果最佳的在线工具，其成功率也不到 50%，如表 I 所示。

- 我们的研究旨在激发对这一关键研究领域的进一步探讨，并推动开发更有效、更准确的人工智能生成文本检测方法。此外，我们的研究结果强调，在评估人工智能检测工具时，进行全面测试和验证至关重要。

II. 相关研究

本节概述了当前有关区分人工智能生成文本与人工撰写文本的研究。为对大多数基于自动化机器学习的合成文本检测方法进行分类，我们采用了 OpenAI 的分类体系 [33]，，该体系将这些方法分为三大类：i）简单分类器 [18], [33],; ii）零样本检测技术 [22], [25], [39], ； 以及 iii）基于微调的检测 [26]。简单分类器属于黑盒检测技术，而零样本检测技术和基于微调的检测技术则属于白盒检测技术。还有其他方法不属于这三类，但同样重要，值得考虑。这些替代方法包括：使用各种抄袭检测工具检测 ChatGPT 生成的文本 [20], ； 设计基于深度神经网络的人工智能检测工具 [6], ； 基于抽样的方法 [16], ； 以及在线检测工具 [1], [3], [10]，[13], [17], [28], [29], [36], [40]。在接下来的章节中，我们将分析上述类别中的现有方法及其他替代方法，重点关注它们检测人工智能生成文本的有效性。

A. 简单分类器

OpenAI [33], 是一家人工智能研究公司，分析了人类检测与基于机器学习的自动检测合成文本的方法。对于合成数据集的人类检测，作者发现，在 GPT-2 数据集上训练的模型往往会提高人们对 GPT-2 生成文本的感知人类特征程度。因此，他们测试了简单逻辑回归模型、零样本检测模型（见第 II-B 节）以及基于微调的检测模型（见第 II-C 节）。简单逻辑回归模型使用 TF-IDF（一元语法、二元语法及词频-逆文档频率）特征进行训练，随后针对不同的生成策略和模型参数进行分析。研究发现，简单分类器的准确率最高可达 97%。然而，对于这些模型而言，检测较短的输出比检测较长的输出更困难。Guo 等人 [18] 开展了人工评估，并比较了数据集这些数据集由 ChatGPT 和人类专家生成。研究人员分析了回答的语言和风格特征，并指出二者之间的差异。随后，作者尝试通过以下方式检测文本是由 ChatGPT 生成还是由人类撰写：首先，基于 GLTR Test-2 特征训练简单逻辑回归模型；其次，采用基于 RoBERTa [24] 的预训练深度分类器模型，用于单文本检测和问答检测（见第 II-C 节）。然而，由于数据集语料极不平衡，且未能涵盖 ChatGPT 的所有文本生成风格，所提出的检测模型无法有效区分 ChatGPT 生成文本与人工撰写文本。Kushnareva 等 [22] 的另一项研究运用拓扑数据分析（TDA），提取了三种可解释的拓扑特征，用于识别人工生成文本，例如图中连通分量、边和环的数量。随后，作者使用这些特征训练逻辑回归分类器，并在 WebText 与 GPT-2、Amazon Reviews 与 GPT-2，以及 RealNews 与 GROVER [39]的数据集上测试该方法。然而，该方法未在 ChatGPT 这一特定模型上进行测试，因此可能不适用于 ChatGPT。

B. 零样本检测技术

OpenAI [33] 还开发了一个 GPT-2 检测器，该检测器使用拥有 15 亿个参数的 GPT-2 模型，能够以 83% 至 85% 的准确率识别排名前 40 的生成结果。然而，当模型针对 Amazonreviews 数据集进行微调时，准确率降至 76%。在另一项研究中， [25],作者探讨了人工智能生成文本的零样本检测，并部署了一款在线检测工具（DetectGPT），用于区分 GPT-2 生成文本与人工撰写文本。他们利用生成模型的对数概率实现了这一目标。作者通过实验表明，人工智能生成文本位于模型对数概率函数的负曲率区域。不过，需要注意的是，作者假设能够评估所研究模型的对数概率，但这并非总是可行。此外，正如作者所指出的，这种方法仅适用于 GPT-2 提示词。

