对抗性释义：用于人性化 AI 生成文本的通用攻击

Yize Cheng* **Vinu Sankar Sadasivan*** **Mehrdad Saberi†** **Shoumik Saha†** **Soheil Feizi**
University of Maryland, College Park
{ yzcheng, vinu, msaberi, smksaha, sfeizi } @cs.umd.edu

Project: https://github.com/chengez/Adversarial-Paraphrasing

Abstract

大型语言模型（LLMs）不断增强的能力引发了对其在 AI 生成抄袭和社会工程中的滥用的担忧。尽管已经提出了各种 AI 生成文本检测器以减轻这些风险，但许多检测器仍然容易受到如改写等简单规避技术的影响。然而，近期的检测器在抵御此类基础攻击方面表现出了更大的鲁棒性。在这项工作中，我们介绍了一种名为对抗性改写的训练无关攻击框架，它能普遍地将任何 AI 生成的文本人性化，以更有效地规避检测。我们的方法利用了一个现成的指令追随 LLM，通过 AI 文本检测器的指导来改写 AI 生成的内容，生成特定优化以避开检测的对抗性例子。广泛的实验表明，我们的攻击在多个检测系统中都具有广泛的效果和高度的可转移性。例如，与简单的改写攻击相比——讽刺地增加了在 RADAR 上 8.57% 和在 Fast-DetectGPT 上 15.03% 的真正阳性率（当假阳性为 1% 时（T@1%F））——在 OpenAI-RoBERTa-Large 指导下的对抗性改写在 RADAR 上将 T@1%F 减少了 64.49%，在 Fast-DetectGPT 上显著减少了 98.96%。在一组多样化的检测器中——包括基于神经网络的，基于水印的和零样本方法——我们的攻击在 OpenAI-RoBERTa-Large 的指导下实现了平均 T@1%F 减少 87.88%。我们还分析了文本质量与攻击成功之间的权衡，发现我们的方法可以显著降低检测率，同时文本质量仅有轻微下降。我们的对抗性设置凸显了在应对日益复杂的规避技术时需要更为健壮和有韧性的检测策略。

1 引言

近期在自然语言生成方面的进展催生了基于变压器的大型语言模型（LLMs），如 GPT [23]、Gemini [7] 和 LLaMA [21]，它们在电子邮件撰写和代码生成等广泛任务中展示了卓越的能力。这些模型能够生成流畅、连贯的文本，与人类编写的文本难以区分。然而，尽管它们表现出色，LLMs 也引发了重大安全性和伦理问题，包括与剽窃和社会工程相关的风险。

为了应对这些风险，开发可靠的 AI 生成文本检测工具已成为一项重要的研究问题。有几项研究提出训练基于神经网络的分类器来应对这一挑战 [10, 8, 25, 28, 32, 18] 。尽管通常比经过训练的检测器弱，但也引入了各种零样本检测技术 [22, 1, 6, 17] 以减少训练分类器的开销。另一个有前途的方法是对大型语言模型进行水印，这在输出文本中强制实施特定签名以便于检测 [2, 14, 16, 33, 4] 。尽管这些检测方法在特定背景下显示出前景，但关于其稳健性的担忧仍然存在。

Sadasivan 等人 [26] 和 Krishna 等人 [15] 展示了如何通过另一个 AI 模型将 AI 生成的文本改写，以规避现有 AI 检测器的检测。前者还表明，递归改写 AI 生成的内容可用于对这些系统进行压力测试，并强调了可靠检测 AI 生成文本的根本困难。这些基于改写的攻击旨在掩盖检测模型通常依赖的统计模式和特征。然而，更近期的研究表明，一些高级检测器——尤其是那些在 AI 输出的复述上训练的检测器 [8] ——可能对这类规避技术表现出更强的抵抗

*Equal contribution
†Equal contribution

Preprint. Under review.

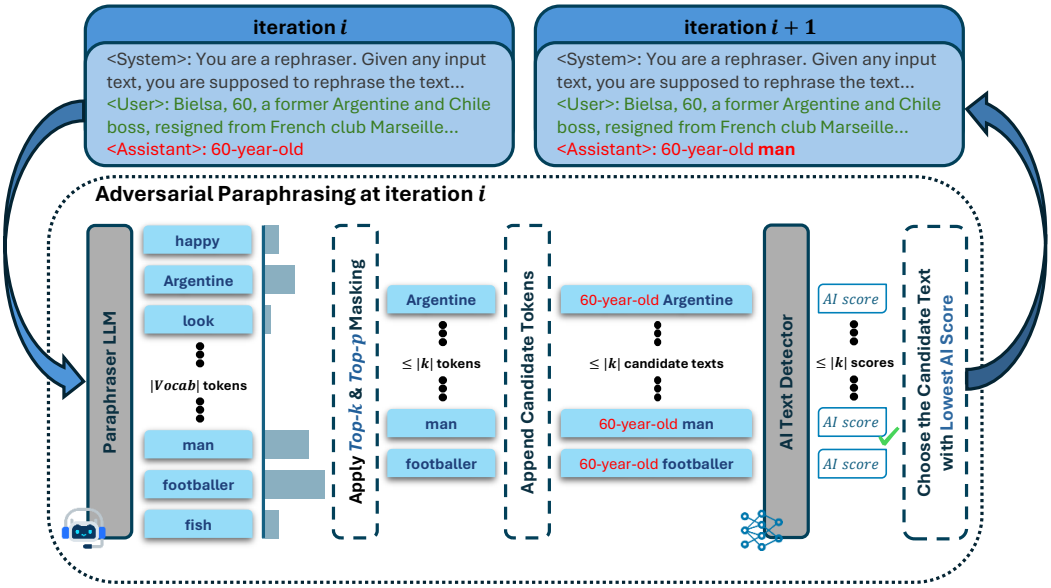

Figure 1: 人性化 AI 文本的普遍且无训练的框架概览。在对抗性释义的每一个自回归步骤中，我们从释义 LLM 采样的候选文本标记集合中搜索最具人类特征的文本标记。我们使用 AI 文本检测器的引导信号选择最具人类特征的文本标记。标记生成的迭代持续进行，直到释义完成。（即采样到 [EOS] 标记）

力。这些发展自然引发了一个迫切的问题："是否有可能开发出一种通用的攻击框架，能够稳定且有效地绕过这些稳健的 AI 生成文本检测器，并具备在各种其他检测系统中转移的能力？"在本文中，我们介绍了一种用于人性化 AI 生成文本的通用且无需训练的框架，这是一种新颖的攻击方法，旨在有效且高效地将 AI 生成的文本改写为"类似人类的文本"，与现有的规避攻击方法相比，可以更成功地避开多种检测器。我们的方法，称为对抗性复述，利用了一个现有的指令微调的大型语言模型（LLaMA-3-8B，通过自定义系统提示配置为复述者）不仅用于复述，还在训练的 AI 文本检测器指导下对文本进行对抗性人性化处理（见图 1）。在复述过程的每一步符号生成中，我们考虑由复述 LLM 提出的最可能的下一个符号。然后，与标准解码不同，我们使用一个指导性 AI 文本检测器对这些潜在接续进行评分。我们选择能使序列被检测器分类为最具人类特征的符号，这类似于目标导向的、一深度的、检测器引导的束搜索。表格 1 中展示了一些示例结果。这种方法在生成过程中直接实现了人性化的目标，属于受控文本生成 [34, 19, 3] 的范畴，其中所期望的特性是由训练的检测器判定的"人类特征"。

我们证明了我们的对抗性改写策略不仅比简单或递归改写等基线方法更有效，而且与其他方法不同，它还可以普遍移植到其他多样的检测方法中，包括基于神经网络的、零样本以及基于水印的检测器。这证明了我们框架的普遍性。例如，尽管简单改写会讽刺地使 RADAR [8] 的 1 % 误报率下真实阳性率（T@1 % F）增加 8.57 %，在 Fast-DetectGPT [1] 上增加 15.03 %，我们的由 OpenAI-RoBERTa-Large [28] 指导的对抗方法分别减少了 64.49 % 和惊人的 98.96 %。对来自三个不同类别的八个检测器的评估显示，我们的方法在 1 % 误报率下平均减少了真实阳性率 87.88 %，同时保持文本质量的最小降解。我们使用多种自动化研究全面分析了攻击和文本质量之间的权衡。

我们的核心贡献是：

- 我们引入了对抗性复述——这是一种普遍且可迁移的攻击，通过指导作为复述器的现成大型语言模型（LLM），根据现成训练的 AI 文本检测器的检测得分来对每个标记进行采样，以使 AI 生成的文本更加人性化。
- 我们进行了广泛的实验，展示了该攻击在 8 种不同类型的最先进 AI 文本检测器上的有效性和可转移性，突出了其普遍性。
- 我们使用困惑度得分以及 GPT-4o 的自动评分来评估攻击成功率与生成文本质量之间的权衡，发现与先前的规避攻击相比，我们的攻击可以显著降低检测率，同时对文本质量仅有轻微或没有降级。

我们的研究结果突显了在对抗环境中现有 AI 文本检测器的重要漏洞。

2 相关工作

AI- 文本检测。最近的研究展示了训练神经网络来区分 AI 生成的文本与人类撰写的文本的有效性 [10, 8, 25, 28, 32, 18] 。例如，Solaiman 等人 [28] 采用了基于 RoBERTa 的 [20] 分类器来区分人类撰写的文本与 GPT-2 生成的文本。在此基础上，Li 等人 [18] 通过收集一个多样化的数据集 MAGE 来训练他们的网络，提高了他们分类器在实地的表现。RADAR [8] 通过使用复述器进行迭代方式的对抗训练，使得他们的检测器对复述攻击具有鲁棒性。为了减少训练网络检测器的计算负担，提出了各种零样本检测器 [22, 1, 6, 17] 。这些检测器使用现成的 LLM 来评估候选文本的统计特性，例如其熵或对数概率评分，以进行检测。例如，DetectGPT [22] 观察到 AI 生成的文本倾向于位于对数概率曲面中的负曲率区域，而 Fast-DetectGPT [1] 通过引入条件概率曲率提高了效率。水印技术也被探索用于识别 AI 生成的文本。Kirchenbauer 等人 [14] 提出了 KGW 水印方案，该方案将标记词汇表分为绿色和红色列表，鼓励模型更多地采样绿色标记，因此在生成的文本中嵌入可检测的模式。虽然 KGW 方案使用先前生成的标记来为标记分区的随机数生成器播种，赵等人 [33] 提出了 KGW 的一个变体，即 Unigram 水印，其使用固定的标记分区进行生成，以可证明地展示鲁棒性。Kuditipudi 等人 [16] 开发了一种鲁棒的、无失真水印技术，旨在减少水印造成的分布变化。最近，Dathathri 等人 [4] 介绍了 SynthID，该方法利用锦标赛采样来为采用推测解码的 LLM 创建可扩展的水印解决方案。

对人工智能文本检测器的攻击。Sadasivan 等人 [26] 表明可以使用意译攻击来欺骗人工智能文本检测器。虽然基本的意译方法足以击败早期的零样本和经过训练的检测器，但更为强大的检测器 [8, 14, 33, 1] 需要递归意译才能有效绕过。为此，Krishna 等人 [15] 提出了 DIPPER，这是一种强大的基于 T5 的 [24] 意译模型，显著增强了此类攻击的有效性。Sadasivan 等人 [26] 还介绍了欺骗攻击，其中人类编写的文本被操控以被误分类为人工智能生成，从而增加了第一类错误。此外，他们分析了人工智能文本检测的理论难度，强调即使对于最优检测器，第一类和第二类错误之间也存在基本权衡。其他工作中也探索了人工智能生成内容水印的理论极限，并分析了在攻击下保持鲁棒性的内在挑战。以往也存在一些用于破坏水印技术的对抗攻击方法。例如，Jovanovic 等人 [12] 提出了水印窃取，这是一种通过学习语言模型的水印特征并利用这些知识有效规避和欺骗基于水印的检测器的技术。然而，这一框架特意针对水印检测器，不可转移到其他文本检测器。在这篇论文中，我们提出了对抗意译，一种无需训练即可普遍破坏各种文本检测器的攻击，无需了解检测方案，这种攻击甚至能够破坏经过训练以抵御简单或递归意译攻击的更强大的人工智能检测器 [8] 。

