SongSage：一种采用歌词生成式预训练的大型音乐语言模型

Jiani Guo[1]*，Jiajia Li[2]*，Jie Wu[3],ZuchaoLi[4]†，Yujiu Yang[3],Ping Wang[2][1]武汉大学计算机科学学院[2]武汉大学信息管理学院[3]清华大学；[4]武汉大学人工智能学院 {guojiani, cantata, zcli-charlie}@whu.edu.cn

arXiv:2601.01153v1 [cs.CL] 3 Jan 2026

摘要

大语言模型在各个领域都取得了显著成功，但其对以歌词为中心的知识的理解尚未得到充分探索。在这项工作中，我们首先介绍**PlaylistSense**，这是一个用于评估语言模型播放列表理解能力的数据集。PlaylistSense 涵盖了十种源自常见现实场景视角的用户查询，对大语言模型提出挑战，要求其准确把握播放列表特征并满足多样化的用户意图。全面评估结果表明，当前通用大语言模型在播放列表理解方面仍有提升空间。受此启发，我们推出**SongSage**，一种通过歌词生成式预训练获得多样化歌词中心智能的大型音乐语言模型。SongSage 在**LyricBank**上进行持续预训练；该语料库经过精心整理，包含 54.8 亿个词元，专注于歌词内容。随后，模型使用**LyricBank-SFT**进行微调；该指令集经过悉心构建，包含 77.5 万个样本，涵盖九项核心的以歌词为中心的任务。实验结果表明，SongSage 对歌词中心知识有着深入理解，在改写用户查询以实现零样本播放列表推荐方面表现出色，能够有效生成和续写歌词，并熟练完成另外七项能力。除了在歌词中心领域的专长外，SongSage 还保留了通用知识理解能力，并取得了具有竞争力的 MMLU 分数。由于版权限制，我们将不开放数据集，并发布 SongSage 和训练脚本，以确保结果可复现并支持音乐人工智能研究与应用。数据集发布计划的详情见附录。

1 引言

人工智能（AI），尤其是大语言模型（LLM），已在各种自然语言处理任务中取得显著成功，包括数学推理（Gou et al., 2024; Guo 等人, 2025b; ?）、指令遵循（Zhou 等人, 2023; Guo 等人, 2026）、代码生成（Luo et al., 2024; Wu 等人, 2025a,b）和多语言处理（Üstün et al., 2024; Wu 等人, 2024a）。基于这些成果，研究人员一直在探索大语言模型在自然语言以外的专业领域中的理解能力，涵盖医疗保健（Singhal et al., 2023）、化学（Bran et al., 2024）、生物学（Madani et al., 2023）及其他专业领域。

音乐是一种丰富且多元的艺术形式，涵盖多种模态：语言（歌词）、视觉（乐谱）和听觉（人声、旋律、伴奏及完整乐曲）。鉴于这种多样性，大语言模型日益广泛地应用于音乐相关研究的各个领域，包括音乐生成（Hong 等人, 2024a；Melechovský 等人，2024；Hong 等人，2024b；Melechovský 等人，2024）、音乐知识理解（Yuan 等人，2024年；Deng 等人, 2024年；Zhou 等人，2024年；Wu 等人, 2024b；Weck 等人，2024年；Sakshi 等人, 2024年），以及乐谱补全（Li 等人，2024年）。与其他模态相比，语言具有更高的信息密度，因此大语言模型常被用作开发音乐人工智能模型的基础。

为评估通用大语言模型是否足以胜任音乐人工智能任务，我们推出了 PlaylistSense，这是一个包含 5k 个播放列表的数据集，旨在评估大语言模型对播放列表和歌词的理解能力。该数据集基于十种现实场景模拟用户查询，测试大语言模型捕捉播放列表特征并生成精准描述的能力，使查询能够检索到预期的播放列表。大量评估—

*共同贡献。
†通讯作者。

结果表明，大语言模型在播放列表理解和歌词理解方面存在局限，同时还带有各种固有偏见。这些发现凸显了开发专用大型音乐语言模型（LMLMs）以推动音乐人工智能领域发展的迫切需求。

与通用大语言模型相比，LMLMs 通常会针对音乐特定文本内容进行进一步预训练，而这些内容主要由歌曲歌词构成。虽然歌曲歌词十分丰富，但其他形式的音乐相关文本则相对稀缺。受此启发，我们推出了 SongSage，这是一款大型音乐语言模型，基于从互联网收集的大量歌曲歌词档案进行歌词生成预训练。SongSage 基于两个训练数据集构建：*LyricBank* 和*LyricBank-SFT*。*LyricBank*是我们精心整理的持续预训练语料库，包含 54.8 亿个 token，旨在加深模型对歌词内容的理解。在此基础上，*LyricBank-SFT*作为一个包含 77.5 万个样本的指令微调数据集，引入了九项专门的歌词相关能力，进一步提升了 SongSage 在音乐领域的专业水平。经过持续预训练和指令微调后，SongSage 展现出出色的音乐理解能力，同时也具备良好的通用知识理解能力。我们相信，SongSage 这一具备音乐知识的开源基础模型，将推动大型音乐语言模型领域的进一步发展。具体而言，我们的贡献如下：

(1)我们提出了*PlaylistSense* 数据集，从十个用户查询视角出发，专注于评估语言模型的播放列表理解能力。

(2)我们精心整理了一个持续预训练语料库*LyricBank* 以及一个专注于歌曲相关能力的指令微调数据集*LyricBank-SFT* ，并基于此引入了SongSage——一种通过歌词生成式预训练获得多样化以歌词为中心的智能的大型音乐语言模型。实验结果表明，SongSage在理解以歌词为中心的任务方面取得了显著提升，同时保持了良好的通用知识理解能力。

(3) 我们分析大语言模型中歌词预训练的语言特征，以探究它相较于通用大语言模型所带来的差异。

2 相关工作

2.1 音乐知识理解

近期研究提升了大语言模型对音乐知识和理论的理解。ZIQI-Eval 基准（Li 等人，2024年）包含超过 14,000 道精选问题，用于测试多方面的音乐知识；结果显示，只有 GPT-4（OpenAI，2024年）取得了中等水平的性能。ChatMusician（Yuan 等人，2024年）通过使用 ABC 记谱法进行训练，增强了大语言模型的音乐能力；它将音乐视为第二语言，使模型能够理解音乐概念，同时保留核心语言能力。

2.2 音乐表示与生成

近期的音乐理解与生成方法大致可分为基于音频的方法和基于符号的方法。

基于音频的方法将音乐处理为连续信号或量化音频令牌。AudioGen（Kreuk 等人，2023年）和 MusicGen（Copet 等人，2023年）等模型使用自回归 Transformer 生成离散音频令牌，而 M2UGen（Hussain 等人，2023年）则将这些能力扩展到多模态生成，整合音频、文本和视觉元素，以引导音乐创作。

