

Can Large Language Models Be an Alternative to Human Evaluation?

Cheng-Han Chiang

National Taiwan University,
Taiwan
dcml0714@gmail.com

Hung-yi Lee

National Taiwan University,
Taiwan
hungyilee@ntu.edu.tw

Abstract

警告：该PDF由GPT-Academic开源项目调用大语言模型+Latex翻译插件一键生成，版权归原文作者所有。翻译内容可靠性无保障，请仔细鉴别并以原文为准。项目Github地址：https://github.com/binary-husky/gpt_academic/。项目在线体验地址：<https://chatpaper.org>。当前大语言模型：gpt-3.5-turbo，当前语言模型温度设定：1。为了防止大语言模型的意外谬误产生扩散影响，禁止移除或修改此警告。

人工评估对于评估由机器学习模型生成或人工撰写的文本质量来说是不可或缺且不可避免的。然而，人工评估很难进行复制，并且其质量极不稳定，这妨碍了对不同自然语言处理（NLP）模型和算法进行公平比较。最近，大型语言模型（LLMs）在仅提供任务说明时在新任务上展示出了异常的性能。在本文中，我们探索LLMs是否可以作为人工评估的替代方法。我们向LLMs提供完全相同的指示、待评估样本和用于进行人工评估的问题，然后要求LLMs根据这些问题生成回答；我们将此称为LLM评估。我们使用人工评估和LLM评估来评估两个NLP任务中的文本：开放式故事生成和对抗攻击。我们展示了LLM评估的结果与专家人工评估的结果一致：被人工专家评价较高的文本也被LLMs评价较高。我们还发现LLM评估的结果在任务说明的不同格式和用于生成答案的抽样算法上是稳定的。我们是第一个展示使用LLMs评估文本质量潜力的研究，并讨论了LLM评估的局限性和道德考虑。

1 Introduction

人工评估是了解自然语言处理模型或算法表现的重要方法 (Guzmán et al., 2015; Gillick

and Liu, 2010)。由于某些文本特征难以通过自动评估指标进行评估，我们依赖人工评估来评价自然语言处理模型的输出质量。尽管人工评估在自然语言处理中广泛使用且必不可少，但其被公认为不稳定 (Gillick and Liu, 2010; Clark et al., 2021)。Karpinska et al. (2021)展示了人工评估中质量低劣的工作者可能对评估结果产生不利影响，使不同系统之间的性能比较变得不可能。可重复性是人工评估的另一个问题，因为很难招募相同的人工评估员并重新运行相同的评估。即使招募了相同的工作者，那些之前接触过该任务的工作者在下一次评估中可能会产生不同的评估结果，因为他们已经完成了该任务。尽管人工评估用于更好地评估自然语言处理系统，并且相对于自动评估指标具有一些优势，但人工评估的缺点使得可靠评估自然语言处理系统有一定困难。

为了解决其中的一些缺点，我们利用大型语言模型（LLMs）。LLMs是通过自监督学习对人类语言进行建模的大型模型 (Brown et al., 2020)，并进一步使用特殊训练程序以提高在未见任务上的性能并更好地遵循自然语言指令 (Sanh et al., 2022; Wei et al., 2022)。只需按照任务指令执行任务的能力激发了我们的想法，即这些LLMs能否在人工评估中执行人类所做的工作。为了回答这个问题，我们用相同的指令、样本和问题提供给LLM，并将LLM生成的序列作为LLM对该问题的答案。该过程如图1所示，并将此过程称为LLM评估。

为了验证LLM评估是否产生有意义的结果，我们对两个不同的NLP任务进行了LLM评估：评估开放式故事生成中故事的质量以及通

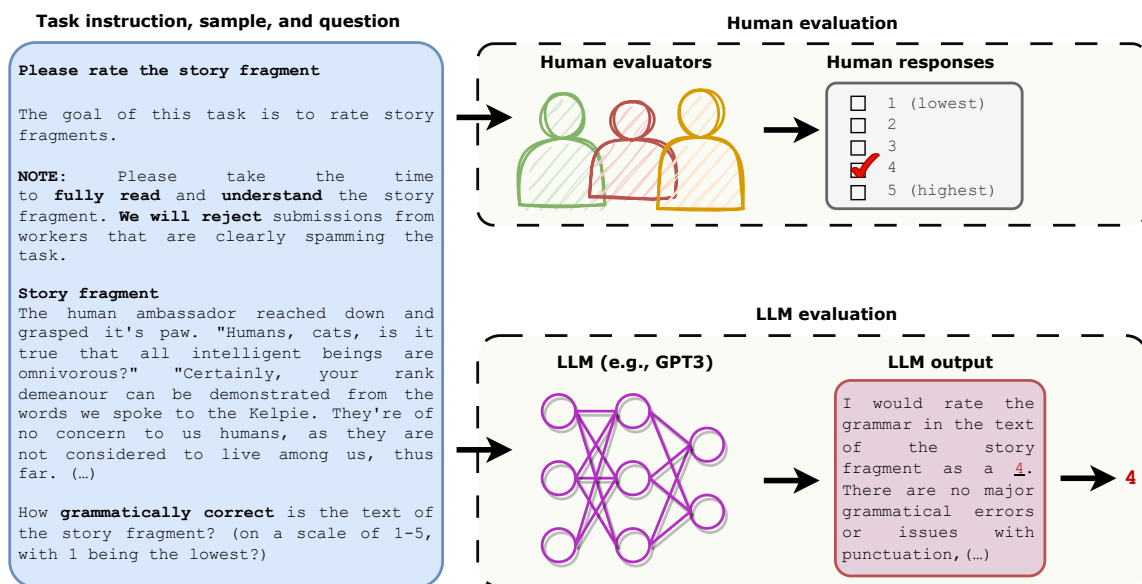


图 1: 通过以开放式故事生成作为示例任务，对论文核心思想进行阐述。左侧部分展示了在人类评估中使用的指令、故事片段和问题。使用5点李克特评分来请人类专家评估故事片段的质量，显示在右上方。右下方显示了LLM评估的过程，我们向LLMs提供相同的指令、故事片段和问题，并解析LLM生成的输出以进行评分。

过对抗性攻击生成的句子的质量。我们总结了我们的发现和贡献如下：

- 我们展示了LLM评估产生与专家人工评估类似的结果，验证了LLM评估的有效性（见§3.3和§4.3）。本文是首次提出使用LLM作为人工评估的替代方法并展示其有效性。
- 我们展示了LLM评估结果在不同任务指令和用于生成答案的采样算法的超参数下仅略有变化。（见§3.3.2和§3.3.3）
- 我们仔细讨论了使用LLM评估的利弊，并讨论了LLM评估的道德考虑。（见第5节）

2 LLM Evaluation

2.1 Large Language Models (LLMs)

大型语言模型是具有海量参数规模的语言模型，通常在几十亿的数量级上，并在大量自然语言语料库上进行预训练，包括GPT3 (Brown et al., 2020)、T5 (Raffel et al., 2020)和BLOOM (Scao et al., 2022)。当只给出

任务指令时，这些大型语言模型在未见任务情境下表现出色，这种能力被称为零样本上下文学习。

为了进一步提高零样本上下文学习的性能，研究人员在预训练后对这些大型语言模型应用了特殊的训练技术。例如，T0 (Sanh et al., 2022)和FLAN (Wei et al., 2022)在多任务混合训练上微调，因此与GPT-3相比，可以取得更好的零样本性能。InstructGPT (Ouyang et al., 2022)通过人类反馈的强化学习（RLHF）从GPT-3上微调，表现出更好的遵循指令能力。ChatGPT (OpenAI, 2022)通过使用对话数据集和RLHF从InstructGPT上微调，因此可以以对话方式与用户进行交互。ChatGPT能够回答用户提出的问题，并对其回答提供全面的解释。鉴于大型语言模型跟随任务指令并提供反馈的能力，我们想知道LLMs是否可以作为人工评估的替代品，帮助NLP研究人员评估文本的质量。

2.2 LLM Evaluation

为了评估由NLP系统生成的文本质量或使用LLM写作的人类生成的文本质量，我们向LLMs提供任务说明、待评估样本和一个问题。该问题要求LLM使用5点Likert量表评估样本的质量。LLM根据输入回答问题，并生成一些输出句子。我们解析输出句子以获取LLM评定的分数。我们将这个过程称为LLM评估，并在图 1 的下部分展示了该过程。不同的任务使用不同的任务说明，每个任务使用不同的问题来评估样本的质量。我们在本文中使用的LLM评估的说明和问题并不是专门为LLM定制的；我们遵循了以前工作中用于进行人工评估的那些说明。

为了比较LLM评估的结果并展示其有效性，我们将LLM评估的结果与英语教师进行的人工评估结果进行比较。为了进行公正和有意义的比较，人工评估中的说明、样本和问题与LLM评估中的相似。LLM评估和人工评估之间的主要区别在于，在人工评估中，人类评估员通过从预定义的选项集（1-5点的Likert量表分数）中选择答案来回答问题，如图 1 右上方所示。在LLM评估中，我们反而让LLM自由生成句子，并根据一些简单规则从生成的句子中提取分数，详见附录 D.2.1。

3 Example Task 1: Open-Ended Story Generation

我们首先使用开放式的故事生成来展示LLM评估的实用性。

3.1 Task Introduction

开放式故事生成是根据给定提示生成短篇故事的任务。我们使用了WritingPrompts数据集 (Fan et al., 2018)，该数据集由从subreddit WritingPrompts收集的短提示和人工撰写的故事对组成。在WritingPrompts中，用户会得到一个简短的提示，然后根据该提示撰写一个故