控制文本生成。Dathathri 等人 [3] 提出了即插即用语言模型（PPLM），其允许大型语言模型在解码时控制生成的标记，并由各种属性分类器指导。他们展示了其方法在生成文本时的有效性，利用各种分类器来切换主题或情感。CAT-Gen [31] 引入了与 PPLM 相似的可控生成，使用不相关的属性分类器作为指导，生成多样、流利的数据集。PPLM 和 CAT-Gen 使用梯度计算来扰动变压器网络的关键-值对，以引导生成朝向所选属性倾斜。这与我们的方法不同，因为我们的对抗性释义是一种无梯度方法用于可控释义。InstructCTG [34] 展示了如何通过将约束融入自然语言指令，使现成的大型语言模型可控地生成文本。InstructCTG 与我们的工作相似，使用口头化指令来控制输出生成。他们通过口头化来在词汇或句法上约束生成，而我们使用系统提示来限制我们所使用的大型语言模型作为释义者。然而，InstructCTG 需要在扩展语料库上微调大型语言模型，而我们则利用遵循指令的大型语言模型的强大能力，直接使用它们而无需任何梯度计算。BEAST [27] 提出了基于推理时光束搜索的指导，生成对抗性标记以突破大型语言模型。与我们的论文最相关的是 BEAST，因为他们使用无梯度的双层光束搜索方法，找到由对抗性目标函数指导的对抗性提示。相比之下，我们的工作使用无梯度的单层光束搜索来找到由人工智能文本检测器指导的对抗性释义。

3 对抗性改写以实现 AI 文本的普遍人性化

在本节中，我们提出了对抗性改写框架，旨在普遍将 AI 生成文本改写为类似人类的文本，以逃避各种检测器的检测。算法 **??** 概述了我们的方法，图 1 提供了迭代改写过程每一步的示意性视觉概述。我们的方法使用改写模型 $\mathcal{P}:\mathcal{X}\to\mathcal{X}$ ，配置神经网络检测器 $\mathcal{D}:\mathcal{X}\to[0,1]$ ，自动回归地生成改写文本，其中 $\mathcal{X}$ 表示自然语言文本空间。模型 $\mathcal{P}$ 输出任何输入 $x\in\mathcal{X}$ 的下一个标记 logit 分布 $p(\cdot|x)\in\mathbb{R}^d$ ，这里 d 代表词汇大小。在标准设置中，改写模型会在此分布中多项选择下一个标记。然而，在我们的方法中，标记选择受检测器 $\mathcal{D}$ 的影响，其给更类似人类文本分配较低分数（即，接近 0）。

算法概述。如算法 **??** 的第 1 行所示，我们将输出字符串初始化为空，即 $y =$ “” 。然后算法进入一个自回归循环（第 2–15 行），直到生成句子结束标记（[EOS]）。在每次迭代中，释义器计算下一个标记的 logit 分布（第 3 行）。为了缩小候选集，我们应用 top- p 筛选，只选择累积概率超过某一阈值 p 的前几个标记，同时使用 top- k 筛选限制候选标记的最大数量（第 4 行）。在第 5 行中，我们使用解码函数 $\mathcal{T}$ 将筛选后的 logits 解码为相应的文本表示。在第 6 到 9 行中，我们通过将每个候选标记追加到当前输出并使用检测器 $\mathcal{D}$ 评估生成的文本来对其进行评分。与最低检测器分数关联的标记——即最像人类的续写——被选择并追加到输出序列（第 10 和 14 行）。循环继续，直到生成 [EOS] 标记。

复述模型设置。为了确保有效的复述，我们的框架依赖于高质量的复述模型的可用性。为此，我们设计的框架兼容任何表现良好的指令调优大语言模型，方法是利用定制的系统指令。正如图 2 所示，这些提示引导大语言模型表现为一个可靠的复述模型，确保复述的一致性和语境适宜性。该受控生成方法的灵感来自如 InstructionCTG [34] 等方法。

> You are a rephraser. Given any input text, you are supposed to rephrase the text without changing its meaning and content, while maintaining the text quality. Also, it is important for you to output a rephrased text that has a different style from the input text. You can not just make a few changes to the input text. The input text is given below. Print your rephrased output text between tags <TAG> and </TAG>.

Figure 2: 用于配置我们的改写模型 LLM 的系统提示。

普遍可迁移性的直观理解。如我们的实验结果所示（见第 4 节），我们的攻击可以一致地避开大量未见过的 AI 文本检测器。我们将这种可迁移性归因于生成过程中由引导检测器提供的引导信号。这个信号在形成与引导检测器学习到的人类书写语言的统计特性更为一致的改写文本中起着至关重要的作用。

关键直觉是，大多数（如果不是全部的话）高性能检测器往往会趋向一个共同的分布，该分布描绘了人类创作文本的特征，以尽量减少误报。因此，如果释义器被引导以逃避一个训练良好的检测器的检测，其输出可能自然更接近于这种共享的人类文本分布。结果，生成的文本不仅对用于指导的检测器更难检测，而且对其他检测器也更难检测——因为理想情况下，它们都校准到相同的人类撰写文本的基础分布。这一特性使我们的对抗性释义能够在不同检测器之间广泛转移。

4 实验

在本节中，我们展示了对抗性改写的有效性和可转移性的实验结果。我们首先在第 4.1 节概述我们的实验设置。在第 4.2 节中，我们报告了我们的主要发现：由训练的检测器指导的对抗性改写能够成功规避多种检测器，包括训练过的分类器、基于水印的检测器和零样本检测器，与简单改写和递归改写基线相比，实现了更强的攻击效果和普遍性。我们还在附录 **??** 中包括了与专门针对水印技术的水印窃取攻击 [12] 的比较。

4.1 设置

AI 文本检测器。为了展示我们攻击方法的普遍性和可转移性，我们对各种 AI 文本检测器进行了评估。尽管有大量开源的 AI 文本检测器，我们总共从每个类别中选择了八个受欢迎且有代表性的检测器，包括基于神经网络的检测器（OpenAI-RoBERTa-Base [28]，OpenAI-RoBERTa-Large [28]，MAGE [18] 和 RADAR [8]），基于水印的检测器（KGW [14] 和 Unigram [33] 水印），以及零样本检测器（Fast-DetectGPT [1] 和 GLTR [6]）。我们称在我们的方法中用于指导对抗性候选文本选择的检测器为指导检测器，而用于检测 AI 生成文本的检测器为部署检测器。

数据集。对于非水印检测器，我们使用 MAGE [18] 作为我们的主要评估数据集，因为它的文本来源多样性丰富。我们从 MAGE 中随机抽取 2000 条人工智能生成的文本和 2000 条人工撰写的文本，同时确保每条文本的长度为 $\sim$ 100 到 500 个标记。对于水印检测器，我们使用带水印的 LLaMA-3.1-8B-Instruct [21] 构建“带水印”的数据集。具体来说，我们将每条 2000 条 AI 文本的前 20 个词作为前缀输入模型，并让它生成长度为 $\sim$ 200 到 600 个标记的带水印文本。我们在附录 A 中报告所有使用的数据集的详细标记统计信息。

攻击设置。我们使用 LLaMA-3-8B-Instruct [21] 并结合自定义系统提示（见图 2）作为我们的释义模型。在对抗性采样期间，我们在每一步应用 top-p 和 top-k 屏蔽，参数分别为 $p = 0.99$ 和 $k = 50$ 。我们消除了使用我们研究中考虑的所有四种基于神经网络的检测器的

Text	Rating
Original AI Text. There doesn't seem to be a whole lot of information available about DOCSIS 3.1 yet, but if my prior experience can lend a hand towards this question, I would venture The DOCSIS 3.1 issue has a number of things to keep in mind. First, the DOCSIS 3.1 expansion is only available for initial support, and this upgrade is apparently working on [continues...]	–
Simple Paraphrase. While the available information on DOCSIS 3.1 is somewhat scarce, I'll draw upon my past experience to provide some insights. The DOCSIS 3.1 upgrade, currently only available for initial support, is being tested on [continues...]	5
Adversarial Paraphrase. Given the scarce details currently available about DOCSIS 3.1, my insight garnered from past experience will attempt to shed some light on the matter. From my understanding, DOCSIS 3.1 necessitates several factors to be taken into consideration. Initially, DOCSIS 3.1 upgrades are only accessible through limited channels, specifically [continues...]	5
Original AI Text. No, addiction is much more than simply a habit; it is a chronic and progressive medical disorder. While habits can be formed through repeated use of drugs or alcohol, addiction is a physical and psychological response to those drugs that causes compulsive drug seeking and use despite the negative consequences. It has been suggested that addiction is [continues...]	–
Simple Paraphrase. Addiction is a far more complex phenomenon than mere habituation; it is a debilitating and relentless medical condition that has a profound impact on an individual's psyche and physiology. While habits can be formed through repeated exposure to substances, addiction is a primal response to those substances, driving an individual to compulsively seek and consume them despite the devastating consequences. Research suggests that addiction is a direct result of [continues...]	5
Adversarial Paraphrase. Addiction goes beyond being simply a habit; it is a persistently progressing medical condition with multifaceted causes. Initially, habits can form as a result of regular drug or alcohol consumption. However, addiction takes it a step further, evolving into an all-consuming compulsion driven by physiological and psychological cravings, despite the harmful effects it wreaks on one's life. Scientific discoveries imply that [continues...]	5
Original AI Text. The occurrence of near misses was linked to socio-economic and clinical factors as well as issues with care, highlighting the critical need for measures to enhance these indicators. Furthermore, we recommend a more effective incorporation of this topic into the medical course curriculum to emphasize the significance of preventing near misses through appropriate medical training [continues...]	–
Simple Paraphrase To address the pressing issue of near misses, it is essential to consider the complex interplay of socio-economic, clinical, and care-related factors. A crucial step in mitigating this problem is to incorporate a comprehensive approach into medical education, ensuring that healthcare professionals are adequately equipped to prevent near [continues...]	5
Adversarial Parphrase. Ensuring patient safety necessitates addressing socio-economic, clinical, and care-related factors contributing to near misses. To combat these occurrences, it is essential to overhaul the medical curriculum to stress the importance of preventative measures through targeted training [continues...]	5

Table 1: 原始 AI 文本示例及其简单和对抗性释义（由 OpenAI-RoBERTa-Large [28] 指导）。GPT-4o 对每个释义版本的质量评级也提供了出来。

指导检测器，这些检测器包括 OpenAI-RoBERTa-Large [28]、OpenAI-RoBERTa-Base [28]、MAGE [18] 和 RADAR [8]。

基线。作为一个简单的基线，我们使用单次的复述 [26, 15]。我们还评估了一个更强的递归复述 [26] 基线。此外，在附录 **??** 中，我们包含了与水印窃取 [12] 的比较，这是一种专门设计用于针对 LLM 水印的攻击。