Zellers等人 [39] 使用了一个与 GPT-2 所用模型相同的 Transformer，但在文本生成过程中选择下一个词时，他们采用了核采样，而不是 Top-k采样。他们开发的模型名为 Grover，可以生成虚假新闻等文本，并检测自身生成的文本。该模型也可在线使用。作者使用 Grover、GPT- 2（124M 或 355M ）、BERT（BERT-Base 或 BERT-Large）和 FastText 验证工具，对 Grover 生成的新闻文章进行分类。他们证明，在验证其自身生成的虚假新闻方面，Grover 优于前文提到的检测器。然而，目前尚不清楚它是否适用于 GPT 模型生成的文本。此外，研究表明，在检测 GPT-2 文本时，双向 Transformer 模型 RoBERTa 的表现优于参数规模相当的 Grover 模型 [33]。

C. 基于微调的检测

在 [33], 中，作者通过基于 RoBERTa$_{BASE}$ 和 RoBERTa $_{LARGE}$构建分类器，开展了微调预训练语言模型以检测人工智能生成文本的实验。他们发现，微调 RoBERTa 的表现始终优于微调容量相当的 GPT-2 模型。然而，如 [31] 所示，该方法无法检测 ChatGPT 生成的文本。在另一项研究中，Mitrovic 等人 [26] 探讨了训练机器学习模型以区分 ChatGPT 生成的查询与人类生成的查询是否可行。作者尝试使用基于 DistilBERT 的框架检测 ChatGPT 生成的两行餐厅评论。DistilBERT 是一种基于 BERT 训练的轻量级模型，并经过了基于 Transformer 的模型微调。此外，作者还使用 SHAP 方法解释模型的预测结果。作者得出结论，机器学习模型无法成功识别 ChatGPT 生成的文本。Guo at al. [18] 还开发了一种基于 RoBERTa 的预训练深度分类器模型，用于单文本检测和问答检测。其局限性与 II-A 节所述相同。

D. 文献中提出的其他方法

大量学术研究探讨了 ChatGPT 对教育的不利影响，指出学生和学者可能利用 ChatGPT 进行抄袭。为解决这一问题，研究人员使用了 RoBERTa、Grover 或 GPT-2 等现有工具，检查教育内容是否与 ChatGPT 生成的文本重复。在[6], 作者提出了一种名为 AICheatCheck 的基于 Transformer 的模型，这是一种基于网络的人工智能检测工具，旨在识别给定文本是由人类还是 ChatGPT 生成的。AICheatCheck 会检查一个句子或一组句子中的模式，以判断其来源。作者使用了 Guo 等人收集的数据 [18]（该数据存在局限性）以及来自教育领域的数据。此外，论文并未说明 AICheatCheck 凭借何种依据或特征取得较高的准确率。

[20], 中的研究评估了两款常用抄袭检测工具 iThenticate 和 Turnitin 在检测 ChatGPT 生成的 50 篇论文是否存在抄袭方面的有效性。作者还将 ChatGPT 的表现与其自身进行比较，发现它比传统抄袭检测工具更有效。Gao 等 [15] 的另一项研究旨在使用 GPT-2 输出检测器 RoBERTa、剽窃检测器和人工审查，对 ChatGPT 生成的学术论文摘要进行比较。作者从五本高影响力医学期刊中收集了十篇研究摘要，随后让 ChatGPT 根据这些摘要的标题和期刊名称生成研究摘要。然而，研究中使用的工具无法在线获取以供验证。在近期的一项研究中，Cotton 等人 [8] 探讨了在学术领域使用 ChatGPT 的利弊，尤其是与抄袭相关的问题。在另一项研究中， [16] 的作者利用了模型所用底层文本的分布特性。他们部署了一款名为 GLTR 的工具，该工具会用不同颜色标示输入文本，以判断其真实性。研究人员使用 GPT-2 15 亿参数模型 [30] 生成的提示词以及社交媒体平台上的人工生成文章测试了 GLTR。作者还开展了一项人工研究，请学生辨别虚假新闻和真实新闻。