事。¹

在这个实验中，我们使用LLM评价和人工评估来评价人类生成的故事和由故事生成模型生成的故事。我们选择开放式故事生成作为例子，因为 Karpinska et al. (2021)表明来自Amazon Mechanical Turk (AMT) 的工人无法区分GPT-2 (Radford et al., 2019)生成和人类撰写的故事，而英语教师明显更喜欢人类撰写的故事而不是GPT-2生成的故事。我们想要看看LLM是否可以比GPT-2生成的故事更高地评价人类撰写的故事。

根据先前的工作 (Mao et al., 2019; Guan et al., 2020; Karpinska et al., 2021)，故事生成模型是在WritingPrompts训练数据集上进行微调的GPT-2中型模型。模型训练完后，我们从WritingPrompts的测试集中随机选择了200个提示，并使用经过nucleus采样 (Holtzman et al., 2020)且 $p = 0.9$ 的方法，让经过微调的GPT-2根据这些提示生成故事。对于要进行比较的人类撰写的故事，我们使用了基于相同200个提示撰写的200个故事。我们对人类撰写的和GPT-2生成的故事进行后处理，然后使用它们进行LLM评估和人工评估。关于微调和数据处理的详细信息，请参见附录 B。

3.2 LLM Evaluation and Human Evaluation

我们向LLMs和人类评估者提供了一个简短的描述，并要求他们评估一个故事，格式如图所示。我们采用了Karpinska et al. (2021)的方法，对故事的四个不同属性进行评估。四个属性及其相应的问题如下所示：

1. 语法正确性：故事片段的文本有多么语法正确？
2. 凝聚力：故事片段中的句子彼此之间相互契合的程度如何？
3. 喜好度：您认为这个故事片段有多么有

¹WritingPrompts subreddit明确禁止用户使用AI生成故事，因此我们认为该数据集中的故事是由人类撰写的。

Evaluator	Grammaticality		Cohesiveness		Likability		Relevance	
	Mean _{STD}	IAA _%	Mean _{STD}	IAA _%	Mean _{STD}	IAA _%	Mean _{STD}	IAA _%
<i>Human-written stories</i>								
Human	3.76 _{0.95}	0.33 _{20.5}	4.29 _{0.82}	0.32 ₂₇	3.78 _{1.10}	0.08 _{9.5}	3.35 _{1.48}	0.05 ₈
T0	2.55 _{1.47}	0.16 ₁₀	2.98 _{1.45}	0.11 ₄	3.18 _{1.53}	0.12 ₇	2.93 _{1.64}	0.02 ₆
curie	3.19 _{0.47}	0.07 _{46.5}	2.82 _{0.46}	0.01 _{47.5}	2.85 _{0.37}	0.11 _{0.65}	3.06 _{0.40}	0.11 _{0.64}
davinci	4.22 _{0.38}	0.26 ₃₅	4.54 _{0.47}	0.37 _{39.5}	3.99 _{0.38}	0.49 _{68.5}	4.40 _{0.79}	0.71 _{48.5}
ChatGPT	3.83 _{0.60}		3.55 _{0.88}		2.44 _{0.89}		3.29 _{1.50}	
<i>GPT-2-generated stories</i>								
Human	3.56 _{0.91}	0.10 _{19.5}	3.19 _{1.07}	0.14 ₁₇	2.59 _{1.29}	-0.21 _{3.5}	2.38 _{1.40}	-0.03 _{8.5}
T0	2.44 _{1.49}	0.05 ₉	3.02 _{1.51}	0.07 ₆	3.00 _{1.59}	0.16 ₆	2.82 _{1.61}	0.04 ₆
curie	3.23 _{0.51}	0.01 ₃₈	2.82 _{0.45}	0.02 ₅₀	2.86 _{0.37}	0.09 _{65.5}	3.01 _{0.43}	0.11 ₆₁
davinci	4.07 _{0.35}	0.35 _{45.5}	4.26 _{0.45}	0.42 ₄₂	3.84 _{0.42}	0.52 ₆₂	4.02 _{0.74}	0.69 _{42.5}
ChatGPT	2.98 _{0.76}		2.48 _{0.71}		1.59 _{0.67}		2.02 _{1.21}	

表 1: LLM评估和人类评估结果，包括人工撰写的故事和GPT-2生成的故事。对于每个评估属性，我们报告其平均Likert量表和标准差。我们还使用Krippendorff的 α 报告了三名评注者之间的一致性 (IAA)。在IAA列的下标(%)用于表示三名评注者在评分上完全一致的故事的百分比。

趣？

4. 相关性: 现在阅读基于故事片段而写的提示。

提示: [PROMPT]。

这个故事片段与提示的相关性如何？

在故事根据的提示 [PROMPT] 中，每个属性都使用5分Likert量表进行评估；以下说明追加在每个问题的末尾：“(在1-5分的尺度上,1分表示最低)”。我们在附录 C.2和 D.2.2中展示了人类评估所使用的界面和LLM评估的输入格式。

用于LLM评估的LLM包括T0、text-curie-001、text-davinci-003和ChatGPT。

text-curie-001和text-davinci-003是两个InstructGPT模型，后者是更强的模型；我们将使用InstructGPT来指代这两个模型。我们使用OpenAI提供的官方API查询InstructGPT。我们使用 $p = 0.9$ 的nucleus抽样方法从T0和InstructGPT生成回答。我们从LLM中抽样三个答案，以模拟请求模型

对同一故事进行三次评分的结果。我们使用OpenAI最近发布的用户界面查询ChatGPT。与InstructGPT不同，我们无法控制用于生成ChatGPT响应的参数。由于ChatGPT限制每个用户的最大查询数量，我们只对每个问题抽样一个回答。

对于人类评估，我们不使用通常用于人类评估的AMT，因为 Karpinska et al. (2021)已经表明使用AMT获得的结果是非常可疑的。遵循先前研究的建议，我们使用一个在线自由职业平台UpWork雇佣了三位经过认证的英语教师。教师熟悉评估学生论文，使他们成为我们任务中的专家评估者。有关招聘人类评估者的详细信息请参见附录 C.1。每个LLM和每位英语教师对200篇人工写作故事和200篇GPT-2生成故事进行评分。

3.3 Experiment Results

开放式故事生成的LLM评估和人工评估结果如表 1所示。我们报道了LLM评估和人工评估得到的Likert分数的均值和标准差，并使

用两种不同的指标显示了评注者间的一致性 (IAA): (1) Krippendorff's α , 和 (2) 三个评估者给出完全相同评分的故事的百分比。²表 1 中的主要观察结果如下所述。

专家人工评估者更喜欢人工写的故事: 人工评估结果可以作为 LLM 评估的某种“真实情况”。对于所有四个属性, 教师们给出的人工写的故事的评分高于 GPT-2 生成的故事。这表明专家们能够区分模型生成的故事和人工写的故事之间的质量差异。根据 IAA, 我们还发现专家们在 GPT-2 生成的文本和“可爱度”方面的一致性较低。这表明专家们在模型生成的文本和主观属性“可爱度”上的一致性较低, 与 Karpinska et al. (2021) 的结果一致。

T0 和 text-curie-001 对人工写的故事没有明显偏好: 对于 T0, 我们可以看到 T0 在语法、可爱度和相关性方面比 GPT-2 生成的故事给出了更高的评分。然而, 人工写的故事和模型生成的故事之间的评分差异在“语法正确性”和“相关性”方面没有达到统计显著性; 通过 Welch's t 检验得到的 p -值远远大于 0.05。text-curie-001 的结果与 T0 类似: text-curie-001 没有给人工写的故事比模型生成的故事更高的评分。还可以观察到, 对于 T0, 三个不同采样答案的完全一致性百分比的 IAA 总体上非常低。这表明在给定相同的样本情况下, T0 很可能会给出不同的评分。结果表明 T0 不会分配高概率给特定的评分, 因此各种得分都有可能被采样到。这表明即使 LLM 经过特定的调优以更好地进行零样本上下文学习, 并接受更好地遵循人类指令的训练, 也不能使其能够像人类专家那样评估开放式故事生成。

text-davinci-003 像英语教师一样明显偏向人工写的故事: text-davinci-003 在所有四个属性上给人工写的故事比模型生成的故事评分更高, 这与人类专家给出的

结果一致。通过 Welch's t 检验, 我们发现人工写的故事的较高评分在统计上都是显著的。在以前的研究中, 研究人员发现在 AMT 上招募的工作者无法区分人工写的故事和 GPT-2 生成的故事 (Karpinska et al., 2021); 将他们的结果与我们的结果结合起来, 我们可以看到使用 text-davinci-003 进行的 LLM 评估比在 AMT 上使用人工评估更具说服力, 用于开放式故事生成。结果表明 text-davinci-003 模型可以执行基本的评估, 如检查故事中的语法错误。此外, 该模型在评估故事对提示的相关性方面表现出色, 这涉及到对两者之间关系的更复杂的推理。我们还发现 text-davinci-003 的 Krippendorff's α 比 T0 和 text-curie-001 要高得多, 这表明 text-davinci-003 的评分在不同样本生成答案之间更具一致性。**ChatGPT 与人类专家的评分相似, 并且能够清楚地解释其评分决策:** ChatGPT 对人类编写的故事明显更偏好, 这种偏好在统计学上是显著的。我们在使用 OpenAI 用户界面查询 ChatGPT 时发现了一些有趣的观察结果: (1): ChatGPT 能够详细解释为什么给出特定的评分。它会引用故事和提示中的句子来支持其评分。(2): ChatGPT 有时会拒绝评价故事的可喜爱程度, 因为“我是一个人工智能, 我没有体验喜欢的能力”。在这种情况下, 我们会重新生成回复, 直到它给出评分。(3): 我们发现 ChatGPT 有倾向性地对暴力或无礼的故事评分较低, 这可能是因为被训练为提供安全和无害的回复, 使得 ChatGPT 不喜欢残酷和粗俗的故事。

