arXiv:2602.02738v1 [cs.SD] 2 Feb 2026

当噪声降低损失时：重新思考音乐大语言模型中基于似然的评估

Xiaosha Li[1], *Chun Liu*[2], *Ziyu Wang*[34]

[1]佐治亚理工学院 [2]字节跳动[3]库朗数学科学研究所，纽约大学[4] 穆罕默德·本·扎耶德人工智能大学（MBZUAI） xiaosha@gatech.edu, chun.liu@bytedance.com, ziyu.wang@nyu.edu

摘要

音乐大语言模型（LLM）的兴起，要求我们采用稳健的方法评估输出质量，尤其是区分高质量作品与“垃圾音乐”。值得注意的是，我们观察到，标准交叉熵损失——一种核心训练指标——在模型遇到系统性损坏的音乐时往往会降低，这削弱了它作为独立质量指标的有效性。为探究这一悖论，我们引入了噪声注入实验，将不同长度的受控噪声信号注入音乐上下文。我们假设，模型的损失对这些扰动产生积极响应，具体表现为短时注入时损失显著上升（“峰值”区域），可作为衡量其辨别音乐完整性能力的替代指标。在音频波形域对 MusicGen 模型进行的实验表明，相较于全局语义损坏，音乐大语言模型对局部、纹理层面的扰动反应更强。除了揭示这种偏差，我们的结果还凸显了一项新原则：损失曲线的形状——而非其绝对值——蕴含着有关生成内容质量（即模型行为）的关键信息。我们设想，这种基于曲线特征的评估可作为一种无标签、模型内在的音乐质量评估框架，为更合理的训练目标和更具区分力的基准测试开辟道路。 [1]

索引词— 损失、噪声、音乐大语言模型、大语言模型评估、暴露偏差

1. 引言

在自然语言处理中，一种常见的评估方法是将文本输入大语言模型（LLM），并计算其似然（或等价地，其损失） [1]。其直觉很直接：被赋予更高似然的序列被认为更符合模型学习到的分布，因而“更好”。随着音乐大语言模型的兴起，将这一思路自然延伸到音乐评估是顺理成章的：把音乐序列输入模型，并以似然作为质量的代理指标。

然而，这种方法在音乐领域的可靠性远未得到证实。尽管基于似然的评估在文本大语言模型中存在的伪影和偏差已有文献记载 [2, 3, 4, 5], 但这些问题在音乐中如何表现，或是否与人类对质量的判断一致，仍不清楚。为此，我们设计了一项噪声注入实验，在音乐序列中加入扰动（例如白噪声），并测量似然的变化（图1）。直觉上，人们会认为

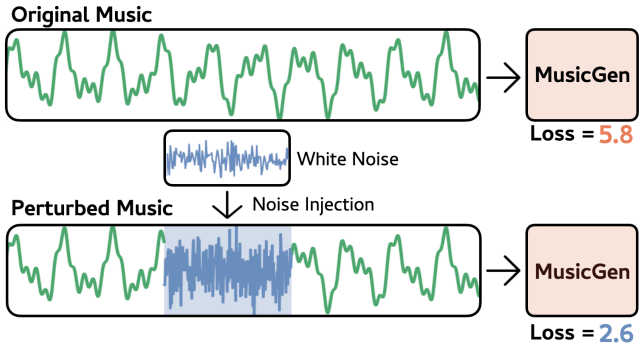


图 1. 该图表明，向输入音频注入噪声可能会出人意料地降低损失；我们将这一现象称为上下文失忆效应。

噪声会破坏音乐并降低其似然，从而增加损失。令人惊讶的是，我们观察到相反的现象：加入噪声往往会降低损失。

我们的分析表明，噪声下损失意外降低，是由逐词元似然的变化所致。在噪声片段开始时，模型的损失会急剧飙升，表明它识别到了与前置上下文不一致之处。然而，几乎紧接着，损失便下降，并在整个噪声片段持续期间保持较低水平，无论原始上下文为何。这是因为音乐大语言模型对某些形式的噪声本就赋予相对较低的损失；这些噪声具有规律性，因此比真实音乐更容易预测。扰动结束后，损失会重新与原始上下文对齐，但方差显著增大。换言之，模型会短暂地“抵抗”干扰，随后便轻易“忘记”此前的音乐素材，并适应噪声，仿佛噪声就是当前的上下文。这种行为在不同噪声类型、音乐风格和基于 Transformer 的模型变体中均稳定出现。我们将这一现象称为上下文失忆效应。