E. 在线工具

下面，我们将考察多种在线工具，并简要介绍其功能。

1) AI生成文本文体计量检测 [21]: 此工具通过识别文本中独特的模式和特征，利用文体计量信号来分析其写作风格。随后从输入文本中提取这些信号，以便对 AI生成的推文进行基于序列的检测。 2) ZeroGPT [40]: 此工具专为检测 OpenAI文本而开发，但检测较短文本的能力有限。 3) OpenAI 文本分类器 [28]: 这个由 OpenAI 开发并经过微调的 GPT 模型，会预测文本由包括 ChatGPT 在内的各种来源生成的可能性。不过，该工具仅适用于 1000 个字符及以上的文本，而且在判断文本是否由人工智能生成方面可靠性较低。 4) GPTZero [29]: 此分类模型用于判断文档是否由大语言模型撰写，并提供句子、段落和文档层面的预测。不过，它主要适用于英语内容，且只接受字符数在 250 至 5000 之间的文本。 5) Hugging Face [13]: Hugging Face 发布了这款用于检测 ChatGPT 生成文本的工具。不过，它往往会过度将文本判定为由 ChatGPT 撰写。 6) Perplexity（困惑度，PPL） [17]: 困惑度（PPL）是评估大语言模型（LLM）效能的常用指标，其计算方法是取文本在大语言模型下平均对数似然相反数的指数。PPL 值越低，表示语言模型对其预测越有把握。大语言模型使用大量文本语料库进行训练，从而能够学习常见的语言模式和文本结构。因此，PPL 可用于衡量给定文本在多大程度上符合这些典型特征。 7) Writefull GPT检测器 [36]: 此工具主要用于检测抄袭，也可以识别文本是否由 GPT-3 或 ChatGPT 生成。不过，该工具通过百分比判断文本是否由 AI 生成，其结果对于人类撰写的样本和 ChatGPT 生成的样本都存在一定的不确定性。 8) Copyleaks [10]: 该工具声称能够检测文本是否由 GPT-3、ChatGPT、人类或人类与 AI 共同生成。该工具仅接受字符数不少于 150 的文本。

9) Content at Scale [3]: 这是一款用于检测 ChatGPT 生成文本的在线工具。不过，该工具只能分析不超过 25000 个字符的样本。

表 I：分析论文摘要

方法	发表时间	目标模型 Grover	GPT-2	GPT-3	ChatGPT*	公开可用	免费/付费	ChatGPT 检测等能力（真正率%）	人类文本检测等能力（真负率%）
Kumarage 等 [21]	2023		✓			✓	Free	23.3	94.7
Bleumink等人 [6]	2023			✓	✓	✓	Paid	13.4	95.4
ZeroGPT [40]	2023				✓	✓	Paid	45.7	92.2
OpenAI分类器 [28]	2023				✓	✓	Free	31.9	91.8
Mitchell等人 [25]	2023		✓			✓	Free	18.1	80.0
GPTZero [29]	2023		✓	✓	✓	✓	Paid	27.3	93.5
Hugging Face [13]	2023				✓	✓	Free	10.7	62.9
Guo 等人 [18]	2023				✓	✓	Free	47.3	98.0
Perplexity（困惑度，PPL） [17]	2023				✓	✓	Free	44.4	98.3
Writefull GPT [36]	2023			✓	✓	✓	Paid	21.6	99.3
Copyleaks [10]	2023			✓	✓	✓	Paid	22.9	92.1
Cotton 等人 [8]	2023			✓	✓	×	-	-	-
Khalil 等人 [20]	2023				✓	×	-	-	-
Mitrovic 等人 [26]	2023		✓		✓	×	-	-	-
Content at Scale [3]	2022		✓	✓	✓	✓	Paid	38.4	79.8
Orignality.ai [1]	2022			✓	✓	×	Paid	7.6	95.0
Writer AI 检测器 [37]	2022			✓	✓	✓	Paid	6.9	94.5
Draft and Goal [12]	2022			✓	✓	✓	Free	23.7	91.1
Gao 等 [15]	2022				✓	×	-	-	-
Fr¨ohling 等 [14]	2021	✓	✓	✓		✓	Free	27.8	89.2
Kushnareva 等 [22]	2021	✓	✓			✓	Free	25.1	96.3
Solaiman 等 [33]	2019		✓			✓	Free	7.2	96.4
Gehrmann 等 [16]	2019		✓			✓	Free	32.0	98.4
Zellers等人 [39]	2019	✓				✓	Free	43.1	91.3

*GPT 3.5 及以上版本。

10) Originality.ai [1]：这款付费工具专为 GPT-3、GPT 3.5（DaVinci-003）和 ChatGPT 模型设计。不过，该工具仅适用于至少 100 个单词的文本，而且容易将 ChatGPT 生成的内容误判为真实内容。11) Writer AI 内容检测器 [37]：这款工具专为 GPT-3 和 ChatGPT 模型设计。不过，它的限制是每次检测最多只能检查 1500 个字符。12) Draft and Goal [12]：这款工具用于检测 GPT-3 和 ChatGPT 模型生成的内容，并支持检测英语和法语文本。不过，要有效运行，该工具要求输入文本至少包含 600 个字符。

