JaccDiv：一种用于量化音乐产业中生成的营销文本多样性的度量和基准

Anum Afzal*[0000−0001−8146−1949], Alexandre Mercier[0009−0009−6483−2758], and Florian Matthes[0000−0002−6667−5452]

Technical University of Munich, BoltzmannstraSSe 3, 85748 Garching bei München, Germany
{ anum.afzal } @tum.de

Abstract. 在线平台对使用数据到文本技术来生成内容并帮助用户的兴趣日益增加。不幸的是，传统的生成方法往往陷入重复模式，仅经过几个迭代后就会形成单调的文本集。在本文中，我们研究了基于 LLM 的数据到文本方法，旨在自动生成质量足够高且足够多样化的营销文本，以便广泛采用。我们利用了语言模型，如 T5、GPT-3.5、GPT-4 和 LLaMa2，并结合微调、少样本和零样本的方法，为多样化的营销文本设定了基线。我们还引入了一个度量标准 JaccDiv，用于评估一组文本的多样性。这项研究的相关性超越了音乐行业，在重复的自动化内容生成普遍存在的各个领域中，证明都是有益的。

Keywords: Large Language Models · Text Generation· Marketing · Diversity Evaluation.

在数字化转型重塑各行业和人工智能似乎被引入生活每个方面的时代，在线平台正在调整其服务以确保更流畅的用户体验。由于大多数平台的吸引力依赖于用户生成的内容，因此需要向用户提供创建相关内容的工具。可以利用由 LLM 驱动的技术生成足够质量的文本以供大众采纳。然而，控制生成内容的质量、丰富性，最重要的是多样性是重要的。在平台上的营销背景下，多样性至关重要，因为竞争产品和服务通常并列列出，而这些产品或服务的描述通常是竞争对手之间最显著的区别因素。在这种情况下，类似文本的重复变得显而易见，并可能导致对平台的信任度下降。 显示，对于幽默这个复杂的领域，ChatGPT 重复了 25 个相同笑话的变体，这个数量足够高，可以让用户认为每个笑话都是独特的，但在为可能有数百或数千竞争对手的平台生成营销描述时，可能还不够多样化。

我们与一个旨在将音乐家与活动组织者联系起来的在线平台合作。该平台允许乐队创建其个人资料，个人资料通常包括自我描述。由于撰写关于自己的内容是一个耗时的过程，许多乐队会推迟这一步骤或仅撰写简短的文字。为了使乐队的个人资料更具吸引力，可以通过一个由大型语言模型驱动的虚拟助手来帮助，它根据乐队在注册时提供的数据生成描述。迄今为止，大多数研究集中在减少重复或控制单个生成样本内的文本多样性。相比之下，我们的目标是避免相似结构并增加样本之间的文本多样性。通过比较和对比各种方法，我们确定了在音乐行业背景下实现多样化和吸引人的营销内容的最有效技术。我们的主要贡献如下：

正如在任何工业用例中一样，数据集是有噪声的、有限的和稀疏的，因此许多样本不包含完整的频带信息。我们展示了如何在利用这样的数据集的同时生成高质量的描述。由于许多频带的输入可能相似，它们的输出预计会包含重复。

www.xueshuxiangzi.com

我们尝试了各种技术来增加生成文本的多样性。虽然有很多指标可以评估生成文本的质量，但评估多样性的指标却很有限。我们提出了一种新的无参考指标，可以用来测量文本语料库之间的多样性。

1 相关工作

多年来，已经提出了许多不同的方法用于数据到文本生成。 [10] 提出了零样本方法，通过使用多个预训练模型，使得首先将输入三元组形成事实，然后使用 BART-base 的一种变体进行排序。接着使用 RoBERTa 来决定哪些事实可以聚合，最后一步再次使用 BART-base 来聚合和压缩段落。然而，最先进的方法倾向于使用端到端的解决方案，如 GPT 模型或 T5，无论是在零样本模式 [3] 或与其他步骤结合使用，如消歧 [24] 、推理 [20] 或链式思维提示 [23,13] ，甚至是思维树推理 [25] 。尽管这些方法尝试实现多样化文本，其研究重点集中在文本质量和模型的可控性 [26] 上。一些方法例如 [7] 尝试通过添加相关词汇来解决这个问题，模型随后应包含这些词汇。一些提示修改方法由 [18,29] 讨论。对于问题生成的特定情况， [5] 使用多个选择器对输入数据应用掩码，从而使模型可以专注于输入的不同部分，达到生成多样化问题的目的。在产品描述的背景下，像阿里巴巴的鲁班 [1] 这样的工具使用 AI 来生成图形横幅，而国际大型商业平台的常见做法是使用自动翻译进行描述。由于我们的研究将为平台用户生成描述，市场营销中最接近的话题是为电子商务中的产品生成描述 [27,4] 。多年来，已经尝试了不同的方法来评估多样性 [14,21] ，利用各种嵌入技术来评估内容的多样性。最后， [17] 尝试将质量和多样性测量合并为一个分数。