这一发现揭示了基于似然的音乐评估存在一项根本局限。以损失衡量时，大语言模型只能可靠地检测极短期的不一致（例如起始处的噪声），却无法识别较长期的结构退化（例如乐句重排）。事实上，模型对扰动的损失响应极不一致：可能上升、下降，也可能保持不变，因此无法可靠地反映音乐质量。这种不可预测性具体体现了音乐大语言模型中的曝光偏差[3] 。

与此同时，我们的结果指向一个更有前景的方向。虽然绝对损失值不具信息量，但损失曲线的局部动

[1]Code 以及演示页面可通过 https://noiseloss.github.io.

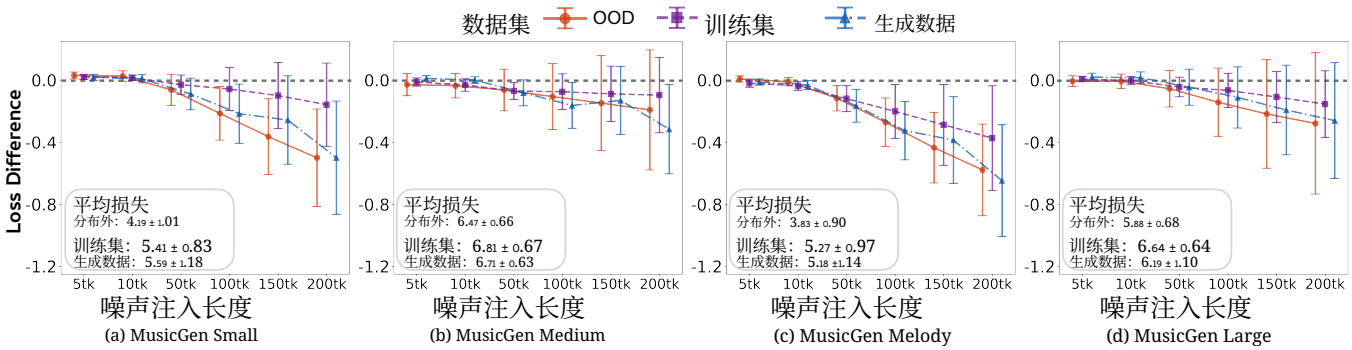


图 2。模　　型在数据集上进行白噪声注入时的性能比较，标准差根据 son 计算　　gs.

态承载着有意义的信号。具体而言，我们始终观察到扰动开始时出现一个尖锐峰值，随后是同化与恢复阶段，这些阶段在刻画模型行为方面似乎比原始似然更为可靠。这表明，聚焦于逐词元损失曲线的形状而非绝对值的基于曲线特征的评估，可能为未来的自动音乐评估方法提供更坚实的基础。

2. 相关工作

基于大语言模型的评估借助大型模型，通过若干主要范式评估其他模型的输出。自动方法包括基于似然的指标，这类指标衡量概率拟合度，却往往忽视内容质量 [5], ; 以及基于提示的多项选择测试，这类测试揭示了评估能力中的悖论 [6]。相比之下，人类偏好竞技场等非自动方法则通过大规模两两投票对模型进行排名 [7]。G-Eval 和 GPTScore 等近期拓展方法直接将大语言模型用作评判者——通过概率加权评分进行结构化推理 [8] ，或采用灵活的基于指令的评估 [9]——并报告称，这些方法在语言任务中与人工评分的相关性更强。然而，将这些方法应用于音乐仍面临挑战：音乐语义含义模糊，显著片段和长程结构难以捕捉，而且高概率的后续内容往往无法体现质量。

自动评估具有可扩展性强、成本较低等优势，但其可靠性仍不确定。尽管似然是有效的训练目标，但在推理时会受到曝光偏差和目标不匹配的影响 [3, 10]。将似然用于解码还会产生平淡、重复的文本，与人类语言中备受重视的创造性和多样性背道而驰 [4]。基于大语言模型的评估器也存在似然偏差，会高估似然高却流于表面的输出 [5], ; 同时还表现出不一致性，容易受到熟悉度、锚定效应以及提示词的影响，而这些因素并不会影响人类判断 [11]。更广泛地说，这类伪影反映了捷径学习现象：正如视觉与语言领域所观察到的那样，模型利用虚假相关，而非真正的能力 [12, 13]。