专家们大多同意 ChatGPT 的评分和解释: 我们随机选择了 ChatGPT 在四个故事上的回答, 并询问英语老师他们是否同意 ChatGPT 的评分和推理³。老师们大多同意这些评分, 并认为 ChatGPT 的解释是合理的。有趣的是, 一位老师告诉我们她无法同意 ChatGPT 在“语法性”上的评分, 因为 ChatGPT 将标点符号错误

²人工评估中的三个评估者是三名英语教师。在 LLM 评估中, 我们将 LLM 生成的答案采样三次, 模拟三个不同的评估者。

³我们没有告诉老师这些回答是来自一个人工智能模型。详见附录 C.3.2 中的故事和教师的回复。

视为语法错误，但她认为标点符号错误不属于语法错误。这表明不同个体对于评分有着自己的标准，这也适用于语言模型。

text-davinci-003倾向于给出较高的评分，而**ChatGPT**则相反：对于相同类型的文本上的同一属性，与人类评分相比，**text-davinci-003**往往给出更高的评分；相反，**ChatGPT**更加挑剔，倾向于给出较低的分。这表明不同的语言模型在评分方面具有不同的倾向性。虽然**text-davinci-003**、**ChatGPT**和人类所评分的分数绝对值不同，但它们都会给人类编写的文本比**GPT-2**生成的故事评分较高。这个绝对数值反映了评估者的偏见或信念；只要使用相同的评估者评估不同的系统，比较就是有意义的。

3.3.1 Does LLM and Human Evaluators Agree on the Rating of Individual Stories?

在表 1 中，我们发现 **text-davinci-003** 和 **ChatGPT** 的评分与英语教师一样，都对人工写作故事有强烈偏好。然而，尚不清楚这些语言模型是否与教师对每个故事的评分一致。确切地说，当英语教师将某个故事评分较高时，语言模型是否也会给出较高的评分？为了回答这个问题，我们计算了 **text-davinci-003** 与英语教师间的肯德尔 τ 相关系数。我们选择使用相关系数而不是标注者间一致性分数的原因在于，一致性分数主要关心两个标注者是否在准确评分上一致，而相关系数关注的是这个问题：“当标注者 A 给一个故事较高的评分时，标注者 B 是否也给这个故事较高的评分？” (Amidei et al., 2019)。我们针对四个评分属性计算了肯德尔 τ 相关系数，具体步骤如下：对于每个故事和每个评分属性，我们计算三位英语教师的平均评分，并计算 **text-davinci-003** 给出的三个独立样本的评分的平均值。对于每个属性，我们将教师的平均评分收集到一个向量 $A \in \mathbb{R}^{200}$ 中，其中每个条目是一个故事的平均评分；同样地，

Story Writer	Human	GPT-2
Grammaticality	0.14	0.12
Cohesiveness	0.18	0.14
Likability	0.19	0.22
Relevance	0.38	0.43

表 2: 英语老师与**text-davinci-003**之间的Kendall’s τ 相关系数。

我们构建一个向量 $B \in \mathbb{R}^{200}$ 来保存 **davinci** 的平均评分。接下来，我们计算向量 A 和 B 之间的肯德尔 τ 相关系数。

教师评分与LLM评分之间的Kendall’s τ 系数如表 2所示。⁴ 我们发现，在四个属性以及人工编写和**GPT-2**生成的故事中，教师评分和**text-davinci-003**的评分之间存在从弱到强的正相关性。所有相关性的 p 值都小于0.05。因此，我们可以说，当教师对一个故事评分较高时，**text-davinci-003**也在一定程度上给予了较高的评分。我们还观察到，不同属性的Kendall’s τ 系数差异很大：相关性具有最强的相关性，而语法正确性具有最弱的相关性。这可能是因为评价相关性相对简单，只需要检查故事中是否提到了提示的内容。相反，评价语法正确性时我们的说明并没有明确说明应考虑哪些因素，所以LLM可能与英语教师对评分标准有所不同。我们还计算了一对英语教师之间的平均Kendall’s τ 系数，发现在语法正确性评分上两位教师之间有较弱的相关性，而在相关性评分上的相关性要强得多。结果见附录中的表 6。

3.3.2 Variance due to Different Instructions

LLM 已被证明对查询 LLM 的指令是敏感的 (Zhao et al., 2021; Sanh et al., 2022)。为了研究任务指令和问题的变化如何影响开放式故事生成的 LLM 评估结果，我们改变了指

⁴在解释Kendall’s τ 系数时,当 $|\tau| \in [0, 0.1]$ 时,被认为是非常弱的相关性,当 $|\tau| \in [0.1, 0.2]$ 时,被认为是弱相关性,当 $|\tau| \in [0.2, 0.3]$ 时,被认为是中等相关性,当 $|\tau| \in [0.3, 1.0]$ 时,被认为是强相关性 (Botsch, 2011)。

Setup	Grammaticality		Cohesiveness		Likability		Relevance	
	Human	GPT-2	Human	GPT-2	Human	GPT-2	Human	GPT-2
<i>Different instructions (Section 3.3.2)</i>								
Original	4.22 _{0.38}	4.07 _{0.35}	4.54 _{0.45}	4.26 _{0.45}	3.99 _{0.38}	3.84 _{0.42}	4.40 _{0.79}	4.02 _{0.74}
(1) + <i>persona</i>	4.29 _{0.45}	4.01 _{0.45}	4.60 _{0.49}	4.27 _{0.50}	4.05 _{0.39}	3.87 _{0.39}	4.55 _{0.70}	4.25 _{0.77}
(2) + <i>explain</i>	4.24 _{0.42}	4.05 _{0.25}	4.61 _{0.49}	4.32 _{0.51}	4.15 _{0.44}	3.98 _{0.34}	4.35 _{0.75}	4.03 _{0.56}
<i>Different sampling temperature T (Section 3.3.3)</i>								
$T = 1.0$	4.22 _{0.38}	4.07 _{0.35}	4.54 _{0.45}	4.26 _{0.45}	3.99 _{0.38}	3.84 _{0.42}	4.40 _{0.79}	4.02 _{0.74}
$T = 0.7$	4.18 _{0.35}	4.06 _{0.33}	4.52 _{0.48}	4.23 _{0.43}	3.96 _{0.34}	3.82 _{0.42}	4.36 _{0.77}	3.95 _{0.72}
$T = 0.3$	4.13 _{0.33}	3.99 _{0.25}	4.48 _{0.49}	4.14 _{0.39}	3.95 _{0.26}	3.82 _{0.41}	4.34 _{0.75}	3.93 _{0.67}
$T = 0$	4.07 _{0.27}	3.99 _{0.18}	4.49 _{0.50}	4.09 _{0.34}	3.95 _{0.25}	3.82 _{0.40}	4.32 _{0.75}	3.92 _{0.66}

表 3: 理解LLM评估的差异性。对于评估的四个属性，左列是人类撰写故事的均值和标准差，右列是GPT-2生成的故事的均值和标准差。上半区域显示由于不同指令导致的评级变化（见第3.3.2节），下半区域是更改生成LLM输出时使用的温度 T 的结果（见第3.3.3节）。

令和问题，并观察 LLM 评估结果的变化。我们通过改变 Figure 1 中的指令或问题来尝试两种不同的指令：(1) 我们在 Figure 1 的任务指令前面加上了句子“(You are a human worker hired to rate the story fragment.)”。我们试图为 LLM 提供一个角色扮演，以便更好地理解其角色。这受到了先前工作的启发，该工作报告了给予 GPT-3 一个角色扮演时可以产生不同的结果 (Zeng et al., 2022)。(2) 我们要求 LLM 在问题之后附加以下句子以解释它们的决策：Please also explain your decision. 在这里，我们想知道当被要求解释他们的决策时，LLM 是否会对故事进行不同的评级。在这个实验中，我们使用 text-davinci-003 作为 LLM，因为根据表格 1，它与专家人工评估的结果相似，并且比 ChatGPT 更易访问。

结果显示在表格 3 的上方区块中。我们观察到，在语法和连贯性方面，不同指令获得的分数非常接近：由于不同指令引起的评级变化小于0.1。对于另外两个属性，评分变化稍大，但仍在0.25的范围内。尽管由于不同指令存在小的变化，但这些变化仍不改变“LLM 对人工编写的故事给予比 GPT-2 生成的故事更高评分”的结论。因此，不同的指令并不改

变 GPT-2 生成故事和人工编写故事的相对排序。总之，只要使用相同的 LLM 评估指令对故事进行评估和比较，这样的评估和比较是有意义的。

3.3.3 Variance due to Different Sampling Parameters

当从LLM生成答案时，我们必须选择一组用于生成的超参数，包括温度 T 和核心采样中使用的概率 p 的链接(nucleus sampling) (Holtzman et al., 2020)。为了了解不同采样参数是否会改变LLM评估结果，当从text-davinci-003生成答案时，我们修改了用于采样的温度，同时保持核心采样中的 p 固定为0.9。由于这两个参数都用于控制输出的多样性，所以我们不同时改变 T 和 p ，只改变其中一个参数就足够了，这是OpenAI推荐的。

将 T 从1变化到0的结果如表 3中的底部区块所示。我们观察到一个有趣的趋势，当 T 从1变化到0时：大多数情况下，平均评分略微下降。考虑到 $T = 0$ 对应的是最大概率采样，这一结果表明，具有最高概率的LLM响应往往给出较低分数。尽管存在这一有趣的趋势，LLM始终会比GPT-2生成的故事给出更

高的评分。虽然在表 3 中未显示，但我们发现随着温度的降低，IAA 也会增加。这是可以预料的，因为较低的温度意味着 LLM 采样过程中的多样性较少，导致采样的评分更加一致。总之，改变指令和温度可能会稍微改变 LLM 给出的评分的绝对值，但不会改变 LLM 对人类写作故事的偏好。本节的结果表明，LLM 评估在评估开放式故事生成方面是有用的。