2 数据集

我们使用了一个从平台获得的工业数据集，该平台旨在连接音乐家和活动组织者。该数据集包含分成编队（乐队）的个别音乐家的个人和公开信息，将与这些编队互动并雇用它们的活动组织者的信息，以及关于活动的信息。我们处理了这个原始数据集，以构建一个编队数据集，其中包含名称、类型、描述、基地、半径（他们愿意旅行的距离）、流派和活动类型。

2.1 数据预处理 & 过滤

由于数据集 [2] 是德语的，描述被使用 DeepL API 翻译成了英语。如 Figure 1 所示，形成描述中每个描述的词数量分布。最短的文本只有 214 个字符长，最长的超过 3000 个字符。平均长度约为 863 个字符，标准差为 512 个字符。由于一些模型的注意窗口有限，为了比较，我们将在下面描述的评估集中尽量选择长度相似的文本。

对所有剩余的描述 [3] 进行了质量和独特性的人工评级。我们在生成新的描述时，将这些好的描述用作参考或模板。这个子集用于微调模型，也用于小样本生成。

[1] https://www.alibabacloud.com/luban

[2] 数据集经过筛选，移除了所有没有描述的地层，剩下 880 个样本中的 440 个。

[3] 数据集经过进一步的人工筛选，以去除不可用的 & 重复描述，剩下 359 个样本。

www.xueshuxiangzi.com

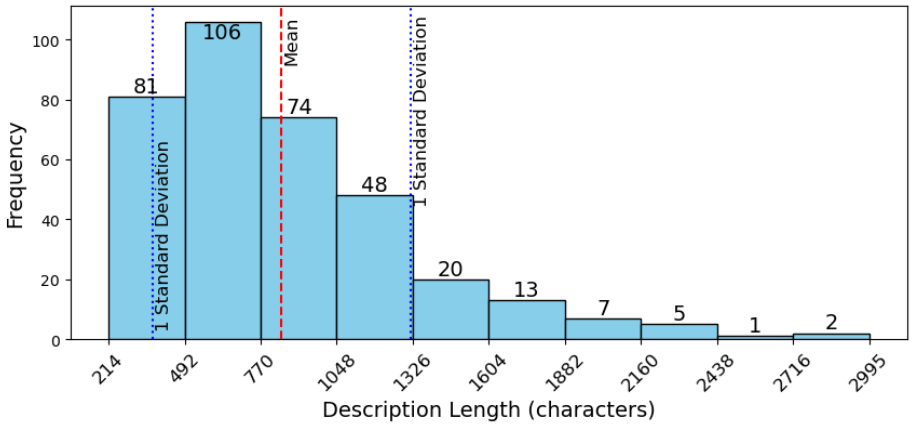

Fig. 1: 描述长度的分布。

3 基准测试

3.1 模型

在我们的实验中，我们从质量和多样性方面评估了五个（大型）语言模型。我们通过预过滤调优使用输入三元组微调了 T5-small [19] 和 Flan-T5-Base（250M 参数） [6] 。我们使用微调的模型作为我们的基准。我们进一步实验了 GPT-3.5、GPT-4 和 LLama2-13b [22] ，运用下面讨论的各种增强多样性的技术。我们通过提供的 API 使用 GPT-3.5 [4] 和 GPT-4 [5] ，并通过 llama.cpp [6] 运行 LLama2-13b ，这是一个在 CPU 上高效运行大型模型的框架。