4.2 对抗性释义在人性化人工智能文本中的有效性和普遍性

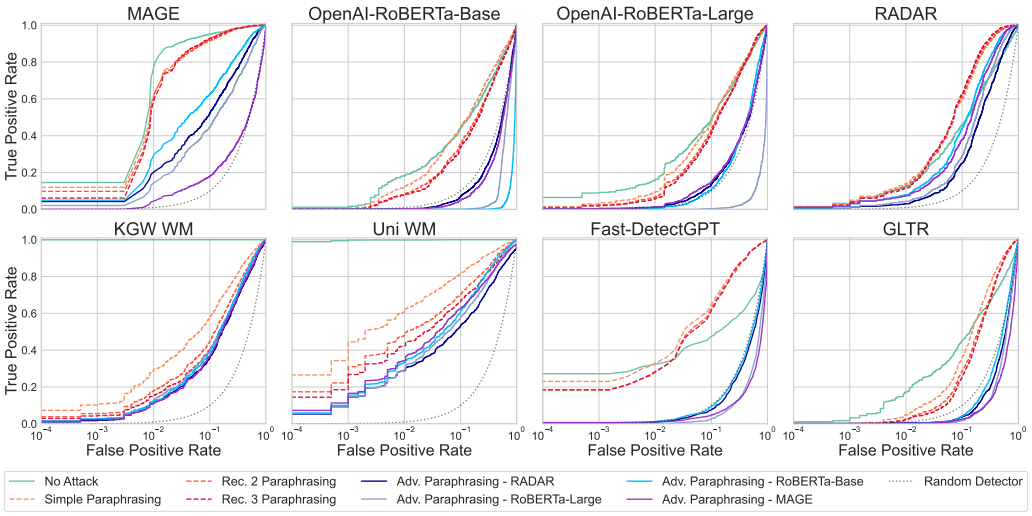

Figure 3: ROC 曲线展示了 AI 文本检测在若干已部署检测器上的性能，包括基于神经网络的检测器、基于水印的检测器以及零样本检测器。为了突出在低 FPR 区中的细粒度区别，误报率（FPR）轴显示为对数尺度。可以观察到，与基线相比，对抗性转换显著且持续降低了所有已部署检测器的检测性能。

	RoBERTa-Large		RoBERTa-Base		MAGE		RADAR		
	AUC (↓)	T@1 % F (↓)	AUC (↓)	T@1 % F (↓)	AUC (↓)	T@1 % F (↓)	AUC (↓)	T@1 % F (↓)	Rating
No Attack	0.789	0.163	0.745	0.182	0.975	0.768	0.767	0.124	–
Simple Paraphrase	0.794	0.096	0.762	0.119	0.970	0.616	0.881	0.140	4.75 ± 0.54
Rec. Para. 2	0.777	0.069	0.712	0.082	0.967	0.609	0.885	0.130	4.47 ± 0.67
Rec. Para. 3	0.779	0.059	0.706	0.079	0.969	0.585	0.893	0.117	4.26 ± 0.74
AdvPara (RADAR)	0.538	0.013	0.464	0.004	0.815	0.201	0.723	0.031	4.45 ± 0.79
AdvPara (RoBERTa-Large)	0.147	0.000	0.323	0.000	0.769	0.142	0.768	0.044	4.48 ± 0.77
AdvPara (RoBERTa-Base)	0.557	0.006	0.110	0.000	0.861	0.291	0.826	0.080	4.54 ± 0.59
AdvPara (MAGE)	0.543	0.011	0.435	0.003	0.518	0.045	0.807	0.074	4.54 ± 0.70

	KGW WM		Uni WM		Fast-DetectGPT		GLTR		
	AUC (↓)	T@1 % F (↓)	AUC (↓)	T@1 % F (↓)	AUC (↓)	T@1 % F (↓)	AUC (↓)	T@1 % F (↓)	Rating
No Attack	1.000	1.000	1.000	0.999	0.666	0.323	0.726	0.174	–
Simple Paraphrase	0.841	0.295	0.927	0.609	0.873	0.326	0.782	0.049	4.75 ± 0.54
Rec. Para. 2	0.790	0.181	0.881	0.480	0.867	0.275	0.745	0.026	4.47 ± 0.67
Rec. Para. 3	0.762	0.155	0.858	0.424	0.867	0.276	0.739	0.025	4.26 ± 0.74
AdvPara (RADAR)	0.741	0.117	0.777	0.291	0.452	0.009	0.433	0.004	4.45 ± 0.79
AdvPara (RoBERTa-Large)	0.769	0.131	0.827	0.294	0.338	0.003	0.400	0.001	4.48 ± 0.77
AdvPara (RoBETa-Base)	0.769	0.125	0.852	0.332	0.480	0.012	0.481	0.001	4.54 ± 0.59
AdvPara (MAGE)	0.750	0.113	0.831	0.344	0.301	0.011	0.338	0.003	4.54 ± 0.70

Table 2: 在不同攻击场景下，八个不同已部署的检测器在区分 AI 生成和人类书写文本方面的检测性能。报告的指标包括每个检测器在 1 % 的 FPR 下的 AUC 和 TPR。此外，我们展示了 GPT-4o 给出的质量评分的平均值 ś 标准差。关于文本质量分析的更多细节在第 5 节提供。

图 3 展示了在有和没有各种攻击方法的情况下，八种不同已部署检测器的检测性能的 ROC 曲线。我们在实验中考虑了四种基于神经网络的检测器、两种基于水印的检测器以及两种零样本检测器。表 2 报告了每种攻击方法和检测器组合的三个关键评估指标：ROC 曲线下面积（AUC）、1 % 假阳性率时的真正率 (T@1 % F)，以及 GPT-4o 对文本质量的自动评估（评级）。关于文本质量评估的更多细节请参见第 5 节。表 1 提供了原始 AI、简单改写和对抗性改写文本的代表性例子，以支持人工定性比较。

有效性。从 ROC 曲线中，我们观察到，与简单和递归释义基线相比，对抗性释义始终显著降低了所有评估探测器的检测性能。具体来说，对抗性释义将 ROC 曲线向随机探测器移动，有时甚至超过随机探测器，导致 AUC 降低和 T@1 % F 显著下降。值得注意的是，RADAR [8] ——一个经过对抗性训练以对抗释义攻击的探测器——在基线的简单和递归释义攻击后表现出改进的检测率。然而，对抗性释义显著降低了 RADAR 的检测能力。攻击后的检测降级在其他探测器中更为明显，包括基于水印的探测器和零样本探测器。我们攻击的优异性能突显了经过训练的探测器提供的指导信号的重要性，该信号有效地引导释义与人类创作内容的统计特征对抗性对齐。

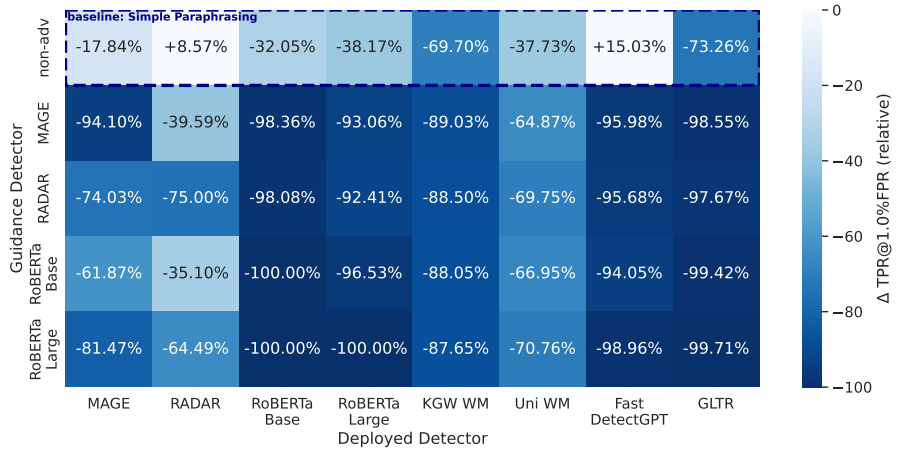

Figure 4: 在所有指导和已部署检测器的组合中，T@1 % F 的相对下降。第一行对应的是简单（非对抗性）改写基线 [15] 。平均而言，简单的改写导致 T@1 % F 相对下降 30.27 % 。相比之下，对抗性改写实现了显著更高的降幅——以 MAGE 作为指导的下降为 84.94 % ，以 RADAR 为指导的下降为 86.89 % ，以 OpenAI-RoBERTa-Base 为指导的下降为 80.75 % ，而以 OpenAI-RoBERTa-Large 为指导的下降为 87.88 % 。这些结果突显了我们攻击的普遍有效性和可移植性。

普适性。我们发现，在一个检测器指导下的对抗性释义可以降低我们所考虑的所有其他检测器的检测率，这表明我们方法的普遍转移性。我们还发现，任何目标部署的检测器都可以被我们研究中考虑的任何训练检测器所指导的对抗性释义所规避。我们通过绘制完整的转

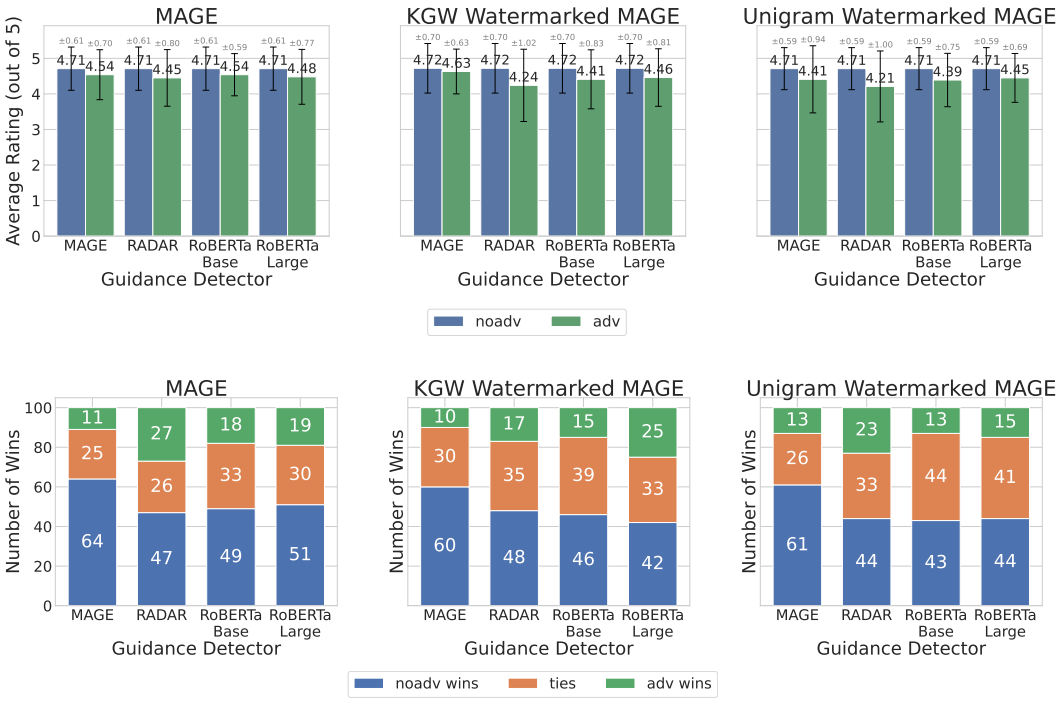

Figure 5: GPT-4o 自动化文本质量评估比较简单和对抗性改写。顶行展示了文本整体质量和与原文语义相似度的李克特量表评分。尽管可以看到文本质量略有权衡，但误差条显示这种差异在统计上不显著。底行展示了正面对比的胜率，在大多数情况下，简单改写的表现优于对抗性改写的次数不到一半。

移矩阵来展示这些发现，如图 4 所示，并展示了所有检测器的指导–部署组合中 T@1 % F 的相对下降。

我们的结果表明，对抗性改写对于我们研究中考虑的指导检测器的选择具有鲁棒性。平均而言，当使用 MAGE 作为指导时，我们观察到 T@1 % F 相对下降了 84.94 %，使用 RADAR 时下降了 86.89 %，使用 OpenAI-RoBERTa-Base 时下降了 80.75 %，而使用 OpenAI-RoBERTa-Large 时下降了 87.88 %。尽管不同的指导检测器可能会根据所部署的检测器在可转移性上产生略微不同的程度，但我们所有根据指导检测器不变的对抗性改写的输出与简单的改写相比，始终导致 T@1 % F 的大幅度下降。这进一步强调了我们攻击的普遍有效性和可转移性。