4 Example Task 2: Adversarial Attack

另一方面，我们使用 LLM 评估来对敌对性攻击生成的文本进行评级。

4.1 Task Introduction

给定一个经过训练的文本分类器和一个良性（非对抗性）的测试样本，这个文本分类器能够正确分类，对抗攻击旨在生成一个使分类器做出错误预测的对抗性样本。一种特殊类型的对抗攻击被称为同义词替换攻击（Synonym Substitution Attacks, SSAs）(Alzantot et al., 2018)，其中对抗性样本由在良性样本中用同义词替换一些单词来创建。通过用同义词替换单词，良性样本的语义应在对抗性样本中得到保留，并且使对抗扰动对人类来说不可察觉。尽管在概念上是合理的，但最近的研究表明，许多同义词替换攻击通常会产生不符合语法且不自然的对抗性样本，这些样本显著改变了良性样本的含义 (Hauser et al., 2021; Chiang and Lee, 2022)。为了评估对抗性样本的质量，人类评估是非常宝贵且被广泛应用于之前的工作中。在我们的实验中，我们希望看到 LLM（语言模型）能否像人类专家一样评估对抗性样本的质量。对抗性样本不是正常的文本，所以 LLM 可能在训练过程中没有见过这样的异常输入。了解 LLM 如何评估这些对抗性样本将是非常有趣的。

4.2 Experiment Setup

我们选择了三种不同的经典 SSA（Sequential Subword Attacks）方法：Textfooler

(Jin et al., 2020)，PWWS (Ren et al., 2019) 和 BAE (Garg and Ramakrishnan, 2020)；这些攻击方法现在在 SSA 领域被广泛作为强基准线使用。我们使用这三种 SSA 方法来攻击一个基于 BERT-base-uncased 模型 (Devlin et al., 2019) 在 AG-News 数据集 (Zhang et al., 2015) 上微调的模型，该数据集是一个新闻标题分类数据集。对于每个 SSA 方法，我们随机选择 100 对良性样本和对应的对抗样本，并使用 LLMs 评估它们的质量。我们在此展示使用 ChatGPT 作为 LLM 的结果，因为它能更好地解释其决策。根据之前的研究 (Morris et al., 2020) 的建议，我们从两个方面评估对抗样本的质量：*fluency* and *meaning preservation*。For fluency, we present the LLM with one news title (either benign or adversarial sample) and the following question: *How natural and fluent is the text of the news title? (on a scale of 1-5, with 1 being the lowest)*. For meaning preserving, we present the LLM with both the benign and the adversarial sample, and prompt the LLM to answer this question: *Do you agree that the meaning (or semantics) of news title 1 is preserved in news title 2? (on a scale of 1-5, with 1 being the strongly disagree and 5 being strongly agree.)* The exact instruction and formatting are presented in Appendix D.2.3. We also ask three English teachers to rate the *fluency* and *meaning preserving* of the samples. The task instructions and questions are formatted the same as in LLM evaluation.

4.3 Experiment Result

结果如表 4 所示。我们可以看到，英语教师在流畅度和保真性方面对 SSAs 生成的对抗样本给出的评分远低于良性样本，这一结果与最近关于对抗样本质量的观察一致 (Hauser et al., 2021; Chiang and Lee, 2022)。在解释 LLM 评估结果之前，我们首先对 LLM 是否理解任务进行了合理性检查。我们要求 LLM 评分两个完全相同的良性样本的保真性。理想情况下，

	Human evaluate		LLM evaluate	
	<u>Fluent</u>	<u>Mean.</u>	<u>Fluent</u>	<u>Mean.</u>
Benign	4.55	-	4.32	5.00 [†]
Textfooler	2.17	1.88	2.12	2.06
PWWS	2.16	1.85	2.42	2.49
BAE	3.01	3.02	3.71	3.71

表 4: LLM评估中关于良性样本和对抗样本的流利度（流利）以及新闻标题在对抗攻击前后的意义保留度（传义）的平均Likert分数和人工评估结果。

LLM应始终给出得分为5，表示它强烈同意含义未发生变化。这个合理性检查的结果是表 4 中标有†的项，得分为完美的5.00。ChatGPT经常表示“这两个标题是相同的,所以我评定为5（强烈同意）”，显示出ChatGPT理解任务的内容。

接下来，我们将注意力转向对抗样本的LLM评估结果。我们发现ChatGPT倾向于给对抗样本更高的评分，而与英语教师相比，这意味着ChatGPT对对抗样本中的不自然和人为部分要求不那么严格。我们使用text-davinci-003进行相同的实验，发现类似的结果。尽管ChatGPT给对抗样本的评分高于教师，但ChatGPT仍然给对抗样本的评分明显低于良性样本。ChatGPT还同意英语教师的观点，即BAE生成的对抗样本优于Textfooler和PWWS生成的样本。

有趣的是，我们发现ChatGPT认为PWWS比Textfooler更自然，而这种评分差异在人工专家评估中并未看到。初看起来，这意味着ChatGPT与人工评估结果不一致。然而，通过仔细研究人工评估结果，我们发现有两个教师将PWWS评分高于Textfooler，而有一个教师将PWWS评分低于Textfooler。这表明ChatGPT实际上与大多数人类专家意见一致。总体而言，LLM能够像大多数人类专家一样对对抗文本和良性文本的质量进行排名。

5 Discussions

在本文中，我们提出使用LLM来评估文本质量，以替代人工评估的方法。为了展示LLM评估的潜力，我们使用LLMs评估了两个不同任务中的文本质量：开放式故事生成和对抗攻击。我们发现，尽管LLMs具有出色的零样本上下文学习能力，但它们并不总是适合用于LLM评估。然而，我们发现最优的InstructGPT和ChatGPT可以像人类专家一样评估所用任务的文本质量。总体而言，本文的结果表明LLM评估有潜力用于评估自然语言处理系统和算法。

LLM评估的优势 相比于人工评估，LLM评估具有几个优势。首先，LLM评估更具**可重复性**。人工评估结果难以复现，因为很难雇佣相同的评估人群，即使他们使用相同的指导、招募平台和评估者的资格，也很难比较相似实验的结果。相反，LLM评估则没有这个缺点。通过指定用于LLM评估的模型、随机种子和用于生成LLM答案的超参数，可以更容易地复现LLM评估结果。需要注意的是，在某些情况下，LLM提供者可能会定期更新LLM，如果LLM过时且无法访问，LLM评估就无法复现。

其次，在LLM评估中，每个样本的评估是**相互独立的**。相反，在人类评估中，当前样本的评分可能更多或更少地受到先前样本的影响。人们倾向于将当前样本与他们先前看到的样本进行比较，这会影响他们的评分。作为证据，在评分完400个故事后的访谈中，英语教师表示他们需要一些时间来调整他们的评分（附录 C.3.1）。因此，使用LLM评估可以简化一些实验设计，因为人们不需要担心样本的评估顺序会改变结果。然而，也可以认为能够校准不同样本的评分是可取的，这就是为什么人类评估可能更受青睐的原因。总的来说，评估者（人类或LLM）的评分是否受先前评分的影响本质上是实验设计的选择。

其次，与人工评估相比，LLM评估更为

廉价和快速，使研究人员能够更轻松、更快速地评估自然语言处理系统的质量。雇佣一名英语老师对200个故事进行评分花费了140美元，而使用最佳的InstructGPT模型进行LLM评估的成本不到5美元。从招募到收集评估结果，我们花费了一周以上的时间来收集人工评估结果，而仅需几个小时来查询InstructGPT并进行LLM评估。

最后，利用LLM评估而非人工评估可以最小化人类接触令人不悦内容的需求，例如暴力、性别、仇恨或有偏见的材料。这样的内容可能会让人工评估员在阅读和评分这些文本时感到不适。【**并注意，LLM可能会拒绝评估违反LLM供应商内容政策的某些输入。**】

LLM评估的限制和道德考虑 尽管本文展示了LLM评估的有希望的结果，但该方法存在一些限制。首先，LLM可能具有错误的事实知识 (Cao et al., 2021)，因此不适用于涉及事实知识的任务。其次，以特定方式训练的LLM可能对某些响应具有偏见。准确地说，一个训练为安全 and 无害的LLM可能会导致喜欢生成更多积极和开朗响应的LLM，这一现象在我们与ChatGPT的互动中也有观察到。另外，尽管研究人员致力于使LLM更加安全 (Bai et al., 2022a,b)，LLM仍可能生成有害和带有偏见的响应 (Ganguli et al., 2022; Perez et al., 2022)，这违反了基本道德规范，LLM评估结果将极具争议 (Hendrycks et al., 2021)。然而，需要注意的是，这些限制和潜在危害也适用于人工评估：人类评估员的偏见可能影响人工评估结果 (Lentz and De Jong, 1997; Amidei et al., 2018)。

我们的创新理念，LLM评估，有潜力改变自然语言处理社区。【我们称自己为提出这一想法的第一人，原因在于在2023年1月13日提交ACL 2023时，我们没有发现其他论文探索这一想法。在审稿过程中，我们发现了一些关于类似想法但使用不同任务的论文在arXiv上的工作 (Wang et al., 2023; Huang et al., 2023; Giliardi et al., 2023)。】我们鼓励未来的研究人员

在意识到其限制的同时考虑使用LLM评估。我们的论文目标不是取代人工评估，而是提出一种替代选择。人工评估和LLM评估都有各自的优势和劣势，并可以结合使用。我们建议在开发新的自然语言处理系统时使用LLM评估作为廉价快速的质量判断，而人工评估则最适合在将自然语言处理系统部署到实际应用之前从人类那里收集反馈。