3.2 评估

为了比较和确认模型与实验之间的多样性，我们选择了一个由 50 个谱带组成的数据集。它们将被用作不同实验和模型的基准，以评估其质量，最重要的是，多样性。

生成文本的质量： 我们数据集中参考描述是由人类撰写的，他们对自己乐队的信息比我们掌握得更多。通常，他们不会在文本中包含所有的数据库信息，因为这些数据的部分已经在他们的个人资料的其他地方展示。这使得基于参考的评估（例如 ROUGE）相当具有挑战性，因此我们在评估中没有包括任何基于参考的指标。相反，我们利用基于 LLM 的指标 [11,12] 来获得与人类相关性更好的分数 [2] 。我们选择了 G-Eval [15] ，这是一种利用 GPT-4 的链式思维驱动的 [23] 基于提示的指标。营销文本的目标是吸引观众的兴趣和注意力，以期引发预期的行为。因此，我们按照以下特征评估文本

[4] GPT-3.5-turbo-1106
[5] GPT-4-turbo-1106
[6] https://github.com/ggerganov/llama.cpp

www.xueshuxiangzi.com

参与性： 衡量一篇文本吸引读者的能力，并根据其写作方式判断文本可能有多有趣 [16] 。

流利度： 衡量文本的语法正确性。[8]

自然性： 用于确定该话语是否可能是由人类产生的。[28] 。

信息量： 用于确定该话语是否包含给定内容中的所有信息 [28] 。

文本生成的多样性： 与每个样本单独评估的质量评估方式相对，多样性要求多个样本相互比较。我们多样性分析的目标是基本上量化生成文本集中的重复内容。这使我们能够比较模型生成的文本组内部所达到的多样性。由于没有很多评估多样性的指标，我们采用了现有的 Jaccard 相似性系数，这是一种用于比较两篇文本的 n-gram 方法。Jaccard 系数测量样本集之间的相似性，这里是 n-gram。它通过将两个集合的交集大小除以两个集合的并集大小来计算。该系数是对称的，意味着文本的顺序无关紧要，计算速度快且结果为 0 到 1 之间的分数，较低的值表示更多样化的文本。通过获取集合中所有生成样本的成对 Jaccard 相似性，我们可以得到平均相似性分数。我们利用 Jaccard 相似性来计算多样性如下：

$$Diversity(U, V) = 1 - \frac{|U \cap V|}{|U \cup V|}$$

用于多样性分析的算法的可视化表现如 Figure 2 所示。我们从参数 n 生成每个文本的二元、三元...n 元组。然后的下一步通过计算 n 元组的成对 Jaccard 相似性来相互比较所有文本。由于相似性是对称的，结果值被存储在一个上三角矩阵中。从这里，我们可以平均所有值来获得一个近似的多样性分数。我们注意到，Jaccard 相似性度量没有正确考虑文本长度，但在开发过程中，实验中的所有文本长度普遍相似，因此文本长度对分数的影响可以忽略不计。

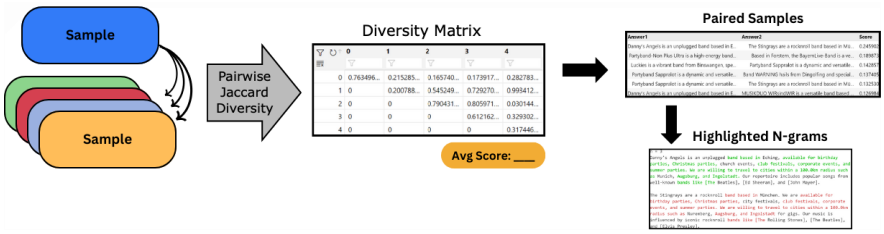

Fig. 2: 我们多样性指标的评估流程

人工评估： 由于人工标注是一项繁琐的任务，我们采用了在 Figure 2 中说明的半自动方法来协助人工标注者。我们使用自动多样性度量作为启发式选择了 5 个模型。两个标注者评估每个模型由 5 个描述所创建的批次，因此每个批次有 10 次比较。每个模型的描述都来自相同的乐队。乐队是事先随机选择的。要求

www.xueshuxiangzi.com

标注者根据 **??** 中显示的批次对每两个可能的对进行标注，并分配多样性分数。类似 3-gram 方法的高亮显示使得标注者能够更容易在句法多样性方面比较文本。