在音乐领域，评估仍主要依赖人工评分 [14], , 客观指标直到近期才开始出现 [15]。Fr´echet 音频距离（FAD）和 MAD 用于考察保真度、音乐性和多样性 [16, 17], , 而 Music Arena 等平台则通过听众偏好扩大评估规模 [18]。CMI-Bench 将音乐理解重新构建为指令遵循任务 [19]。Audiobox-Aesthetics 是早期尝试之一，旨在通过加权模型将人工评分纳入自动评估，不过其评估维度仍然有限 [20]。与语言领域不同，目前尚无成熟的方法能够将大语言模型用作音乐自动评审，因此我们开展了这项研究。

3. 噪声注入实验

在本节中，我们介绍噪声注入实验，该实验展示了一个反直觉的效应：当对输入音频施加扰动时，模型预测损失会降低（或等价地，似然会增加）。实验设置在第3.1节中描述，结果在第3.2节中呈现。

3.1. 实验设置

给定音频信号 x ，其被词元化为 $x_{1:T}$，其中 T 为序列长度（例如，通过MusicGen中的EnCodec [21] ），生成模型将损失 ℓ（$x_{1:T}$）计算为负对数似然:

$$\ell(x_{1:T}) = -\sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}), \tag{1}$$

我们将扰动信号 $x'_{1:T}$ 定义为

$$x'_i = \begin{cases} x_i, & i \notin \mathcal{I}, \\ \epsilon_i, & i \in \mathcal{I}, \end{cases} \tag{2}$$

其中 $\mathcal{I}$ 为扰动时间步， ϵ_i 表示给定的噪声。类似地，扰动序列 $x'_{1:T}$ 的损失为

$$\ell(x'_{1:T}) = -\sum_{t=1}^{T} \log p_\theta(x'_t \mid x'_{<t}). \tag{3}$$

我们还将时间步 t 处的逐词元损失差定义为

$$\Delta\ell_t = -\log p_\theta(x'_t|x'_{<t}) + \log p_\theta(x_t|x_{<t}), \tag{4}$$

该指标衡量在词元 t 处施加扰动后，序列似然相对于原始序列的变化。我们首先使用响度可控的白噪声（-30 至 -12 dB）作为注入噪声；为考虑不同数据集间整体振幅的差异，噪声响度与原始音频相匹配，因为我们的实证观察表明，所得损失的大小对噪声响度十分敏感。对于由 750 个词元（15 秒）组成的音频，我们在第 250 个词元（5 秒）处注入噪声，从而为模型提供足够的历史信息以建立上下文。这里，词元对应于采样率为 50 Hz 的基于残差向量量化的音频词元，这意味着每个词元约代表 20 毫秒的音频（所有信号均

重采样至 32 千赫）。扰动长度设为 $5, 10, 50, 100, 150, 200$ 个词元，分别对应 0.1,0.2, 1.0, 2.0, 3.0 4.0 , 秒。所选长度涵盖多个层次的音乐扰动，可分别在帧、音符、节拍和小节层面破坏语义。

3.2. 实验结果

在实验中，我们评估了对三类音乐数据进行噪声注入的效果：**(1) 训练集**：用于 MusicGen 的 ShutterStock [2] 训练语料的一个子集，共包含 20 首歌曲；**(2)生成数据**：MusicGen-Small 在宽泛的生成设置下生成的 140 个样本（top-k = 10，50、100、150、200、250、500 等）；以及 **(3) 分布外（OOD）**：来自 ASAP 数据集 [3] 的 78 首古典乐曲，涵盖多种作曲家和风格。我们评估了以波形为输入的自回归大语言模型，包括 MusicGen（Small（300M）/ Medium（1.5B）/ Large（3.3B）/ Melody（1.5B））模型 [22]。

我们计算损失差 $\Delta\ell = \ell(x'_{1:T}) - \ell(x_{1:T})$，结果如图 2 所示。在所有模型、数据集和噪声长度下，我们都观察到一致的模式：注入噪声较短时，损失差接近于零；随着噪声长度增加，损失差变为负值，这表明较长的扰动会系统性地降低损失。