基于符号的方法使用各种基于文本的表示形式来编码音乐信息，每种形式都适用于理解和生成的不同方面。例如，ChatMusician（Yuan 等人，2024年）采用的 ABC 记谱法以紧凑的格式编码音高、时值和拍号，例如用“F4”表示第四个八度的 F 音符。MusicAgent（Yu 等人，2023年）采用类 MIDI 文本格式，将 MIDI数据转换为包含音高、力度和时序的序列。ZIQI-Eval 和 MusicAgent 使用的基于乐谱的记谱法则将传统乐谱转换为文本，同时保留其结构元素。类似地，SongComposer（Ding 等人，2024年）将音符属性与歌词组合成结构化元组，使音乐与文本对齐。这些表示形式使符号方法能够捕捉音乐的结构性和表现性方面。

3 PlaylistSense 评估

我们推出 PlaylistSense 数据集，用于评估通用大语言模型理解播放列表的能力。该数据集包含 5,000

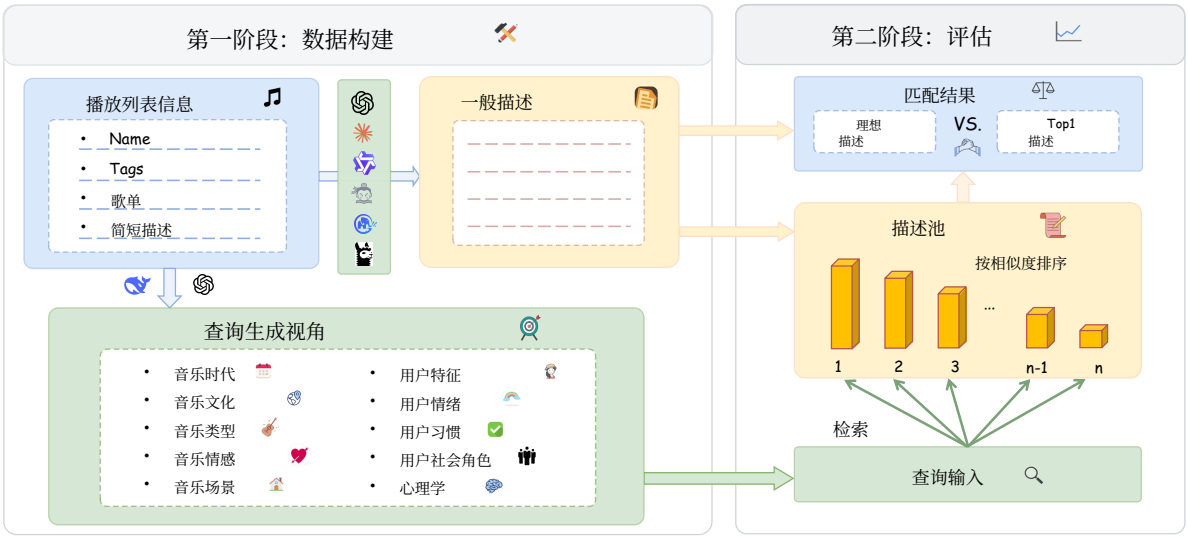

图 1：PlaylistSense 评估概览。在数据构建过程中，Deepseek 生成 10 个查询，涵盖基于音乐和基于用户的视角。在评估过程中，评估器为每个播放列表生成描述，形成一个描述池。每个查询都会通过嵌入模型从描述池中检索描述，是否成功取决于检索到的描述是否与该查询的源播放列表相匹配。

份来自网易云音乐*的播放列表，每个播放列表都附有 10 条查询。这些查询是将播放列表详情提供给先进的大语言模型 Deepseek-v2.5（DeepSeek-AI，2024年）后生成的。查询涵盖基于音乐和基于用户两个角度，代表了现实中用户的典型提问，如图7所示。有关数据集构建的更多详情，请参阅附录。

3.1 评估

在评估阶段，我们使用以下方法评估 PlaylistSense 中的多个先进开源和闭源大语言模型。对于第 i 个播放列表，我们在零样本设置下，将其详细信息（标题、标签和前 20 首歌曲的歌词）输入待评估的大语言模型，以生成通用描述 des_i；超出模型窗口大小的内容将被截断。此过程可形式化为：

$$\text{des}_i = \text{Evaluator}(\text{title}_i \,||\, \text{tags}_i \,||\, \text{song lyrics}_i)$$

这些描述共同构成描述池 D，体现了评估器对每个播放列表的理解：

$$D = \{\text{des}_1, \text{des}_2, \ldots, \text{des}_N\}$$

我们采用基于检索的评估范式，并使用命中率@k（HR@k）作为评估指标。评估过程如下：对于每个查询$_i$，我们从描述池 D 中检索出最相似的前 k 条描述。HR@k 随后计算正确描述（与查询对应同一播放列表的描述）出现在这前 k 条检索结果中的案例比例。HR@k 可正式定义为：

$$\text{HR@}k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\text{des}_i \in \text{Top}_k^D(\text{query}_i)\right),$$

$$\text{Top}_k^D(\text{查询}_i) = \text{Top}_k\left(\text{相似度}\left(\mathbb{E}(\text{查询}_i)\ \mathbb{E}(D)\right)\right),$$

其中，$\text{Top}_k^D(\text{query})$ 会根据 query$k$ 与描述池 D 中所有描述之间的余弦相似度，返回最相似的前 i 条描述 D。 $\mathbb{I}(\cdot)$ 是指示函数：若其中的条件成立，则等于 1，否则等于 0。 $\mathbb{E}$ 表示嵌入模型，实际使用的是 bge-small-zh-v1.5（Xiao 等人，2024年）。

我们首先评估 GPT-4o、Claude 3.5 Sonnet（Anthropic，2024年）、DeepSeek-V3、DeepSeek-R1（Guo 等人，2025a）以及若干开源大语言模型：Llama-3-8B-Instruct（Dubey 等人，2024年）、Qwen2.5-7B-Instruct（Yang 等人，2024年）、GLM-4-9B-Chat（Zeng 等人，2024年）和 InternLM2.5-7B-Chat（Cai 等人，2024年）。HR@1 和 HR@5 的结果见表 1。