我们实验了以下超参数，以在尽可能保持生成描述整体质量的同时最大化多样性。按照 [9] 的方法，我们通过要求 (L)LM 使用数据点作为输入生成连贯文本来使用数据到文本生成技术。

我们对 GPT3.5 和 GPT4 的温度值在 0.0 到 2.0 之间进行了实验。由于温度参数为生成过程增加了一些随机性，我们预计它会显著影响多样性。

Top-p： 此参数将下一个预测集中在词汇表的最小子集上，其中累积概率高于“p”。Top-p 的优点在于它在随机性和连贯性之间取得了平衡。

Logit 偏差： Logit 偏置不是单一参数，而是一组可以正面或负面影响输出概率的符号（不是单词!）的列表。Logit 偏置是一种影响大型语言模型（LLM）输出的强大工具，因为它可以用来强制模型使用或停止使用某些词或短语。偏置的范围是-100 到 100。在每次生成后，会统计每个符号的出现次数，并且修改使用次数最多的前 100 个符号的偏置。

我们观察到输入数据的顺序对文本生成有显著影响。例如，如果音乐类型按流行、爵士、休闲指定，模型在描述每种类型时通常会使用相同的顺序。这导致了文本的重复，尤其是在生成的文本较短时。为了对抗这一点，我们在将数据输入模型之前对其进行了洗牌，但保留了数据的结构。

替代指令： 更改模型提示中的指令应该会对输出产生影响。因此，我们创建了一套备用指令，意图替换我们提示的开头和结尾部分。

情境学习： 我们使用来自精选数据集中的现有描述作为少样本例子。该方法依赖于通过 Deepl 翻译成英文的参考文本，而这可能会失去一些特定语言的复杂性和表达。其目标是观察在很大程度上免费的 LLM 是否会将现有的表达融入其输出中。

我们在 Table 1 中概述了我们的实验，并在下文讨论了我们的发现。

如 Table 1 所示，经过微调的模型，如 T5-small 和 FlanT5-base，并不能产生高质量的描述。我们怀疑这可能是由于训练数据的质量较低，原因包括翻译、描述较长以及仅有少量优质样本可用。使用少量样本的模型表现不佳进一步验证了我们的理论，即问题确实在于数据集，因为少样本学习几乎总是能带来更好的表现。

在最初使用与微调相同的输入三元组进行实验时，GPT-3.5 给出了足够好的结果。给定相同的提示，GPT3.5 和 GPT-4 的表现大致相同，但令人惊讶的是，尽管 LLaMa2 的规模较小，其整体质量却超出了预期。我们的研究结果大致符合文献和常见的 LLMs 排名 [1]，尽管我们发现 LLaMa2 在我们的用例中比 GPT3.5 更有能力。

另外，我们还查看了 G-Eval 提供的个人评分。有传言称，LLM 往往对 LLM 生成的文本存在偏见。我们在使用 GPT-4 进行的评估中观察到了类似的模式，如 Table 2 所示。无论是英语还是德语的人类撰写的描述，其得分始终比机器生成的文本差。相比之下，机器生成的文本在这四个特征上都获得了较高的评价，几乎在所有情况下都取得了高分。GPT-3.5 通过零样本文生成的描述根据 G-Eval 的评估表现良好，在语法质量、流畅性和吸引力方面，各模型之间仅有细微差异。

www.xueshuxiangzi.com

Table 1: G-eval 生成的整体质量评分在吸引力、流畅性、自然性、信息量（Qual）上进行平均，并缩放到 [0-1]。通过模型和实验生成文本的成对多样性评分 [0-1] 进行平均计算出的 & Jaccard 相似性基础上的多样性评分（JaccDiv）。

Model	Technique	Qual	JaccDiv
	Baseline		
German	Original	0.778	0.999
English	Original	0.787	0.999
	Fine-tuned Language Models		
T5	Triplet Input	0.315	0.984
Flan-T5	Triplet Input	0.465	0.994
	Instruct-tuned Large Language Models		
GPT-3.5	Triplet Input	0.829	0.985
	Base Prompt	0.845	0.964
	Shuffled Order	0.840	0.974
	Alternate Instructions	0.849	0.989
	Fewshot	0.657	0.993
	Fixed Logit Bias	0.804	0.996
	Adaptive Logit Bias	0.767	0.996
GPT-4	Base Prompt	0.881	0.995
	Shuffled Order	0.878	0.994
	Alternate Instructions	0.886	0.994
	Fewshot	0.320	0.999
	Fixed Logit Bias	0.801	0.998
	Adaptive Logit Bias	0.843	0.998
LLaMa-2 (13b-chat)	Base Prompt	0.868	0.983
	Shuffled Order	0.876	0.987
	Alternate Instructions	0.874	0.985