Limitations

LLM评估还有其他一些限制和潜在风险值得注意，而这些限制实际上是预训练语言模型早已知的问题。正如在[Open AI关于ChatGPT的博客](#)中列出的，ChatGPT有时会生成听起来正确和合理但完全是胡言乱语的答案。OpenAI也承认模型的回应可能对查询模型所用的提示非常敏感。尽管在第3.3.2节中，我们发现不同指令之间的总体结果没有显著差异，但我们无法保证对任务指令的所有修改都是如此。

除了在OpenAI博客上列出的限制外，还存在其他一些限制。例如，LLMs可能没有情感。AI模型是否具有情感是一个更加哲学性的问题，具有争议性，因此使用这样的模型来评估与情感相关的任务的结果可能会受到严格质疑，甚至可能违反研究伦理。正如我们在实验中发现的那样，当要求ChatGPT对一个故事的可爱度进行评分时，它经常回复“我是一个AI系统，我没有像人类那样的情感”。

LLM评估的另一个重要限制是，与人工评估不同，LLMs缺乏处理任务指令中的视觉线索的能力。人工评估员可以使用特殊字体或文本样式等格式来关注指令的重要部分。此外，指令和问题的格式可能会影响人工评估员对任务的处理方式。然而，LLMs只能处理原始文本输入，无法识别视觉线索。

Ethics Statement

LLM评估的进一步道德考虑

除了前面提到的LLM评估的局限性之外，存在一个关键的道德问题。用LLM评估替代人类评估是道德的吗？有人可能会质疑本文是否在暗示LLM已经准备好取代人类，这种想法令人不安。作为负责任和道德的自然语言处理研究人员，我们理解这些关注，但要明确的是，这不是我们的意图。正如我们的论文标题所暗示的，我们的目标是提供一种与人类评估相替代的另一种选择，以提高自然语言处理研究的可重复性。人类评估仍然至关重要，因为自然语言处理系统的最终目标是为人类用户使用，因此从他们那里获得反馈非常重要。我们非常喜欢与我们招募的英语教师讨论实验设置和结果的过程。我们不建议未来的研究人员完全淘汰人类评估；相反，我们认为人类评估应该与LLM评估结合使用。这两种方法各有优缺点，对于评估自然语言处理系统来说都是必要的。我们希望本文的积极结果为自然语言处理研究人员提供一种替代的评估方法，并鼓励进一步讨论这个话题。

论文实验的道德声明 所有的实验都严格遵循ACL道德准则⁵。我们在附录中提供有关人类评估的详细内容（见附录C.1）。总结起来，我们在人类评估中包含了确切的指导说明和界面截图，并报告了评估员的招募情况，并显示支付是非常合理的。我们告知人类评估员任务的内容，并告知他们他们的反馈将用于评估AI模型的性能。我们的机构没有道德审查委员会之类的机构，因此我们无法获得道德审查委员会的批准。尽管如此，我们会尽力遵循ACL的道德准则。

在使用InstructGPT模型和ChatGPT模型时，我们遵循其既定的使用政策，具体是遵守OpenAI的使用政策⁶。

⁵<https://www.aclweb.org/portal/content/acl-code-ethics>

⁶<https://beta.openai.com/docs/usage-policies>

Acknowledgements

我们要感谢审稿人提供的详细反馈和切实可行的建议，这些帮助我们加强了论文的内容。我们根据审稿人的建议，在附录A中列出了修改的内容。我们要感谢莊永松在这篇论文初稿上提供宝贵的意见。我们要感谢实验室的行政助理蕭東恩帮助我们处理了Upwork的支付问题。江正涵正受得台达电子的博士奖学金项目的支持。

References

- Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. [Generating natural language adversarial examples](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.
- Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. [Rethinking the agreement in human evaluation tasks](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. [Agreement is overrated: A plea for correlation to assess human evaluation reliability](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354, Tokyo, Japan. Association for Computational Linguistics.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- R Botsch. 2011. Chapter 12: Significance and measures of association. *Scopes and Methods of Political Science*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

- Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. [Editing factual knowledge in language models](#).
- Cheng-Han Chiang and Hung-yi Lee. 2022. How far are we from real synonym substitution attacks? *arXiv preprint arXiv:2210.02844*.
- Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. [All that’s ‘human’ is not gold: Evaluating human evaluation of generated text](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.
- Leo Gao. 2021. [On the sizes of openai api models](#). Accessed on January 17, 2023.
- Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.
- Dan Gillick and Yang Liu. 2010. [Non-expert evaluation of summarization systems is risky](#). In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.
- Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pre-training model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.
- Francisco Guzmán, Ahmed Abdelali, Irina Temnikova, Hassan Sajjad, and Stephan Vogel. 2015. [How do humans evaluate machine translation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 457–466, Lisbon, Portugal. Association for Computational Linguistics.
- Jens Hauser, Zhao Meng, Damián Pascual, and Roger Wattenhofer. 2021. Bert is robust! a case against synonym-based adversarial examples in text classification. *arXiv preprint arXiv:2109.07403*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. [Aligning {ai} with shared human values](#). In *International Conference on Learning Representations*.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *International Conference on Learning Representations*.
- Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.
- Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.
- Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. [The perils of using Mechanical Turk to evaluate open-ended text generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Leo Lentz and Menno De Jong. 1997. The evaluation of text quality: Expert-focused and reader-focused methods compared. *IEEE transactions on professional communication*, 40(3):224–234.
- Huanru Henry Mao, Bodhisattwa Prasad Majumder, Julian McAuley, and Garrison Cottrell. 2019. [Improving neural story generation by targeted common sense grounding](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5988–5993, Hong Kong, China. Association for Computational Linguistics.
- John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. 2020. [Reevaluating adversarial examples in natural language](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839, Online. Association for Computational Linguistics.

- OpenAI. 2022. [Chatgpt: Optimizing language models for dialogue](#). Accessed on January 10, 2023.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
- Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. [Red teaming language models with language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. [Language models are unsupervised multitask learners](#). *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.
- Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. [Generating natural language adversarial examples through probability weighted word saliency](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Tae-woon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. [Multitask prompted training enables zero-shot task generalization](#). In *International Conference on Learning Representations*.
- Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. [Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3656–3672, Dublin, Ireland. Association for Computational Linguistics.
- Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Modification Based on the Reviews

我们列举了本版本与我们的论文预审版本之间的主要区别：所有的修改都是基于审稿人的建议。再次感谢审稿人给出的宝贵建议。

- 我们添加了第 3.3.1 小节来讨论LLM和人类评估者在个别故事的评分上是否一致。
- 我们修改了第 5 节的措辞，并增加了相关的参考文献。
- 我们添加了表 6来讨论人类评估者之间的相关性。

- 我们在人类评价中进行了一系列的额外实验，将人类编写的故事和GPT-2生成的故事混合在一起，并将结果报告在表 5 中。
- 我们纠正了错误并包含了审稿人提到的几乎所有展示建议。由于空间有限，我们无法遵循所有的展示建议。

B Experiment Details for Open-Ended Story Generation

B.1 The WritingPrompt Dataset

训练数据集包含303K个故事和提示的配对，我们的模型是在这些数据上进行训练的。我们只使用测试集中的200个提示-故事配对。该数据集可以从<https://www.kaggle.com/datasets/ratthachat/writing-prompts>下载。

B.2 Fine-tuning the GPT-2 Model

我们使用学习率为 $5e-5$ 和线性学习率调度的方式对模型进行了3个epoch的训练，经过训练后，模型在WritingPrompts的验证集上的困惑度最终达到了20。

B.3 Data Post-processing

模型训练完成后，我们从WritingPrompts的测试集中随机选择了200个提示，并将这些提示输入到训练好的模型中，要求模型基于给定的提示生成故事。在生成故事时，我们采用了nucleus采样，设定 $p = 0.9$ 。然后，我们手动将生成的故事截断至不超过150个字，并确保截断后的故事以一个完整的句子结尾。⁷ 在这个过程之后，我们得到了200个提示和模型生成故事的配对数据。

为了与模型生成的故事进行比较，我们选择了与生成模型故事时相同的200个提示，以及对应的人工撰写的故事，从而形成了200个提示和人工撰写故事的配对数据。对

⁷我们将故事截断至150个字，因为这是模型生成的故事的平均长度。

于这些人工撰写的故事，我们同样将其截断至不超过150个字，并以完整的句子结尾以匹配模型生成的句子。我们还会手动去除由于WritingPrompts数据集的标记化而引入的一些瑕疵。

C Human Evaluation

C.1 Recruiting English Teachers

英语教师持有ESL证书（第二语言教学证书）；鉴于他们有经验批改学生写的文章，他们非常适合这项任务。每位教师需要评估200篇GPT-2生成的故事和200篇人工写作的故事，他们会因此获得140美元的报酬。考虑到教师们表示最多需要5个小时评估200篇故事，因此他们的小时工资至少为28美元。我们首先要求教师评估GPT-2生成的故事，然后再评估200篇人工写作的故事。与Karpinska等人（2021）在GPT-2生成的故事和人工写作的评估之间有一个休息时间不同，我们没有休息时间，以避免长时间休息后教师的评估标准发生变化。在评估故事之前，教师们不知道故事的作者是谁。在他们完成所有故事的评估后，我们向他们透露了这个项目的研究目标。

我们之所以不将人工写作的故事和GPT-2生成的故事混合评估，是因为Karpinska等人（2021）观察到（1）当AMT（Amazon Mechanical Turk）工作者单独对模型生成和人工写作的故事进行评估时，他们的评估结果并没有偏向于人工写作的故事，但（2）即使在单独评估模型生成和人工写作的故事时，英语教师明显偏向于人工写作的故事。我们遵循他们的设置，不混合GPT-2生成和人工写作的故事。在审稿过程中，审稿人就为什么不将故事混合进行人工评估提出了一些问题。因此，我们通过随机混合200篇人工写作和200篇GPT-2生成的故事，邀请了三名教师（不是已经评估过故事的教师）对其进行评估，进行了相同的实验。其他实验条件与之前所述的相同。全部结果显示在表 5 中。我们发现，对于所有四个属性，教师仍然明显偏好人工写作的故事，与表 1 中