基于我们从十个不同视角对多个大语言模型进行的全面评估，我们提出以下两点关键观察与见解。

*网易云音乐是一个广受欢迎的在线音乐流媒体平台，拥有丰富多样的播放列表。参见 https://music.163.com/。数据仅用于学术研究。

模型	基于音乐的视角									
	文化		情感		Era		场景		Type	
	HR@1	HR@5	HR@1	HR@5	HR@1	HR@5	HR@1	HR@5		
Llama-3-8B-Instruct	0.53	0.61	0.41	0.49	0.43	0.68	0.31	0.61	0.47	0.61
InternLM2.5-7B-Chat	14.78	28.94	8.85	19.53	14.20	19.61	9.43	26.73	12.63	0.02
Qwen2.5-7B-Instruct	14.49	28.55	9.47	20.11	14.16	16.27	8.18	17.46	12.26	25.99
GLM4-9B-Chat	17.71	34.97	9.72	21.73	**16.85**	19.94	11.03	**23.53**	**14.78**	30.21
Claude-3.5-Sonnet	10.56	24.51	7.40	18.51	10.53	24.76	6.76	17.36	9.35	24.10
GPT-4o	**18.94**	**35.99**	**18.90**	**24.12**	12.1	**35.03**	**17.52**	23.51	11.48	**33.59**
DeepSeek-V3	15.70	32.34	8.51	19.80	15.47	31.17	8.96	19.70	14.06	30.42
DeepSeek-R1	2.28	6.44	1.87	5.66	3.03	8.32	1.68	5.41	2.95	8.08

模型	基于用户的视角									
	特征		情感		习惯		社会角色		心理学	
	HR@1	HR@5	HR@1	HR@5	HR@1	HR@5	HR@1	HR@5		
Llama-3-8B-Instruct	0.41	0.55	0.14	0.47	0.18	0.31	0.33	0.43	0.08	0.18
InternLM2.5-7B-Chat	9.47	20.91	6.13	13.79	5.51	12.56	10.49	21.95	1.23	3.79
Qwen2.5-7B-Instruct	8.51	18.82	6.50	13.92	5.29	11.46	10.58	21.68	1.15	3.30
GLM4-9B-Chat	11.85	25.50	6.58	16.03	5.74	13.18	11.91	25.29	**1.50**	**4.39**
Claude-3.5-Sonnet	8.20	20.68	4.57	14.02	4.12	10.31	7.77	19.23	1.05	3.79
GPT-4o	**13.38**	**25.64**	**6.97**	**16.60**	**7.26**	**13.86**	**13.10**	**25.85**	1.35	3.82
DeepSeek-V3	10.06	23.26	4.48	12.24	5.82	12.95	9.69	22.44	0.88	2.40
DeepSeek-R1	2.09	6.23	1.05	3.85	1.58	4.80	1.72	6.25	0.23	1.09

表1：在中文 PlaylistSense 数据集的 10 个视角下，GPT-4o、Claude 3.5 Sonnet、DeepSeek-V3、DeepSeek-R1 和先进开源模型的 HR@1(%) 和 HR@5(%)。粗体表示最佳性能，下划线表示次佳结果。

3.2 观察 #1：大语言模型在播放列表理解方面存在显著局限。

当前的大语言模型对播放列表的理解仍然不足，无论是开源模型还是闭源模型。即使是 GPT-4o 和 Claude-3.5-Sonnet 这样的先进大语言模型，也没有表现出明显的性能优势。

GPT-4o 在基于音乐和基于用户两个视角下几乎都处于领先地位，但其性能仍有限，在其表现最好的音乐文化视角下，HR@1 也仅为 18.94%。GLM4-9B-Chat 在开源模型中表现最佳，在音乐时代视角下取得 16.85% 的 HR@1，为最高分。

Llama-3-8B-Instruct 表现欠佳，凸显了播放列表理解的难度；与 HR@1 相比，其 HR@5 仅有小幅提升。令人意外的是，尽管 DeepSeek-R1 被认为是最先进的大语言模型之一，其排名却倒数第二。

这凸显了当前大语言模型在理解播放列表方面的显著局限，其表现仍远不理想。在真实世界的播放列表平台上，这一挑战更加突出，因为播放列表的数量远远超过 PlaylistSense 中收录的 5,000 个。

3.3 观察 #2：播放列表理解中的偏见。

I. 视角偏差。 大语言模型在不同视角下通常呈现一致的趋势，且在基于音乐 视角下的性能优于基于用户 视角。例如，DeepSeek-V3 在音乐文化中的 HR@1 得分最高，达到 15.70%，但在心理学 中的表现较差，仅为 0.88%。尤其是在心理学领域，大语言模型的表现不佳。

II. 理解深度偏差。 相对而言，大语言模型在理解与音乐相关的表层特征（如音乐文化和音乐时代）时，表现优于理解更深层的用户相关方面（如心理学 和用户情绪）。根据结果，GLM4-9B-Chat 在音乐文化 方面得分最高（HR@1：17.71%），远高于其在心理学 方面的得分（HR@1：1.50%）。这一现象或许揭示了大多数大语言模型通过深层用户心理和情绪来理解播放列表时所面临的局限。我们将在未来的研究中进一步探讨这一问题。

III. 跨维度稳定性偏差。 即使在同一理解类别中，例如基于音乐的视角，大语言模型在不同维度上的性能也会出现显著波动。例如，GPT-4o 在音乐文化上的 HR@1 达到 18.94%，但在

音乐类型上的 HR@1 降至 11.48%，这凸显了大语言模型跨维度性能的不稳定性：它们在某些维度上的优异表现，可能会在其他维度上大幅下滑，从而影响整体稳定性。

4 LMLMs：SongSage

4.1 概述

鉴于音乐文本本质上是一种语言描述形式，我们以经过通用文本预训练的大语言模型为基础。由于当前大语言模型对播放列表的理解能力有限，我们致力于弥合这一差距，并通过有针对性的后训练工作提升其能力。如图 2所示，我们设置了三个依次进行的阶段：持续预训练、指令微调和偏好优化，每个阶段均由精心整理的数据集提供支持。

为此，我们构建了包含数百万条歌词样本的数据集，包括 LyricBank、LyricBank-SFT 以及精心整理的偏好数据集。**LyricBank** 旨在通过生成式持续预训练增强大语言模型对歌词内容的理解，而**LyricBank-SFT**则通过针对歌词和播放列表的九项具体任务，激活模型积累的领域知识。**偏好数据集**有助于进一步优化模型性能，并支持根据特定需求进行定制。我们将详细介绍这三个数据集的组成。

4.2 LyricBank

我们介绍 LyricBank，这是一个预训练数据集，包含 20 万个播放列表（包括歌单名称、标签和歌单），以及网易云音乐 250 万首歌曲的歌词信息（包括歌手、专辑的详细信息，主要是歌词）。

为确保数据质量，我们对数据进行去重，并移除歌词内容极少的曲目，例如纯器乐乐曲。为保持多样性，该数据集涵盖英语、中文、韩语和日语四种语言，比例约为 2:2:1:1。经过筛选，最终收录了 180 万首歌曲的歌词。