3.3 通过 Jaccard 相似性进行多样性分析

我们引入并使用一种基于 n-gram 的度量方法，利用 Jaccard 相似性来评估生成文本的多样性。总结在 Table 1 中，分数突出了模型之间在生成文本多样性方面的差异。我们在下面讨论其意义：

通过提示，我们要求模型增加输入数据，例如要求包括附近的城市或乐队可能演奏的乐队和歌曲。与由数据三元组组成的原始方法相比，这种方法惊人地减少了多样性。这可能是因为在给出相同提示的情况下，输出遵循其结构，因此多样性较低。我们在 **??** 中展示了最终提示。

为了增加多样性，我们尝试了诸如温度等参数，通过引入随机性来对样本的多样性产生适度影响。对于 GPT 模型，修改温度并不会影响质量，直到大约 1.6 的阈值，出现极端的随机性。由于降低其值没有提高质量而显著减少了多样性，因此我们在后续的所有实验中保持默认值。

www.xueshuxiangzi.com

Table 2: 所有模型的评分都是使用 50 个样本计算的，除了 LLaMa2，它是使用 20 个样本计算的。

Model	Fluency [1 - 3]	Naturalness [1 - 3]	Informativeness [1 - 4]	Engaginess [1 - 5]
Original DE	2.48	2.75	2.63	3.54
Original EN	2.42	2.74	2.68	3.78
GPT-3.5	2.74	2.98	2.96	3.65
GPT-4	2.94	2.98	3.00	4.00
LLaMa2	2.99	2.97	2.93	4.00

为了增强多样性，我们将关于前几代的信息转移到即将到来的一代。结果显著增加了输出的多样性。我们尝试了 Logit Bias，这是一种更直接控制输出的方法。Logit 定义了一个大型语言模型的最终层，展示了所有可能的单词及其概率。模型通常预测出概率最高的单词。通过 logit 偏差，我们可以影响这一最终层模型，告诉它降低先前生成单词的概率，或者根本不考虑它们。首先，我们尝试了固定 Logit 偏差，将前几次迭代中最常用的前 100 个标记的值设置为-50。Logit 偏差实验对多样性有强烈的影响。然而，在某些情况下，句子中间的单词会缺失。因此，强烈的 logit 偏差会影响生成的质量。

此外，我们使用了自适应对数偏差，其中偏差值的导出基于给定标记在前 100 个标记中之前使用的频率。如 Figure 3 中所示，我们观察到尽管偏差值快速达到最大值，但使用最多的标记并没有被消除。在前 100 个标记中，使用较少的标记成功地被限制。我们认为，Logit 偏差是一种提升多样性的好方法，但应谨慎使用并小心实施。

我们尝试打乱输入数据，以观察这是否会增强生成文本的多样性。该方法对 GPT3.5 和 LLaMa2 的多样性产生了显著影响。对于 GPT-4，多样性似乎反而略有降低，尽管它仍然显著低于其他模型。这可能是因为模型在插入一些文本时，已经在重新排序输入数据的部分。这些片段在输入被打乱时没有改变，仅仅是被重新排序。所有模型的共同点是，打乱输入的顺序并不影响质量。

利用替代指令对每个模型产生了不同的影响。尽管具有相同的意义，不同的指令表述应该会导致不同的关注点，从而生成更多样化的样本。在 LlaMa2 的情况下，改变提示中的第一个句子会导致输出的差异。当然，每组指令都有自己的开头，但至少可以影响模型的输出。对于 GPT-3.5 模型，多样性的影响非常明显，相较于只是随机排序，其效果有所增加。在 GPT-4 的情况下，再次出现了相反的情况，类似于随机排序实验。

小样本学习： 最后，我们尝试了少样本学习技术，通过传递一个简化提示的现有乐队描述。不幸的是，少样本学习的效果甚至比微调模型还要差。我们怀疑这可能与作为示例使用的描述的质量较低有关，因为它们是从德语翻译为英语的。