III. 评估

本节评估能够区分人工智能生成和人工生成回复的公开可用文献、工具或代码。我们的重点是那些声称能够检测 ChatGPT 生成内容的工具。不过，我们也尽可能评估其他人工智能生成文本检测工具的性能，这些工具并未明确声称能够检测针对 ChatGPT 提示词生成的 ChatGPT 内容。为评估这些工具的有效性，我们采用了一个基准数据集（第 III-A 节），其中包含来自 ChatGPT 和人类的提示词。随后，我们测量这些工具对 ChatGPT 生成内容和人工生成内容的检测能力，并在表 I 中展示结果。

A. 基准数据集

我们通过 OpenAI API，使用 Guo 等人提出的查询提示词 [18] 来生成基准数据集。该数据集包含 58,546 条人工生成的回复和 72,966 条由 ChatGPT模型生成的回复，共计 131,512 个唯一样本，涵盖来自医学、开放领域和金融等领域的 24,322 个不同问题。此外，该数据集还纳入了来自热门社交网络平台的回复，提供了广泛的用户生成观点。为评估人工生成回复与 ChatGPT 生成回复之间的相似度，我们使用了句子转换器 all-MiniLM-L6-v2[2]。随后，我们选取相似度最高和最低的回复，组成一个基准数据集；该数据集缩减至我们生成的原始数据集的约 10%。这个基准数据集为评估不同技术检测 ChatGPT 生成内容的能力提供了标准化参照。

B. 评估指标

为衡量和比较各方法的有效性，我们采用了以下指标：

- 真正率（TPR）：此指标表示该工具检测 ChatGPT 生成文本的灵敏度。真正例（TP）是正确识别的样本总数，而假负例

[1]https://openai.com/blog/introducing-chatgpt-and-whisper-apis [2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

（FN）是未被归类为生成文本或被错误识别为人工文本的样本数。因此，真正率（TPR）$=\frac{\text{真正例}}{\text{真正例}+\text{假负例}}$。

- 真负率（TNR）：该指标表示工具检测人工撰写文本的特异性。真负例（TN）是被正确识别的样本总数，而假正例（FP）是被错误分类为由 ChatGPT 生成的样本数。因此，TNR $=\frac{TN}{TN+FP}$。

C. 评估的工具和算法

我们评估了第二节中概述的若干工具和算法。表 I 从真正率（TPR）和真负率（TNR）的角度，概述了这些工具检测 ChatGPT生成的内容的能力。我们可以看到，所评估的方法都无法持续、可靠地检测 ChatGPT生成的文本。分析表明，如表 I 所示，用于检测生成文本的效果最好的在线工具，成功率也不到 50%。

IV. 结论

本研究深入探讨了用于检测 ChatGPT生成文本的各种方法。通过全面回顾文献并考察现有方法，我们评估了这些技术区分 ChatGPT 生成的回复与人类生成的回复的能力。此外，我们还使用基准数据集测试并验证了在线检测工具和算法。该数据集涵盖金融、医学等多个主题，以及来自热门社交网络平台的用户生成的回复。我们的实验凸显了 ChatGPT 欺骗检测器的卓越能力，并进一步表明，大多数受分析的检测器倾向于将任何文本归类为人工撰写，整体真负率较高（90%），而真正率较低。这些发现对提升在线讨论的质量和可信度具有重要意义。总之，我们的结果强调，面对日益复杂的人工智能生成的内容，仍需持续努力提升文本检测技术的准确率和稳健性。

参考文献

[1] Originality.ai. https://originality.ai/，2022年。 [在线；访问日期：28-Mar-2023]。[2] 在线心理健康咨询中隐蔽使用 ChatGPT 引发伦理担忧。https://www.psychiatrist.com/news/hidden-use-of-chatgpt-in-online-mental-health-counseling-raises-ethical-concerns/，2023年。在线；访问日期：24-Mar-2023]。[3] Content at Scale。AI 检测器。https://contentatscale.ai/ai-content-detector/，2023年。 [在线；访问日期：23-Mar-2023]。[4] Yejin Bang、Samuel Cahyawijaya、Nayeon Lee、Wenliang Dai、Dan Su、Bryan Wilie、Holy Lovenia、Ziwei Ji、Tiezheng Yu、Willy Chung 等。《对 ChatGPT 在推理、幻觉和交互性方面的多任务、多语言、多模态评估》。arXiv 预印本 arXiv:2302.04023，2023年。[5] Som Biswas。《ChatGPT 与医学写作的未来》，2023年。[6] Arend Groot Bleumink 和 Aaron Shikhule。《在教育中确保 AI 诚实：识别 GPT 生成文本》，2023年。https://www.aicheatcheck.com/。