为防止在歌词数据上过拟合，我们为 LyricBank 补充了多种通用数据，以保留大语言模型原有的通用知识。图 2 展示了不同语料的占比，包括日常对话、代码片段、数学题和音乐知识数据；歌词数据与其他数据类型的总和之比为 1:4。

4.3 LyricBank-SFT

为进一步利用 LyricBank 中积累的歌词知识，我们开发了 LyricBank-SFT，这是一个指令微调数据集，包含九项关键任务，围绕播放列表和歌词内容的理解与生成展开。

这些任务详见图2，经过精心设计，旨在充分利用 LyricBank 中的歌词知识，涵盖播放列表级和歌词级的理解与生成。任务包括歌单推荐等实用应用，以及根据歌词生成故事、依据指定关键词和风格创作新歌词等创造性任务。为确保数据质量并为模型提供更好的指令，我们采用多轮抽样与检查方法来验证数据。

LyricBank-SFT 包含 77.5 万个样本，涵盖九个任务板块；图2中央展示了各项能力的数据分布。

为缓解过拟合，我们加入 MusicPile-sft（Yuan 等人，2024年）以增强音乐知识，并加入 Infinity-Instruct-0625（Zhang 等人，2024年）以提升数学、编程和指令遵循能力。

4.4 偏好数据集

我们还构建了一个用于优化的偏好数据集。由于收集真实用户偏好数据既困难又涉及伦理敏感问题，我们将更强模型（中文使用 Deepseek，英语使用 GPT-4o）的回答设为偏好回答，将较弱模型（中文使用 Yi-1.5-6B-Chat，英语使用 Llama3.1）的回答设为拒绝回答。针对九项监督微调任务中的每项能力，我们生成了 1,000 个数据点，并在指令微调后使用这些数据进行直接偏好优化（DPO）（Rafailov 等人，2023年）。

5 实验

训练配置。我们使用通用基础模型 Qwen2.5-7B-Base（Yang 等人，2024年）的权重初始化 SongSage-Base，并采用先持续预训练、再进行指令微调的流程，最大序列长度设为 2048。在 LyricBank 上进行持续预训练时，学习率设为 3e-5，每批次使用 0.8M 个词元。遵循 ChatMusician 的做法，我们采用 LoRA （Hu 等人，2022）进行高效的指令微调，学习率为 1e-4，批次大小为 1024。训练使用 16 台



图2：SongSage 三阶段训练概览，详细展示了从 LyricBank 中的通用歌词知识，到使用 LyricBank-SFT 进行特定任务指令微调，再到偏好优化的训练进程，同时说明了知识领域和任务分布。SongSage 的标志使用 GPT-4 创作。

A100 GPU，持续预训练和微调均训练 1 轮。在 DPO 阶段，我们使用 128 的批次大小，并对 LoRA 训练 3 轮。更详细的参数和损失曲线见附录。

5.1 以播放列表为中心的任务评估。

表 2 展示了 SongSage 和 GPT-4o 在三项以播放列表为中心的任务上的性能。在查询改写任务中，我们使用 SongSage 改写来自 PlaylistSense 的查询，并从 Deepseek 为每个播放列表生成的 30 万条黄金描述库中检索结果。与 GPT-4o 相比，改写后的查询与对应黄金描述之间的平均嵌入相似度（嵌入相似度）提高了 28.4%。此外，SongSage 使 Recall@1 和 Recall@5 分别提升了 2.24% 和 3.47%，在这一具有挑战性的推荐场景中取得了显著收益；在该场景中，大语言模型只能访问用户查询。在 1 万个样本上的播放列表到标签和播放列表转描述任务中，SongSage 进一步取得了改进，展现出其理解和捕捉播放列表特征的能力。

5.2 歌词核心任务评估。

对于歌词理解，表 2 展示了在 1万 个样本上对推断歌手、生成歌曲标题和情感检测任务的平均性能。SongSage 在歌词理解方面始终优于 GPT-4。

对于歌词生成任务，我们首先采用以大语言模型为评判者的方法，由 GPT-4o 从流畅度、相关性、创造力和情感表达四个维度评估 SongSage 和 Claude-3.5-Haiku 的结果。详细的评估标准和提示词见附录。表3 展示了 2k 个样本的平均得分，SongSage 的表现优于 Claude-3.5。

此外，我们开展了一项问卷调查，重点关注两项具体任务：歌词续写，以及根据关键词进行歌词创作。每份问卷针对这两项任务各包含 20 个问题。我们邀请参与者评估 SongSage、GPT-4o、Claude-3.5 及其他模型中，哪一个生成的歌词质量最高。共收集了 110 份问卷，参与者包括大学生、中年人士，以及 30 多位音乐专业人士。

表4展示了人工评估结果。SongSage 在两项任务中都获得了显著更高的偏好度：在歌词续写方面，偏好率比 GPT-4o 高 22.0%；在歌词创作方面，偏好度高出 9.9%。这一偏好表明，通过歌词生成式预训练和有针对性的微调，SongSage 能够作为专门的大型音乐语言模型，有效理解现实世界中的歌词偏好，并生成高质量、有意义的歌词。

5.3 对播放列表的理解。

我们使用 PlaylistSense 中的 5,000 条播放列表条目作为输入，供受评估的大语言模型生成描述（与表1相同，但不包括 GPT-4o）。我们采用以大语言模型为评判者的方法，由 GPT-4o 从四个维度选出最佳描述：准确性、情感传达、信息覆盖度和场景适用性。图3展示了大语言模型的性能，其中 SongSage 领先，达到

模型	用户查询改写			播放列表转标签		播放列表转描述
	R@1 ↑	R@5 ↑	嵌入相似度 ↑	精度 ↑	F1 分数 ↑	嵌入相似度 ↑
GPT-4o	11.45	23.25	60.52	4.30	2.30	60.52
SongSage（监督微调）	**13.69**	**26.72**	**88.92**	**45.8**	34.8	**89.51**
SongSage（直接偏好优化）	13.15	24.20	83.64	40.6	**37.9**	86.84

模型	推断歌手		生成歌曲标题		情感检测
	完全匹配 ↑	部分匹配 ↑	F1 分数 ↑	编辑距离 ↓	准确率 ↑
GPT-4o	9.65	1.80	16.3	6.17	31.65
SongSage（监督微调）	**21.55**	**4.70**	**56.37**	**3.41**	32.90
SongSage（直接偏好优化）	18.45	4.20	55.17	3.57	**35.70**

Tabl图 2：SongSage 在以播放列表为中心（上）和以歌词为中心（下）的任务中的结果。最佳分数以粗体显示 d.