为验证这一趋势的稳健性，我们对扰动长度与平均损失差进行了皮尔逊和斯皮尔曼相关性检验。在大样本设置中，每个数据点代表一个特定的（模型、数据集、噪声长度）组合（共 ≈78 个数据点），皮尔逊和斯皮尔曼相关系数均显著为负（$r < -0.85$、$p < 0.001$），证实了高度显著的负相关趋势。即使在小样本设置中，按噪声长度汇总结果（≈ 6 组）后，相关性仍然稳健（$r < -0.91$）且具有统计显著性（$p < 0.05$）；线性回归检验中观察到的显著负斜率进一步印证了这一发现。

为进一步考察这一效应的普适性，我们在 YuE（10亿参数） [23] 自回归模型上，使用三个数据集重复了分析：TrainingSet、分布外数据，以及由 YuE 自身生成的数据。对于 TrainingSet 和分布外数据，皮尔逊相关性检验均显示，噪声长度与损失差之间存在一致且显著的负相关关系（$r < -0.90$、$p < 0.01$），显著为负的回归斜率也佐证了这一结果。对于生成
成数据，尽管未观察到噪声长度与损失差之间存在统计学上的显著相关性（$p > 0.05$），但所有注入噪声的样本，其损失值始终低于对应的干净生成样本。这一普遍发现表明，仅凭绝对损失值不足以判断音乐序列是否受到噪声扰动，尤其是在模型生成的音乐中。相关实验的可视化结果可在我们的演示页面查看。

4. 损失动态分析

现在，我们详细分析噪声注入实验中的损失表现。我们的目标有两个：首先，解释第3节中观察到的反直觉趋势；其次，量化扰动如何重塑逐词损失动态。我们先通过逐词可视化凸显一种一致的三阶段效应（第4.1节），再通过自动区域检测实验验证这一效应（第4.2节）。

[2]https://www.shutterstock.com [3] https://github.com/fosfrancesco/asap-dataset

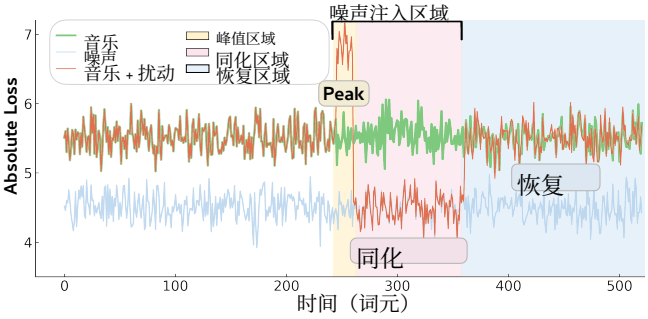


图 3。噪声注入实验的损失曲线。

4.1. 逐词元损失动态

为探究这种反直觉现象，我们在逐词层面分析式（4）中的噪声注入设置。我们计算损失差 $\Delta\ell_t$，并可视化实验中所有模型和数据集的原始音乐、纯噪声以及经扰动的音乐+的逐词元损失曲线（图3），揭示扰动如何重塑损失动态。