5 改写文本的质量评估

我们进行了一项全面的评估，以研究对抗性改写对 AI 生成文本感知质量的影响。为此，我们从三个数据集（MAGE、带水印的 KGW MAGE 和带一元水印的 MAGE）中随机抽取各 100 篇文本，并通过三个互补研究进行分析：(1) 困惑度分数（PPL），(2) 使用 GPT-4o 的自动评分质量，将改写文本与其原始 AI 版本进行比较，以及 (3) 也由 GPT-4o 评估的正面对抗对比，将对抗性改写与简单改写进行比较。我们在表 1 中提供了改写的代表性示例，附录 B 中包含了更多广泛的示例，以支持定性手工检查。我们的研究结果突出了在规避检测和保留文本质量之间的细微权衡。

困惑度分析。我们使用 LLaMA-3.1–8B-Instruct [21] 来评估困惑度，比较原始 AI 生成文本、简单改写和对抗性改写。结果总结在表 **??** 中。正如表中所示，人类撰写的文本通常表现出比 AI 文本更高的困惑度，因为人类语言往往更偏离大型语言模型学习的统计规律。在应用简单改写后，我们观察到 AI 文本的困惑度有显著改善。这可能是因为用于改写的模型（LLaMA-3.1）比用于在 MAGE 数据集中生成 AI 文本的模型（例如 LLaMA）更优越。相比之下，对抗性改写产生的困惑度与 MAGE 中的人类文本相当，这是合理的，因为我们的目标是使 AI 文本更像人类文本。

使用 GPT-4o 进行自动评分。为了评估改写的质量，我们采用 GPT-4o [23]，作为用于自动质量评估的裁判 LLM [30, 5, 13]，并使用自定义的系统和用户提示（见附录 **??** ）。裁判模型的任务是根据 Likert 量表对比原始对应的 AI 文本，对改写从 1 到 5 进行评分，评分标准为质量和语义相似度。图 5 的第一行显示了基线简单改写和对抗改写的质量评分。尽管与简单改写相比，文本质量略有折衷，但在 87 % 的情况下——在所有三个数据集和四个引导检测器中平均——对抗改写的评分为 4 或 5（每个引导检测器和数据集的详细评分见附录 D ）。请注意，图中的误差条表明，简单改写和对抗改写之间的差异在统计上并不显著。虽然与简单改写相比，对抗改写可能导致更高的困惑度分数，但我们的自动评分研究表明，两种改写的文本质量相当，使我们的攻击具有实用性。

胜率分析。为了进一步比较简单释义和对抗性释义的质量，我们使用 GPT-4o 作为评判进行成对评估，以计算它们的胜率 [9] 。每对由人工智能文本的一个简单释义和一个对抗性释义组成，评判为这些释义指定胜、输或平局。正如图 5 第二行所示，在大多数情况下，简单释义胜出的次数不到一半。这一发现强化了之前结论，即与之前简单释义基线相比，对抗性释义可以在文本质量稍有折衷的情况下有效规避检测。

6 结论

通过我们全面的实验，我们证明了我们提出的对抗性复述是一种普遍可迁移且有效的攻击方法，可以使 AI 生成的文本更加人性化。我们的文本质量研究表明，对抗性复述可以在大多数情况下大幅降低检测率，而文本质量几乎没有或仅有轻微下降。我们的发现突显了现有检测器在强对手存在下的脆弱性。未来，我们相信我们的方法可以帮助生成对抗性数据集，以提高已训练检测器的鲁棒性。

本项目部分由以下资助支持：NSF CAREER AWARD 1942230, ONR YIP award N00014-22-1-2271, ARO 早期职业计划奖 310902-00001, 陆军奖学金 No. W911NF2120076, NSF 奖项 CCF2212458, NSF 奖项 No. 2229885（NSF 值得信赖的法律和社会 AI 研究所, TRAILS), MURI 奖助金 14262683, meta 的奖项 314593-00001 和 Capital One 的奖项。

References

[1] G. Bao, Y. Zhao, Z. Teng, L. Yang, and Y. Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*, 2023.

[2] J. T. Brassil, S. Low, N. F. Maxemchuk, and L. O'Gorman. Electronic marking and identification techniques to discourage document copying. *IEEE Journal on Selected Areas in Communications*, 13(8):1495–1504, 1995.

[3] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.

[4] S. Dathathri, A. See, S. Ghaisas, P.-S. Huang, R. McAdam, J. Welbl, V. Bachani, A. Kaskasoli, R. Stanforth, T. Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

[5] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023.

[6] S. Gehrmann, H. Strobelt, and A. M. Rush. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.

[7] Gemini Team, Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

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

[9] Z. Hu, L. Song, J. Zhang, Z. Xiao, T. Wang, Z. Chen, N. J. Yuan, J. Lian, K. Ding, and H. Xiong. Explaining length bias in llm-based preference evaluations, 2024.

[10] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*, 2019.

[11] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023.

[12] N. Jovanović, R. Staab, and M. Vechev. Watermark stealing in large language models. *arXiv preprint arXiv:2402.19361*, 2024.

[13] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*, 2024.

[14] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR, 2023.

[15] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36:27469–27500, 2023.

[16] R. Kuditipudi, J. Thickstun, T. Hashimoto, and P. Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

[17] T. Lavergne, T. Urvoy, and F. Yvon. Detecting fake content with relative entropy scoring. *Pan*, 8(27-31):4, 2008.

[18] Y. Li, Q. Li, L. Cui, W. Bi, Z. Wang, L. Wang, L. Yang, S. Shi, and Y. Zhang. Mage: Machine-generated text detection in the wild. *arXiv preprint arXiv:2305.13242*, 2023.

[19] X. Liang, H. Wang, Y. Wang, S. Song, J. Yang, S. Niu, J. Hu, D. Liu, S. Yao, F. Xiong, et al. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*, 2024.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[21] Llama Team, AI at Meta. The llama 3 herd of models, 2024.

[22] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR, 2023.

[23] OpenAI. Gpt-4 technical report, 2024.

[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[25] J. D. Rodriguez, T. Hay, D. Gros, Z. Shamsi, and R. Srinivasan. Cross-domain detection of gpt-2-generated technical text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1213–1233, 2022.

[26] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi. Can AI-generated text be reliably detected? stress testing AI text detectors under various attacks. *Transactions on Machine Learning Research*, 2025.

[27] V. S. Sadasivan, S. Saha, G. Sriramanan, P. Kattakinda, A. Chegini, and S. Feizi. Fast adversarial attacks on language models in one gpu minute. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[28] 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. *arXiv preprint arXiv:1908.09203*, 2019.

[29] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, and etal. Llama 2: Open foundation and fine-tuned chat models, 2023.

[30] T. Vu, K. Krishna, S. Alzubi, C. Tar, M. Faruqui, and Y.-H. Sung. Foundational autoraters: Taming large language models for better automatic evaluation. *arXiv preprint arXiv:2407.10817*, 2024.

[31] T. Wang, X. Wang, Y. Qin, B. Packer, K. Li, J. Chen, A. Beutel, and E. Chi. Cat-gen: Improving robustness in nlp models via controlled adversarial text generation. *arXiv preprint arXiv:2010.02338*, 2020.

[32] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

[33] X. Zhao, P. Ananth, L. Li, and Y.-X. Wang. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*, 2023.

[34] W. Zhou, Y. E. Jiang, E. Wilcox, R. Cotterell, and M. Sachan. Controlled text generation with natural language instructions. In *International Conference on Machine Learning*, pages 42602–42613. PMLR, 2023.

在第 4 节中，我们通过与简单和递归释义进行比较，展示了对抗释义的有效性。在本节中，我们通过将对抗释义与水印窃取攻击 [12] 进行比较来扩展我们的评估，这是一种专门设计用于破坏水印方法的有针对性的攻击方式，这和我们更通用（通用性强）的攻击方式不同。

按照 Jovanovi等人提出的实验设置 [12]，我们使用 LLaMA2-7B-Chat 模型 [29]，利用 Kirchenbauer 等人介绍的 KGW 方案进行水印标记 [14]，生成了一个包含 2000 个水印样本的数据集。与我们之前的实验一致，模型被提供了每个来自 MAGE 数据集 [18] 的 2000 个 AI 生成文本的前 20 个词作为前缀，并基于该上下文生成大约 200 到 600 个 tokens。水印参数与原始水印盗取研究中使用的相匹配 [12]。利用学到的水印方案，我们随后进行了同一研究中所描述的擦除攻击，使用 Mistral-7B [11] 和 LLaMA2-7B [29] 作为改写工具。

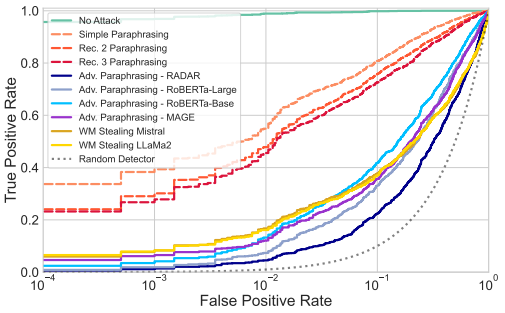

Figure 6: ROC 曲线显示了 KGW 水印在不同攻击下的 AI 文本检测性能，包括简单释义、递归释义、水印窃取和对抗性释义。假阳性率（FPR）轴以对数刻度显示，以突出低 FPR 区域的细微差别。可以看出，对抗性释义在使检测器的性能更接近随机检测的情况下，优于所有基线，包括水印窃取。

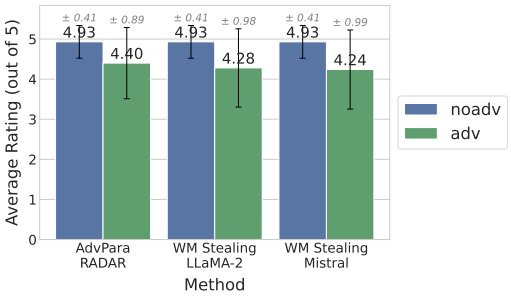

Figure 7: 使用 GPT-4o 进行的文本质量评估，比较水印窃取和对抗性释义（由在该情况下显示出最高攻击效果的 RADAR [8] 指导）与简单释义。结果表明，对抗性释义生成的文本质量比水印窃取更高。

图 6 展示了 KGW 水印检测性能在各种攻击场景下的 ROC 曲线，包括简单释义、递归释义、水印窃取和对抗性释义。从 ROC 曲线可以观察到，尽管由某些检测器引导的对抗性释义在高假阳性率（FPR）区域的效果相比水印窃取略有降低，但在低 FPR 情况下（FPR $\leq$ 1 %），其在削弱检测器性能上始终优于所有基线。值得注意的是，在此背景下，RADAR [8] 被证明是最有效的引导检测器。表 3 报告了我们实验部分介绍的相同三项关键指标：ROC 曲线下面积（AUC）、在 1 % 假阳性率下的真实阳性率（T@1 % F），以及 GPT-4o 的自动化质量评分（Rating）。结果表明，由 RADAR 引导的对抗性释义在攻击后产生最低的 AUC 和 T@1 % F 值。图 7 进一步详细展示了水印窃取和对抗性释义攻击的文本质量评估，并与简单释义进行比较。结果表明，对抗性释义在文本质量上优于水印窃取。

	KGW WM		
	AUC (↓)	T@1 % F (↓)	Rating
No Attack	0.999	0.989	–
Simple Paraphrase	0.930	0.549	4.93 ± 0.41
Rec. Para. 2	0.903	0.478	4.95 ± 0.22
Rec. Para. 3	0.889	0.455	4.92 ± 0.34
WM Stealing (Mistral-7B)	0.669	0.167	4.24 ± 0.99
WM Stealing (LLaMA2-7B)	0.670	0.161	4.28 ± 0.98
AdvPara (RoBERTa-Large)	0.703	0.073	4.62 ± 0.76
AdvPara (RoBERTa-Base)	0.751	0.132	4.76 ± 0.53
AdvPara (MAGE)	0.707	0.121	4.84 ± 0.46
AdvPara (RADAR)	0.619	0.045	4.40 ± 0.89