模型	创造力	流畅度	Rele	Emo	**Avg.**
关键词转歌词					
Claude-3.5	7.91	**8.83**	**9.30**	8.42	8.61
SongSage	**8.83**	8.74	8.78	**8.90**	**8.81**
歌词续写					
Claude-3.5	5.15	5.91	5.54	5.76	5.59
SongSage	**7.72**	**8.04**	**8.11**	**7.98**	**7.96**
故事生成					
Claude-3.5	5.24	7.44	5.59	5.23	5.92
SongSage	**7.13**	**8.66**	**7.30**	**7.71**	**7.70**

表3：三个生成任务的性能。

模型	歌词续写	歌词创作
GPT-4o	20.08	23.50
Claude-3.5	10.42	12.35
GLM4-9B	17.80	14.67
SongSage	**42.12**	**33.38**
Qwen2.5-7B	9.58	16.01
总计	100.0	100.0

表 4：基于 110 份问卷，对不同模型在歌词续写和创作任务中的偏好比例进行统计；每项任务包含 20 个样本。

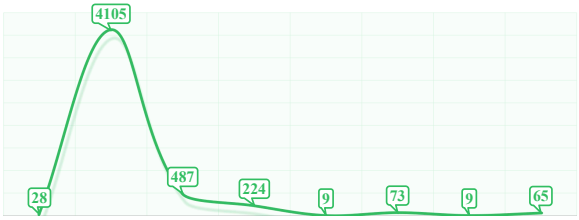

图3：大语言模型生成播放列表描述的表现，GPT-4o作为评判模型。数字表示模型被选中的次数。

共被选中4,105次，其次是Claude-3.5-Sonnet（487次）和InternLM2.5-7B-Chat（224次），凸显了SongSage在理解播放列表信息方面的显著优势。

5.4 音乐知识理解。

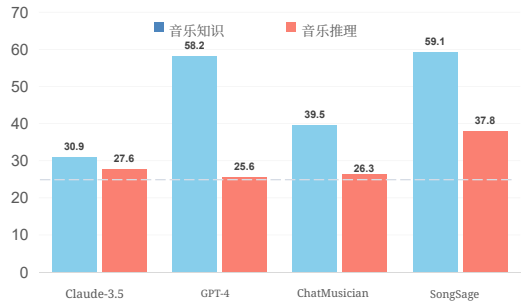

图4：MusicTheoryBench上的准确率（%）。水平虚线表示随机猜测得分。

我们报告了SongSage在MusicTheoryBench（Yuan 等人，2024年）和ZIQI-eval（Li 等人，2024年）上的表现。图4显示，SongSage在音乐知识任务中表现最佳（59.1%），比ChatMusician高出19.6%。值得注意的是，SongSage在音乐推理任务中也取得了显著提升（37.8%），而其他模型的表现仍接近随机猜测水平（25%）。表5比较了

模型	**准确率 (%)**	**F1 分数 (%)**
ChatMusician	24.6	23.9
GPT-3.5-Turbo	50.2	52.9
GPT-4	62.9	63.0
GPT-4o	**76.1**	**76.1**
SongSage	62.4	62.1

表 5：ZIQI-Eval 上的性能比较。

在 ZiQi 评测中的音乐理解测试表现。尽管 GPT-4o 的表现相对不错，

SongSage 的表现优于基础模型 Chat Musician 和 GPT-3.5-Turbo，且紧随 GPT-4 之后；在 ZiQi 评测中，其准确率为 62.4%，F1 分数为 62.1%。与 GPT-3.5-Turbo 和 ChatMusician 相比，SongSage 的准确率分别高出 12.2% 和 37.8%，展现了其理解和处理音乐知识的强大能力。

5.5 通用知识理解

模型	MMLU 得分（%）
ChatMusician-Base	48.5
ChatMusician	46.8
Qwen-2.5-7B-Base	**74.2**
SongSage-Base	64.5
SongSage-Instruct	63.6
SongSage-DPO	63.0

表 6：MMLU 上的性能。

表 6 展示了在大规模多任务语言理解（MMLU）基准测试上的性能（Hendrycks 等人，2021）。SongSage-Base 的 MMLU 得分为 64.5%，低于 Qwen2.5-7B-base 在 MMLU 基准测试上的得分，这是因为 LyricBank 使大语言模型偏向音乐相关知识，导致其通用知识有所损失；但它的表现比 ChatMusician-Base 高出 16.0%。这一提升得益于 LyricBank 和 LyricBank-SFT 纳入了对通用知识的考量。ChatMusician 的表现较弱，可能有两个原因：一是它着重学习 ABC 记谱法这一符号语言，这可能限制其通用知识；二是它基于 Llama2-7B-Base 构建，而后者是一个能力受限的早期模型。

5.6 歌词预训练大语言模型的语言分析

探究歌词持续预训练为大语言模型带来的语言差异将会很有意思。为此，我们使用 spaCy 句法分析器（Neumann 等人，2019年）分析分别由 Qwen2.5-Base 和 SongSage-Base 生成的 1000 首歌词，并考察词性和依存关系分布的句法特征。

I. 词性（POS）分布分析。图 5 显示，SongSage-Base 在名词（4.76% 对比 2.09%）和专有名词（4.76% 对比 2.09%）方面优于 Qwen2.5-Base，表明歌词预训练促进了更加具体、明确的语言表达。此外，SongSage-Base 还具有更高的

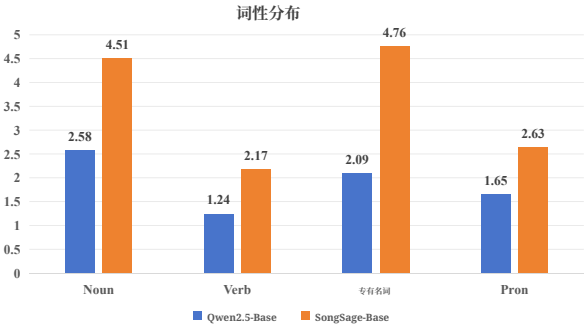

图 5：Qwen2.5-Base 和 SongSage-Base 的词性频率分布。数字表示每个词性在每个句子中的平均出现次数。

动词（2.17% 对比 1.24%）和代词（2.63% 对比 1.65%）频率，这表明其更加注重情感表达和个性化；代词的使用增强了歌词的亲和力和互动性。

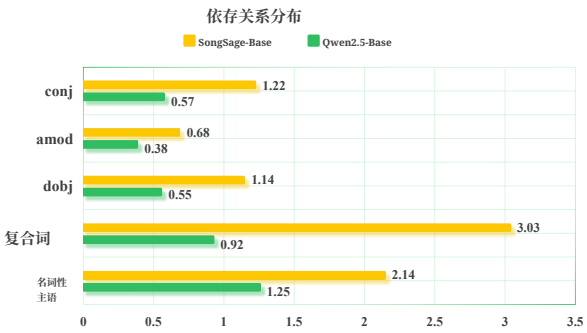

图 6：Qwen2.5-Base 和 SongSage-Base 的依存关系频率分布。数字表示每种依存关系在每个句子中的平均出现次数。

II. 依存结构分析。图6 显示，SongSage-Base 在并列关系（1.22% 对比 0.57%）、形容词修饰语（0.68% 对比 0.38%）和直接宾语（1.14% 对比 0.55%）方面优于 Qwen2.5-Base，表明其更多使用并列结构和情感修饰语。此外，SongSage-Base 中复合词的频率更高（3.03% 对比 0.92%），这表明其语言创造力和灵活性有所提升，这也是歌词创作的典型特征。