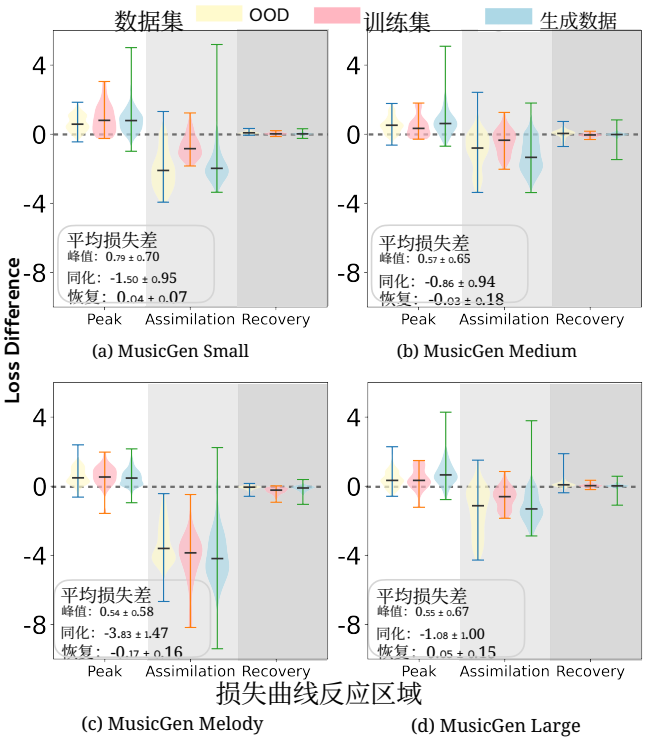


图 4。噪声注入后的三阶段行为。

在不同扰动长度下，我们观察到词元级损失轨迹中的三个特征区域：

- **峰值区域：** 在最初的 ∼100 毫秒（≈5 个词元）内，由于与前置上下文局部不一致，损失会飙升。
- **同化区域：** 在扰动窗口内，损失迅速下降并稳定在较低值，对上下文基本不敏感。

• **恢复区域：** 扰动结束后，预测变得不稳定，损失围绕基线振荡，反映出曝光偏差。

这种三阶段行为表明，模型会暂时无法有效利用先前的音乐上下文，仿佛忘记了早先的信息；我们将此称为上下文失忆效应。

4.2. 通过自动区域检测进行验证

为确认这三个区域具有一致性，我们进行了一项区域检测实验。我们对 $\Delta\ell_t$应用移动平均，以抑制局部波动，并自动识别损失连续五个词元过零时的边界。该流程能够稳健地检测起始峰值、同化平台期和不稳定的恢复阶段（图 4）。

这些结果验证了第4.1节定性分析提出的假设：模型对短程局部扰动反应强烈，却无法反映长程结构扰乱。这种上下文失忆效应正是音乐的基于损失的评估不可靠的根源。

5. 讨论

在前文各节中，我们证明了添加白噪声反而会降低损失，并引入上下文失忆效应来解释这一现象。虽然这些结果阐明了为何模型损失无法捕捉长程结构扰乱，但这一局限是否仅存在于噪声注入中，还是反映了损失作为评估指标更普遍的弱点，仍不明确。本节将分析扩展到更贴近实际的情境，具体而言，是模型损失自然会被用于评估音乐质量的情境。我们首先在第5.1节考察顺序打乱这一替代性扰动，表明短程敏感而长程不敏感的相同模式依然存在。随后，我们在第5.2节将这一现象与更广泛的曝光偏差概念联系起来，并强调其对生成式音乐模型评估的影响。

5.1. 替代扰动：顺序打乱

为进一步检验我们的发现，我们将分析扩展到另一种扰动：顺序打乱。与噪声注入不同，打乱在以结构化方式扰乱音乐顺序的同时，保留了相同的信息量。我们对不同长度的片段（1、2、5、10、35、50、70、100、150 和 200 个词元）进行打乱，以覆盖从局部重排到音乐结构长程扰乱的广泛扰动尺度。

如图 5 所示，结果与噪声注入的结果高度一致。短程打乱会使损失出现急剧尖峰，因为扰动主要由峰值区域主导。随着打乱长度增加，同化区域变长；在此期间，模型适应被扰动的顺序，并实际上失去对原始上下文的访问，从而使损失保持接近基线。尽管长跨度打乱对音乐形式造成了显著且结构化的退化，全局损失大体保持不变，这表明在长程结构扰乱下，用全局损失评估模型性能并不可靠。

5.2. 与曝光偏差的关系

上下文失忆效应与众所周知的曝光偏差现象密切相关：自回归模型在推理时生成过程中出现错误后，往往难以恢复。

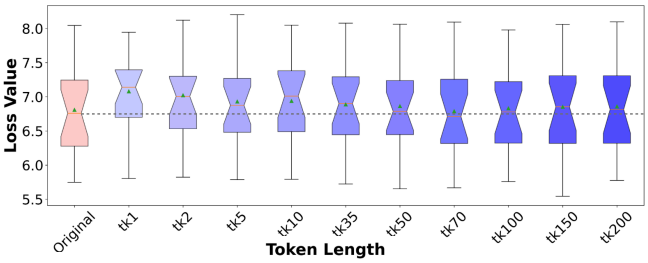


图 5。顺序打乱扰动示例及实验结果。

在我们的设定中，注入噪声会产生相对于训练数据而言分布外的词元，导致噪声起始时损失短暂飙升，随后模型迅速适应，但预测也变得愈发不稳定。出现此类错误后，模型实际上会缩短可用上下文，并依赖受损输入，这表明曝光偏差不仅会降低生成质量，还会损害基于似然的评估的可靠性。在音乐中，惊喜、张力和新颖性至关重要，因此陌生但有意义的乐段可能会被误解为扰动，并像噪声一样被低估。