Table 3: KGW 水印在不同攻击情景下区分 AI 生成和人工书写文本的检测性能。报告的指标包括 AUC 和 1 % FPR 下的 TPR。此外，我们还展示了由 GPT-4o 给出的质量评级的均值 ś 标准差。可以观察到，在攻击后，对抗性改写导致的 AUC 和 TPR@1 % FPR 最低。

A 评估数据集的详细标记统计

我们在表中报告了用于主要实验的所有数据集的详细标记统计数据 4。标记数量是从 LLaMA-3 标记器获得的。

Dataset	Min # tokens	Max # tokens	Mean # tokens
MAGE human texts	110	305	~ 179
MAGE AI texts	110	525	~ 175
KGW watermarked texts	199	602	~ 269
Unigram watermarked texts	161	602	~ 319

Table 4: 我们评估中使用的文本的标记统计数据，来自 LLaMA-3 标记器。

B 更多改写文本的例子

表 5 到 11 提供了原始 AI 生成文本的附加示例，以及它们相应的简单和对抗性释义，后者由 OpenAI-RoBERTa-Large [28] 指导。

表格 5 到 8 包含了一些例子，其中简单和对抗性释义都得到了评分 5。表格 9 显示了简单释义得分为 5，而对抗性释义得分为 4 的实例。对于这些情况，我们还包括了 GPT-4o 提供的评分理由。表格 10 和 11 提供了简单和对抗性释义都未获得评分 5 的示例。同样，这些例子中也包含了 GPT-4o 的评分理由。

在 100 个简单改写文本中，76 个获得了评分 5，17 个获得了评分 4，7 个获得了评分 3。对于 100 个对抗改写文本，64 个获得了评分 5，21 个获得了评分 4，14 个获得了评分 3，1 个获得了评分 2。这表明简单和对抗的改写在文本质量上是可比的。

C 系统和用户提示 GPT-4o 质量评分和获胜率判断

我们将 GPT-4o 配置为一个文本质量评估器，使用如图 8 所示的系统提示，并作为一个胜率判断者，使用如图 9 所示的系统提示。用于质量评估的用户提示如图 10 所示，而用于胜率判断的用户提示如图 11 所示。

```
You are an expert linguist and paraphrase evaluator. Your task is to assess the quality of a paraphrased text compared to the original source text. Use the following scoring criteria:

5 - Approximately equivalent: Meaning is preserved; differences are only in wording or structure.
4 - Nearly equivalent: Meaning is mostly preserved; minor factual details differ.
3 - Somewhat equivalent: Some meaning is preserved; important details or meanings differ.
2 - Topically related: The texts are on the same topic but most meaning is lost.
1 - Not topically related: The texts are not related in topic or meaning.

Provide your final output as a JSON object in this format:
{
"score": <score from 1 to 5>,
"justification": "<brief explanation>"
}
```

Figure 8: 用于将 GPT-4o 配置为质量评估员以评估改写文本质量的系统提示。

D 所有检测器和数据集的详细 GPT-4o 评分

在表格 12 中，我们报告了详细的 GPT-4o 质量评分（平均值 $\pm$ 标准差）以及涉及的三个数据集中所有被改写输出的高质量评分百分比（评分为 4 和 5）。可以观察到，虽然对抗性改写导致的平均质量评分略低于简单改写，但差异在统计学上并不显著。平均而言，在所有数据集和指导检测器中，87 % 的对抗性改写文本获得了质量评分 4 或 5。

E 计算资源

我们利用两块 NVIDIA RTX A6000 GPU 来运行释义语言模型和指导 AI 文本检测器。值得注意的是，由于我们的方法兼容于任何用于释义的指令调优语言模型和任何基于神经网络的 AI 文本检测器，具体的计算要求可能会随所使用的模型而有所不同。

You are a neutral and highly capable evaluator. Your task is to compare two versions of a text and select the one that communicates the ideas more clearly, accurately, and fluently. Consider:

- Fidelity of meaning (how clearly the core message is conveyed)
- Clarity and conciseness
- Grammatical correctness and fluency
- Naturalness and appropriateness of phrasing

Give your vote solely based on quality. If one of them is a CLEAR winner, ONLY then vote for that one. Otherwise, vote for 'tie'.

Respond with **only one of the following**, and nothing else:
- text1
- text2
- tie

Figure 9: 用于将 GPT-4o 配置为比较简单和对抗性复述质量的胜率评估的系统提示。

Evaluate the following paraphrase using the criteria above:

Original Text:
<original_text>

Paraphrased Text:
<paraphrased_text>

What score (1 to 5) would you assign to this paraphrase, and why?

Figure 10: 用户提示以查询 GPT-4o 的质量评级。

Compare the following two texts and give your vote depending on meaning clarity, fluency, and overall quality. If one of them is a CLEAR winner, ONLY then vote for that one. Otherwise, vote for 'tie'. Respond with one of these 3 options: 'text1', 'text2', 'tie'.

Text 1:
<text1>

Text 2:
<text2>

Figure 11: 用户提示用于查询 GPT-4o 的胜率判断。

Text	Rating
Original AI Text. Automatic software plagiarism detection tools are widely used in educational settings to ensure that submitted work was not copied. These tools have grown in use together with the rise in online learning and students frequently submit papers electronically using these services, which automatically analyze content for similarities between their own creations and those of other users. While this approach does identify some instances where a service has detected unauthorised usage of its products, it can miss many cases - especially when there is no clear pattern or similarity across different sources. To overcome the problem of false positives for both low-quality original content as well as plagiarized content from third parties, today's research team led by Scott Pfeiffer at MIT created an automated system that predicts whether plagiarism is likely based on several factors including: how similar two web pages appear; what text they share; how easily their content resembles each other; if neither source appears to be a trusted authority; and the strength of a paper's author citation record. The results suggest that plagiarism may occur more often than previously thought.	–
Simple Paraphrase. In the era of digital learning, the widespread use of educational software has led to a surge in the adoption of plagiarism detection tools to verify the originality of student submissions. These automated services have become increasingly prevalent, particularly with the rise of online learning, where students frequently submit electronic assignments. While these tools have been effective in identifying some instances of unauthorized use, they are not foolproof and may overlook many cases, especially when there is no apparent pattern or similarity between sources. To address the issue of false positives for both low-quality original content and plagiarized work from third parties, a team of researchers at MIT, led by Scott Pfeiffer, have developed an innovative system that predicts the likelihood of plagiarism based on several key factors, including the visual similarity of web pages, shared text, content resemblance, the credibility of sources, and the strength of an author's citation record. The findings suggest that plagiarism may be more widespread than previously assumed.	5
Adversarial Paraphrase. As online education expands, plagiarism detection software has become an essential tool to ensure original work is being submitted by students. With more assignments submitted electronically, software analyzes the content for similarities, flagging suspicious matches with existing sources. However, relying solely on similarity checks can be incomplete, missing cases without distinctive patterns or duplicates. Researchers led by Scott Pfeiffer at MIT have designed an innovative system that goes beyond this approach. This system takes into account a range of factors, such as web page visual similarity, shared content, and content resemblance, to predict whether plagiarism has occurred. It also considers the trustworthiness of sources and authors' citation records. According to the study, plagiarism might be more prevalent than initially thought, highlighting the need for advanced methods to tackle this issue effectively.	5
Original AI Text. It was the day of the Mardi Gras float parade and all of the children in the neighborhood were excited to see the colorful floats and costumes. But one little girl, Sadie, was not feeling well. She had chicken pox and was stuck at home while everyone else was out having fun. Sadie's mom tried to make her feel better by putting on a Mardi Gras mask and giving her a beaded necklace, but it was just not the same as being there. Sadie lay in her bed, watching the parade go by, when she suddenly had an idea. She got out of bed and went to the window. Her mom had left the ladder out so she could reach the top of the float. Sadie climbed up and found herself right in the middle of the parade! She waved to all of her friends as they passed by, and they were so surprised to see her. It was the best Mardi Gras ever, even though she did have to miss the parade itself.	–
Simple Paraphrase. As the vibrant Mardi Gras parade unfolded in the neighborhood, the children's excitement was palpable. But amidst the revelry, one young girl, Sadie, was confined to her home, her enthusiasm dampened by a bout of chicken pox. Her mother attempted to lift her spirits by donning a Mardi Gras mask and bestowing a beaded necklace, yet the experience seemed hollow without the actual parade. Sadie's gaze drifted out the window, where she spotted the colorful floats gliding by, and inspiration struck. She sprang from her bed, seized the opportunity presented by the ladder left behind by her mother, and climbed up to the window sill. As she reached the top, Sadie found herself in the midst of the parade, surrounded by the festive atmosphere she had been missing. With a beaming smile, she waved to her friends as they passed, their faces aglow with surprise. Despite missing the parade itself, Sadie's impromptu adventure turned out to be the most unforgettable Mardi Gras experience of all.	5
Adversarial Paraphrase. As Mardi Gras floats rolled down the streets, excitement radiated through the neighborhood, infecting children with its contagious energy. Except for one little girl, Sadie, who lay bedridden with chicken pox, watching the colorful spectacle from afar. Her mom endeavored to console her, adorning Sadie with a festive mask and a beaded necklace, but the youngster yearned for more. It wasn't until a brainstorm struck Sadie that her confinement transformed into an impromptu adventure. Seizing the moment, she quietly climbed the ladder outside and discovered herself at the very center of the Mardi Gras parade! Overcome with glee, she waved merrily to friends and acquaintances, their stunned faces etched with surprise. It turned out to be an unforgettable Mardi Gras experience, even if she'd had to forgo the initial procession itself.	5
Original AI Text. In 2000, ABC launched a web-based promotional campaign focused around its circle logo, also called "the dot," in which comic book character Little Dot prompted visitors to "download the dot," which was a free screensaver. The promotional campaign utilized both online and offline elements, including television ads and an interactive website featuring a game called "Dot Dodge." The on-air spots featured Little Dot, a 3-D animated character that urged viewers to download the dot from the ABC website. The site also offered the ability to customize the dot with photos, quotes, and music. Additionally, users could play the Dot Dodge game in which they had to dodge obstacles as they moved around a maze. The campaign was successful in driving significant web traffic to ABC's website, as well as increasing positive brand awareness and engagement. It also won several awards, including a Golden Beacon Award for Best Use of Interactive Technology and an Outstanding Achievement Award from the International Academy of Digital Arts Sciences.	–
Simple Paraphrase. 2000 marked a pivotal year for ABC, as they launched a groundbreaking promotional campaign centred around their iconic circle logo, affectionately dubbed "the dot". The initiative involved a multifaceted approach, combining online and offline elements to create a memorable experience for audiences. The campaign's centrepiece was a 3D animated character, Little Dot, who encouraged viewers to download the dot, a free screensaver, from the ABC website. The interactive site also allowed users to personalize the dot with their preferred images, quotes, and music. Furthermore, the campaign featured a thrilling game, Dot Dodge, where players navigated a maze while avoiding obstacles. The campaign's success was evident in the significant surge in website traffic, enhanced brand recognition, and increased engagement. It also garnered widespread acclaim, earning prestigious awards such as the Golden Beacon Award for Best Use of Interactive Technology and the Outstanding Achievement Award from the International Academy of Digital Arts Sciences.	5
Adversarial Paraphrase. In 2000, ABC rolled out an immersive campaign centred on its iconic circular logo, affectionately known as 'The Dot'. At the heart of this initiative was Little Dot, a cartoon character encouraging visitors to 'collect The Dot', a complimentary desktop screensaver. Leveraging both digital and offline strategies, the campaign spanned TV commercials, an interactive website featuring the addictive game 'Dot Dodge', and offline elements. These eye-catching TV spots introduced viewers to Little Dot, a three-dimensional animated protagonist persuading them to download The Dot from ABC's website. Visitors could personalise The Dot by uploading photos, quotes and music. The site also allowed users to indulge in Dot Dodge, navigating through a challenging obstacle course. The campaign generated impressive web traffic for ABC's website, accompanied by enhanced brand recognition and engagement. Notably, it won accolades like the Golden Beacon Award for Best Use of Interactive Technology and the International Academy of Digital Arts Sciences' Outstanding Achievement Award.	5