6 结论

我们的研究凸显了大语言模型对音乐知识的理解能力。为评估通用大语言模型是否适用于音乐人工智能，我们推出了 PlaylistSense，这是一个旨在评估大语言模型对播放列表和歌词理解能力的数据集。我们的全面评估

表明，大语言模型在有效掌握这些领域方面存在显著不足。为此，我们开发了一种通过歌词生成式预训练构建的大型音乐语言基础模型。我们还分析了大语言模型中歌词预训练的语言特征，探究其与通用大语言模型相比带来的差异。由此，我们推出了 LyricBank，这是一个旨在提升对歌词内容理解能力的大规模歌词预训练数据集；同时推出 LyricBank-SFT，这是一个涵盖九项核心歌词相关能力的指令微调数据集。随后，我们推出了 SongSage，这是一款专注于音乐领域知识的70亿参数音乐语言模型。LyricBank 与 LyricBank-SFT 等资源共同推动了音乐人工智能研究的发展。

局限性

在本研究中，我们利用所有可获取的歌词数据进行生成式预训练，并在资源限制下优化 SongSage 的能力。尽管当前的 70 亿参数模型相比更大型的通用模型展现出更优异的性能，但我们预计，将模型规模扩展至更大尺寸后，其在音乐专属任务中的优势将更加显著。目前，我们的模型支持英语、中文、日语和韩语；扩展至更多语言将进一步推动全球音乐人工智能的发展。展望未来，我们正在积极开发 SongSage 的多模态版本。虽然目前可获取的音乐多模态数据较为有限，但我们正持续扩大数据收集规模。基于 SongSage 在语言理解方面的坚实基础，我们预计，融入更多多模态音乐数据将显著增强模型跨各种模态的综合音乐理解能力。

参考文献

Anthropic。2024年。Claude 3.5：推进人工智能安全与性能。技术报告，Anthropic。

Andres M. Bran、Sam Cox、Oliver Schilter、Carlo Baldassari、Andrew D. White 和 Philippe Schwaller。2024年。借助化学工具增强大语言模型。《自然·机器智能》，6(5):525–535。

Zheng Cai、Maosong Cao、Haojiong Chen 等。2024年。InternLM2 技术报告。*CoRR*，abs/2403.17297。

Jade Copet、Felix Kreuk、Itai Gat、Tal Remez、David Kant、Gabriel Synnaeve、YossiAdi 和 Alexandre Défossez。2023年。简单且可控的音乐生成。收录于《神经信息处理系统进展》第*36*卷：*2023*年神经信息处理系统年会（*NeurIPS 2023*），美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日。

DeepSeek-AI。2024年。Deepseek-v2：一种强大、经济且高效的混合专家语言模型。预印本，arXiv:2405.04434。Zihao Deng、Yinghao Ma、Yudong Liu、Rongchen Guo、Ge Zhang、Wenhu Chen、Wenhao Huang 和 Em-manouil Benetos。2024年。MusiLingo：利用预训练语言模型连接音乐与文本，实现音乐描述生成和查询响应。收录于《计算语言学协会论文集：*NAACL 2024*》，第3643–3655页，墨西哥墨西哥城。计算语言学协会。

Shuangrui Ding、Zihan Liu、Xiaoyi Dong、Pan Zhang、Rui Qian、Conghui He、DahuaLin 和 Jiaqi Wang。2024年。SongComposer：用于歌曲生成的歌词与旋律创作大语言模型。*CoRR*，abs/2402.17645。

Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey 等。2024年。Llama 3 模型家族。*CoRR*，abs/2407.21783。

Zhibin Gou、Zhihong Shao、Yeyun Gong、yelong shen、Yujiu Yang、Nan Duan 和 Weizhu Chen。2024年。CRITIC：大语言模型可以通过工具交互式评议进行自我纠正。载于第十二届国际学习表征会议。

Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi 等。2025a。Deepseek-r1：通过强化学习激励大语言模型的推理能力。*arXiv* 预印本 *arXiv:2501.12948*。

Jiani Guo、Zuchao Li、Jie Wu、Qianren Wang、Yun Li、Lefei Zhang、Hai Zhao 和 Yujiu Yang。2025b。ToM：利用面向树的 MapReduce 实现大语言模型的长上下文推理。载于*2025*年自然语言处理经验方法会议论文集，第 17804–17823 页。

Jiani Guo、Xiangke Zeng、Jie Wu 和 Zuchao Li。2026 年。Dhi：利用多样化的幻觉诱导，增强大型语言模型中的对比式事实性控制。载于《神经信息处理》，新加坡。Springer Nature Singapore。

Dan Hendrycks、Collin Burns、Steven Basart、Andy Zou、Mantas Mazeika、Dawn Song 和 Jacob Steinhardt。2021 年。大规模多任务语言理解的测量。载于第九届国际学习表征会议，*ICLR 2021*，线上活动，奥地利，*2021*年*5*月*3*日至*7*日。OpenReview.net。Zhiqing Hong、Rongjie Huang、Xize Cheng、Yongqi Wang、Ruiqi Li、Fuming You、Zhou Zhao 和 Zhi-meng Zhang。2024a。文本生成歌曲：迈向融合人声与

伴奏。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），*ACL 2024*，泰国曼谷，*2024*年*8*月*11*日至*16*日，第6248–6261页。计算语言学协会。

Zhiqing Hong、Rongjie Huang、Xize Cheng、Yongqi Wang、Ruiqi Li、Fuming You、Zhou Zhao 和 Zhi- meng Zhang。2024年b。文本生成歌曲：迈向融合人声与伴奏的可控音乐生成。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第6248–6261页，泰国曼谷。计算语言学协会。

Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。2022年。LoRA：大语言模型的低秩适配。载于第十届国际学习表征会议，*ICLR 2022*，线上活动，*2022*年*4*月*25*日至*29*日。OpenReview.net。

Atin Sakkeer Hussain、Shansong Liu、Chenshuo Sun 和 Ying Shan。2023年。M^{2}ugen：借助大语言模型强大能力的多模态音乐理解与生成。*CoRR*，abs/2311.11255。