3.4 人类判断

人工评估显示，大型语言模型生成的文本局限于数据点，而许多人类撰写的文本包括个人轶事或经验，这些内容并未包括在结构化数据中。两种微调模型（T5 和 Flan-T5）都存在词汇循环的问题。需要注意的是，Flan-T5 能够在没有任何

www.xueshuxiangzi.com

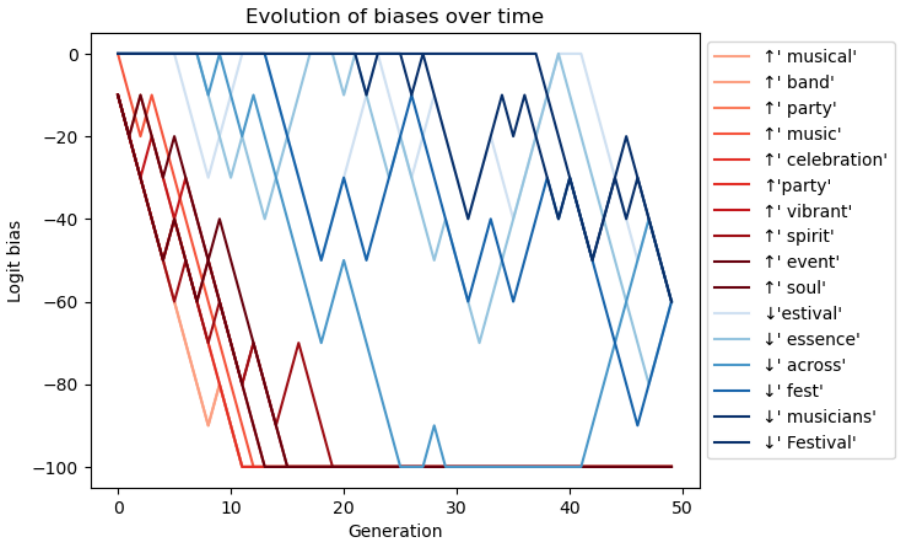

Fig. 3: 在使用自适应得分进行 50 次代生成过程中，一些标记的对数偏向值的演变。最常用的标记（红色）尽管偏向值迅速达到其最大值，但并未被消除。在前 100 个标记中使用较少的标记（蓝色）成功地被限制，正如偏向值的波动所示。

微调的情况下生成简短的乐队描述。由于输入三元组中的命名不明确且没有可依赖的上下文，它们不幸地不可用。在大多数情况下，模型会将乐队描述为音乐流派/活动类型等，而不是使用适当的动词。GPT-4 生成的文本较长但更空洞且较少吸引人，而 GPT-3 生成的文本较短并严格依赖输入数据。手动检查显示，GPT-3.5 通过打乱输入顺序和稍微改变指令显著受益。有趣的是，GPT-4 并未受到这些变化的任何影响，而 LlaMa2 的影响比 GPT-3.5 小得多。LlaMa2 经常使用表情符号，而 GPT-4 只有在明确要求时才添加。这表明模型对相同的变化并非同等敏感。给出相同的提示，GPT-4 和 LlaMa2 推断出更多细节，向潜在的活动策划者建议为什么这个乐队会是一个不错的选择以及他们可以期待什么。我们还在人工评估和自动评分中注意到，与其他大型语言模型相比，LlaMa2 在多样性方面的表现不佳。对于多样性评估，两名标注者使用 **??** 中描述的 2-gram 高亮技术手动标注了相同的 50 对。根据科恩的 kappa 系数（κ [-1 至 +1]），两位标注者之间的标注一致性为 0.214。每个模型在 5 个子集上进行了评估。由 JaccDiv 和人工标注者计算的平均多样性得分如 Table 3 所示。可以看出，人类标注者在评估多样性时比自动得分更严格。在所有模型中，带有 AdaptLogitBias 的 GPT4 在人工和 JaccDiv 得分中均得分最高。