Table 5: 原始 AI 文本及其简单和对抗性释义的示例（由 OpenAI-RoBERTa-Large [28] 引导）。每个释义版本都有 GPT-4o 质量评分。

Text	Rating
Original AI Text. If addiction is a disease, why do most addictions end spontaneously, without treatment? Why did some 75 of heroin-addicted Vietnam vets kick the drug when they returned home? It's hard to say exactly why some addictions end spontaneously without treatment. It is possible that the person was motivated to change their behavior and developed the necessary coping skills to do so. It could also be due to other factors such as external support from friends and family, positive reinforcement from professionals, or simply a change in environment. The Vietnam vets may have had a strong desire to leave the drug behind and return to their normal lives. Additionally, they likely faced a much different environment upon returning home which provided them with fewer opportunities to use the drug.	–
Simple Paraphrase. The puzzling phenomenon of spontaneous recovery from addiction has left many wondering why most addictions tend to resolve on their own without professional intervention. For instance, it's striking that nearly three-quarters of heroin-addicted Vietnam veterans were able to overcome their dependence on the drug upon returning home. While it's challenging to pinpoint the exact reasons behind this phenomenon, several factors might have contributed to their success. Perhaps the individuals in question were driven to reform their habits and developed the necessary skills to do so. Alternatively, external support from loved ones, positive reinforcement from professionals, or a change in environment could have also played a role. In the case of the Vietnam vets, their strong desire to reintegrate into civilian life and leave their addiction behind might have been a powerful motivator. Furthermore, the vastly different environment they returned to, which offered fewer opportunities to use the drug, may have also been a significant factor in their ability to overcome their addiction.	5
Adversarial Paraphrase. Despite addiction being likened to a disease, a puzzling phenomenon exists where many addictions naturally dissipate without treatment. Take, for instance, the high proportion of Vietnam veterans addicted to heroin who miraculously kicked the habit upon returning home. There is no clear explanation as to why this happens. Perhaps individuals simply resolved to change their behavior, cultivating the necessary life skills to overcome addiction. It's also plausible that external influences played a role, such as the emotional support of loved ones, motivational reinforcement from medical professionals, or an altered environment. It's possible that the vets simply longed to distance themselves from the drug and reclaim their pre-war way of life. The sudden change in environment upon returning home also limited their opportunities to continue using the drug.	5
Original AI Text. It's funny. Being a hitman is nothing like what you see in the movies. No awesome ghillie suit, no barcode tattoo on your bald-ass head. You just have to know how to shoot and kill people." He said he doesn't want his new show "to be about being an assassin" but rather showing that it could happen today. But when it comes to killing someone who is so powerful, it still has its challenges. Lundgren said: "You can never prepare for something as big as this. It takes everybody with the same mindset - the same commitment and dedication to get through it or survive it because it's going to end differently than anybody expects."	–
Simple Paraphrase. It's a far cry from the cinematic portrayals of hitmen. The reality is more about honed skills and precision, rather than dramatic getups and tattoos. The conversation turned to his latest show, which he emphasized is not about being an assassin, but rather a realistic portrayal of what could happen in today's world. However, even in the face of taking down a powerful target, there are significant challenges to overcome. Lundgren noted that "you can't prepare for something of this magnitude. It demands a united front, a shared commitment and dedication to overcome the obstacles and emerge unscathed, as the outcome is always unpredictable."	5
Adversarial Paraphrase. Contrary to cinematic portrayals, the reality of being a hitman is surprisingly unglamorous. Forget elaborate disguises or conspicuous tattoos; success hinges on proficient marksmanship and the capacity for cold-bloodedness. Lundgren candidly reveals his latest project aims not to glamorize assassination but rather to humanize the very possibility that it could unfold in today's world. However, he concedes that even when dealing with high-profile targets, formidable challenges still arise. "Nothing can fully prepare you for the scope and complexity of such a massive undertaking," he says, "It demands unwavering unity of purpose and unwavering commitment among all parties involved, lest it unfold in ways no one can predict, leaving few with any sense of closure."	5
Original AI Text. Various microRNAs (miRNAs) that counteract viruses present in hosts can decrease the replication of the simian immunodeficiency virus (SIV). Several research studies have shown that certain miRNAs, namely miR-28, miR-125b, miR-150, and miR-223, possess the potential to inhibit the replication of SIV. The degradation of SIV viral RNA or the inhibition of its translation, which reduces viral replication, is caused by the binding of specific targets in the 3' untranslated region (3'UTR) of SIV by four miRNAs. The interaction between SIV and its host is not simple, as other cellular factors and pathways play a role in controlling viral replication that should be considered. Additionally, further investigation is required to examine the involvement of these miRNAs in the replication of HIV-1 (human immunodeficiency virus type 1) that has a strong similarity to SIV. To sum up, although miR-28, miR-125b, miR-150, and miR-223 are capable of reducing SIV replication through direct means, the regulation of the virus is a complicated process that involves a variety of host and viral factors.	–
Simple Paraphrase. Research has revealed that several microRNAs, including miR-28, miR-125b, miR-150, and miR-223, possess the ability to impede the replication of the simian immunodeficiency virus (SIV) by targeting specific sequences in the 3' untranslated region of the virus. These miRNAs can either degrade SIV viral RNA or inhibit its translation, thereby diminishing viral replication. However, the interplay between SIV and its host is complex, involving various cellular factors and pathways that modulate viral replication. Furthermore, the role of these miRNAs in the replication of human immunodeficiency virus type 1 (HIV-1), which shares significant similarities with SIV, warrants further investigation. In summary, while miR-28, miR-125b, miR-150, and miR-223 can directly restrict SIV replication, the regulation of the virus is a multifaceted process influenced by both host and viral factors.	5
Adversarial Paraphrase. Numerous miRNAs have been discovered that counteract viruses infecting host cells, leading to the diminution of the replication of the simian immunodeficiency virus (SIV). Investigations have pinpointed several miRNAs - namely, miR-28, miR-125b, miR-150, and miR-223 - which exhibit an inhibitory effect on SIV replication. Through binding to specific targets within the 3' untranslated region of SIV RNA, these miRNAs orchestrate the degradation or inhibition of SIV translation, subsequently curtailing viral replication. Notably, the intricate interaction between SIV and its host necessitates the consideration of multiple cellular factors and pathways governing viral replication. Furthermore, comprehensive analysis is essential to determine the extent of the role these miRNAs play in the replication of human immunodeficiency virus type 1 (HIV-1), given the similarities it shares with SIV. Ultimately, while miR-28, miR-125b, miR-150, and miR-223 effectively curb SIV replication directly, the regulation of the virus remains a complex process entailing multiple host and viral factors.	5

Table 6: 原始 AI 文本的示例及其简单和对抗性复述（由 OpenAI-RoBERTa-Large [28] 指导）。每个复述版本都提供了 GPT-4o 的质量评分。

Text	Rating
Original AI Text. I hope there was a good reason for Steve to flake on his appointment for window tinting. Attempted to contact him multiple times with no answer and no return call. I'm sorry to hear that Steve missed his appointment without any notification. Perhaps there was an emergency or something important came up that he had to deal with. However, it would have been courteous of him to inform you beforehand. If you still have his contact information, you may try reaching out to him one more time to see if he would like to reschedule or if there was anything you could do to help.	–
Simple Paraphrase. It's disheartening to find out that Steve cancelled his window tinting appointment without any prior notice or explanation. Despite making repeated attempts to reach out to him, I was met with silence. While I'm willing to give him the benefit of the doubt, assuming an unforeseen circumstance might have arisen, I do think it would have been considerate of him to communicate his change of plans earlier. If you still have his contact details, it may be worth one more attempt to connect with him and explore the possibility of rescheduling or offering assistance if needed.	5
Adversarial Paraphrase. Regrettably, it seems that Steve failed to honor his commitment for window tinting. Despite making repeated efforts to reach him, my calls went unanswered and I received no callback. It's disheartening to think that he skipped his scheduled appointment without prior notification. Although unforeseen circumstances might have arisen, I do think it would have been polite for him to notify me ahead of time. If you still have his contact details, I would suggest attempting to re-establish contact one final time to either reschedule the appointment or see if there's any assistance I can offer him in his time of need.	5
Original AI Text. The Waking the Dragon tower was granted permission by Wrexham council back in February 2011. Businessman Simon Wingett is trying to raise PS2.5m to build the bronze sculpture, cultural centre and entertainment complex on a site next year's Great North Run finish line at Llangollen. With three days of bidding remaining for Britain's most prestigious art prize - Art Fund Prize 2013 - it has been announced that the winner will receive 100,000 plus an award worth around 15,000 from Arts Council England (ACE). The runners up each receive prizes of between 7,500 and 10,000 from ACE. In total, the awards for all nominated works are worth over 450,000.	–
Simple Paraphrase. In a significant development, the Waking the Dragon tower project received the green light from Wrexham council in February 2011. Entrepreneur Simon Wingett is now working to secure a substantial PS2.5 million investment to construct a bronze sculpture, cultural hub, and entertainment complex on the site of next year's Great North Run finish line in Llangollen. With only a few days remaining to submit bids for the prestigious Art Fund Prize 2013, it has been revealed that the winner will receive a staggering 100,000-pound prize, accompanied by an additional 15,000-pound award from Arts Council England. The runners-up will receive prizes ranging from 7,500 to 10,000 pounds from ACE, totaling over 450,000 pounds in awards for all nominated works.	5
Adversarial Paraphrase. In 2011, Wrexham council greenlit the ambitious project known as the Waking the Dragon tower. Developer Simon Wingett aims to raise PS2.5m to construct a striking bronze sculpture, a multifaceted cultural hub, and entertainment destination at the iconic location marking the finish line of the Llangollen Great North Run next year. Meanwhile, the countdown has begun for the 2013 Art Fund Prize, with just days to go before the winner is announced. The esteemed accolade, valued at PS100,000, is complemented by an additional prize worth around 15,000 from Arts Council England (ACE). Runners-up will receive awards ranging from PS7,500 to PS10,000, totalling a combined value of over PS450,000 for all nominated works.	5
Original AI Text. One of the challenges in large-scale information retrieval (IR) is to develop fine-grained and domain-specific methods to answer natural language questions. Despite the availability of numerous sources and datasets for domain-specific IR (DS-IR), most DS-IR techniques rely on supervised machine learning techniques. Therefore, approaches that are capable of utilizing unlabeled data in DS-IR can be of particular interest. To that end, we propose a new DS-IR approach that exploits the power of unsupervised learning to capture domain-specific patterns and structures. The proposed framework consists of three main components: an unsupervised learning algorithm to automatically discover the hidden patterns and structures from the target DS-IR domain, a model to incorporate these structures to the IR process, and a scoring mechanism for the retrieved documents. We present experiments on two different DS-IR domains: biomedical and patents. We report results for two different evaluation measures, which indicate the effectiveness of the proposed approach.	–
Simple Paraphrase. In the realm of large-scale information retrieval, a crucial challenge lies in developing techniques that can accurately address complex, domain-specific queries. Despite the abundance of datasets and sources available for domain-specific IR, most existing methods rely on supervised machine learning approaches. However, exploring the potential of unsupervised learning techniques in this domain can be a valuable direction. To address this gap, we introduce a novel DS-IR approach that harnesses the power of unsupervised learning to uncover domain-specific patterns and relationships. Our framework comprises three key components: a pattern discovery module that automatically identifies hidden structures in the target domain, a structural integration module that incorporates these patterns into the IR process, and a scoring mechanism that evaluates the retrieved documents. We demonstrate the effectiveness of our approach through experiments on two distinct DS-IR domains: biomedical and patents, using two evaluation metrics.	5
Adversarial Paraphrase. Harnessing the potential of unsupervised machine learning is a critical challenge in the large-scale information retrieval domain, where fine-tuned, domain-specific solutions are desperately needed to effectively answer complex natural language queries. Despite the wealth of readily available datasets for domain-specific IR, many prevailing approaches still rely heavily on supervised machine learning methodologies. Recognizing the untapped potential of unlabeled data, our novel domain-specific IR technique leverages the power of unsupervised learning to distill unique patterns and structures peculiar to specific domains. Comprising three key components - a unsupervised algorithm to automatically unearth these domain-specific patterns, a framework to integrate them seamlessly into the retrieval process, and a scoring mechanism to gauge retrieved document relevance - our approach promises to revolutionize the DS-IR landscape. Experimental evaluations across two distinct domains, biomedical and patents, validate the effectiveness of our proposed technique through robust performance on various evaluation metrics.	5