[7] Paul F Christiano、Jan Leike、Tom Brown、Miljan Martic、Shane Legg 和 Dario Amodei。基于人类偏好的深度强化学习。《神经信息处理系统进展》，第 30 卷，2017年。[8] Debby RE Cotton、Peter A Cotton 和 J Reuben Shipway。聊天与作弊：确保 ChatGPT 时代的学术诚信。《国际教育与教学创新》，第 1–12 页，2023年。[9] Luigi De Angelis、Francesco Baglivo、Guglielmo Arzilli、Gaetano Pier-paolo Privitera、Paolo Ferragina、Alberto Eugenio Tozzi 和 Caterina Rizzo。ChatGPT 与大型语言模型的兴起：人工智能驱动的信息疫情对公共卫生构成的新威胁。SSRN 收录，编号 4352931，2023年。[10] Copyleaks AI 内容检测器。Copyleaks。https://copyleaks.com/ai-content-detector，2023年。[在线；访问日期：23-Mar-2023]。[11] Michael Dowling 和 Brian Lucey。ChatGPT 用于（金融）研究：Bananarama 猜想。《金融研究快报》，第 103662 页，2023年。[12] Draft and Goal。ChatGPT - GPT-3 内容检测器。https://detector.dng.ai/，2023年。 [在线；访问日期：23-Mar-2023]。[13] Hugging Face。Hugging Face ChatGPT 检测器。https://huggingface.co/spaces/imseldrith/ChatGPT-Detection，2023年。[在线；访问日期：23-Mar-2023]。[14] Leon Fr¨ohling 和 Arkaitz Zubiaga。基于特征的自动化语言模型检测：应对 GPT-2、GPT-3 和 Grover。《PeerJ 计算机科学》，7:e443，2021年4月。https://peerj.com/articles/cs-443/#supplementary-material。[15] Catherine A Gao、Frederick M Howard、Nikolay S Markov、Emma C Dyer、Siddhi Ramesh、Yuan Luo 和 Alexander T Pearson。使用人工智能生成内容检测器、抄袭检测器和盲法人工审阅者，将 ChatGPT 生成的科学摘要与原始摘要进行比较。bioRxiv，第 2022–12 页，2022年。[16] Sebastian Gehrmann、Hendrik Strobelt 和 Alexander M Rush。GLTR：生成文本的统计检测与可视化。arXiv 预印本 arXiv:1906.04043，2019年。http://gltr.io/。[17]Biyang Guo、Xin Zhang、Ziyuan Wang、Minqi Jiang、Jinran Nie、Yuxuan Ding、Jianwei Yue 和 Yupeng Wu。使用语言特征的 ChatGPT 检测器。https://huggingface.co/spaces/Hello-SimpleAI/chatgpt-detector-ling，2023年。 [在线；访问日期：23-Mar-2023]。[18] Biyang Guo、Xin Zhang、Ziyuan Wang、Minqi Jiang、Jinran Nie、Yuxuan Ding、Jianwei Yue 和 Yupeng Wu。ChatGPT 与人类专家有多接近？对比语料库、评估与检测。arXiv 预印本 arXiv:2301.07597，2023年。https://github.com/Hello-SimpleAI/chatgpt-comparison-detection。[19] Philipp Hacker、Andreas Engel 和 Marco Mauer。监管 ChatGPT 及其他大型生成式人工智能模型。arXiv 预印本 arXiv:2302 02337，2023年。[20] Mohammad Khalil 和 Erkan Er。ChatGPT 会让你被抓到吗？重新思考抄袭检测。arXiv 预印本 arXiv:2302 04335，2023年。[21]Tharindu Kumarage、Joshua Garland、Amrita Bhat-tacharjee、Kirill Trapeznikov、Scott Ruston 和 Huan Liu。Twitter 时间线中人工智能生成文本的文体计量检测。arXiv 预印本 arXiv:2303.03697，2023年。https://github.com/TSKumarage/Stylo-Det-AI-Gen-Twitter-Timelines。[22] Laida Kushnareva、Daniil Cherniavskii、Vladislav Mikhailov、Ekate-rina Artemova、Serguei Barannikov、Alexander Bernstein、Irina Pio-ntkovskaya、Dmitri Piontkovski 和 Evgeny Burnaev。通过考察注意力图的拓扑结构进行人工文本检测。arXiv 预印本 arXiv:2109.04825，2021年。https://github.com/danchern97/tda4atd。[23] Nathan Lambert、Louis Castricato、Leandro von Werra 和 Alex Havrilla。基于人类反馈的强化学习（RLHF）图解。Hugging Face 博客，2022年。https://huggingface.co/blog/rlhf。[24] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike [Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。RoBERTa：一种稳健优化的 BERT 预训练方法。arXiv 预印本 arXiv:1907.11692，2019 年。[25] Eric Mitchell、Yoonho Lee、Alexander Khazatsky、Christopher D Manning 和 Chelsea Finn。DetectGPT：利用概率曲率进行零样本机器生成文本检测。arXiv 预印本 arXiv:2301.11305，2023年。https://github.com/eric-mitchell/detect-gpt https://detectgpt.ericmitchell.ai/。[26] Sandra Mitrovi´c、Davide Andreoletti 和 Omran Ayoub。ChatGPT 还是人类？检测并解释。解释用于检测 ChatGPT 生成短文本的机器学习模型的决策。arXiv 预印本 arXiv:2301.13852，2023年。