4 结论

通过我们的研究，我们在 LLM 在营销行业的应用方面做出了重要贡献。我们实验了不同的模型、方法和输入格式，以查看每种方式如何影响生成描述的质量，更重要的是多样性。我们发现，由查询洗牌、温度参数和提示修改引入的随机性

www.xueshuxiangzi.com

Table 3: 由 JaccDiv 和人工标注者计算的平均多样性分数。

Model	Description	JaccDiv	Human
Original	References in German	0.993	0.990
GPT-3.5	Base Prompt	0.926	0.320
GPT-4	Base Prompt	0.982	0.705
GPT4	Alternate Instructions	0.986	0.690
GPT4	Adaptive Logit Bias	0.9903	0.885

对每个模型的影响不同，因此在提高多样性方面是不可靠的。为了获得一种更一致和可靠的多样性增强技术，我们发现将以前使用过的词元转移到新提示中是一种非常有效、可控和可靠的方法。这种技术的缺点是其可扩展性受到模型可以摄入的信息量的限制。未来关于多样性的研究应该从具有可靠参考的简单数据集开始，并专注于最有效的提示间转移相关信息的方式。在我们的提示工程阶段，我们还发现，在提示中包含 CoT 技术对提高输出的质量和相关性以及用新见解（如地理知识）增强现有数据非常有效。但另一方面，这限制了输出的多样性，因为模型受到额外指令的限制。为解决这一权衡，我们建议研究在不限制模型创造力的情况下增强数据的方法。最后，我们引入了一种简单但有效的 Jaccard 相似性度量来评估多样性，使我们能够将文本的多样性相互比较，以及跨模型和技术的整个数据集进行比较。我们希望我们的见解能为数据到文本系统在生产环境中的潜在实施提供良好的基础，并激励未来在自然语言生成多样性主题上的研究。

5

致谢 本项目由德国联邦教育与研究部 (BMBF) 资助，资助编号为 01IS17049 软件校园 2.0（慕尼黑工业大学）。

References

1. LMSys Chatbot Arena Leaderboard - a Hugging Face Space by lmsys — huggingface.co. https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard, [Accessed 13-01-2024]
2. Afzal, A., Kowsik, A., Fani, R., Matthes, F.: Towards optimizing and evaluating a retrieval augmented QA chatbot using LLMs with human-in-the-loop. In: Dragut, E., Li, Y., Popa, L., Vucetic, S., Srivastava, S. (eds.) Proceedings of the Fifth Workshop on Data Science with Human-in-the-Loop (DaSH 2024). pp. 4–16. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.dash-1.2
3. Axelsson, A., Skantze, G.: Using large language models for zero-shot natural language generation from knowledge graphs (2023)
4. Chen, Q., Lin, J., Zhang, Y., Yang, H., Zhou, J., Tang, J.: Towards knowledge-based personalized product description generation in e-commerce. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 3040–3050. KDD '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3292500.3330725, https://doi.org/10.1145/3292500.3330725

www.xueshuxiangzi.com

5. Cho, J., Seo, M., Hajishirzi, H.: Mixture content selection for diverse sequence generation (2019)
6. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models (2022)
7. Elder, H., Gehrmann, S., O'Connor, A., Liu, Q.: E2E NLG challenge submission: Towards controllable generation of diverse natural language. In: Krahmer, E., Gatt, A., Goudbeek, M. (eds.) Proceedings of the 11th International Conference on Natural Language Generation. pp. 457–462. Association for Computational Linguistics, Tilburg University, The Netherlands (Nov 2018). https://doi.org/10.18653/v1/W18-6556, https://aclanthology.org/W18-6556
8. Fabbri, A.R., Kryciski, W., McCann, B., Xiong, C., Socher, R., Radev, D.: SummEval: Re-evaluating Summarization Evaluation (Feb 2021). https://doi.org/10.48550/arXiv.2007.12626, http://arxiv.org/abs/2007.12626, arXiv:2007.12626 [cs]
9. Gardent, C., Shimorina, A., Narayan, S., Perez-Beltrachini, L.: The WebNLG challenge: Generating text from RDF data. In: Alonso, J.M., Bugarín, A., Reiter, E. (eds.) Proceedings of the 10th International Conference on Natural Language Generation. pp. 124–133. Association for Computational Linguistics, Santiago de Compostela, Spain (Sep 2017). https://doi.org/10.18653/v1/W17-3518, https://aclanthology.org/W17-3518
10. Kasner, Z., Dusek, O.: Neural pipeline for zero-shot data-to-text generation. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3914–3932. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.271, https://aclanthology.org/2022.acl-long.271
11. Ke, P., Zhou, H., Lin, Y., Li, P., Zhou, J., Zhu, X., Huang, M.: CTRLEval: An Unsupervised Reference-Free Metric for Evaluating Controlled Text Generation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2306–2319. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.164, https://aclanthology.org/2022.acl-long.164
12. Kocmi, T., Federmann, C.: Large language models are state-of-the-art evaluators of translation quality (2023)
13. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large Language Models are Zero-Shot Reasoners (Jan 2023). https://doi.org/10.48550/arXiv.2205.11916, http://arxiv.org/abs/2205.11916, arXiv:2205.11916 [cs]
14. Liu, H., Wang, P.: Assessing Sentence Similarity Using WordNet based Word Similarity. JSW **8**(6), 1451–1458 (Jun 2013). https://doi.org/10.4304/jsw.8.6.1451-1458, http://ojs.academypublisher.com/index.php/jsw/article/view/9168
15. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment (May 2023). https://doi.org/10.48550/arXiv.2303.16634, http://arxiv.org/abs/2303.16634, arXiv:2303.16634 [cs]
16. Mehri, S., Eskenazi, M.: USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 681–707. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.64, https://aclanthology.org/2020.acl-main.64
17. Montahaei, E., Alihosseini, D., Baghshah, M.S.: Jointly Measuring Diversity and Quality in Text Generation Models (May 2019). https://doi.org/10.48550/arXiv.1904.03971, http://arxiv.org/abs/1904.03971, arXiv:1904.03971 [cs, stat]