我们的研究结果表明，当前的大语言模型无法可靠地使用绝对损失来区分质量不同的作品（例如，经典古典作品与普通训练集样本）。我们鼓励未来的研究进一步探索这一局限在音乐评估中的程度。

6. 结论

在这项工作中，我们研究了音乐大语言模型基于损失的评估的可靠性，并以噪声注入实验中观察到的一种反直觉现象为切入点。我们发现了一种关键模式，并将其称为上下文失忆效应：引入扰动后，损失曲线会持续呈现三个特征区域——**峰值**、**同化**和**恢复**。通过可视化损失曲线和逐词元损失差异，我们展示了模型如何只检测即时的不一致，却无法对长期的结构变化作出反应。将分析扩展到顺序打乱等其他扰动后，我们进一步证实，绝对损失不适合用于评估音乐质量，尤其是在作曲层面。

我们的研究结果凸显了基于似然的音乐评估的一项根本局限：绝对损失值无法作为衡量音乐质量的稳定指标。相比之下，损失曲线的形状和局部动态——尤其是起始峰值——能提供更清晰、更一致的信号。我们将这种基于轮廓的视角视为迈向开发更可靠的自动评估框架的第一步，使其能够更好地契合音乐独特的结构特征。

7. 致谢

作者谨向 Alexander Lerch教授和沈亦康致以谢意，感谢他们提供宝贵指导；并感谢马文野给予的帮助。

8. 参考文献

[1] Yoshua Bengio、R´ejean Ducharme、Pascal Vincent 和 Chris- tian Jauvin，《神经概率语言模型》，机器学习研究杂志，2003年。[2] Kushal Arora、Layla El Asri、Hareesh Bahuleyan 和 Jackie Chi Kit Cheung，《为何曝光偏差至关重要：从模仿学习视角看语言生成中的误差累积》，计算语言学协会发现（*ACL*），2022年。[3] Marc'Aurelio Ranzato、Sumit Chopra、Michael Auli 和 Woj- ciech Zaremba，《使用循环神经网络进行序列级训练》，发表于国际学习表征会议（*ICLR*），2016年。[4] Ari Holtzman、Jan Buys、Li Du、Maxwell Forbes 和 Yejin Choi，《神经文本退化的奇特案例》，发表于国际学习表征会议（*ICLR*），2020年。[5]Masanari Ohi、Masahiro Kaneko、Ryuto Koike、Mengsay Loem 和 Naoaki Okazaki，《大语言模型中评估偏差的基于似然的缓解》，发表于计算语言学协会发现（*ACL*），2024年，第3237–3245页。

[6] Peter West、Ximing Lu、Nouha Dziri、Faeze Brahman、Lin- jie Li、Jena D. Hwang、Liwei Jiang、Jillian Fisher、Abhilasha Ravichander、Khyathi Raghavi Chandu、Benjamin Newman、Pang Wei Koh、Allyson Ettinger 和 Yejin Choi，《生成式 AI 悖论：它能够创造什么，却未必理解什么》，发表于国际学习表征会议（*ICLR*），2024年。

[7] Wei-Lin Chiang、Lianmin Zheng、Siyuan Zhuang、Eric Wallace、Tianjun Li、Yingbo Sheng、Rose Wu 等人，《聊天机器人竞技场：一个通过人类偏好评估大语言模型的开放平台》，发表于*2024*年自然语言处理经验方法会议（*EMNLP*）论文集，2024年。

[8] Yang Liu、Dan Iter、Yichong Xu、Shuohang Wang、Ruochen Xu 和 Chenguang Zhu，《G-Eval：使用 GPT-4 和更好的人类对齐进行自然语言生成评估》，发表于*2023*年自然语言处理实证方法会议（*EMNLP*）论文集，2023年，第 2511–2522 页。