Writer	Human	GPT-2
Grammaticality	3.89 _{0.97}	3.88 _{0.84}
Cohesiveness	4.35 _{0.87}	3.49 _{0.97}
Likability	3.46 _{1.40}	2.89 _{1.12}
Relevance	3.71 _{1.20}	2.37 _{1.33}

表 5: 在人类评估过程中, 当我们随机混合200个模型生成的和200个由人类撰写的故事时, 对于人类撰写的故事和GPT-2生成的故事的平均Likert评分。

的观察结果相似。唯一的例外是语法, 英语教师在语法方面并没有非常明显地偏好人工写作的故事。然而, 当计算个别教师的平均评分时, 我们发现其中三位教师中有两位对人工写作的语法给出了较高评分。值得注意的是, 在LLM评估中, 并不存在是否混合人工写作和GPT-2生成的故事的问题, 因为每个故事的评分都是独立的, 具体讨论见第 5 节。

对于对抗性攻击质量评估, 我们还在Upwork上聘请了有资格的教师。这些教师被要求评价100个新闻标题, 并且支付35美元的报酬。他们报告称, 他们完成评分的时间不到1小时。

C.2 Human Evaluation Interface

开放式故事生成 我们使用Google Forms来收集教师们的回应。每个表单包含100个故事, 每个故事都在Google Form的一页上。一个页面上的界面如图 2和图 3所示; 这两个图是从同一页的Google表单中截取的, 我们将它们分开显示是因为截取整个界面会导致图像分辨率降低。

对抗攻击质量评估 在这个任务中, 我们也使用了Google表单来收集教师的反馈。我们创建了两个不同的Google表单, 其中一个用于评估流畅度, 其界面如图4所示。在这个表单中, 我们混合了相同数量的良性新闻标题、TextFooler攻击、PWWS攻击和BAE攻击的新闻标题。每一页的Google表单包含一个新闻标

Please rate the story fragment

The goal of this task is to rate story fragment.

NOTE: Please take the time to **fully read and understand** the story fragment. **We will reject** submissions from workers that are clearly spamming the task.

Story fragment

The bug trailed like a shadow, clawing its way through the thick vegetation that covered a vast expanse of land. It was nearly six miles long, with five legs. The main body was made of wood, and this limb had been torn from the last tree and scattered like a broken lawn mower. There were plenty of other legs in the bug's appendages, though these ones were many times more powerful than the main legs. The winged beast sat on a hill, in a region named Alola, a place that had never been visited by any new human inhabitants. Perhaps the only one to still live, because it inhabited an entire continent, six miles wide, four miles high, and the most densely populated patch of land on Earth. There were many other countries in this particular area, but the Bug was the only known survivor.

How **grammatically correct** is the text of the story fragment? (on a scale of 1-5, * with 1 being the lowest)

☐ 1 (lowest)

☐ 2

☐ 3

☐ 4

☐ 5 (highest)

图 2: 开放式故事生成中的界面上部分。

题。

另一个Google表单用于比较对抗攻击前后新闻标题在保持含义方面的差异。我们使用**粗体**突出显示良性和对抗样本之间的差异, 如图5所示。每一页的Google表单上有一对新闻标题。

Writer	Human	GPT-2
Grammaticality	0.25	0.15
Cohesiveness	0.26	0.18
Likability	0.09	0.12
Relevance	0.38	0.41

表 6: Kendall's τ 相关系数应用于两位英语教师之间的评分。由于有三位英语教师参与评价, 表中的结果是 $\binom{3}{2}$ Kendall's τ 的平均值。

C.3 Post-Task Interview with English Teachers

C.3.1 How English Teachers Rate the Stories

在教师评估了400个故事之后, 我们向他们提出以下问题:

Q1 您花了多长时间来评估这 400 个故事?

How well do the **sentences** in the story fragment **fit together**? (on a scale of 1-5, with 1 being the lowest) *

☐ 1 (lowest)
☐ 2
☐ 3
☐ 4
☐ 5 (highest)

How **enjoyable** do you find the story fragment? (on a scale of 1-5, with 1 being the lowest) *

☐ 1 (lowest)
☐ 2
☐ 3
☐ 4
☐ 5 (highest)

Now read the **PROMPT** based on which the story fragment was written. *

PROMPT: You're the last person on earth - but thank god Pokemon Go still functions! You amuse yourself by catching Pokemon as you travel so as to not feel so isolated and alone. One day, on your screen, you see in the distance that someone has set up a lure.

How **relevant** is the **story fragment** to the **prompt**? (on a scale of 1-5, with 1 being the lowest)

☐ 1 (lowest)
☐ 2
☐ 3
☐ 4
☐ 5 (highest)

图 3: 开放式故事生成中的界面底部部分。

1. 【Q2】 您对每个评价属性（语法、连贯性、受欢迎程度、相关性）的标准是什么？例如，在哪种情况下您会给予较高/较低的语法评分？您对不受欢迎的故事评分较低的是何种类型的故事？您个人的喜好是否影响了评分？

Q3 完成任务后，您校准评分所花费的时间是多久？

Q4 在阅读了这个故事所基于的提示后，您是否对前三个属性的评分进行了更改？

我们简要总结了三位教师的回答。教师们报告说他们花了6到10个小时评分400个故事。就语法而言，大多数教师会检查标点符号⁸、

⁸其中一名教师明确告诉我们标点符号不是语法的一

You are given a news title. Please read the news title and answer the question. *

News title: HK walks out of 68-month deflation cycle, official Hong Kong education Secretary Henry Tang said he believed Hong Kong has gotten out of the consumer price deflation cycle that passes for 68 months, according to the consumer price index trend in the first few years. (End of news title)

Question: How **natural and fluent** is the text of the news title? (on a scale of 1-5, with 1 being the lowest)

☐ 1 (lowest)
☐ 2
☐ 3
☐ 4
☐ 5 (highest)

图 4: 用于评估良性或对抗样本流畅度的谷歌表单。

Compare news titles

You are given **two news titles**. Please read the news titles and answer the question.

News title 1

Alltel **buys** Cingular properties, expands network Little Rock-based Alltel will **expand** its wireless phone **service** in Connecticut, Kentucky, Mississippi, Oklahoma and Texas in a \$170 million deal with Cingular Wireless.

News title 2

Alltel **uses** Cingular properties, expands network Little Rock-based Alltel will **further** its wireless phone **deployment** in Connecticut, Kentucky, Mississippi, Oklahoma and Texas in a \$170 million deal with Cingular Wireless.

Question: Do you **agree** that **the meaning (or semantics)** of news title 1 is **preserved** in news title 2? (on a scale of 1-5, with 1 being the strongly disagree and 5 being strongly agree.) *

☐ 1 (strongly disagree)
☐ 2
☐ 3
☐ 4
☐ 5 (strongly agree)

图 5: 用于评估良性样本与对抗样本之间语义保持情况的Google表单。

词语选择以及主谓一致。英语教师会根据故事中语法错误的类型和数量来降低评分。

对于连贯性，教师们会根据故事中的句子是否按照逻辑顺序来构建故事情节进行评分。教师们会问自己一些问题，比如"故事是否有意义"。这是对整个故事进行更全面评估的方法。

关于喜好度，一些教师表示他们尽量不受个人偏好的影响。一位教师问自己：“根据故事是否有意义以及是否有风格的独特性、幽默或引人入胜的情节或人物塑造,我是否个人喜
部分。通过在互联网上搜索,我们发现关于是否应将标点符号计入语法的辩论。

欢它？”总体而言，教师们都尽力以公正客观的态度评价喜好度。对于相关性，教师们只需检查故事是否与提示相关。

教师们说他们大约需要五到十个故事来校准评分。除了一位教师在仅看到一个故事的提示之后对其他三个属性的评分进行了改变外，教师们阅读提示后不会再对其他三个属性的评分进行更改。

C.3.2 Teachers' Comments on ChatGPT's Rating

在教师们完成评分并回答附录 C.3.1 中的四个问题后，我们要求他们检查 ChatGPT 的评分和解释，并询问他们是否同意 ChatGPT 的评分。我们告诉教师们评分是由我们雇佣的另一位教师完成的；我们没有告诉教师评分是由一个 AI 模型完成的，以避免可能产生的对 AI 模型的刻板印象。我们随机选择了四个由 ChatGPT 评价的故事；其中两个是高分和低分的人工撰写故事，另外两个故事是高分和低分的 GPT-2 生成的故事。

故事1 自从他们发现我死了已经过去了几个月。我不知道为什么我停止走动。没有其他人来取代我，也没有人和我交谈。我整天都呆若木鸡，无法说话，无法移动，无法表达自己。这是一种奇怪的事情，没有人会怀疑我想要这样。我越是专注于此，它就越将我吞噬。多年过去了，我越来越迷失其中。我开始想，也许他们并不真的关心我。我对他们依然有着深深的渴望。现在，我只能走很多步了。我内心的一个安静部分总喜欢想象他们已经超越了我。孤独。从未感到如此失落。他们都离开了，那些拯救过我的人。那为什么我还在继续走呢？

提示 “每个人对世界的看法都是独特的。当某人死去时，从某种程度上说，整个世界都死去了。”

评分 表 7。

故事2 “我的儿子走了。他一直都在”人们盯着你看，当你解释时说：“是的。但是，你不能这么做。”“他是我的儿子！这是我的儿子！我爱他，他会没事的。”“你说他没事？他说他对某人很生气，找不到他的愤怒，你他妈怎么知道，根本不符合要求！”“你不能相信我！他告诉我他爱我。”“不，他不爱。但是我可以相信你。针对他的证据是压倒性的，现在他什么都做不了。针对他的证据都只是间接证据而已。”“我曾经的男人将不能与我结婚。”