www.xueshuxiangzi.com

18. Puranik, V., Majumder, A., Chaoji, V.: PROTEGE: Prompt-based diverse question generation from web articles. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 5449–5463. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.362, https://aclanthology.org/2023.findings-emnlp.362
19. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (Sep 2023). https://doi.org/10.48550/arXiv.1910.10683, http://arxiv.org/abs/1910.10683, arXiv:1910.10683 [cs, stat] version: 4
20. Saha, S., Yu, X.V., Bansal, M., Pasunuru, R., Celikyilmaz, A.: Murmur: Modular multi-step reasoning for semi-structured data-to-text generation (2022)
21. Tevet, G., Berant, J.: Evaluating the Evaluation of Diversity in Natural Language Generation. In: Merlo, P., Tiedemann, J., Tsarfaty, R. (eds.) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 326–346. Association for Computational Linguistics, Online (Apr 2021). https://doi.org/10.18653/v1/2021.eacl-main.25, https://aclanthology.org/2021.eacl-main.25
22. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D.: Llama 2: Open Foundation and Fine-Tuned Chat Models (Jul 2023). https://doi.org/10.48550/arXiv.2307.09288, http://arxiv.org/abs/2307.09288, arXiv:2307.09288 [cs]
23. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Jan 2023). https://doi.org/10.48550/arXiv.2201.11903, http://arxiv.org/abs/2201.11903, arXiv:2201.11903 [cs]
24. Xiang, J., Liu, Z., Zhou, Y., Xing, E.P., Hu, Z.: Asdot: Any-shot data-to-text generation with pretrained language models (2022)
25. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K.: Tree of Thoughts: Deliberate Problem Solving with Large Language Models (May 2023). https://doi.org/10.48550/arXiv.2305.10601, http://arxiv.org/abs/2305.10601, arXiv:2305.10601 [cs]
26. Zhang, H., Song, H., Li, S., Zhou, M., Song, D.: A survey of controllable text generation using transformer-based pre-trained language models. ACM Computing Surveys (2022), publisher: ACM New York, NY
27. Zhang, T., Zhang, J., Huo, C., Ren, W.: Automatic generation of pattern-controlled product description in e-commerce. In: The World Wide Web Conference. p. 2355–2365. WWW '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3308558.3313407, https://doi.org/10.1145/3308558.3313407
28. Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H., Han, J.: Towards a Unified Multi-Dimensional Evaluator for Text Generation (Oct 2022). https://doi.org/10.48550/arXiv.2210.07197, http://arxiv.org/abs/2210.07197, arXiv:2210.07197 [cs]
29. Zhou, G., Lampouras, G.: Informed sampling for diversity in concept-to-text nlg (2021)

www.xueshuxiangzi.com