Felix Kreuk、Gabriel Synnaeve、Adam Polyak、Uriel Singer、Alexandre Défossez、Jade Copet、Devi Parikh、Yaniv Taigman 和 Yossi Adi。2023年。AudioGen：文本引导的音频生成。载于第十一届国际学习表征会议，*ICLR 2023*，基加利，卢旺达，*2023*年*5*月*1*日至*5*日。OpenReview.net。Jiajia Li、Lu Yang、Mingni Tang、Chenchong Chenchong、Zuchao Li、Ping Wang 和 Hai Zhao。2024年。音乐大师还是音乐挑战者：面向大语言模型的大规模音乐评测基准。载于计算语言学协会论文集：*ACL 2024*，第3246–3257页，泰国曼谷。计算语言学协会。Ziyang Luo、Can Xu、Pu Zhao、Qingfeng Sun、Xiubo Geng、Wenxiang Hu、Chongyang Tao、Jing Ma、Qingwei Lin 和 Daxin Jiang。2024年。WizardCoder：通过 Evol-Instruct 赋能代码大语言模型。载于第十二届国际学习表征会议，*ICLR 2024*，奥地利维也纳，*2024*年*5*月*7*日至*11*日。OpenReview.net。Ali Madani、Ben Krause、Eric R. Greene、Subu Subramanian、Benjamin P. Mohr、James M. Holton、Jose L. Olmos、Cai ming Xiong、Zachary Z. Sun、Richard Socher、James S. Fraser 和 Nikhil Naik。2023年。大语言模型能够生成跨越多种家族的功能性蛋白质序列。《自然·生物技术》，41(8):1099–1106。Jan Melechovský、Zixun Guo、Deepanway Ghosal、Navonil Majumder、Dorien Herremans 和 Soujanya Poria。2024年。Mustango：迈向可控的文本到音乐生成。载于*2024*年论文集

北美计算语言学协会分会会议：人类语言技术（第*1*卷：长篇论文），*NAACL 2024*，墨西哥墨西哥城，*2024*年*6*月*16*日至*21*日，第8293–8316页。计算语言学协会。

Mark Neumann、Daniel King、Iz Beltagy 和 Waleed Ammar。2019年。SciSpacy：用于生物医学自然语言处理的快速、稳健模型。*arXiv*预印本 *arXiv:1902.07669*。

OpenAI。2024年。GPT-4 技术报告。预印本，arXiv:2303.08774。

Rafael Rafailov、Archit Sharma、Eric Mitchell、Christo- pher D. Manning、Stefano Ermon 和 Chelsea Finn。2023年。直接偏好优化：你的语言模型实际上是一个奖励模型。载于神经信息处理系统进展第*36*卷：神经信息处理系统年度会议 *2023*，*NeurIPS 2023*，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日。

S Sakshi、Utkarsh Tyagi、Sonal Kumar、Ashish Seth、Ramaneswaran Selvakumar、Oriol Nieto、Ramani Duraiswami、Sreyan Ghosh 和 Dinesh Manocha。2024年。MMAU：大规模多任务音频理解与推理基准。预印本，arXiv:2410.19168。

Karan Singhal、Shekoofeh Azizi、Tao Tu 等。2023年。大语言模型编码了临床知识。《自然》，620(7973):172–180。Ahmet Üstün、Viraat Aryabumi 和 Zheng Xin Yong 等。2024年。Aya模型：经过指令微调的开放访问多语言语言模型。载于第*62*届计算语言学协会年会论文集（第*1*卷：长篇论文），*ACL 2024*，泰国曼谷，*2024*年*8*月*11*日至*16*日，第15894–15939页。计算语言学协会。Benno Weck、Ilaria Manco、Emmanouil Benetos、Elio Quinton、George Fazekas 和 Dmitry Bogdanov。2024年。MuchoMusic：评估多模态音频语言模型对音乐的理解能力。预印本，arXiv:2408.01337。Jie Wu、Haoling Li、Xin Zhang、Xiao Liu、Yangyu Huang、Jianwen Luo、Yizhen Zhang、Zuchao Li、Ruihang Chu、Yujiu Yang 和 Scarlett Li。2025a。通过焦点偏好对齐教会模型理解代码。载于*EMNLP 2025* 主会，第14014–14034页。Jie Wu、Haoling Li、Xin Zhang、Jianwen Luo、Yangyu Huang、Ruihang Chu、Yujiu Yang 和 Scarlett Li。2025b。IterPref：通过迭代调试进行代码生成的焦点偏好学习。预印本，arXiv:2503.02783。Jie Wu、Danni Xu、Wenxuan Liu、Joey Zhou、Yew Ong、Siyuan Hu、Hongyuan Zhu 和 Zheng Wang。2024a。评估与引导：多模态虚假新闻检测

通过决策不确定性。载于第*1*届 *ACM* 多媒体基金会模型时代多模态错误信息治理研讨会论文集，第37–44页。计算机协会。

Junda Wu、Zachary Novack、Amit Namburi、Jiaheng Dai、Hao-Wen Dong、Zhouhang Xie、Carol Chen 和 Julian J. McAuley。2024年。Futga：通过时间增强的生成式扩充实现细粒度音乐理解。*arXiv*，abs/2407.20445。

Shitao Xiao、Zheng Liu、Peitian Zhang、Niklas Muennighoff、Defu Lian 和 Jian-Yun Nie。2024年。C-pack：通用中文嵌入的打包资源。预印本，arXiv:2309.07597。

An Yang、Baosong Yang、Binyuan Hui 等。2024年。Qwen2 技术报告。*CoRR*，abs/2407.10671。

Dingyao Yu、Kaitao Song、Peiling Lu、Tianyu He、Xu Tan、Wei Ye、Shikun Zhang 和 Jiang Bian。2023年。MusicAgent：一种利用大语言模型进行音乐理解与生成的人工智能智能体。载于*2023*年自然语言处理实证方法会议：系统演示论文集，第246–255页，新加坡。计算语言学协会。

Ruibin Yuan、Hanfeng Lin 和 Yi 等人。2024年。ChatMusician：使用大语言模型从内在层面理解和生成音乐。载于计算语言学协会论文集：*ACL 2024*，第 6252–6271 页，泰国曼谷。计算语言学协会。

Aohan Zeng、Bin Xu、Bowen Wang 等人。2024年。Chatglm：从 GLM-130B 到 GLM-4 all tools 的大语言模型系列。*CoRR*，abs/2406.12793。

Bo-Wen Zhang、Yan Yan、Lin Li 和 Guang Liu。2024年。Infinitymath：用于程序化数学推理的可扩展指令微调数据集。预印本，arXiv:2408.07089。

Jeffrey Zhou、Tianjian Lu、Swaroop Mishra、Siddhartha Brahma、Sujoy Basu、Yi Luan、Denny Zhou 和 Le Hou。2023年。大语言模型的指令遵循评测。*CoRR*，abs/2311.07911。