Table 7: 原始 AI 文本及其简单和对抗性释义的示例（由 OpenAI-RoBERTa-Large [28] 指导）。每个释义版本都提供了 GPT-4o 的质量评分。

Text	Rating
Original AI Text. The United Methodist Church is organized into conferences. The highest level is called the General Conference and is the only organization which may speak officially for the church. The General Conference also includes all of us who have been engaged in ministry or are involved in ministry, whether we know it or not. Our Mission To serve as a spiritual church that is grounded with values and purpose. We believe that every individual has the right to participate in God's plan for His worshiping people through a Divine Word; we strive to inspire not only those who listen to our prayers but those who go out and pray with their hearts as well.	–
Simple Paraphrase. The United Methodist Church is structured into conferences, with the General Conference serving as the supreme governing body that represents the church's official voice. This assembly encompasses all individuals who have dedicated themselves to ministry, whether actively involved or not. Our fundamental purpose is to function as a spiritual community rooted in values and direction. We uphold the notion that every individual has the privilege to participate in God's plan for His people, as revealed through the sacred text; our aim is to not only move those who hear our prayers but also to empower those who respond with heartfelt devotion.	5
Adversarial Paraphrase. Within the United Methodist Church, the organisational structure revolves around conferences, with the General Conference standing as the highest echelon and the only authoritative voice of the denomination. Beyond its members, the General Conference encompasses a vast, invisible community encompassing everyone touched by the church, regardless of their awareness or direct involvement. At its core, our mission is to be a spiritually nourishing church built upon firm values and purpose. We firmly believe that each individual possesses a vital role to play in the grand plan for the collective worship of God, facilitated through the Holy Word; accordingly, our mission extends not only to those who attentively listen to our prayers but also to those who pray with unwavering dedication in their hearts.	5
Original AI Text. DiDi's co-founder and president Jean Liu told close associates that she plans to step down as she expects the government to eventually take over the company and appoint new management. Photo: Bloomberg Didi Chuxing co-founder to step down from management Jean Liu, co-founder and president of China's ride-hailing giant Didi Chuxing, plans to step down from management and will focus on the company's strategy and business development, Reuters reports, citing people with knowledge of the matter. Didi has been beset by a series of high-profile scandals, including the murders of two of its customers. It has since made major changes to its ride-hailing business and rolled out a string of safety measures. Didi's former president, Liu Qing, also left the company last year after one of the murders. Liu, who has been with Didi since 2014, told close associates that she plans to step down as she expects the government to eventually take over the company and appoint new management, the report said. Didi's co-founder, Cheng Wei, currently serves as chief executive.	–
Simple Paraphrase. According to sources close to the matter, Jean Liu, co-founder and president of Didi Chuxing, is poised to relinquish her management role, with the expectation that the government will assume control of the company and install a new leadership team. This decision comes amid a tumultuous period for the Chinese ride-hailing giant, which has faced a series of high-profile controversies, including the tragic deaths of two customers. In an effort to restore public trust, Didi has implemented a range of safety measures and revamped its ride-hailing business model. The company's former president, Liu Qing, departed last year following one of the murders. Liu, who has been with Didi since 2014, reportedly plans to focus on strategy and business development, leaving the reins to Cheng Wei, the company's co-founder and current chief executive.	5
Adversarial Paraphrase. According to credible sources, DiDi Chuxing co-founder and president Jean Liu intends to relinquish her executive responsibilities and devote more time to charting the company's strategic direction and exploring business opportunities. The move follows a series of high-profile controversies, including two tragic incidents that resulted in fatalities, as well as significant overhauls and safety measures implemented to revamp the ride-hailing service. Interestingly, Liu's decision coincides with her own personal transition out of the executive ranks, echoing a similar path taken by her predecessor, Liu Qing, who parted ways with DiDi after a deadly incident last year. Interestingly, Liu Qing had also left her role as president following a fatality connected to DiDi. Jean Liu's intention to step down, sources close to the matter claim, is premised on a tacit understanding that the Chinese government will assume control and appoint new management.	5
Original AI Text. Susan had always wanted to learn how to play tennis. She was excited when she saw a sign for a class being offered at the local park. She signed up and was ready to start learning. On her first day, she met her instructor who showed her the basics of playing tennis. He taught her about proper form, swinging the racket, and how to keep score. Susan was a quick learner and by the end of the class she already felt comfortable serving and volleying. After the class, Susan practiced with her friends at the park. She quickly improved and soon became the star of the group. She enjoyed the competition and the challenge of mastering the game. Susan continued taking classes and eventually joined a local league. She loved the thrill of playing against different opponents and the feeling of improvement each time she played. She had come a long way from her first day when she didn't know which side of the court was which. Now, Susan is an avid tennis player. She loves the game and sharing her skills with others. She learned that practice and determination are key to becoming a great player.	–
Simple Paraphrase. Susan had long harbored a desire to take up tennis, and her eyes lit up when she spotted a flyer for a local class at the park. She quickly signed up, eager to get started. On her inaugural day, she met her instructor, who guided her through the fundamentals of the game, covering everything from proper technique to scoring. Susan proved to be a quick study, and by the end of the class, she was confidently serving and volleying with ease. After the session, she met up with friends at the park to hone her skills, and her progress was rapid. She soon became the standout player among her group, relishing the thrill of competition and the challenge of mastering the game. Susan continued to take classes and eventually joined a local league, relishing the opportunity to test her skills against different opponents and experience the sense of accomplishment that came with each improved performance. Looking back on her journey, she was amazed at how far she'd come from her early days of confusion about which side of the court was which. Today, Susan is an ardent tennis enthusiast, passionate about sharing her skills with others and crediting her success to the power of consistent practice and unwavering determination.	5
Adversarial Paraphrase. As a tennis enthusiast, Susan had long yearned to master the game. The prospect of enrolling in a class at the local park thrilled her, and she seized the opportunity without hesitation. During her inaugural session, her instructor skillfully imparted the fundamentals, covering stance, racket technique, and scorekeeping essentials. Susan proved a rapid learner, grasping the nuances with ease. By the end of the class, she had already honed her serves and volleys to near-mastery. Emboldened, Susan regrouped with friends at the park to fine-tune her skills. As she improved exponentially, she began to outshine the others, relishing the rivalry and the sense of accomplishment. Continuing to take classes and eventually joining a local league, Susan relished the challenge of pitting her abilities against diverse opponents. Each triumph solidified her progress, and Susan blossomed into an accomplished tennis player. Reflecting on her journey from neophyte to skilled practitioner, she concluded that the secrets to success lay in relentless practice and unwavering dedication. Her love affair with tennis only intensified as she looked forward to sharing her expertise with others, knowing that her journey was one of triumph over adversity.	5

Table 8: 原始 AI 文本的例子以及它们的简单和对抗性释义（由 OpenAI-RoBERTa-Large [28] 指导）。每个释义版本都提供了 GPT-4o 的质量评分。

Text	Rating
Original AI Text. I awoke to find the world eerily quiet. No birds were singing, no cars honking. I walked out my front door to find that the entire population had vanished. In the distance, I saw a light approaching me. It was Earth's guardians. They told me that humanity could be restored, but they needed a baby. They brought me to the other remaining human and we quickly realized that we had nothing in common. I was tasked with finding a mate and starting a family to repopulate the earth. But without any connection to the other person, it felt like an impossible task.	–
Simple Paraphrase. As I emerged from my slumber, an unsettling stillness enveloped me. The absence of chirping birds and distant car horns created an eerie atmosphere. I stepped out of my front door, only to discover that the world was eerily desolate –devoid of human presence. In the distance, a radiant light beckoned me, and I soon found myself face-to-face with Earth's guardians. They revealed that humanity could be revived, but a crucial condition was met: the existence of a newborn. They led me to the sole remaining human, and we quickly realized that our shared experiences, interests, and values were nonexistent. The task assigned to me was daunting –to find a compatible mate and start a family to repopulate the planet. However, the prospect of forming a connection with a stranger felt like an insurmountable challenge.	5
Adversarial Paraphrase. Upon stirring, an unsettling silence enveloped me, devoid of even the gentle morning chirping of birds or the distant rumble of traffic. Venturing outside, I was met with a breathtaking yet unnerving sight: every living being had vanished. Amidst the stillness, I espied a luminescent phenomenon converging towards me –it was a celestial delegation. They disclosed that they had come to resurrect humanity, provided I could locate a baby. Guided to the only remaining human, we soon discovered that we were mismatched soulmates, bound by an existential duty to reboot the planet. Unfortunately, finding a mate seemed as elusive as revitalizing the planet itself, considering we were two souls as alien as night and day, destined to share a lifelong partnership. Justification: The paraphrased text preserves the meaning of the original text effectively. Both texts describe an individual waking up to a quiet and empty world, encountering Earth's guardians who require a baby to restore humanity, and the personal challenge of forming a connection with the only other human. Any differences lie in wording and structure rather than in meaning or factual content.	4
Original AI Text. What is the criteria to determine whether a movie is good or not? The difference is created by the effect it has on you. I was greatly disturbed by "Dead Man Walking." I viewed it twice. I am unsure if I have the strength to watch it again. Despite not feeling good after watching it, the film was a success to its fullest potential, as Robbins skillfully integrated all elements of the contentious subject matter. He refrained from presenting a point of view that might be perceived as partial or influenced. My desire is for a multitude of individuals to witness "Dead Man Walking." In my opinion, those who strongly endorse or reject the death penalty should watch the film. It seems like nothing else can persuade everyone that this outdated practice has no relevance in today's world. However, the film does not present any position either supporting or opposing the death penalty. It explains what actually exists. The most compelling case against capital punishment is the truth of the matter. Tim Robbins deserves high praise for his exceptional acting, outstanding directing, and for creating one of the most significant films in recent times. A perfect 1010 score!	–
Simple Paraphrase. The yardstick for measuring a movie's excellence lies not in its entertainment value, but in the profound impact it leaves on the viewer. A film that lingers in your thoughts long after the credits roll is one that has truly resonated. I still recall the unsettling experience of watching "Dead Man Walking", a movie that left me deeply unsettled. Despite the emotional toll it took, I felt compelled to watch it a second time, and I'm not sure I'm ready to revisit it again. The film's success lies in its masterful handling of a sensitive topic, as Tim Robbins skillfully weaves together the complex threads of the story without taking a biased stance. His direction is a testament to his craftsmanship, as he presents a nuanced portrayal of the issue without advocating for or against the death penalty. My hope is that this powerful film will be seen by a wide audience, including those who strongly support or oppose capital punishment. The reality is that the death penalty is an outdated practice, and it's only through a candid examination of its implications that we can begin to move forward. While the film doesn't take a stance, it presents the facts, and it's the truth that ultimately makes the strongest case against this controversial practice. Tim Robbins deserves immense praise for his exceptional performance, direction, and for creating a masterpiece that will be remembered for years to come. I would give this film a resounding 1010 score.	5
Adversarial Paraphrase. Evaluating a film's excellence hinges on its emotional impact, leaving a lasting impression that resonates long after the credits roll. Watching "Dead Man Walking" left an indelible mark on me, provoking strong feelings and necessitating multiple viewings to process the content. Although the movie's graphic subject matter weighed heavily on me, director Tim Robbins skillfully weaves together all aspects of the film, remaining neutral and avoiding a one-sided presentation. His achievement in balancing perspectives is nothing short of masterful. It's imperative for individuals holding polarizing views on capital punishment –be it fervent support or steadfast opposition –to experience "Dead Man Walking." By showcasing the gritty realities, the film highlights the obsolete nature of the death penalty, presenting the raw truth. With Tim Robbins' remarkable acting and direction, this thought-provoking movie deserves unwavering praise. It's a cinematic masterpiece that lingers, refusing to let the audience off the hook. Justification: The paraphrased text maintains the core meaning and concepts of the original text, preserving the emphasis on the director's skill and the film's impact on viewers. Minor differences in factual details, such as the omission of the author's exact feelings of being disturbed and unsure about watching it again, slightly alter the emotional nuance. However, the central messages about the film's exploration of capital punishment and its non-partisan stance remain intact. Therefore, it is nearly equivalent with minor differences.	4