[27] Reham Omar、Omij Mangukiya、Panos Kalnis 和 Essam Mansour。ChatGPT 与传统知识图谱问答：现状及知识图谱聊天机器人的未来发展方向。arXiv 预印本 arXiv:2302.06466，2023年。[28] OpenAI。https://beta.openai.com/ai-text-classifier，2023年1月。[29] Edward Tian（普林斯顿大学）。GPTZero。https://gptzero.me/，2023年。 [在线；访问日期：23-Mar-2023]。[30] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。语言模型是无监督的多任务学习者。OpenAI 博客，1(8):9，2019年。[31] J¨urgen Rudolph、Samson Tan 和 Shannon Tan。ChatGPT：胡说八道的生成器，还是高等教育中传统评估的终结？《应用学习与教学杂志》，6(1)，2023年。[32] Yiqiu Shen、Laura Heacock、Jonathan Elias、Keith D Hentel、Beatriu Reig、George Shih 和 Linda Moy。ChatGPT 和其他大型语言模型是一把双刃剑，2023年。[33] Irene Solaiman、Miles Brundage、Jack Clark、Amanda Askell、Ariel Herbert-Voss、Jeff Wu、Alec Radford、Gretchen Krueger、Jong Wook Kim、Sarah Kreps 等。语言模型的发布策略及其社会影响。arXiv 预印本 arXiv:1908.09203，2019年。https://openai-openai-detector.hf.space/ https://github.com/openai/gpt-2-output-dataset/tree/master/detector。[34] Teo Susnjak。在科学文献中应用 BERT 和 ChatGPT 对莱姆病进行情感分析。arXiv 预印本 arXiv:2302.06474，2023年。[35] Ruixiang Tang、Yu-Neng Chuang 和 Xia Hu。检测大语言模型生成文本的科学方法。arXiv 预印本 arXiv:2303.07205，2023年。[36] writefull。GPT 检测器。https://x.writefull.com/gpt-detector，2023年。[在线；访问日期：23-Mar-2023]。[37] Writer.com。AI 内容检测器。https://writer.com/ai-content-detector/，2023年。 [在线；访问日期：23-Mar-2023]。[38] Yee Hui Yeo、Jamil S Samaan、Wee Han Ng、Peng-Sheng Ting、Hirsh Trivedi、Aarshi Vipani、Walid Ayoub、Ju Dong Yang、Omer Liran、Brennan Spiegel 等。评估 ChatGPT 在回答有关肝硬化和肝细胞癌问题时的表现。medRxiv，第 2023–02 页，2023年。[39] Rowan Zellers、Ari Holtzman、Hannah Rashkin、Yonatan Bisk、Ali Farhadi、Franziska Roesner 和 Yejin Choi。抵御神经网络生成的假新闻。《神经信息处理系统进展》，第 32 卷，2019年。https://rowanzellers.com/grover/。[40] ZeroGPT。https://www.zerogpt.com，2023年1月。