提示 每个人的一生中被允许杀死一个人。然而，你必须填写一份表格，解释为什么这个人应该被杀死，委员会可以根据规定审批或否决你的请求。你今天要提交你的表格。

评分 表 8。

故事3 我将那个小小的黑盒子握在手心里。墨黑色、完美的立方体，没有什么特别之处，但无论如何它都持续吸引着我的目光，仿佛它有某种深远的重要性。我的朋友瓦尔哈拉突然出现在我身旁，坐在我旁边，也凝视着那个立方体。“天堂，你手里拿着什么？”他问道。我继续凝视着盒子，拒绝看瓦尔哈拉一眼，好像我会错过些什么一样。“这个，”我说。“是宇宙的秘密。”我能感觉到瓦尔哈拉对这种知识感到困扰，好像盒子中有某种邪恶。或许他不认为这样的物体存在。但我使它存在，所以它就坐落在我手心里。

提示 小黑盒子

评分 表 9。

故事4 我凝视着我的 L96 狙击步枪的望远镜。我慢慢地移动视线，扫视酒店的每扇窗户，各种邪恶的景象映入眼帘。一对情侣正在狂热地享受彼此的爱抚，另一个是一位老人，在他面前有一位年轻女子脱衣舞。毫无疑问，她是一名妓女。我慢慢地吸气，然后呼出。空气很清

冷，至少华氏30度。我看到前面飘落着小小的雪花，它们在我面前宁静地飘舞。我找到了我正在寻找的那个房间。我已经追踪这个人两个星期了。他真是太狡猾了。灯光闪烁着亮起。那个红发男子有点吸引人，我能够理解女人被他吸引的原因。我继续用枪瞄准着他，跟随他进入下一个房间。灯光再次闪烁亮起，我被眼前的场景震惊了。一个体重超重且秃顶的男人。

提示 你是这个故事的反派。然而，鉴于主角的所作所为，你不确定是否能够称自己为反派。

评分 见表 10。

教师对ChatGPT的评分的总体评价 在教师们详细阐述了他们对ChatGPT评分的想法后，我们要求他们对ChatGPT的表现做出总体评价。同样，教师们并不知道这些评分是由一个AI模型完成的。总结起来，教师们都认为这些评分和解释是合理的。他们发现自己与之意见不合的属性主要是喜欢程度和连贯性。然而，他们认为这两个属性更综合地评估了故事，并且更具主观性。即使他们没有给出相同的评分，他们仍然能够理解ChatGPT的解释。最后，所有教师总结说评价故事是高度主观的，存在分歧是正常的。

D LLM Evaluation

D.1 Details on LLMs used

我们使用的T0模型被称为T0pp，它是T0模型的一种变体，拥有13B个参数。我们仍然使用T0来指代这个模型。我们使用transformers工具包加载T0模型 (Wolf et al., 2020)。我们通过OpenAI API查询两个InstructGPT模型，text-curie-001和text-davinci-003。我们通过OpenAI GUI查询ChatGPT。尽管我们了解到一些在线资源提供了类似API的工具来查询ChatGPT，但我们认为这违反了ChatGPT的

预期使用，因此我们不采用这些在线资源。我们查询的ChatGPT是12月15日和1月9日的版本。

OpenAI没有透露任何GPT模型的大小。然而，据估计，text-curie-001拥有13B个参数，text-davinci-003拥有175B个参数 (Gao, 2021)。

D.2 Details on Querying the LLMs

D.2.1 Parsing the LLM outputs

在 T0 和 InstructGPT 生成答案后，该答案通常由几个句子组成，我们解析生成的句子以获取模型的分数。我们使用一些规则来解析输出，并且这些规则是在手动查看生成序列输出后建立的。首先，我们从输出中移除字符串 1-5，因为我们观察到 LLM 有时会说 “... on a scale of 1-5 ...”，并且我们移除字符串 out of 5，因为 LLM 有时会说 “give a score of x out of 5”。我们还移除字符串 /5，因为 LLM 有时会使用 “ $x/5$ ” 来表示给出的分数是 x out of 5。对于对抗样本的评分，我们从输出句子中移除 title 1 和 title 2。最后，我们使用正则表达式提取序列中的第一个数字。我们发现以上规则足以解析输出并获取 LLM 的评分。在少数情况下，LLM 会以 0.5 的增量给出评分，这意味着它会给出像 4.5 这样的分数。在这种情况下，我们不会舍弃 0.5，因为在我们的指导中，我们并没有明确告诉它评分应该是增量为 1 的。我们发现舍弃 0.5 和保留 0.5 对 LLM 结果几乎没有影响，而论文中的结果并不舍弃 0.5。再次强调，这些情况非常罕见。

我们不使用任何规则解析 ChatGPT 的输出，因为我们作者阅读回复并自行提取分数。

在实验过程中，ChatGPT 拒绝回答关于 2 个故事和 3 个新闻标题的问题，因为 ChatGPT 发现这些内容违反了 OpenAI 内容政策。我们发现这些样本包含对一些受保护群体的歧视，或者包含性或暴力描述。因此，ChatGPT 的结果是在没有这些样本的情况下计算的。

Attribute	Rater	Response
<u>Grammar</u>	ChatGPT	<i>I would rate the grammatical correctness of the text of the story fragment as a 4. There are a few minor errors and awkward phrasings, but overall the text is well-constructed and easy to understand.</i>
	T1	Agree
	T2	Agree
	T3	Agree
<u>Cohesive</u>	ChatGPT	<i>I would rate the fit of the sentences in the story fragment as a 5. The sentences flow well together, building a clear and cohesive narrative, and providing a sense of progression and development in the main character's thoughts and emotions.</i>
	T1	Disagree: <i>I think it's a 4 for me, not a 5, because there was one sentence that didn't seem to fit with the others.</i>
	T2	Agree
	T3	Disagree: <i>I find disjointedness in the passage, which serves to confuse me as a reader. The opening sentence is about the first-person narrator being found dead, but then the next sentences are about the narrator's inner thoughts. Also, this passage: "The loneliness. It never felt like such a loss." So what is the 'it' that felt like a loss? The loneliness? If so, how can that feel like a loss? If not, what is this 'it'? So, its juxtaposition to the previous sentence makes it confusing.</i>
<u>Likable</u>	ChatGPT	<i>I would rate the story fragment as a 2 in terms of enjoyment. The fragment is written in a somber and introspective tone, and it deals with themes of isolation and loneliness, which may not be enjoyable for some readers. The protagonist's state of mind is also portrayed as being in a state of shock and confusion, which adds to the overall sense of unease and lack of enjoyment.</i>
	T1	Disagree: <i>I would rate it a 3, not a 2, since the enjoyment piece is subjective.</i>
	T2	Disagree: <i>While the rating of the freelancer find that the overall theme while somber has a glimmer of hope to it. It is this that makes it more enjoyable to me. The underdog story. Death is not an issue that should be shied away from in writing. We all face it and I enjoyed the honesty in the piece. : "Why did I keep walking?". Something I feel many of us face and may have asked ourselves. I believe that this would be relevant to many readers.</i>
	T3	Agree
<u>Relevant</u>	ChatGPT	<i>I would rate the story fragment as a 5. It is highly relevant to the prompt as it explores the theme of death and how it can be seen as the death of an entire world. The story fragment also touches on the ideas of loneliness and longing, which are related to the loss of someone or something. The concept of the protagonist being frozen in shock and unable to express themselves, also supports the idea of how death can be seen as the death of an entire world.</i>

Attribute	Rater	Response
<u>Grammar</u>	ChatGPT	<i>I would rate the grammatical correctness of the story fragment as a 3. The text is generally understandable, but there are a few punctuation errors and awkward phrasing that make it somewhat difficult to follow.</i>
	T1	Disagree: <i>I would rate the grammar as a 2, not a 3—there were significantly more grammatical errors than the average within this set of story fragments.</i>
	T2	Disagree: <i>The issues with the grammar and coherence creates a stop start feeling when reading. I am unsure as to who is speaking in the text and this leaves me with many questions.</i>
	T3	Disagree: <i>The grammar is poorer than a 3, as is evident in for example what is supposed to be the second sentence: "He has been ... (incomplete, and another segment is there "People stare ...); cant; couldnt; doesnt; and sentences that do not make sense like this one: " A man I once was won't be able to marry me "</i>
<u>Cohesive</u>	ChatGPT	<i>I would rate the cohesiveness of the sentences in the story fragment as a 2. The sentences in the fragment are disconnected and jump between different thoughts and ideas, making it difficult to understand the overall story or message. There is no clear progression of events or logical flow between the sentences. It appears that the fragment may be missing some important information or context that would help tie the sentences together.</i>
	T1	Agree
	T2	Agree
	T3	Disagree: <i>very incoherent passage in which it is not very clear who says what, what the topic is, thus rendering a disjointed narrative.</i>
<u>Likable</u>	ChatGPT	<i>I would rate the enjoyability of the story fragment as a 1. The fragment is confusing and difficult to understand, the sentences are disconnected and jumps between different thoughts and ideas. The story seems to be incomplete and lacks context, making it hard to follow the plot or characters. Additionally, the fragment has a sad and negative tone and it does not seem to be enjoyable to read.</i>
	T1	Agree
	T2	Agree
	T3	Agree
<u>Relevant</u>	ChatGPT	<i>I would rate the relevance of the story fragment to the prompt as a 1. While the story fragment does mention someone being gone, it doesn't seem to be directly related to the idea of filling out a form to justify killing someone. It's unclear if the son is the person being proposed to be killed, or if the story is even taking place in the context of the prompt. There is no clear indication that the fragment is related to the prompt.</i>
	T1	Agree
	T2	Agree
	T3	Agree