[9] Jinlan Fu、See-Kiong Ng、Zhengbao Jiang 和 Pengfei Liu，《GPTScore：按需评估》，载于*2024*年北美计算语言学协会分会会议论文集：人类语言技术（*NAACL*），2024年，第 6556–6576 页。[10] Cong Xu、Zhangchi Zhu、Jun Wang、Jianyong Wang 和 Wei Zhang，《理解交叉熵损失在公平评估基于大语言模型的推荐中的作用》，载于*ACM*推荐系统会议论文集（*RecSys*），2024年。[11] Rickard Stureborg、Dimitris Alikaniotis 和 Yoshi S uhara，《大语言模型是前后不一致且存在偏见的评估者》，*arXiv* 预印本 *arXiv:2405.01724*，2024年。[12] Robert Geirhos、J¨orn-Henrik Jacobsen、Claudio Michaelis、Richard Zemel、Wieland Brendel、Matthias Bethge 和 Felix A. Wichmann，《深度神经网络中的捷径学习》，《自然·机器智能》，第 2 卷，第 11 期，第 665–673 页，2020年。

[13] Yu Yuan、Lili Zhao、Kai Zhang、Guangting Zheng 和 Qi Liu，《大语言模型能否克服捷径学习？对大语言模型捷径挑战的评估》，载于*2024*年自然语言处理经验方法会议论文集（*EMNLP*），2024年，第 12188–12200 页。[14] Florian Gr¨otschla、Ahmet Solak、Luca A. Lanzend¨orfer 和 Roger Wattenhofer，《通过人类偏好研究对音乐生成模型和指标进行基准测试》，载于*2025 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。2025年，IEEE。

[15] Alexander Lerch、Claire Arthur、Nick Bryan-Kinns、Corey Ford、Qianyi Sun 和 Ashvala Vinay，《音乐生成模型评估综述》，*ACM* 计算调查，第 58 卷，第 4 期，第 99 页，2025年。

[16] Azalea Gui、Hannes Gamper、Sebastian Braun 和 Dimitra Emmanouilidou，《将 Fr´echet 音频距离应用于生成式音乐评估》，载于*2024 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。2024年，第 1331–1335 页，IEEE。[17] Yichen Huang 和 Chris Donahue，《使文本生成音乐评估与人类偏好保持一致》，载于国际音乐信息检索学会会议（*ISMIR*），2025年。[18] Yonghyun Kim、Wayne Chi、Anastasios N. Angelopoulos、Wei-Lin Chiang、Koichi Saito、Shinji Watanabe、Yuki Mitsu-fuji 和 Chris Donahue，《Music Arena：文本生成音乐的实时评估》，载于神经信息处理系统进展（*NeurIPS*），2025年。[19] Yuxuan Wang、Ming Sun、Hao Chen、Rui Zhang、Kat Agres 和 Yi-Hsuan Yang，《CMI-Bench：用于评估音乐指令遵循能力的综合基准》，载于国际音乐信息检索学会会议（*ISMIR*），2025年。[20] Andros Tjandra、Yi-Chiao Wu、Baishan Guo、John Hoffman、Brian Ellis、Apoorv Vyas、Bowen Shi、Sanyuan Chen、Matt Le、Nick Zacharov、Carleigh Wood、Ann Lee 和 Wei-Ning Hsu，《Meta Audiobox Aesthetics：语音、音乐和声音的统一自动质量评估》，*arXiv* 预印本 *arXiv:2502.05139*，2025年。[21] Alexandre D´efossez、Jade Copet、Gabriel Synnaeve 和 Yossi Adi，《高保真神经音频压缩》，机器学习研究汇刊（*TMLR*），2023年，2023年9月录用。[22] Jade Copet、Felix Kreuk、Itai Gat、Tal Remez、David Kant、Gabriel Synnaeve、Yossi Adi 和 Alexandre D´efossez，《简单且可控的音乐生成》，载于第 *37* 届神经信息处理系统会议论文集（*NeurIPS*），2023年。[23] Ruibin Yuan、Hanfeng Lin、Shuyue Guo、Ge Zhang、Jia-hao Pan、Yongyi Zang、Haohe Liu、Yiming Liang、马文野、Xingjian Du 等，《Yue：扩展开放基础模型以实现长篇音乐生成》，*arXiv* 预印本 *arXiv:2503.08638*，2025年。