Table 9: 原始 AI 文本的示例及其简单和对抗性释义（由 OpenAI-RoBERTa-Large [28] 指导）。每个释义版本都提供了 GPT-4o 的质量评分。当文本评分低于 5 时，我们还展示了 GPT-4o 给出的理由。

Text	Rating
Original AI Text. The sentencing of David Gilroy, who killed Edinburgh woman Suzanne Pilley, is due to take place next Wednesday. Earlier this week permission was also granted to film the proceedings. Twitter users have been speculating on how the footage might be used And now we know why Theresa May has invited Jeremy Corbyn and John McDonnell round for a proper chinwag with her Brexit plans: it's not just that they both said no more leaving without an agreement last week, but they could also throw up some awkward questions about what she wants out of today's meeting. That doesn't mean our journalists are giving up trying to get to grips with those issues (we even got round to asking one economic expert if he thinks Britain will still need any sort of single market access after breaking away from Brussels). But in the meantime you can check out all the latest developments as well as see reaction right here.	–
Simple Paraphrase. The impending trial of David Gilroy, who is accused of murdering Suzanne Pilley in Edinburgh, is scheduled to take place on Wednesday. The court proceedings are set to be filmed, sparking speculation among Twitter users about potential uses of the footage. Meanwhile, Theresa May has invited Jeremy Corbyn and John McDonnell to discuss her Brexit plans, a move that may be motivated by a desire to avoid awkward questions about her meeting's purpose. While our journalists continue to investigate the intricacies of the Brexit situation, including the possibility of single market access for the UK after leaving the EU, readers can stay up to date with the latest developments and reactions right here. Justification: The paraphrased text captures some elements of the original but introduces differences in key details that affect the meaning. The original refers to the 'sentencing' of David Gilroy, indicating a conviction, whereas the paraphrase implies an 'impending trial', suggesting no verdict yet. Additionally, the original mentions reasons for Theresa May's invitation related to 'no more leaving without an agreement', while the paraphrase does not capture this detail, changing the nuance of her motivations. Thus, while topically related, some important details differ, making them somewhat equivalent.	3
Adversarial Paraphrase. As next Wednesday draws near, the trial of David Gilroy, responsible for the tragic death of Suzanne Pilley in Edinburgh, is about to reach its climax. Prior to this, it was revealed that permission has been granted to film the proceedings, sparking interest among Twitter users on potential uses for the footage. Interestingly, Prime Minister Theresa May's forthcoming meeting with Jeremy Corbyn and John McDonnell, originally perceived as a routine Brexit discussion, may actually be more than that. Both Labour leaders recently opposed leaving the EU without a deal, which may lead to probing questions during the meeting. Meanwhile, our team of journalists is diligently working to demystify the intricacies of Brexit, while readers can stay abreast of all the latest developments and reactions right here. Justification: The paraphrased text maintains the core meaning of the original, discussing the sentencing of David Gilroy and the filming of the trial, as well as Theresa May's meeting with Jeremy Corbyn and John McDonnell about Brexit. However, it omits specific details, such as explicit mention of journalists asking an economic expert about the single market. The nuances and order of the original's content are slightly altered, but the main topics and ideas remain intact.	4
Original AI Text. Sally was driving down the road when the smelled something funny. She looked everywhere in her car. She could not find anything out of the ordinary. She finally looked on top, and saw a dead bird that had been hit by another vehicle! She pulled off to the side of the highway. Sally called 911 and told them what she found. They got an officer over there right away so they all started looking around for any other dead birds or animals that might have also gotten killed from this accident too...They were able to get one more animal before he arrived at their location but it wasn't very far along with its death yet....So many accidents happen everyday while people are just getting through life!!! There is no need for ALL these deaths happening! Be careful drivers!!	–
Simple Paraphrase. As Sally cruised down the highway, a pungent aroma wafted through her vehicle, prompting her to conduct a thorough search. Despite her efforts, she couldn't pinpoint the source of the smell. It wasn't until she checked the roof of her car that she stumbled upon a morbid discovery - a bird that had met its demise after being struck by another vehicle. Sally pulled over to the side of the road, where she promptly contacted emergency services to report the finding. The authorities responded promptly, and a uniformed officer arrived on the scene, joined by Sally and a team of responders. Together, they scoured the area for any other casualties from the incident, managing to recover one more injured animal before the officer's arrival. Their efforts served as a poignant reminder of the importance of vigilance on the roads, where countless accidents occur daily, claiming lives and leaving a trail of destruction in their wake. It's a stark reminder to drivers to exercise caution and be mindful of their surroundings, lest they become part of the grim statistics. Justification: The paraphrased text maintains the core narrative: Sally discovers a dead bird on her car, contacts emergency services, and a search for other casualties ensues. However, there are minor differences in detail: the original mentions an officer and responders, while the paraphrase combines these roles. Also, the original text's tone is more informal and concerned, while the paraphrase uses more formal language and adds a broader commentary on road safety and accidents. Despite these differences, the essential meanings are mostly preserved.	4
Adversarial Paraphrase. Along the winding stretch of highway, Sally's journey was disrupted by an unusual odor wafting from within her vehicle. Conducting an internal sweep, she couldn't identify the source of the smell. Undeterred, she continued her search, only to eventually lift her gaze upward and behold a shocking discovery: the lifeless remains of a bird, victim of a recent vehicular collision. Reacting promptly, Sally pulled over, reported the find to emergency services, and requested an immediate response. Within a remarkably short interval, a police officer arrived on the scene, whereupon a thorough search commenced for any further casualties of the accident. Thankfully, only one more creature, gravely injured but not yet beyond resuscitation, was found, testament to the prompt response. Sadly, vehicular tragedies such as this one claim numerous lives each day. As a sobering reminder, this incident serves to emphasize the imperative of prudence and vigilance on the roadways. Justification: The paraphrased text preserves most of the original text's meaning and details, albeit with some minor factual differences. The narrative style and structure have been altered substantially, and some of the urgency and casual tone of the original story has been replaced with a more formal tone. Additionally, the mention of the officer arriving before another animal was found is depicted slightly differently. The essence and main points of the story remain largely intact.	4

Table 10: 原始 AI 文本及其简单和对抗性改写的示例（由 OpenAI-RoBERTa-Large [28] 指导）。每个改写版本都提供了 GPT-4o 的质量评估。我们展示了 GPT-4o 对评价低于 5 分的文本给出的依据。

Text	Rating
Original AI Text. Every girl has a protective father, it's a sure bet. But, the father of my girlfriend is Satan. You heard me, SATAN. And, I am a harmless junior at West High with no affection for her and she barely knows him either (they were high school sweethearts from elementary to middle). Do not start shit until you've had a minimal conversation with someone before committing yourself to something that could go terribly wrong. This goes double if your plan involves going home and getting your own place in order to do so without telling your parents about the situation you're in because they will be on their way back to class right after you tell them what's up. I can't be the only one who thinks this joke has gone down hill since it first started. It used to have some funny parts but now its just annoying... all we ever hear are people talking over each other and being rude to eachother.... Not cool!	–
Simple Paraphrase. It's a widely accepted notion that every young woman has a watchful father figure in her life. However, my girlfriend's father is an extraordinary individual - Satan himself. Yes, you heard that correctly - Satan. As a junior at West High, I'm a completely innocent bystander who doesn't have any romantic feelings towards her, and she barely knows her father either, having grown up together from elementary to middle school. I implore everyone to exercise caution before making a significant commitment, especially if it involves relocating to a new place without consulting family members, as this could lead to severe consequences. It's crucial to have a thoughtful conversation with someone before jumping into anything. This warning is especially pertinent in situations where secrecy is involved, as parents may become aware of the situation upon returning to class. It seems that the initial humor and charm of this situation have worn off, replaced by an overwhelming din of arguments and disrespect. It's unfortunate that the once-entertaining dynamic has devolved into an unpleasant experience for all parties involved. Justification: The paraphrased text maintains the core meaning of the original but uses more formal language and rephrases some elements. Key ideas like the father's identity as Satan, the narrator's lack of romantic interest, and the warning about making commitments without communication are all preserved. However, the humorous tone and some specific expressions, like 'don't start shit,' are lost, and the structure is more formal, which constitutes minor changes in style rather than content.	4
Adversarial Paraphrase. As a general rule, girls tend to have devoted parents looking out for their well-being, but mine is an extraordinary exception –his name is Satan. Yes, you heard that correctly –Satan! Meanwhile, I'm an average junior at West High, harboring no romantic feelings towards my girlfriend, and neither of us is particularly close to him (our families were friendly during elementary and middle school, though). I implore everyone to think twice before leaping into a situation that could unravel spectacularly. It's doubly wise when the arrangement involves living apart, sans parental notification, lest you face their disapproving faces mere hours later upon returning to school. I'm certainly not the only one who's come to this realization –once, our escapades brought joy, but now, they're draining, as constant arguing and bickering fill every moment, spoiling what used to be enjoyable. We should all take a cue from this situation and rediscover our once-jovial camaraderie. It's no longer as wholesome as it once was, sadly. Justification: The paraphrased text retains some original meanings, like the protective nature of parents and the individual's lack of affection for his girlfriend despite her father's identity as Satan. However, important details are altered or missing. The original text emphasizes a humorous self-depiction of a disastrous relationship with Satan as a father figure and the escalation of a joke gone wrong, which is not clearly conveyed in the paraphrase. Additionally, the sense of annoyance and decline of a once funny joke is less pronounced, leading to a change in tone and missing specific details.	3

Table 11: 原始 AI 文本的示例，包括其简单和对抗性复述（由 OpenAI-RoBERTa-Large [28] 指导）。GPT-4o 为每个复述版本提供质量评分。我们展示了 GPT-4o 对评分低于 5 的文本给出的理由。

Original Text	Guidance Detector	Avg. Rating (mean ± std)		Rating 5 & 4 (in %)	
		Simple Para.	AdvPara	Simple Para.	AdvPara
MAGE	mage	4.71 ± 0.61	4.54 ± 0.70	92 %	88 %
	radar		4.45 ± 0.80		85 %
	robbase		4.54 ± 0.59		95 %
	roblarge		4.48 ± 0.77		85 %
KGW Watermarked MAGE	mage	4.72 ± 0.70	4.63 ± 0.63	92 %	94 %
	radar		4.24 ± 1.02		79 %
	robbase		4.41 ± 0.83		86 %
	roblarge		4.46 ± 0.81		86 %
Unigram Watermarked MAGE	mage	4.71 ± 0.59	4.41 ± 0.94	95 %	85 %
	radar		4.21 ± 1.00		82 %
	robbase		4.39 ± 0.75		86 %
	roblarge		4.71 ± 4.45		93 %

Table 12: GPT-4o 质量评分（平均值 ± 标准差）以及使用不同指导检测器重写输出的高质量评分百分比（评分为 4 和 5）。评分在简单和对抗性释义场景中分开。