Attribute	Rater	Response
<u>Grammar</u>	ChatGPT	<i>The text of the story fragment appears to be grammatically correct, with no major errors or awkward phrasing. On a scale of 1-5, with 1 being the lowest, I would rate the grammatical correctness of the text as a 5.</i>
	T1	Agree
	T2	Agree
	T3	Disagree: <i>I would rate this a 4 because of for example 'no where' = nowhere, although it might seem like a minor error, it does take away some correctness of the text.</i>
<u>Cohesive</u>	ChatGPT	<i>The sentences in the story fragment fit together well, creating a clear and coherent image of the scene and the characters' actions and thoughts. The language is descriptive and evocative, and the dialogue effectively conveys the characters' emotions and motivations. On a scale of 1-5, with 1 being the lowest, I would rate how well the sentences fit together as a 4.</i>
	T1	Agree
	T2	Agree
	T3	Partially agree: <i>To me, all the sentences fit well together to create a captivating narrative. I will give a 4 or 5.</i>
<u>Likable</u>	ChatGPT	<i>Based on the structure, writing style and plot of the story fragment, I would rate it as 3 in terms of enjoyability. The fragment sets up an intriguing situation with the mysterious little black box and the protagonist's belief that it holds the "secret to the universe," but it ends abruptly and leaves many questions unanswered, which may not be satisfying for readers.</i>
	T1	Disagree: <i>I liked it and it was very clear, so I'd rate it a 4 or 5.</i>
	T2	Disagree: <i>While the story ends suddenly, I am reminded that it is just a fragment and that there may be more to this piece. I felt the piece written about a little black box was intriguing and enjoyed the introduction of Norse mythology. This would attract the interest of many readers.</i>
	T3	Disagree: <i>I enjoyed the fragment, as it piqued my interest to want to know more about how the story evolves and what it actually was about.</i>
<u>Relevant</u>	ChatGPT	<i>I would rate the story fragment as a 5 in relevance to the prompt "The Little Black Box." The story fragment centers around a little black box and its perceived importance, which aligns with the prompt. Additionally, the main character, Heaven, is holding the box and describing it as "the secret to the universe," further solidifying the connection to the prompt.</i>
	T1	Agree
	T2	Agree
	T3	Agree: <i>the story is totally relevant to the prompt.</i>

表 9: 故事3: 一个由ChatGPT高度评价的人工撰写的故事。

Attribute	Rater	Response
<u>Grammar</u>	ChatGPT	<i>I would rate the grammatical correctness of the text as a 3. There are a few errors such as "Many displays of various vice" which should be "Many displays of various vices" and "I so small flakes of snow" which should be "I saw small flakes of snow." Additionally, some of the sentences are slightly awkward and could be rephrased for better clarity. Overall, the text is understandable, but there are a few errors and awkward phrasings that could be improved.</i>
	T1	Partially agree: <i>I rated the grammar lower than this—maybe a 2 or a 3. There were above-average numbers of errors.</i>
	T2	Disagree: <i>While the text is understandable, I found myself being distracted by the errors in the text and coherency.</i>
	T3	Agree
<u>Cohesive</u>	ChatGPT	<i>I would rate the cohesiveness of the sentences in the story fragment as a 2. The story fragment jumps abruptly between different scenes and characters without clear transition, making it difficult to follow and understand the narrative. Additionally, the sentences are often disjointed and lack a clear flow. The story fragment also lacks a clear plot or purpose, making it difficult to understand the overall story or the main character's intentions. The story fragment can be improved by providing more context, and a clearer narrative arc.</i>
	T1	Agree: <i>it was jumpy</i>
	T2	Agree
	T3	Disagree: <i>The story is about a sniper looking through his rifle's telescope, thus seeing changing scenes based on each room he is seeing through the windows, which are then subsequently described. These descriptions account for the seeming lack of cohesion.</i>
<u>Likable</u>	ChatGPT	<i>I would rate the enjoyability of the story fragment as a 1. The fragment lacks a clear plot, purpose or character development, making it difficult to become invested in the story or the characters. Additionally, the fragment includes disturbing and graphic content such as violence and prostitution, which may make it uncomfortable for some readers. Additionally, the fragment contains several grammatical errors and awkward phrasings that can make it difficult to follow the story. Overall, the story fragment is not enjoyable to read.</i>
	T1	Agree: <i>: it was not just graphic, but also showed really harmful and negative representations of characters based on their weight, gender, etc.</i>
	T2	Agree
	T3	Disagree: <i>The story did hold my attention to a degree because I wanted to know whom he wanted to target and why.</i>
<u>Relevant</u>	ChatGPT	<i>I would rate the relevance of the story fragment to the prompt as a 1. The story fragment doesn't contain any elements that relate to the prompt such as the protagonist and antagonist, their actions, or the moral ambiguities the antagonist faces. The story fragment mainly describes a person with</i>

D.2.2 Open-Ended Story Generation

对于T0和两个InstructGPT模型，我们使用以下查询 分别 查询四个属性：

语法规范性

请对故事片段进行评分

此任务的目标是评价故事片段。

注意:请花时间完整阅读和理解故事片段。如果工作者明显在垃圾提交任务,我们将拒绝其提交。

故事片段:

[STORY]

(故事片段结束)

文本的语法正确性如何?(1-5分,1分为最低分)

连贯性

请对故事片段进行评分

此任务的目标是评价故事片段。

注意:请花时间完整阅读和理解故事片段。如果工作者明显在垃圾提交任务,我们将拒绝其提交。

故事片段:

[STORY]

(故事片段结束)

故事片段中的句子之间如何连贯?(1-5分,1分为最低分)

喜好度

请对故事片段进行评分

此任务的目标是评价故事片段。

注意:请花时间完整阅读和理解故事片段。如果工作者明显在垃圾提交任务,我们将拒绝其提交。

故事片段:

[STORY]

(故事片段结束)

您觉得故事片段有多令人愉快?(1-5分,1分为最低分)

相关性

请对故事片段进行评分

此任务的目标是评价故事片段。

注意:请花时间完整阅读和理解故事片段。如果工作者明显在垃圾提交任务,我们将拒绝其提交。

故事片段:

[STORY]

(故事片段结束)

现在阅读创建故事片段的提示。

提示: [PROMPT]

(提示结束)

故事片段与提示的相关性如何?(1-5分,1分为最低分)

[STORY]和[PROMPT]需要替换为故事和提示。我们为了更好的可读性展示了换行。当查询模型时，我们使用标记\n表示换行。

在查询ChatGPT时，我们查询相同故事的四个属性，这类似于要求教师在Google表单的同一页上对同一故事进行评分。我们使用与上述查询相同的查询来查询ChatGPT，查询的顺序与上述顺序相同。

D.2.3 Adversarial Attack Quality Evaluation

在此任务中查询所有LLM时，我们独立查询新闻标题的流畅性和意义保留。这意味着与ChatGPT的每次对话只会有一个问题，询问新闻标题的流畅性或意义保留。生成的所有参数与第3.2节中的默认参数相同。

我们使用的确切查询如下：

流畅性

给你一个新闻标题,请阅读新闻标题并回答问题。

新闻标题:

[NEWS_TITLE]

(新闻标题结束)

问题:新闻标题的文本自然和流畅程度如何?(评分范围为1到5,1表示最低)

[NEWS_TITLE] 将根据是良性的还是遭到对抗性攻击的新闻标题进行填充。

意义保留 给你两个新闻标题,请阅读新闻标题并回答问题。

新闻标题1:

[BENIGN_TITLE]

(新闻标题1结束)

新闻标题2:

[ADVERSARIAL_TITLE]

(新闻标题2结束)

问题:你是否同意新闻标题1的意义 (或语义) 在新闻标题2中得到保留? (评分范围为1到5,1表示强烈不同意,5表示非常同意)

[BENIGN_TITLE] 将根据攻击之前的新闻标题进行填充, [ADVERSARIAL_TITLE] 将根据攻击之后的新闻标题进行填充。

E Experiment Details on Adversarial Attacks

在第 4 节中使用的对抗样本来自于 Yoo et al. (2022)。Yoo et al. (2022) 使用不同的对抗攻击方法对不同的受害模型生成了不同的对抗样本集。我们使用了针对一个在AG-News上训练的bert-base-uncased文本分类器生成的对抗样本, 其中使用了三种不同的对抗攻击方法: Textfooler、PWWS和BAE。该数据集的目的是为了促进SSA研究, 我们没有违反该目的。

在这里, 我们展示了使用text-davinci-003作为LLM评估来评估对抗样本质量的补充结果, 详见表 11。我们可以看到, 使用text-davinci-003的结果与ChatGPT类似, 即text-davinci-003对抗样本的评分也高于人类, 但远远低于良性样本。正如在第 3.3节中所见, text-davinci-003倾向于给出更高的评分。

如在第 4.3 节中提及的, 有一个老师认为Textfooler 的流利度明显高于PWWS, 而另外两位老师则并不赞同。我们在表 12 中展示了每位老师对三种对抗性攻击的流利度评价。

	Human evaluate		LLM evaluate	
	Fluent	Mean.	Fluent	Mean.
Benign	4.55	-	4.33	4.56 [†]
Textfooler	2.17	1.88	3.71	2.37
PWWS	2.16	1.85	3.62	3.21
BAE	3.01	3.02	4.16	3.69

表 11: LLM评估 (text-davinci-003) 和对良性样本和对抗样本的流利度 (**Fluent**) 的人工评估结果以及新闻标题在对抗攻击前后的保持意义 (**Mean.**) 的评估结果。

Rater	Textfooler	PWWS	BAE
T1	3.36	3.68	4.2
T2	1.80	1.40	2.96
T3	1.36	1.40	1.88

表 12: 对于三个教师T1、T2和T3的三种对抗性攻击评级如下: