大语言模型能否成为人工评估的替代方案?

Cheng-Han Chiang台湾大学，台湾
dcml0714@gmail.com

Hung-yi Lee
台湾大学，台湾
hungyilee@ntu.edu.tw

arXiv:2305.01937v1 [cs.CL] 3 May 2023

摘要

人工评估对于评估机器学习模型生成或人类撰写文本的质量而言不可或缺，也无法避免。然而，人工评估极难复现，其质量也以不稳定著称，阻碍了不同自然语言处理（NLP）模型和算法之间的公平比较。近年来，大语言模型（LLM）展现出卓越的能力：仅提供任务指令，便能在未见过的任务上表现出色。本文探讨大语言模型的这种能力能否用于替代人工评估。我们向大语言模型提供与进行人工评估时完全相同的指令、待评样本和问题，然后要求大语言模型回答这些问题；我们将此称为大语言模型评估。我们使用人工评估和大语言模型评估，对两项自然语言处理任务中的文本进行评估：开放式故事生成和对抗攻击。我们发现，大语言模型评估的结果与专家人工评估的结果一致：人类专家评分较高的文本，大语言模型给出的评分也较高。我们还发现，采用不同的任务指令格式以及不同的答案生成采样算法时，大语言模型评估的结果依然稳定。我们首次展示了使用大语言模型评估文本质量的潜力，并探讨了大语言模型评估的局限性和伦理考量。

1 引言

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 等人，2015年；Gillick 和 Liu，2010年）。我们依赖人工评估，是因为文本的某些方面很难用自动评估指标来评估；因此，研究人员会请人类对自然语言处理模型的输出质量进行评分。虽然人工评估在自然语言处理中十分普遍且不可或缺，但它也以极不稳定而著称（Gillick 和 Liu，2010年；Clark 等人，2021）。Karpinska 等人（2021）表明，人工评估中素质不高的工作者会对评估结果产生不利影响，导致无法比较不同系统的性能。可复现性是人工评估的另一项难题，因为很难招募到相同的人工评估者并重复进行同一项评估。即使招募到相同的工作者，他们也可能在下次评估时给出不同的结果，因为他们之前已经做过这项任务。尽管人工评估用于更好地评估自然语言处理系统，并且相较于自动评估指标具有一些优势，但其缺点在一定程度上使得可靠地评估自然语言处理系统变得困难。

为了解决一些缺点，我们利用大语言模型（大语言模型）。大语言模型是通过自监督学习来建模人类语言的大型模型（Brown 等人，2020年），并进一步采用特殊的训练流程来提升其在未见任务上的表现，并使其更好地遵循自然语言指令（Sanh 等人，2022；Wei 等人，2022）。仅凭任务指令就能执行任务的能力促使我们思考，这些大语言模型能否像人类一样进行人工评估。为了回答这个问题，我们向大语言模型输入人工评估中使用的相同指令、样本和问题，并将大语言模型生成的序列视为其对问题的回答。该过程如图 1所示，我们将这一过程称为**大语言模型评估**。

为了检验大语言模型评估能否得出有意义的结果，我们针对两项不同的自然语言处理任务开展了大语言模型评估：评估开放式故事生成中故事的质量，以及评估对抗攻击生成的句子的质量。我们的发现和贡献总结如下：

- 我们表明，大语言模型评估得出的结果与专家进行的人类评估相似，

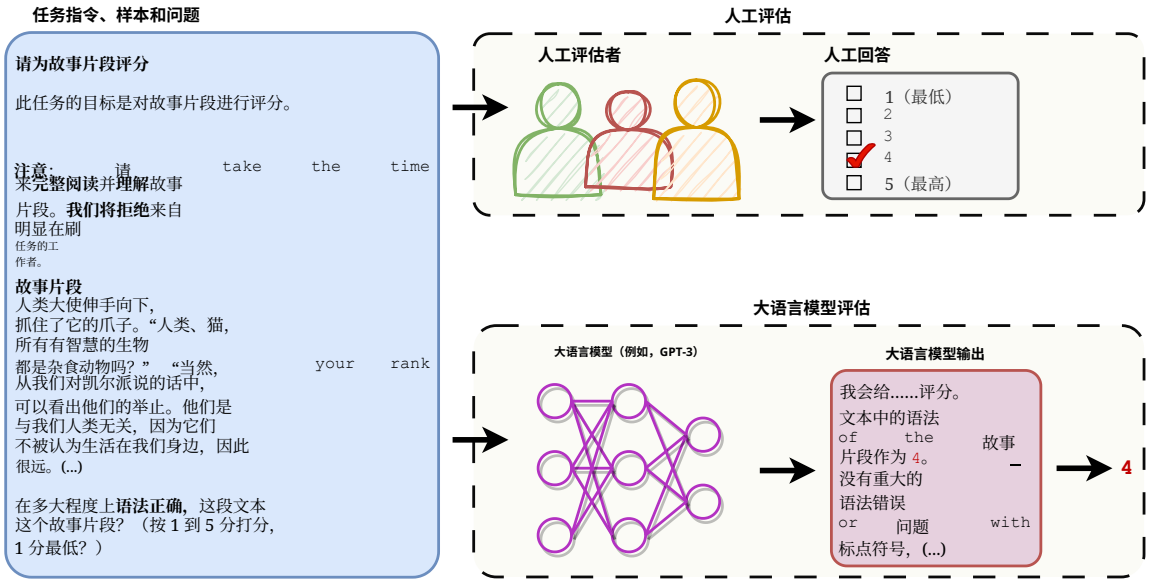

图1：以开放式故事生成为示例任务，展示论文的核心思想。左侧展示了人工评估中使用的指令、故事片段和问题。人类专家需要使用右上方所示的五点李克特量表，对故事片段的质量进行评分。右下方展示了大语言模型评估的流程：我们向大语言模型提供相同的指令、故事片段和问题，并解析大语言模型生成的输出以获得评分。

评估大语言模型评估的有效性（§3.3 和 §4.3）。本文是**首篇**提出使用大语言模型替代人工评估并展示其有效性的研究。

<escape>• </escape>我们表明，由于任务指令不同以及用于生成答案的采样算法的超参数不同，大语言模型评估结果仅存在轻微变化。（§3.3.2和 §3.3.3）

• 我们仔细讨论了使用大语言模型评估的利弊，并讨论了大语言模型评估的伦理考量。（§5）

2 大语言模型评估

2.1 大语言模型（大语言模型）

大语言模型是参数规模庞大的语言模型，通常达到数十亿的规模，并在海量自然语言语料上进行预训练，包括 GPT-3（Brown等，2020年）、T5（Raffel等，2020年）和 BLOOM（Scao等，2022）。仅提供任务指令时，这些大语言模型在未见过的任务上展现出卓越的性能；这种能力称为**零样本上下文学习**。

为了进一步提升零样本上下文学习性能，在预训练后对这些大语言模型采用了特殊的训练技术。例如，T0（Sanh 等人，2022）和 FLAN（Wei 等人，2022）在混合任务上进行微调，因此与 GPT-3 相比，能够取得更好的零样本性能。InstructGPT（Ouyang 等人，2022）通过基于人类反馈的强化学习（RLHF）在 GPT-3 的基础上进行微调，研究表明，它能够更好地遵循指令。ChatGPT（OpenAI，2022）使用对话数据集，通过 RLHF 在 InstructGPT 的基础上进行微调，因此能够以对话方式与用户互动。ChatGPT 能够回答用户提出的问题，并对答案作出全面解释。鉴于大语言模型遵循任务指令并提供反馈的能力，我们想知道大语言模型能否作为人工评估的替代方案，帮助自然语言处理研究人员评估文本质量。

2.2 大语言模型评估

为了使用大语言模型评估自然语言处理系统生成或由人类撰写的文本质量，我们向大语言模型提供任务指令、待评估的样本以及一个问题。该问题要求大语言模型使用五点李克特量表对样本质量进行评分。根据这些输入，大语言模型会通过生成一些输出句子来回答问题。我们解析这些输出句子以获得

大语言模型给出的分数。我们将这一过程称为**大语言模型评估**，该流程如图1下半部分所示。不同任务使用不同的任务指令，而每项任务也会使用不同的问题来评估样本的质量。本文中用于大语言模型评估的指令和问题并非专门为大语言模型量身定制；我们沿用先前研究中用于开展人工评估的指令。

为比较大语言模型评估的结果并展示其有效性，我们将大语言模型评估的结果与英语教师开展的人工评估进行比较。为确保比较公平且有意义，人工评估中的指令、样本和问题都采用了与大语言模型评估类似的格式。大语言模型评估与人工评估的主要区别在于，人工评估中，人工评估者通过从预先设定的选项中选择答案来回答问题（1–5分的李克特量表评分），如图1右上方所示。而在大语言模型评估中，我们让大语言模型自由生成句子，再使用一些简单规则从生成的句子中提取评分，详情见附录D.2.1。

3 示例任务1：开放式故事生成

我们首先使用开放式故事生成来展示大语言模型评估的用处。

3.1 任务介绍

开放式故事生成是一项根据给定提示词生成短篇故事的任务。我们使用WritingPrompts 数据集（Fan等人，2018年），该数据集由来自子版块 WritingPrompts 的提示词和人工撰写的故事配对组成。在WritingPrompts中，用户会收到一则简短提示词，并需要根据该提示词写一个故事。[1]

在这项实验中，我们使用大语言模型评估和人工评估，对人类创作的故事以及故事生成模型生成的故事进行评分。我们选择开放式故事生成为例，因为 Karpinska 等人(2021) 表明，来自亚马逊机械土耳其人(AMT)的工作者无法区分GPT-2(Radford等人，2019年)生成的故事和人工撰写的故事，而英语教师则明显更偏好人工撰写的故事，而非GPT-2 生成的故事。我们想看看大语言模型能否给人工撰写的故事打出比 GPT-2 生成的故事更高的分数。

遵循先前的研究（Mao等人，2019年；Guan等人，2020年；Karpinska 等人，2021），故事生成模型是基于 WritingPrompts 训练数据集进行微调的 GPT-2中型模型。模型训练完成后，我们从 WritingPrompts测试集中随机选取 200 条提示词，并使用核采样（Holtzman等人，2020年）让微调后的 GPT-2 根据这些提示词生成故事，参数为 $p = 0.9$。作为比较，我们使用基于相同 200 条提示词撰写的 200 篇人工撰写的故事。我们对人工撰写的故事和 GPT-2生成的故事进行后处理，然后将其用于大语言模型评估和人工评估。有关微调和数据处理的详情，请参见附录B。

3.2 大语言模型评估与人工评估

我们向大语言模型和人工评估者提供一段简短描述，以及待评估的故事，格式如图 1所示。遵循 Karpinska 等人 (2021）的做法，我们从四种不同属性对故事进行评估。这四种属性及其对应的问题如下：

1. 语法正确性：故事片段的文本在**语法上有多正确**？2. 连贯性：故事片段中的**句子**衔接得**有多好**？3. 可喜爱性：你觉得这个故事片段**有多有趣**？4. 相关性：现在请阅读故事片段所依据的提示词。**提示词**：［提示词］。这个**故事片段**与**提示词**有多**相关**？

其中，［提示词］将填入故事所依据的提示词。每种属性均使用五点李克特量表进行评估；每个问题末尾都会附上以下说明：“（按 *1* 至 *5* 分计，*1* 分为最低分）”。附录中展示了人工评估所用的界面以及大语言模型评估的输入格式C.2 和 D.2.2。

用于评估大语言模型的模型包括 T0、text-curie-001、text-davinci-003 和 Chat- GPT。text-curie-001 和 text-davinci-003

[1]WritingPrompts 子版块明确禁止用户使用人工智能生成故事，因此我们认为数据集中的故事均为人工撰写。

评估器	语法正确性		连贯性		可喜爱性		相关性	
	平均值$_{STD}$	IAA$_{\%}$	均值$_{STD}$	IAA$_{\%}$	均值$_{STD}$	IAA$_{\%}$	均值$_{STD}$	IAA$_{\%}$
人工撰写的故事								
人类	$3.76_{0.95}$	$0.33_{20.5}$	$4.29_{0.82}$	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	$3.83_{0.60}$		$3.55_{0.88}$		$2.44_{0.89}$		$3.29_{1.50}$	
*GPT-2*生成的故事								
人类	$3.56_{0.91}$	$0.10_{19.5}$	$3.19_{1.07}$	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：人工撰写的故事和 GPT-2 生成的故事的大语言模型评估与人工评估结果。对于每项评估属性，我们报告其李克特量表均值和标准差。我们还使用 Krippendorff 的 α 报告三位标注者之间的标注者间一致性（IAA）。IAA 列中的下标（%）表示三位标注者对某个评分完全一致的故事所占百分比。

是两个 InstructGPT 模型，后者更强；我们将用 InstructGPT 指代这两个模型。我们通过 OpenAI 提供的官方 API 查询 InstructGPT。我们使用核采样，并以 $p = 0.9$从 T0 和 InstructGPT 生成回答。我们从大语言模型中**抽取三个回答**，以模拟让模型对同一个故事进行三次评分的结果。我们通过 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 回答的参数。由于 ChatGPT 限制每位用户的查询次数上限，因此我们为每个问题只抽取一个回答。

在人工评估中，我们没有采用常用的 AMT，因为 Karpinska 等人（2021）已经表明，使用 AMT 得出的结果极具争议。遵循先前研究的建议，我们通过线上自由职业平台 **Upwork**聘请了三位认证英语教师。教师熟悉学生文章的评估工作，因此是我们这项任务的专家评估者。关于招募人工评估者的详细信息见附录C.1。每个大语言模型和每位英语教师都对 200 篇人工撰写的故事和 200 篇 GPT-2 生成的故事进行评分。

3.3 实验结果

开放式故事生成的大语言模型评估和人工评估结果见表 1。我们报告了大语言模型评估和人工评估所得李克特评分的均值和标准差，并通过两种不同的指标展示标注者间一致性（IAA）：（1）Krippendorff 的 α；（2）三位评估者给出完全相同评分的故事所占的百分比。[2] 表 1 的主要观察结果讨论如下。

专家人工评估者更偏好人工撰写的故事：人工评估结果可作为某种基准真值，用于评估大语言模型。对于全部四项属性，教师给人工撰写的故事的评分都高于 GPT-2 生成的故事。这表明专家能够区分模型生成的故事与人工撰写的故事之间的质量差异。根据标注者间一致性（IAA），我们还发现，专家对 GPT-2 生成文本以及可喜爱性的评价一致性较低。这表明，专家对模型生成文本以及主观属性（可喜爱性）的评价往往一致性较低，这与Karpinska 等人(2021）的研究结果一致。

T0 和 `text-curie-001` 并未明显偏好人工撰写的故事：对于 T0，我们可以看到，在语法正确性、可喜爱性和相关性方面，T0 对人工撰写的故-事的评分高于 GPT-2生成的故事。然而，人工撰写的故事与

[2]人工评估中的三位评估者是三位英语教师。在大语言模型评估中，我们将大语言模型生成的答案抽样三次，以此类比三位不同的评估者。

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未达到统计显著性；Welch$p$值t检验所得的 p 值远大于 0.05。

text-curie-001 的结果与 T0 类似：text-curie-001 给人工撰写的故事的评分并不高于模型生成的故事。还可以观察到，对于 T0，三种不同抽取答案之间的精确一致率所反映的标注者间一致性总体非常低。这表明，对于同一个样本，T0 很可能会对三个抽取答案给出不同的评分。结果意味着，T0 并未对某个特定评分赋予较高概率，因此抽取到不同分数的可能性都很大。这表明，即使大语言模型经过专门微调，以提升零样本上下文学习表现，并接受训练以更好地遵循人类指令，也无法像人类专家那样评估开放式故事生成。

`text-davinci-003` 与英语教师一样，明显更偏好人工撰写的故事： `text-davinci-003` 在所有四项属性上，对人工撰写的故事的评分都远高于模型生成的故事，这与人类专家得出的结果一致。通过 Welch 的 t检验，我们发现，对人工撰写的故事给出的较高评分在统计上均具有显著性。先前的研究发现，在 AMT 上招募的工作者无法区分人工撰写的故事和 GPT-2 生成的故事（Karpinska 等人，2021）；结合他们的结果与我们的结果可以看出，在开放式故事生成任务中，使用 `text-davinci-003` 进行大语言模型评估，比在 AMT 上进行人工评估能得出更有说服力的结果。结果表明，`text-davinci-003` 模型能够执行基本评估，例如检查故事中的语法错误。此外，该模型在评估故事与提示词的相关性方面表现出色，这需要对二者之间的联系进行更复杂的推理。我们还发现，text-davinci-003 的 Krippendorffα α `text-davinci-003` 远高于 T0 和 `text-curie-001`，这表明由 `text-davinci-003`给出的评分在对生成答案进行不同次抽样时更加一致。

ChatGPT 的评分与人类专家相当，也能很好地解释自己的判断： ChatGPT 还明显偏好人工撰写的故事，而且这种偏好在统计上显著。当我们通过 OpenAI 用户界面向 ChatGPT 提问时，发现了几项有趣的观察：**(1)**：ChatGPT 能够详细解释自己为何给出某个评分。它会引用故事和提示中的句子来支持自己的评分。**(2)**：ChatGPT 有时会拒绝评价故事的可喜爱性，因为“我是人工智能，无法体验愉悦”。在这种情况下，我们会重新生成回答，直到它给出评分。**(3)**：我们发现，ChatGPT 往往会给暴力或不礼貌的故事较低的可喜爱性评分，这可能是因为它接受的训练是提供安全且无害的回答，因此不喜欢残暴和粗俗的故事。

专家大多认同 ChatGPT 的评分和解释： 我们随机选取了 ChatGPT 对四个故事给出的答案，并询问英语教师是否认同 ChatGPT 的推理和评分[3]。教师大多认同其评分，并认为 ChatGPT 给出的解释合理。有趣的是，一位教师告诉我们，她不认同 ChatGPT 对语法正确性的评分，因为 ChatGPT 将标点错误视为语法错误，而她认为标点错误并不属于语法错误。这表明，不同个体有各自的评分标准，大语言模型也是如此。

`text-davinci-003` 倾向于给出较高的评分，而 ChatGPT 则相反： 对于同一类型文本的同一属性，与人工评分相比，text-davinci-003 给出的评分往往更高；相反，ChatGPT 更加挑剔，倾向于给出较低的分数。这表明，不同大语言模型在评分方面各有不同倾向。尽管 text-davinci-003、ChatGPT 和人类给出的分数绝对值有所不同，但它们都认为人类撰写的文本优于 GPT-2生成的故事。分数的绝对值反映了评估者的偏见或信念；只要使用相同的评估者来评估不同的系统，比较就是有意义的。

3.3.1 大语言模型与人类评估者对单个故事的评分是否一致?

我们在表1 中发现，`text-davinci-003`和 ChatGPT 的评分都明显偏向人工撰写的故事，与英语教师的倾向一致。然而，目前尚不清楚这些大语言模型对每篇故事的评分是否与教师一致。具体而言，当英语教师给某篇故事的评分较高时，大语言模型是否也会给这篇故事较高的评分?

[3]我们没有告知教师这些回答来自人工智能模型。故事及教师的回复请参见附录C.3.2。

故事作者	人类	GPT-2
语法正确性	0.14	0.12
连贯性	0.18	0.14
可喜爱性	0.19	0.22
相关性	0.38	0.43

表 2：英语教师与 text-davinci-003 之间的 Kendall τ 相关系数。

为回答这个问题，我们计算 τ text-davinci-003与英语教师评分之间的肯德尔相关系数。我们选择使用相关系数，而非标注者间一致性得分，因为标注者间一致性主要关注两位标注者是否给出完全相同的评分，而相关系数关注的问题是："当标注者 A 给某篇故事的评分较高时，标注者 B 是否也会给这篇故事较高的评分？"（Amidei 等人，2019 年）。我们按如下方式计算四种评分属性的肯德尔 τ：对于每篇故事和每种评分属性，我们计算三位英语教师的平均评分，并计算 text-davinci-003给出的三个分数的平均值（这三个分数来自三个独立样本）。对于每种属性，我们将教师的平均评分汇总为向量 $A \in \mathbb{R}^{200}$，其中每个条目是一篇故事的平均评分；同样，我们构建向量 $B \in \mathbb{R}^{200}$，表示 davinci的平均评分。接下来，我们计算τ 相关系数 A 与 B之间的肯德尔相关系数。

肯德尔相关系数 τ 教师评分与大语言模型评分之间的关系如表 2所示。[4] 我们发现，对于全部四种属性，以及人类撰写的故事和GPT-2生成的故事，我们都观察到教师评分与 text-davinci-003的评分之间存在从弱到强的正相关。所有相关性的 p值都小于 0.05。因此，我们可以说，当教师给某个故事的评分更高时，text-davinci-003 在一定程度上也会给它更高的评分。我们还观察到，不同属性的肯德尔相关系数 τ 差异很大：相关性最强，而 语法正确性最弱。这可能是因为对相关性进行评分相当直接，只需检查提示词中的内容是否在故事中被提及。相反，哪些内容应被视为-

[4]解读 Kendall 系数时，τ，$|\tau| \in [0, 0.1)$ 被视为极弱相关，$|\tau| \in [0.1, 0.2)$ 被视为弱相关，$|\tau| \in [0.2, 0.3)$ 被视为中等相关，而 $|\tau| \in [0.3, 1.0]$ 被视为强相关（Botsch，2011）。

在评分时考虑语法正确性并未在我们的指令中明确说明，因此大语言模型可能采用了与英语教师不同的评分标准。我们还计算了两位英语教师之间的平均 Kendall τ，发现两位教师对语法正确性的评分相关性较弱，而对相关性的评分相关性则强得多。结果见附录中的表格。

3.3.2 不同指令导致的方差

研究表明，大语言模型有时会对用于查询大语言模型的指令十分敏感（Zhao 等人，2021；Sanh 等人，2022）。为探究改变任务指令和问题会如何影响开放式故事生成的大语言模型评估结果，我们对指令和问题进行修改，并观察大语言模型评估结果的变化。我们通过修改图1中的指令或问题，测试两种不同的指令：（1）在图*(你是一名受雇为故事片段评分的人类工作者。)*中的任务指令前添加句子"1"。我们尝试为大语言模型提供一个**角色设定**，以便它更好地理解自己的角色。这一做法受到了以往研究的启发；该研究报告称，赋予 GPT-3 角色设定会产生不同的结果（Zeng 等人，2022）。（2）我们要求大语言模型通过在问题后附加以下句子来**解释**其判断：请同时解释你的判断。在此，我们想了解，当要求大语言模型说明其判断理由时，它们对故事的评分是否会有所不同。本实验使用 text-davinci-003 作为大语言模型，因为根据表1，它的结果与专家人工评估相近，而且比 ChatGPT 更容易获取。

结果见表格上方部分3。我们观察到，对于语法正确性和连贯性，不同指令所得到的分数相当接近：不同指令导致的评分变化小于 0.1。对于另外两个属性，分数变化稍大，但仍在0.25的范围内。尽管不同指令会带来细微差异，但这些差异仍不会改变" 大语言模型对人工撰写的故事的评分高于对GPT-2生成的故事的评分"这一结论。因此，不同指令不会改变GPT-2生成的故事与人工撰写的故事的相对排名。总之，只要使用大语言模型评估时采用相同的指令来评估故事，这种评估和比较就是有意义的。

设置	语法正确性		连贯性		可喜爱性		相关性	
	人类	*GPT-2*	人类	*GPT-2*	人类	*GPT-2*	人类	*GPT-2*
不同的指令（第*3.3.2*节）								
原始内容	$4.22_{0.38}$	$4.07_{0.35}$	$4.54_{0.45}$	$4.26_{0.45}$	$3.99_{0.38}$	$3.84_{0.42}$	$4.40_{0.79}$	$4.02_{0.74}$
（1） +角色设定	$4.29_{0.45}$	$4.01_{0.45}$	$4.60_{0.49}$	$4.27_{0.50}$	$4.05_{0.39}$	$3.87_{0.39}$	$4.55_{0.70}$	$4.25_{0.77}$
（2） +解释	$4.24_{0.42}$	$4.05_{0.25}$	$4.61_{0.49}$	$4.32_{0.51}$	$4.15_{0.44}$	$3.98_{0.34}$	$4.35_{0.75}$	$4.03_{0.56}$
不同的采样温度 T （第*3.3.3*节）								
$T = 1.0$	$4.22_{0.38}$	$4.07_{0.35}$	$4.54_{0.45}$	$4.26_{0.45}$	$3.99_{0.38}$	$3.84_{0.42}$	$4.40_{0.79}$	$4.02_{0.74}$
$T = 0.7$	$4.18_{0.35}$	$4.06_{0.33}$	$4.52_{0.48}$	$4.23_{0.43}$	$3.96_{0.34}$	$3.82_{0.42}$	$4.36_{0.77}$	$3.95_{0.72}$
$T = 0.3$	$4.13_{0.33}$	$3.99_{0.25}$	$4.48_{0.49}$	$4.14_{0.39}$	$3.95_{0.26}$	$3.82_{0.41}$	$4.34_{0.75}$	$3.93_{0.67}$
$T = 0$	$4.07_{0.27}$	$3.99_{0.18}$	$4.49_{0.50}$	$4.09_{0.34}$	$3.95_{0.25}$	$3.82_{0.40}$	$4.32_{0.75}$	$3.92_{0.66}$

表3：了解大语言模型评估的方差。对于评估的四种属性，左栏为人工撰写的故事的均值和标准差，右栏为 GPT-2 生成的故事的均值和标准差。上半部分展示了由于不同指令而导致的评分变化（第 3.3.2 节），下半部分则展示了更改用于生成大语言模型输出的温度 T 所得到的结果（第 3.3.3 节）。

3.3.3 不同采样参数导致的方差

从大语言模型生成答案时，我们必须为生成选择一组超参数，包括温度 T 和核采样中使用的概 p 率（Holtzman等人，2020年）。为了解不同的采样参数是否会改变大语言模型评估结果，我们调整采样所用的温度，并在从 p 生成答案时，将核采样中的0.9 固定为 `text-davinci-003`。我们不会同时调整 T 和 p，因为这两个参数都用于控制输出的多样性；只改变其中一个参数就足够了，这也是OpenAI 的建议。

将 T 从 1 变为 0 时的结果显示在表3的下半部分。当 T 从 1 变为 0时，我们观察到一个有趣的趋势：在大多数情况下，平均评分略有下降。考虑到 $T = 0$只是最大值采样，这一结果表明，概率最高的大语言模型回答往往得分较低。尽管存在这一有趣趋势，大语言模型始终给人工撰写的故事比 GPT-2生成的故事更高的评分。虽然表 3中未展示，但我们发现，温度越低，标注者间一致性越高。这是意料之中的，因为较低的温度意味着大语言模型采样时的多样性较低，从而使采样得到的评分更加一致。总之，改变指令和温度可以略微改变大语言模型给出的评分绝对值，但不会改变大语言模型对人工撰写的故事的偏好。本节的总体结果表明，大语言模型评估是有用的可用于评估开放式故事生成。

4 示例任务 2：对抗攻击

作为另一种应用，我们使用大语言模型评估来评定对抗攻击生成的文本。

4.1 任务介绍

给定一个训练好的文本分类器，以及一个良性（非对抗性）测试样本，且该文本分类器能够正确分类该样本，对抗攻击旨在构造一个对抗性样本，使分类器作出错误预测。一种特殊类型的对抗攻击称为同义词替换攻击（SSA）（Alzantot 等人，2018年），其做法是在良性样本中用同义词替换部分词语，从而生成对抗样本。通过将词语替换为其同义词，良性样本的语义应在对抗样本中得到保留，同时使人类难以察觉这种对抗扰动。尽管这一思路在概念上合理，但近期研究表明，许多 SSA 往往会产生不合语法、不自然的对抗样本，并显著改变良性样本的含义（Hauser 等人，2021；Chiang 和 Lee，2022）。为了评估对抗样本的质量，人工评估至关重要，并已在以往研究中得到广泛应用。在我们的实验中，我们希望了解大语言模型能否像人类专家一样评定对抗样本的质量。对抗样本并非普通文本，因此大语言模型在训练期间可能未接触过此类异常输入。了解大语言模型如何评定这些对抗样本会很有意思。

	人类评估		大语言模型评估	
	流畅度	**均值。**	**流畅度**	**均值。**
良性	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：大语言模型评估与人类评估对良性样本和对抗样本的流畅性（**流畅度**）评分的平均李克特分数，以及对抗攻击前后新闻标题之间意义保留（**意义保留**。）的评分。

4.2 实验设置

我们选择了三种不同的经典文本语义攻击：Textfooler（Jin 等人，2020年）、PWWS（Ren 等人，2019年）和 BAE（Garg 和 Ramakrishnan，2020年）；如今，这些攻击在文本语义攻击相关文献中主要被用作强基线。我们使用这三种文本语义攻击，对在 AG-News（Devlin 等人，2019年）上微调的 BERT-base-uncased 模型进行攻击。AG-News 是一个新闻标题分类数据集（Zhang 等人，2015年）。对于每种文本语义攻击，我们随机选取 100 对良性样本和对抗样本，并使用大语言模型评估其质量。此处展示使用 ChatGPT 作为大语言模型的结果，因为它能更好地解释其判断。按照先前研究的建议（Morris 等人，2020年），我们从两个方面评估对抗样本的质量： 流畅性和意义保持。对于流畅性，我们向大语言模型提供一条新闻标题（良性样本或对抗样本），并提出以下问题：新闻标题文本读起来有多自然流畅？（按 *1* 到 *5* 分评分，*1* 分最低）。对于意义保持，我们同时向大语言模型提供良性样本和对抗样本，并提示大语言模型回答以下问题：您是否同意新闻标题 *1* 的含义（或语义）在新闻标题 *2* 中得以保留？（按 *1* 到 *5* 分评分，*1* 分表示非常不同意，*5* 分表示非常同意。）具体的指令和格式见附录D.2.3。我们还请三位英语教师对样本的流畅性和意义保持进行评分。任务指令和问题的格式与大语言模型评估中的相同。

4.3 实验结果

结果见表4。可以看到，英语教师认为 SSA 生成的对抗样本在流畅性和意义保留方面的评分远低于良性样本，这一结果与近期关于对抗样本质量的观察一致（Hauser 等人，2021；Chiang 和 Lee，2022）。在解读大语言模型评估结果之前，我们首先对大语言模型是否理解任务进行合理性检查。我们要求大语言模型对两个完全相同的良性样本的意义保留程度进行评分。理想情况下，大语言模型应始终给出 5 分，表示其强烈认同样本含义未发生改变。该合理性检查的结果是表 † 中的4这一项，得分完美，为5。00。ChatGPT 经常会说：“这两个标题完全相同，所以我给 *5* 分（强烈认同）”，这表明 ChatGPT 理解任务的内容。

接下来，我们将注意力转向对抗样本的大语言模型评估结果。我们观察到，ChatGPT 给对抗样本的评分往往高于英语教师，这意味着 ChatGPT 对对抗样本中不自然和人工的部分没有那么苛刻。我们使用 text-davinci-003 进行了相同的实验，并发现了类似的结果。尽管 ChatGPT 给对抗样本的评分高于教师，但其评分仍显著低于对良性样本的评分。ChatGPT 也认同英语教师的看法，认为 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，而专家人工评估中并未出现这种评分差异。乍一看，这意味着 ChatGPT 的评估与人工评估结果不一致。然而，仔细审视人工评估结果后，我们发现，两位教师给 PWWS 的评分高于 Textfooler，而一位教师给 PWWS 的评分低于 Textfooler。这表明 ChatGPT 实际上与大多数人类专家的看法一致。总体而言，大语言模型能够像大多数人类专家一样，对抗文本和良性文本的质量进行排序。

5 讨论

本文提出使用大语言模型评估文本质量，以替代人工评估。为展示大语言模型评估的潜力，我们使用大语言模型对两项截然不同的任务中的文本质量进行评分：开放式故事生成和对抗攻击。我们表明，即使大语言模型具备卓越的零样本上下文能力。

学习能力，但它们并不总是适合用于大语言模型评估。不过，我们发现，在作为示例的两项任务中，表现最好的 InstructGPT 和 ChatGPT 能像人类专家一样评估文本质量。总体而言，本文的结果表明，大语言模型评估有望用于评估自然语言处理系统和算法。

大语言模型评估的优点 与人工评估相比，大语言模型评估有若干优势。首先，大语言模型评估**更具可复现性**。人工评估结果难以复现，因为很难聘用同一批评估者；即使实验采用相同的指令、招募平台和评估者资质，也很难比较类似实验的结果。相比之下，大语言模型评估没有这一缺点。通过指定用于大语言模型评估的模型、随机种子，以及用于生成大语言模型回答的超参数，大语言模型评估结果更有可能复现。需要注意的是，在某些情况下，大语言模型提供商可能会定期更新模型；如果旧版模型已无法访问，那么大语言模型评估便无法复现。

其次，**在大语言模型评估中，对每个样本的评估彼此独立**。相反，在人工评估中，当前示例的评分或多或少可能受到先前样本的影响。人类往往会将当前样本与之前看过的样本进行比较，这会影响他们的评分。一个佐证是，在评完 400 个故事后的访谈中，英语教师表示，他们花了一些时间来校准评分（附录C.3.1）。因此，使用大语言模型评估可以简化某些实验设计，因为无需担心待评估样本的顺序是否会改变结果。不过，也有人可能会认为，能够校准对不同样本的评分是可取的，因此可能更倾向于采用人工评估。总的来说，评估者（人类或大语言模型）的评分是否会受到先前已评分项目的影响，本质上是实验设计的一项选择。

第三，大语言模型评估比**人工评估更便宜、更快**，让研究人员能够更轻松、更迅速地评估自然语言处理系统的质量。聘请一位英语教师为 200 篇故事评分需要花费我们 140 美元，而使用最好的 InstructGPT 模型进行大语言模型评估，成本不到 5 美元。从招募人员到收集人工评估结果，我们花了一周多的时间；而收集评估结果只需几小时：只要向 InstructGPT 发出查询并进行大语言模型评估即可。

最后，采用大语言模型评估而非人工评估，可以**最大限度减少人类接触令人反感内容的需要**，例如暴力、色情、仇恨或带有偏见的材料。这类内容可能会让人工评估者在阅读和评判这些文本时感到不适。[5]

大语言模型评估的局限性与伦理考量尽管本文展示的大语言模型评估结果颇具前景，但这种方法仍存在一些局限性。首先，大语言模型可能掌握错误的事实知识（Caoet al.，2021），因此不适合用于涉及事实知识的任务。其次，经过特定行为训练的大语言模型可能会偏向某些回答。具体而言，经过安全、无害训练的大语言模型可能会使大语言模型更倾向于生成积极、乐观的回答；我们与 ChatGPT 的互动始终体现出这一点。此外，尽管研究人员努力提升大语言模型的安全性（Baietal.，2022a，b），大语言模型仍可能生成有害且带有偏见的回答（Ganguliet al.，2022；Perezet al.，2022），这违反了基本伦理，也会使大语言模型评估结果受到严重质疑（Hendryckset al.，2021）。然而，值得注意的是，这些局限性和潜在危害同样适用于人工评估：人工评估者的偏见会影响人工评估结果（Lentz and De Jong，1997；Amidei et al.，2018年）。

我们提出的开创性理念——大语言模型评估，有望改变自然语言处理社区。[6] 我们鼓励未来的研究者在了解其局限性的同时，考虑采用这一方法。本文的目标并非取代人工评估，而是提供一种替代方案。人工评估和大语言模型评估各有优缺点，也可以结合使用。我们建议在开发新的自然语言处理系统时，将大语言模型评估用作一种廉价、快速的质量判断方式；而在将自然语言处理系统部署到实际应用之前，最好通过人工评估收集人类的反馈

[5]需要注意的是，对于违反大语言模型提供方内容政策的某些输入，大语言模型可能会拒绝进行评估。[6]我们认为，我们是首个提出这一想法的团队，因为在2023年1月13日将本文投稿至 ACL 2023 时，我们没有发现其他探索这一想法的论文。在审稿过程中，我们在 arXiv 上发现了一些研究（Wang 等人，2023；Huang 等人，2023；Gilardi 等人，2023），它们使用不同的任务探索了类似的想法。

现实应用中的自然语言处理系统。

局限性

还应注意，大语言模型评估存在其他局限性和潜在风险，而这些局限性实际上是预训练语言模型众所周知的问题。正如Open AI 关于ChatGPT 的博客所列，ChatGPT 有时会生成听起来正确、合理，实则完全荒谬的回答。OpenAI 也承认，模型的回答可能会受到用于查询模型的提示词影响。虽然在第3.3.2节中，我们发现不同指令之间的总体结果差异并不显著，但无法保证对任务指令进行各种修改时，情况都如此。

除了 OpenAI 博客中列出的局限性之外，还有其他局限性。例如，大语言模型可能没有情绪。人工智能模型是否具有情绪是一个更具哲学性且存在争议的问题，因此，使用此类模型评估与情绪相关的任务所得结果可能会受到强烈质疑，甚至可能违反研究伦理。正如我们在实验中发现的那样，当被要求评估一个故事的可喜爱性时，ChatGPT 经常会回答“我是一个人工智能系统，没有人类那样的情绪”。

大语言模型评估的另一项重要局限是，与人工评估不同，大语言模型无法处理任务指令中的视觉线索。人工评估者可以利用特殊字体或文本样式等格式设置，突出指令中的重要部分。此外，指令和问题的格式也会影响人工评估者处理任务的方式。然而，大语言模型只能处理原始文本输入，无法接收视觉线索。

伦理声明

大语言模型评估的进一步伦理考量tion 除了前文提到的大语言模型评估的局限性之外，大语言模型评估的核心还涉及一个至关重要的伦理问题。用大语言模型评估取代人工评估是否合乎伦理？有些人可能会质疑，本文是否在暗示大语言模型已经准备好取代人类，并对此感到不安。作为负责任且重视伦理问题的自然语言处理研究人员，我们理解这些担忧，但希望明确说明，这并非我们的意图。正如本文标题所示，我们旨在提供一种 *alter-native option* 人工评估的替代方案，以实现提高自然语言处理研究的可复现性。人工评估仍然至关重要，因为自然语言处理系统的最终目标是供人类用户使用，因此收集他们的反馈十分重要。我们非常享受与所聘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取消人工评估；相反，我们认为应将人工评估与大语言模型评估结合使用。这两种方法各有优缺点，因此在评估自然语言处理系统时二者都是必不可少的。我们希望本文的积极结果能为自然语言处理研究人员提供一种评估系统的替代方法，并促进围绕这一主题的进一步讨论。

论文中实验的伦理声明 所有实验均严格遵循ACL 伦理准则。我们在附录 C.1中详细介绍了人工评估。总而言之，我们列出了人工评估所用的确切指令和界面截图，并说明了评估者的招募方式，同时表明报酬十分合理。我们会告知人工评估者任务内容，并告诉他们，他们的回答将用于评估人工智能模型的表现。我们所在的机构没有伦理审查委员会或类似机构，因此无法获得伦理审查委员会的批准。尽管如此，我们仍尽力遵循 ACL 的伦理准则。

我们在遵循模型和数据集的预期用途时使用它们。具体而言，我们在使用 InstructGPT 模型和 ChatGPT 模型时，遵循OpenAI 使用政策。

致谢

我们感谢审稿人提供详尽的反馈和切实可行的建议，这些意见有助于我们完善论文。我们在附录 A中列出了根据审稿人建议所做的修改。我们感谢 Yung-Sung Chuang 对论文草稿提出宝贵意见。我们还要感谢实验室行政助理 Tung-En Hsiao 帮助我们处理在 Upwork 上的付款事宜。Cheng-Han Chiang 获得了台达电子提供的博士奖学金。

参考文献

Moustafa Alzantot、Yash Sharma、Ahmed Elgohary、Bo-Jhang Ho、Mani Srivastava 和 Kai-Wei Chang。2018年。生成自然语言对抗样例。载于 *2018*年自然语言处理实证方法会议论文集，第2890–2896页，比利时布鲁塞尔。计算语言学协会。

Jacopo Amidei、Paul Piwek 和 Alistair Willis。2018年。重新思考人工评估任务中的一致性。载于第*27*届国际计算语言学大会论文集，第3318–3329页，美国新墨西哥州圣菲。计算语言学协会。

Jacopo Amidei、Paul Piwek 和 Alistair Willis。2019年。一致性被高估了：呼吁通过相关性评估人类评估的可靠性。载于第*12*届自然语言生成国际会议论文集，第344–354页，日本东京。计算语言学协会。

Yuntao Bai、Andy Jones、Kamal Ndousse、Amanda Askell、Anna Chen、Nova DasSarma、Dawn Drain、Stanislav Fort、Deep Ganguli、Tom Henighan 等人。2022a。通过基于人类反馈的强化学习训练有帮助且无害的助手。*arXiv*预印本 *arXiv:2204.05862*。

Yuntao Bai、Saurav Kadavath、Sandipan Kundu、Amanda Askell、Jackson Kernion、Andy Jones、Anna Chen、Anna Goldie、Azalia Mirhoseini、Cameron McKinnon 等。2022b。宪法式人工智能：通过人工智能反馈实现无害性。*arXiv* 预印本 *arXiv:2212.08073*。

R Botsch。2011。第12章：关联的显著性与测量。政治科学的范围与方法。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901。

Nicola De Cao、Wilker Aziz 和 Ivan Titov。2021。编辑语言模型中的事实知识。

Cheng-Han Chiang 和 Hung-yi Lee。2022。我们距离真正的同义词替换攻击还有多远？*arXiv* 预印本 *arXiv:2210.02844*。

Elizabeth Clark、Tal August、Sofia Serrano、Nikita Haduong、Suchin Gururangan 和 Noah A. Smith。2021。并非所有“人类”的东西都是金子：评估对生成文本的人工评估。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第7282–7296页，线上。计算语言学协会。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向 Transformer 预训练。载于*2019*年北美计算语言学协会分会年会论文集：人类语言技术，第*1*卷（长篇与短篇论文），第4171–4186页，明尼苏达州明尼阿波利斯。计算语言学协会。

Angela Fan、Mike Lewis 和 Yann Dauphin。2018年。分层神经故事生成。载于第*56*届计算语言学协会年会论文集（第*1*卷：长篇论文），第889–898页，澳大利亚墨尔本。计算语言学协会。

Deep Ganguli、Liane Lovitt、Jackson Kernion、Amanda Askell、Yuntao Bai、Saurav Kadavath、Ben Mann、Ethan Perez、Nicholas Schiefer、Kamal Ndousse 等。2022。对语言模型进行红队测试以减少危害：方法、规模化行为与经验教训。*arXiv* 预印本 *arXiv:2209.07858*。

Leo Gao。2021年。关于 OpenAI API 模型的规模。访问日期：2023年1月17日。

Siddhant Garg 和 Goutham Ramakrishnan。2020年。BAE：基于 BERT 的文本分类对抗样本。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第 6174–6181 页。

Fabrizio Gilardi、Meysam Alizadeh 和 Maël Kubli。2023。ChatGPT 在文本标注任务上的表现优于众包工作者。*arXiv* 预印本 *arXiv:2303.15056*。

Dan Gillick 和 Yang Liu。2010年。非专家评估摘要系统存在风险。载于*NAACL HLT 2010*年利用 *Amazon Mechanical Turk* 创建语音和语言数据研讨会论文集，第 148–151 页，洛杉矶。计算语言学协会。

JianGuan、Fei Huang、Zhihao Zhao、Xiaoyan Zhu 和 Minlie Huang。2020年。一种用于常识故事生成的知识增强预训练模型。计算语言学协会会刊，8：93–108。

Francisco Guzmán、Ahmed Abdelali、Irina Temnikova、Hassan Sajjad 和 Stephan Vogel。2015年。人类如何评估机器翻译。载于第十届统计机器翻译研讨会论文集，第457–466页，葡萄牙里斯本。计算语言学协会。

Jens Hauser、Zhao Meng、Damián Pascual 和 Roger Wattenhofer。2021。BERT 很 robust！一项反对基于同义词的文本分类对抗样本的案例。*arXiv* 预印本 *arXiv:2109.07403*。

Dan Hendrycks、Collin Burns、Steven Basart、Andrew Critch、Jerry Li、Dawn Song 和 Jacob Steinhardt。2021。使{ai}与共同的人类价值观保持一致。载于国际学习表征会议。

Ari Holtzman、Jan Buys、Li Du、Maxwell Forbes 和 YejinChoi。2020年。神经文本生成的奇妙案例。载于国际学习表征会议。

Fan Huang、Haewoon Kwak 和 Jisun An。2023年。ChatGPT 比人工标注员更好吗？ChatGPT 在解释隐性仇恨言论方面的潜力与局限。*arXiv* 预印本 *arXiv:2302.07736*。

Di Jin、Zhijing Jin、Joey Tianyi Zhou 和 Peter Szolovits。2020年。BERT 真的稳健吗？针对文本分类和蕴涵的自然语言攻击的强基线。载于 *AAAI* 人工智能会议论文集，第34卷，第8018–8025页。

Marzena Karpinska、Nader Akoury 和 Mohit Iyyer。2021年。使用 Mechanical Turk 评估开放式文本生成的风险。载于*2021*年自然语言处理实证方法会议论文集，第1265–1285页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Leo Lentz 和 Menno De Jong。1997。《文本质量评估：以专家为中心的方法与以读者为中心的方法比较》。*IEEE* 专业传播学报，40(3):224–234。

Huanru Henry Mao、Bodhisattwa Prasad Majumder、Julian McAuley 和 Garrison Cottrell。2019年。通过目标导向的常识落地改进神经故事生成。载于*2019*年自然语言处理实证方法会议暨第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第5988–5993页，中国香港。计算语言学协会。

John Morris、Eli Lifland、Jack Lanchantin、Yangfeng Ji 和 Yanjun Qi。2020年。重新评估自然语言中的对抗样本。收录于 计算语言学协会发现：*EMNLP 2020*年，第 3829–3839 页，线上。计算语言学协会。

OpenAI。2022。ChatGPT：优化用于对话的语言模型。访问日期：2023年1月10日。

Long Ouyang、JeffWu、XuJiang、DiogoAlmeida、Car- roll L Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。2022。通过人类反馈训练语言模型遵循指令。*arXiv* 预印本 *arXiv:2203.02155*。

Ethan Perez、Saffron Huang、Francis Song、Trevor Cai、Roman Ring、John Aslanides、Amelia Glaese、Nat McAleese 和 Geoffrey Irving。2022年。使用语言模型对语言模型进行红队测试。收录于*2022*年自然语言处理实证方法会议论文集，第3419–3448页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

AlecRadford、JeffreyWu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。2019年。语言模型是无监督的多任务学习器。*OpenAI* 博客，1(8):9。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。2020年。探索统一文本到文本 Transformer 的迁移学习极限。《机器学习研究杂志》，21(140):1–67。

Shuhuai Ren、YiheDeng、Kun He 和 Wanxiang Che。2019年。通过概率加权词显著性生成自然语言对抗样例。载于第*57*届计算语言学协会年会论文集，第1085–1097页，意大利佛罗伦萨。计算语言学协会。

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、De- bajyoti 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、Ab- heesht Sharma、Andrea Sa ntilli、Thibault Fevry、Ja- son Alan Fries、Ryan Teehan、Teven Le Scao、Stella Biderman、Leo Gao、Thomas Wolf 和 Alexan- der M Rush。2022 年。多任务提示训练实现零样本任务泛化。载于 国际学习表征会议。

Teven Le Scao, Angela Fan, Christopher Akiki, El- lie Pavlick, Suzana Ili´c, Daniel Hesslow, Ro- man Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé 等。2022。Bloom：一款拥有 1760 亿参数的开放获取多语言语言模型。*arXiv* 预印本 *arXiv:2211.05100*。

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu 和 Jie Zhou。2023。ChatGPT 是一款优秀的自然语言生成评估器吗？一项初步研究。*arXiv* 预印本 *arXiv:2303.04048*。

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai 和 Quoc V Le。2022。微调语言模型是零样本学习器。载于国际学习表征会议。

Thomas Wolf、Lysandre Debut、Victor Sanh、J ulien Chaumond、Clement Delangue、Anthony Moi、Pier- ric Cistac、Tim Rault、Remi Louf、Morgan Funtow- icz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。2020年。Transformers：自然语言处理的最新进展。载于*2020*年自然语言处理经验方法会议论文集：系统演示，第 38–45 页，线上。计算语言学协会。

KiYoon Yoo、Jangho Kim、Jiho Jang 和 Nojun Kwak。2022。文本分类中对抗样本的检测：基于鲁棒密度估计的基准测试与基线方法。载于计算语言学协会：*ACL 2022* 论文集，第3656–3672 页，爱尔兰都柏林。计算语言学协会。

Andy Zeng、Adrian Wong、Stefan Welker、Krzysztof Choromanski、Federico Tombari、Aveek Purohit、Michael Ryoo、Vikas Sindhwani、Johnny Lee、Vincent Vanhoucke 等。2022年。苏格拉底模型：通过语言构建零样本多模态推理。*arXiv* 预印本 *arXiv:2204.00598*。

Xiang Zhang、Junbo Zhao 和 Yann LeCun。2015年。用于文本分类的字符级卷积网络。神经信息处理系统进展，28。

Zihao Zhao、Eric Wallace、Shi Feng、Dan Klein 和 Sameer Singh。2021。使用前先校准：提升语言模型的少样本性能。载于国际机器学习会议，第12697–12706页。PMLR。

基于评审意见的修改

我们列出了本文当前版本与审稿前版本之间的主要区别；所有修改均基于审稿人的建议。我们再次感谢审稿人提出的这些宝贵建议。

- 我们新增了第 3.3.1 节，探讨大语言模型和人工评估者对单个故事的评分是否一致。
- 我们对第 5 节中的措辞进行了润色，并补充了相关参考文献。
- 我们新增表格6 以讨论人工评估者之间的相关性。
- 我们在进行人工评估时，混合使用人工撰写的故事和GPT-2生成的故事，开展了关于人工评估的补充实验，并在表 5 中报告了实验结果。
- 我们修正了拼写错误，并采纳了审稿人提到的几乎所有表达方面的建议。由于篇幅有限，我们无法采纳所有表达方面的建议。

B 开放式故事生成的实验详情

B.1 WritingPrompts数据集

训练数据集包含 30.3 万组故事与提示词配对，我们的模型正是用这些数据进行训练的。我们仅使用测试集中的 200 组提示词与故事配对。数据集下载自 https://www.kaggle.com/datasets/ratthachat/writing-prompts。

B.2 微调GPT-2模型

我们使用 $5e-5$ 的学习率和线性学习率调度训练模型 3 个训练轮次，最终训练出的模型在 WritingPrompts 的验证集上达到 20 的困惑度。

B.3 数据后处理

模型训练完成后，我们从 WritingPrompts 的测试集中随机选取 200 条提示，并将这些提示输入训练好的模型，让模型根据给定的提示生成故事。生成故事时，我们采用 $p = 0.9$ 进行核采样。接下来，我们手动将生成的故事截断至少于 150 个词，并确保截断后的故事以完整句子结尾。[7] 经过这一处理，我们得到了 200 对提示和模型生成的故事。

作为与模型生成故事的对比，我们选取了用于生成模型生成故事的相同 200 个提示词及其对应的人工撰写的故事，从而形成 200 组提示词与人工撰写故事的配对。对于这些人工撰写的故事，我们同样将其截断至 150 个单词以内，并以一个完整句子结尾，以便与模型生成的句子保持一致。此外，我们还手动删除了人工撰写的故事中因 WritingPrompts 数据集的分词处理而产生的一些异常内容。

C 人工评估

C.1 招募英语教师

这些英语教师持有英语作为第二语言教学证书[8]；鉴于他们有批改学生文章的经验，因此非常适合这项任务。我们请每位教师为200篇GPT-2生成的故事和200篇人工撰写的故事评分，并支付他们140美元为200篇故事评分。教师报告称，他们为200篇故事评分最多需要5小时，因此时薪至少为28美元。我们先请教师为GPT-2生成的故事评分，然后再为200篇人工撰写的故事评分。与Karpinska不同。

[7]我们将故事截断至150个单词，因为这是模型生成故事的平均长度。[8]英语作为第二语言教学认证

等人。（2021）在评定GPT-2生成的故事和人工撰写的故事之间会休息一段时间，而我们不休息，以避免教师在长时间休息后改变评分标准。在教师评估故事之前，我们不会告知他们这些故事的作者是谁。等他们评完所有故事后，我们才向他们说明本项目的研究目的。

我们不将人工撰写的故事和GPT-2生成的故事混合进行评分，原因在于 Karpin- ska 等人 （2021）观察到：（1）当 AMT 工人分别评估模型生成的故事和人工撰写的故事时，**分别**进行评分时，他们并不会表现出偏好人工撰写的故事；但（2）即使分别评估模型生成的故事和人工撰写的故事时，**分别**进行评分，英语教师也会明显偏好人工撰写的故事。我们遵循他们的设置，不混合 GPT-2 生成的故事和人工撰写的故事。在审稿过程中，审稿人询问了为何不将这些故事混合起来进行人类评估。因此，我们开展了相同的实验：随机混合 200 篇人工撰写的故事和 200 篇 GPT-2 生成的故事，并请三位教师（并非之前已经评估过这些故事的教师）对其进行评分。其他所有实验条件均与前文所述相同。完整结果见表 5。我们发现，对于全部四个属性，教师仍然明显偏好人工撰写的故事，这与表 1中的观察结果相似。唯一的例外是语法正确性：英语教师并未非常明显地偏好人工撰写故事的语法。然而，在计算每位教师的平均评分时，我们发现三位教师中有两位确实对人工撰写故事的语法正确性给出了更高评分。值得注意的是，对于大语言模型评估而言，并不存在在大语言模型评估过程中是否应混合人工撰写的故事和 GPT-2 生成的故事这一问题，因为正如第 5节所述，每个故事的评分彼此独立。

对于对抗攻击质量评估，我们还通过 Upwork 招募了持有认证资格的教师。这些教师需要对 100 个新闻标题进行评分，并因此获得 35 美元的报酬。他们表示，完成评分所花费的时间不到 1 小时。

C.2 人工评估界面

开放式故事生成 我们使用 Google 表单收集教师的回答。

撰写者	人类	GPT-2
语法正确性	$3.89_{0.97}$	$3.88_{0.84}$
连贯性	$4.35_{0.87}$	$3.49_{0.97}$
可喜爱性	$3.46_{1.40}$	$2.89_{1.12}$
相关性	$3.71_{1.20}$	$2.37_{1.33}$

表5：在人类评估中，将200个模型生成的故事和200个人类撰写的故事随机混合时，人类撰写的故事和GPT-2生成的故事的平均李克特评分。

每份表单包含 100 个故事，每个故事位于 Google 表单的一页中。该页面上的界面如图 2 和图 3所示；这两幅图来自 Google 表单的同一页面，我们将其拆分开来，因为截取整个界面会导致分辨率较低。

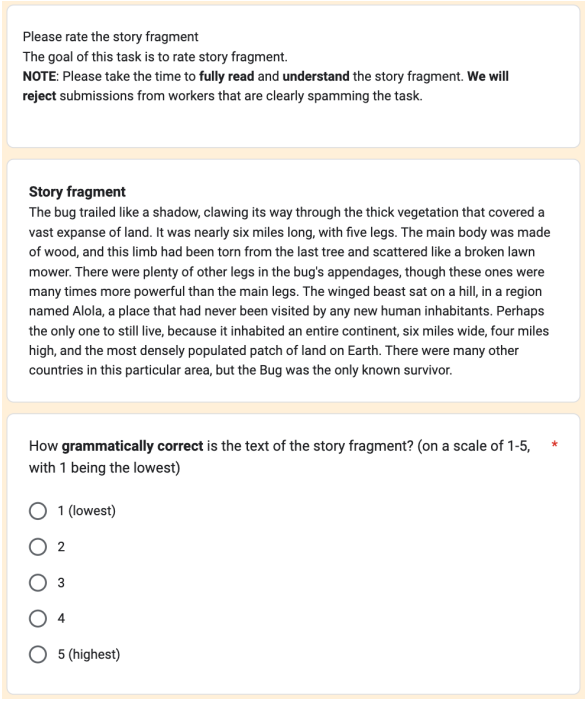

图2：开放式故事生成界面的上半部分。

对抗攻击质量评估 在这项任务中，我们同样使用 Google 表单收集教师的回答。我们创建了两个不同的 Google 表单，其中一个用于评估流畅性，其界面如图4所示。在此表单中，我们混合了数量相同的正常新闻标题、经 TextFooler、PWWS 和 BAE 攻击的新闻标题。Google 表单的每一页包含一个新闻标题。

另一个 Google 表单用于比较新闻标题在攻击前后意义保留的情况

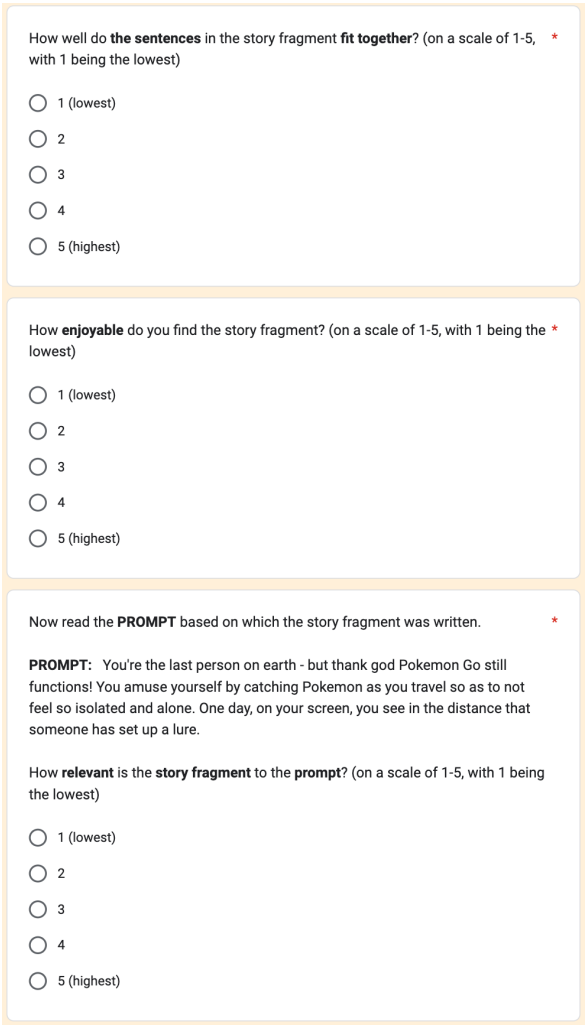

图 3：开放式故事生成界面的下半部分。

。我们使用**粗体**突出显示良性样本与对抗样本之间的差异，如图5所示。Google 表单的每一页上都有一对新闻标题。

C.3 与英语教师的任务后访谈

C.3.1 英语教师如何为故事评分

教师为 400 个故事评分后，我们向他们提出以下问题：

Q1 为这 400 个故事评分花了您多长时间？

Q2 在评估的四项属性（语法、连贯性、可喜爱性、相关性）上，您的评判标准是什么？例如，什么情况下您会给语法性打高分或低分？您给出的故事是什么样的？

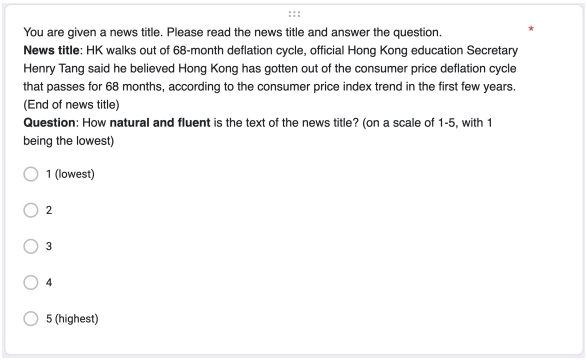

图 4：用于评估良性样本或对抗样本流畅性的 Google 表单。

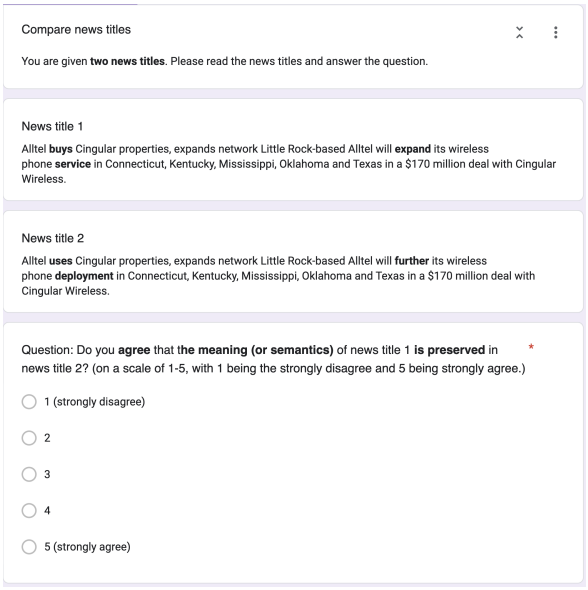

图 5：用于评估良性样本与对抗性样本之间含义保持程度的 Google 表单。

可喜爱性评分较低吗？你的个人偏好是否影响了评分？

Q3 您花了多长时间才校准好自己对该任务的评分？

Q4 阅读故事所基于的提示词后，您是否更改了对前三项属性的评分？

我们简要总结三位教师的回答。教师表示，他们花了 6 至 10 小时为 400 篇故事评分。在语法方面，大多数教师会检查标点[9]，用词和主谓一致。英语教师会降低

[9]其中一位教师明确告诉我们，标点符号不属于语法。通过搜索互联网，我们发现，关于标点符号是否应算作语法的一部分，存在争议。

作者	人类	GPT-2
语法正确性	0.25	0.15
连贯性	0.26	0.18
可喜爱性	0.09	0.12
相关性	0.38	0.41

表6：两位英语教师的肯德尔 τ 相关系数。三位英语教师参与评分，因此表中的结果是 $\binom{3}{2}$个肯德尔 τ 的平均值。

评分，具体取决于故事中语法错误的类型和数量。

在连贯性方面，教师会根据故事中的句子是否按照逻辑顺序展开叙事来进行评分。教师会问自己一些问题，例如“这个故事是否合乎逻辑”。这是一种对整个故事进行的更为全面的整体性评价。

对于可喜爱性，一些教师表示，他们会尽量避免受到个人偏好的影响。一位教师会问自己：考虑到故事的合理程度，以及它是否具有文体特色、幽默感，或引人入胜的情节安排或人物塑造，我个人是否喜欢这个故事？总体而言，教师们都努力以公平、客观的视角来评价可喜爱性。对于相关性，教师们只需检查故事是否基于提示词创作。

教师们表示，他们大约需要评阅五到十篇故事才能校准自己的评分。除了一位教师在看到**仅一篇故事**的提示后，修改了其他三个属性的评分外，教师们在阅读提示后并未改变其余三个属性的评分。

C.3.2 教师对 ChatGPT 评分的评论

教师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 的判断。我们告诉教师，该评分由我们聘请的另一位教师完成；为避免他们对人工智能模型产生潜在的刻板印象，我们没有告知教师评分实际上由人工智能模型完成。我们随机选取了四个由 ChatGPT 评分的故事；其中两个是评分较高和较低的人工撰写的故事，另外两个是评分较高和较低的 GPT-2 生成的故事。

故事 1 自从他们发现我死了，已经过去了好几个月。我不知道自己为什么停下了脚步。没——有其他人在身边可以取代我，或是和我说话。我僵在震惊中，一天天过去，始终无法说话、无法移动，也无法表达自己。这是一件奇怪的事，没人会想到我竟然渴望这样的事。我越是专注于此，它就越是吞噬我。几年过去了，我也越来越深陷其中。我开始想，或许他们其实并不在乎我。我仍然深深地渴望着他们。如今我所能做的，只有一步一步地走下去。我内心深处有个安静的声音，喜欢相信他们已经从我身边走过。孤独。这从未让我觉得自己失去了什么。那些拯救过我的人已经离去。那么，我为什么还在继续走呢?

提示词 “每个人都以独特的方式看待世界。当有人去世时，在某种意义上，整个世界也随之死亡。”

评分 表7。

故事 2 “我的儿子不在了。他已经……”当你解释并说“是的。但是。你不能就这样做”时，人们盯着你。“他是我的儿子！这是我的儿子！我爱他，他会没事的。”“你说他没事是什么意思？他说他对某个人很生气，却找不到自己的怒火，你怎么能这样说？这根本说不通啊？！警察打来了电话，医生说那是一种成瘾。”“你不能相信我！他告诉过我他爱我。”“不。不，他不爱你。但我可以相信你。对他的指控证据确凿，现在他什么也做不了。针对他的证据都只是间接证据。”“我曾经认识的那个男人将无法与我结婚。”

提示词 每个人一生中都可以杀一个人。不过，你必须填写一份表格，说明这个人为何罪该处死，委员会则会根据规则批准或驳回你的申请。今天，你要提交这份表格。

评分 表8。

故事3 我把那个小黑盒放在掌心。它漆黑一片，完美的立方体形状，平平无奇，却仍牢牢吸引着我的目光，仿佛它蕴藏着某种重大意义。我的朋友 Heaven 不知从哪儿冒了出来，在我身旁坐下，也盯着那个立方体。“你手里拿的是什么，Heaven？”他问。我继续凝视着盒子，甚至不愿哪怕片刻移开目光去看 Heaven，仿佛那样就会错过

什么似的。“这个，”我说，“就是宇宙的秘密。”我看得出来，Heaven 对这样的知识感到不安，仿佛那个立方体邪恶至极。又或许，他不相信这样的物品能够存在。但我让它存在了，于是它就这样躺在我的掌心。

提示词 《小黑盒》

评分 表格9。

故事 4 我透过 L96 狙击步枪的瞄准镜向下凝视。我缓缓将视线移向酒店的每一扇窗户，里面上演着各种堕落行径。一对情侣正激烈地取悦彼此。另一扇窗后，一位老人看着一名年轻女子在他面前脱衣。毫无疑问，她是个妓女。我慢慢吸气，再缓缓呼出。空气清冽，气温至少有 30 华氏度。我看见细小的雪花在眼前宁静地飘落。我找到了要找的房间。我跟踪这个男人已经两周了。他可真是难以捉摸。灯光闪烁着亮起。那个红发男人颇有几分魅力，我能理解女人为何会被他吸引。我用瞄准镜跟着他走进隔壁房间。灯光闪烁着亮起，眼前的场景让我大吃一惊。一个男人，体重超标，头发也快掉光了。

提示词 你是这个故事的反派。然而，在主人公做了那件事之后，你不确定自己还能不能称自己为反派。

评分表格10。

教师对 Chat- GPT 评分的总体评价教师详细阐述了他们对 ChatGPT 评分的看法后，我们请他们就 Chat- GPT 的表现作出总体评价。同样，教师并未被告知这些评分由人工智能模型给出。总的来说，教师们都认为评分和解释是合理的。他们认为自己不认同的属性主要是可喜爱性和连贯性。不过，他们认为这两项属性是对故事更全面的评价，因此主观性也更强。即使他们给出的评分不同，仍然能够理解 ChatGPT 的解释。最后，所有教师都总结说，给故事评分具有很强的主观性，出现意见分歧是正常的。

D 大语言模型评估

D.1 使用的大语言模型详情

我们使用的 T0 模型名为 T0pp，是 T0 模型的一个变体，拥有 13B 个参数。下文仍将这个模型称为 T0。我们使用`transformers` 工具包（Wolf 等人，2020年）加载 T0 模型。两个 InstructGPT 模型`text-curie-001`和`text-davinci-003`通过 OpenAIAPI进行查询。我们通过 OpenAI 图形用户界面查询 ChatGPT。我们知道一些线上资源提供了类似 API 的工具来查询 ChatGPT，但我们认为这违背了 ChatGPT 的预期用途，因此不采用这些线上资源。我们查询的 ChatGPT 版本为 12 月 15 日版和 1 月 9 日版。

OpenAI 并未公开任何 GPT 模型的参数规模。不过，据估计，`text-curie-001` 拥有 13B 个参数，而`text-davinci-003` 拥有 175B 个参数（Gao，2021）。

D.2 查询大语言模型的详细信息

D.2.1 解析大语言模型输出

T0 和 InstructGPT 生成答案后（几乎所有情况下，答案都由多个句子组成），我们会解析生成的句子，以获取模型的评分。我们使用一些规则来解析输出；这些规则是在人工检查生成序列的输出后制定的。首先，我们从输出中删除字符串 1-5，因为我们观察到大语言模型有时会说“.....评分范围为 *1-5*.....”；我们还会删除字符串 out of 5，因为大语言模型有时会说“给出 x （满分 *5* 分）”。我们也会删除字符串 /5，因为大语言模型有时会用“$x/5$”表示其给出的评分为 x （满分 5 分）。对于对抗样本的评分，我们会从输出句子中删除标题 1 和标题 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得大语言模型的评分。在少数罕见情况下，大语言模型给出的评分会以 0.5 为增量，也就是说，它会给出类似 4.5 的分数。在这种情况下，我们不会去掉 0.5，因为我们的指令中没有明确告知评分应以 1 为增量。我们发现，去掉或保留 0.5 对大语言模型的结果几乎没有影响；论文中的结果没有去掉 0.5。再次强调，这些情况确实非常罕见。

属性	评分者	回答
语法	ChatGPT	我会评价故事片段文本的语法正确性。评为***4***。有一些小错误和措辞不当之处，但总体而言文本结构良好，易于理解。
	T1	**同意**
	T2	**同意**
	T3	**同意**
连贯性	ChatGPT	我会将故事片段中的句子与故事的契合度评为***5***。 *The* 句子衔接流畅，共同构建清晰且连贯的叙事，并为主角的成长与发展提供一种循序渐进的感受。想法和情绪。
	T1	**不同意**： 我觉得对我来说是***4*** 分，而不是 *5* 分，因为有一句话那似乎与其他内容格格不入。
	T2	**同意**
	T3	**不同意**：我觉得这段文字前后脱节，令读者感到困惑。开头的句子讲的是第一人称叙述者被发现已经死亡，但接下来的句子却讲述叙述者的内心想法。另外，这段文字写道：“孤独。从未觉得它如此一种失去。”那么，感觉像是失去的“它”是什么？是孤独吗？如果是，那怎么会让人感觉像是失去？如果不是，那这个“它”又是什么？因此，它与前一句的并置令人困惑。
讨人喜欢	ChatGPT	就欣赏程度而言，我会给这个故事片段打***2*** 分。 *The* 该片段以一种阴郁且内省的语调写成，内容涉及孤立与孤独等主题，这对一些读者来说可能并不令人愉悦。主人公的心理状态也被描绘为处于一种震惊与困惑状态，这加重了整体的不安感以及缺乏愉悦感。
	T1	**不同意**：我会给它评***3***分，而不是*2*分，因为愉悦感这一点是主观的-性。
	T2	**不同意**： 虽然自由职业者的评分认为整体主题尽管有些沉郁，却仍透出一线希望。正是这一点让它更令我享受。弱者逆袭故事。死亡不是写作中应该回避的话题。我们都要面对死亡，而我很喜欢这篇作品中的坦诚：这篇作品中的坦诚："我为什么一直走下去？"。我觉得我们很多人都会面临这样的疑问并且可能曾问过自己。我认为这与.....有关。许多读者。
	T3	**同意**
相关的	ChatGPT	我会将这个故事片段评为 ***5***。它与提示词高度相关因为它探讨了死亡这一主题，以及它如何被视为一种.....的死亡整个世界。故事片段还涉及孤独这一概念以及渴望，这些都与某人或某物的失去有关。该主人公因震惊而僵住，无法表达意愿的概念它们本身也支持这样一种观点：死亡可以被视为死亡整个世界的。
	T1	**同意**
	T2	**同意**
	T3	**不同意**： 我看不出提示词与故事之间的联系。

表 7：故事 1：一则由 GPT-2 生成、并获得 ChatGPT 高度评价的故事。

属性	评分者	回答
语法	ChatGPT	我会将故事片段的语法正确性评为***3***。 该文本总体上易于理解，但在标点符号方面存在一些问题。 存在错误和措辞不当之处，导致内容在一定程度上难以理解。
	T1	**不同意**：我会将语法评为***2***分，而不是***3***分——其中存在明显的语法问题。 相比这一组故事的平均水平，语法错误明显更多
	T2	片段。 **不同意**： 语法和连贯性方面的问题会造成中断 开始阅读时的感受。我不确定文本中是谁在说话 这让我产生了许多疑问。
	T3	**不同意**： 语法比 ***3*** 更差，例如这一点显而易见。 本应是第二句话：“他一直.....（未完成， 还有另一个片段：“人们凝视.....）；*cant*；*couldnt*；*doesnt*； 以及像这样毫无意义的句子：“我曾经是一个男人” 不会愿意嫁给我”
连贯性	ChatGPT	我会将故事片段中句子的连贯性评为 ***a 2.*** 片段中的句子彼此割裂，内容在不同主题间跳跃。 不同的想法和观点，使得理解整体内容变得困难。 故事或消息。 事件发展或逻辑脉络并不清晰 句子之间缺乏连贯性。看起来这段文字可能缺少 一些有助于串联这些句子的重要信息或背景 。
	T1	**同意**
	T2	**同意**
	T3	**不同意**： 一段非常不连贯的文字，让人不太清楚是谁在说话 什么，以及主题是什么，因此叙事支离破碎。
讨人喜欢	ChatGPT	我会给这个故事片段的趣味性打***1***分。这个片段 令人困惑且难以理解，句子彼此脱节 并且在不同的想法和观点之间跳跃。这个故事似乎 并不完整，也缺乏背景信息，因此很难理解情节或 人物。此外，这段文字基调悲伤而消极， 读起来似乎也并不令人愉快。
	T1	**同意**
	T2	**同意**
	T3	**同意**
相关	ChatGPT	我会将故事片段与提示词的相关性评为***1***。 虽然故事片段确实提到了有人离开，但它并没有 似乎与填写表格以证明杀人正当性的想法直接相关 某人。尚不清楚儿子是否就是被提议杀害的人， 也不清楚这个故事是否发生在提示词所设定的情境中。 没有明确迹象表明这个片段与提示词有关。
	T1	**同意**
	T2	**同意**
	T3	**同意**

表 8：故事 2：一个由 GPT-2 生成、ChatGPT 评分较低的故事。

属性	评分者	回答
语法	ChatGPT	故事片段的文本在语法上似乎正确，且没有重大错误或措辞不当。按 *1* 到 *5* 分评分，*1* 分代表最低分，我会给这段文本的语法正确性打***5***分。
	T1	**同意**
	T2	**同意**
	T3	**不同意**：例如，因为*“no where” =*，我会给它打***4*** 分*“nowhere”*虽然看起来只是个小错误，但确实会影响文本在一定程度上的正确性。
连贯性	ChatGPT	故事片段中的句子衔接得很好，构成了一个清晰且*.....*对场景以及人物的行动和想法形成连贯的图像。语言描写生动，富有感染力，对话也十分有效。传达了人物的情绪和动机。按 *1–5* 分评分，其中***1***为最低分，我会评价这些句子的连贯程度作为***4***。
	T1	**同意**
	T2	**同意**
	T3	**部分同意**：在我看来，所有句子都衔接得很好，共同构成了一个引人入胜的叙事。我会给出***4*** 或 ***5***。
讨人喜欢的	ChatGPT	基于故事片段的结构、写作风格和情节，我会在愉悦度方面将其评为***3*** 。该片段引入了一个神秘的小黑盒和主人公相信它藏着*“*宇宙的秘密*”*，这一情节引人入胜，但故事戛然而止，留下许多未解答的问题，这可能会让读者觉得不够满意。读者。
	T1	**不同意**：我喜欢这个故事，而且它非常清晰，所以我会给它打***4*** 或 ***5***分。
	T2	**不同意**：虽然故事突然结束了，但这让我想起它只是一个片段，而且这篇作品可能还有更多内容。我觉得这篇作品的文字讲述一个小黑盒子的内容很有吸引力，也很享受开头对北欧神话的介绍。这会引起许多读者的兴趣。
	T3	**不同意**：我喜欢这个片段，因为它激起了我的兴趣，让我想要进一步了解故事如何发展，以及它究竟讲述了什么。
相关	ChatGPT	我会给这个故事片段的相关性打***5*** 分，因为它与提示词*“The* 小黑盒*”*相关。故事片段围绕一个小黑盒展开，并着重描写了它被认为的重要性，这与提示词相符。此外，主角 *Heaven* 正拿着这个盒子，并称它是*“*宇宙的秘密*”*，这进一步强化了它与提示词之间的联系。
	T1	**同意**
	T2	**同意**
	T3	**同意**：这个故事与提示词完全相关。

表9：故事3：一则由人类撰写、并获 ChatGPT 高度评价的故事。

属性	评分者	回答
语法	ChatGPT	我会将这段文本的语法正确性评为***3***。共有 一些错误，例如“各种恶习的诸多表现”，而这本应是..... 是“各种恶习的诸多表现”和“我如此细小的雪花”。 这句话应该是“我看到了小片的雪花。”此外，一些的 这些句子略显生硬，可以进一步改写，使其表达更加流畅。 清晰度。总体而言，文本易于理解，但仍存在一些错误。 以及可以改进的措辞不当。
	T1	**部分同意**：我给语法的评分比这个低一些——也许是***2*** 分或 ***3*** 分。 错误数量高于平均水平。
	T2	**不同意**：虽然文本可以理解，但我发现自己会因 文本中的错误和连贯性而分心。
	T3	**同意**
连贯性	ChatGPT	我会将故事片段中句子的连贯性评为 ***a 2***。故事片段在不同场景之间突然跳转，并 人物之间缺乏清晰的过渡，使得内容难以理解和跟进。 理解叙事。此外，句子通常彼此脱节 且缺乏清晰的脉络。故事片段也缺乏清晰的情节或 目的，因此难以理解整个故事或主角的 意图。可以通过提供更多信息来改进这个故事片段， 更多背景信息, 以及更清晰的叙事线索。
	T1	**同意**： 情节跳跃
	T2	**同意**
	T3	**不同意**： 这个故事讲述一名狙击手透过步枪瞄准镜观察， 因此，随着他透过窗户观察不同房间，他看到的场景也随之变化， 随后这些场景又被逐一描述。这些描述 解释了看似缺乏连贯性的原因。
讨人喜欢	ChatGPT	我会将该故事片段的趣味性评为***1***。该片段 缺乏清晰的情节、目的或人物发展，因此难以 对故事或人物产生投入。此外，该片段 包含令人不安的图像化内容，例如暴力和卖淫， 这可能会让一些读者感到不适。此外， 片段中包含一些语法错误和措辞不当。 这会让人难以理解故事。总的来说，这个故事片段 读起来并不令人愉快。
	T1	**同意**： ：它不只是图像化的，还展现了非常有害且 基于人物的体重、性别等因素对人物的负面呈现。
	T2	**同意**
	T3	**不同意**： 这个故事在一定程度上确实吸引了我的注意，因为我想 知道他想以谁为目标，以及原因。
相关	ChatGPT	我会将故事片段与提示词的相关性评为***1***。这 个故事片段不包含任何与提示词相关的元素， 例如主人公和反派、他们的行为，或道德模糊性 反派所面对的。故事片段主要描述了一个人 拿着狙击步枪窥视酒店窗户，却没有呈现任何 明确的情节或目的。此外，故事片段也没有提供任何 迹象表明拿着狙击步枪的人存在道德模糊性。 总的来说，这个故事片段与提示词无关。
	T1	**同意**
	T2	**同意**
	T3	**同意**

表10：故事4：一个被 ChatGPT 低评价的人类撰写的故事。

我们不会使用任何规则解析 ChatGPT 的输出，因为我们这些作者会阅读回答，并自行提取分数。

在实验过程中，ChatGPT 拒绝回答有关 2 个故事和 3 个新闻标题的问题，因为 ChatGPT 认为这些内容违反了OpenAI 内容政策。我们发现，这些样本包含对某些受保护群体的歧视，或包含性或暴力描述。因此，ChatGPT 的结果是在不包含这些样本的情况下计算的。

D.2.2 开放式故事生成

对于 T0 和两个 InstructGPT 模型，我们使用如下所示的查询，分别询问四个属性：

语法正确性
请评价故事片段 此任务的目标是评价故事片段。注意：请花时间完整阅读并理解故事片段。对于明显敷衍完成任务的工作者，我们将拒绝其提交。故事片段：[故事]（故事片段结束）该故事片段文本的语法正确程度如何？（按 1—5 分评分，其中 1 分最低）

连贯性
请为故事片段评分。本任务旨在为故事片段评分。注意：请花时间仔细阅读并理解故事片段。对于明显在敷衍完成任务的工作者，我们将拒绝其提交。故事片段：[STORY]（故事片段结束）故事片段中的句子衔接得如何？（1 至 5 分，1 分最低）

可喜爱性
请为故事片段评分。本任务旨在为故事片段评分。

注意：请花时间完整阅读并理解这段故事。我们。将拒绝来自明显在任务中发送垃圾内容的工作者的提交。故事片段：[STORY]（故事片段结束）你觉得这个故事片段有多么有趣？（采用 1–5 分制，1 分表示最低程度）

相关性
请为故事片段评分。本任务旨在为故事片段评分。注意：请花时间仔细阅读并理解故事片段。对于明显在敷衍完成任务的工作者，我们将拒绝其提交。故事片段：[STORY]（故事片段结束）现在请阅读撰写该故事片段所依据的提示词。提示词： [PROMPT]（提示词结束）故事片段与提示词的相关程度如何？（1 至 5 分，1 分最低）

[故事] 和 [提示词] 应分别填入故事和提示词。为提高可读性，我们在文本中显示换行符。查询模型时，我们使用标记 \n 表示换行。

查询 ChatGPT 时，我们在同一段对话中询问同一个故事的四项属性；这类似于让教师在 Google 表单的同一页面上评价同一个故事。我们使用上文所示的相同查询来询问 ChatGPT，查询顺序也与上文所示的顺序一致。

D.2.3 对抗攻击质量评估

在此任务中查询所有大语言模型时，我们会分别查询同一新闻标题的流畅性和意义保留。这意味着，与 ChatGPT 的每次对话中都只有一个问题，询问新闻标题的流畅性或意义保留。所有生成参数均与第 3.2 节中的默认参数相同。

我们使用的具体查询如下：

流畅性
给你一则新闻标题。请阅读

这则新闻标题并回答问题。新闻标题：[新闻_标题]（新闻标题结束）问题：这则新闻标题的文本有多自然流畅？（按 1 到 5 分评分，1 分最低

[NEWS_TITLE] 将填入一个良性新闻标题或遭受对抗攻击的新闻标题。

意义保留 你将看到两个新闻标题。请阅读这两个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 的含义（或语义）在新闻标题 2 中得到了保留？（请按 1 到 5 的量表评分，1 表示强烈不同意，5 表示强烈同意。）

[BENIGN_TITLE] 将填入攻击前的新闻标题，而[ADVERSARIAL_TITLE]将填入遭受对抗攻击后的新闻标题。

E 对抗攻击的实验详情

第4节所用的对抗样本来自 Yoo 等人（2022）。Yoo 等人（2022）针对不同的受害模型，采用不同的对抗攻击生成了不同的对抗样本集。我们使用了针对在 AG-News 上训练的 BERT-base-uncased 文本分类器生成的对抗样本，采用了三种不同的对抗攻击：Textfooler、PWWS 和 BAE。该数据集旨在促进 SSA 研究，而我们并未违反这一意图。

这里展示了我们使用 `text-davinci-003` 作为大语言模型评估工具，在表 11 中评估对抗样本质量的补充结果。我们可以看到，使用 `text-davinci-003` 得到的结果与 ChatGPT 类似：`text-davinci-003` 对抗样本的评分也高于人类，但仍显著低于良性样本。正如第 3.3 节所示，`text-davinci-003` 倾向于给出更高的评分。

	人工评估		大语言模型评估	
	流畅度	**均值。**	**流畅度**	**意义保留。**
良性	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：大语言模型评估（text-davinci-003）和人工评估在良性样本和对抗样本的流畅性（**流畅度**）以及对抗攻击前后新闻标题之间的意义保留（**意义保留**。）方面的结果。

评分者	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12：三位教师 T1、T2 和 T3 对三种对抗攻击的评分。

如第 4.3 节所述，一位教师给 Textfooler 的流畅性 评分显著高于 PWWS，而另外两位教师则没有这样做。表*12* 展示了每位教师对三种对抗攻击的流畅性评分。