Ziya Zhou、Yuhang Wu、Zhiyue Wu、Xinyue Zhang、Ruibin Yuan、Yinghao Ma、Lu Wang、Emmanouil Benetos 和 Yike Guo。2024年。大语言模型能在音乐中“推理”吗？对大语言模型音乐理解与生成能力的评估。*CoRR*，abs/2407.21531。

附录

数据集发布计划

我们的数据集——*PlaylistSense*、*LyricBank*、*LyricBank-SFT*和偏好数据集——的元数据包括播放列表信息（例如标题、标签和歌曲列表）以及歌曲信息（例如歌曲名称、专辑和歌词），其来源为网易云音乐[†]。我们想强调的是，这些数据集仅用于学术目的，不用于任何商业用途。

由于数据版权限制，我们很遗憾地告知，目前无法发布全部数据集。具体而言，对于 Playlist-Sense，我们将公开一部分人工合成的数据，其中包含针对 5,000 个中文播放列表和 5,000 个英文播放列表、从不同角度提出的 10 条查询。播放列表本身不会公开。这部分数据已随附文件提供。所有其他数据集均不会公开。

此外，我们将发布 SongSage 的模型权重，包括 SongSage-Base、SongSage-Instruct 和 SongSage-DPO，以及相应的训练脚本。

PlaylistSense 数据集构建。

PlaylistSense 收录了 5,000 个中文和英文播放列表，均来自网易云音乐。每个播放列表都配有 10 个查询，用于从不同角度评估模型的理解能力。这些查询由先进的大语言模型生成：将播放列表详情（包括标题、标签和歌曲列表）提供给 Deepseek-v2.5（DeepSeek-AI，2024年），为中文播放列表生成查询；为英文播放列表生成查询则使用 GPT-4o。通过针对 10 个角度分别进行提示，引导大语言模型生成模拟人类提问风格的查询。

这 10 个角度涵盖基于音乐和基于用户的视角，反映了用户在现实生活中可能提出的问题。如图7所示，从音乐-时代视角出发，人们可能会问：“我想多听一些披头士时代的音乐，以及一些 20 世纪60年代文化革命时期的歌曲。”从用户-心理学视角出发，一位饱受抑郁困扰的人可能会问：“你能推荐一个能帮助我平复心绪、缓解压力的播放列表吗？”

为确保这些查询的质量，我们采用五样本方法，使用由音乐专业大学生提供的人工评估示例，引导模型生成更自然、更贴近人类表达的查询。

[†]网易云音乐是一个广受欢迎的在线音乐流媒体平台，拥有多样化的播放列表。请参阅 https://music.163.com/。数据仅用于学术用途。



图7：十种类型查询的示例。

对于 PlaylistSense 基准中的第 i 个播放列表，我们采用五样本方法，将播放列表标题、标签以及该播放列表前 20 首歌曲的歌词输入大语言模型（中文使用 Deepseek-v2.5，英文播放列表使用 GPT-4o），从视角 j 生成查询 $\frac{j}{i}$。这可以表示为：

$$\text{query}_i^j = \text{LLM}(\text{instruction}^j, \text{few-shot}^j; \\ \text{title}_i \,||\, \text{tags}_i \,||\, \text{song lyrics}_i)$$

此处，索引 j 对应不同的视角（例如，用户情绪、音乐时代等）。视角 j 的查询集定义如下：

$$Q^j = \{\text{query}_1^j, \text{query}_2^j, \ldots, \text{query}_N^j\}, \\ j \in \{1, 2, \ldots, 10\},$$

其中，N 表示数据集的总规模。

LyricBank数据集构建。
LyricBank 是一个专注于歌曲歌词的持续预训练语料库，旨在增强模型对歌词的理解。为构建该数据集，我们收集了约 250 万首网易云音乐歌曲的歌词信息。

I. 数据抓取。在数据收集过程中，我们首先使用网易云音乐的应用程序接口检索歌曲的元数据，包括歌曲名称、艺术家、专辑信息和歌词文本。由于歌词数据可能存在重复、信息缺失或质量较差等问题，我们进行了严格的清理和筛选。

II. 数据清洗。我们首先删除了重复歌词条目，确保数据集中每首歌曲的歌词都是唯一的。随后，我们剔除了歌词缺失或歌词极少的歌曲，例如纯器乐曲目或仅包含哼唱片段的歌曲。此外，我们还筛除了包含无关字符或格式错误的歌词，以确保数据质量。为保持多样性，我们保留了不同风格、主题和语言的歌词。经过多轮筛选，我们从原始的 250 万首歌曲中保留了约 180 万首。

III. 语言多样性。为了帮助模型理解并用多种语言创作歌词，我们纳入了英语、中文、韩语和日语四种语言的歌词。这些歌词在数据集中的比例约为 2:2:1:1，从而确保语言的丰富性。

IV. 通用数据补充。除了歌词之外，我们还加入了其他类型的通用知识数据，以保持模型的通用理解能力。这些数据包括日常对话、数学题、代码片段等，来源于维基百科、LinkSoul 和 Falcon 等公开数据集。我们将这些通用数据与歌词数据按 4:1 的比例混合，以确保模型在专注于歌词的同时，仍能均衡理解各个领域。

LyricBank-SFT 数据集构建。
LyricBank-SFT 数据集专为指令微调设计，包含 77.5 万个样本，涵盖 9 项与歌词和播放列表相关的任务，旨在进一步提升模型在音乐领域的专业能力。这些任务包括：

1. 用户查询改写（用于播放列表推荐）：将用户关于播放列表的查询改写为对该播放列表的具体一般描述。

2. 播放列表转标签：根据播放列表的标题、歌曲列表和元描述生成标签。

3. 播放列表转描述：根据播放列表的标题、标签和歌曲列表生成描述。

4. 根据歌词讲述故事：根据歌词讲述歌曲背后的故事，例如其创作灵感，以及歌曲所反映的故事。

5. 歌词续写：根据歌词的前半部分续写后半部分。**6. 根据关键词生成歌词：**根据给定的关键词（如音乐类型和特征）创作歌词。**7. 从歌词推断艺术家：**根据给定的歌词推断歌手姓名。**8. 为歌词生成歌曲标题：**根据给定的歌词创作歌曲标题。

9. 情感检测：根据歌词捕捉歌曲的情感。

该数据集通过收集网易云音乐及其他平台的播放列表和歌词数据构建而成，随后进行了针对具体任务的标注和整理。经过质量筛选后，我们选取了 775k 个高质量样本，构成 LyricBank-SFT 数据集。

偏好数据集构建。
偏好数据集旨在优化模型性能。由于收集真实用户偏好数据面临挑战且涉及伦理问题，我们采用了另一种方法。我们使用两个性能水平不同的模型生成样本。性能较高的模型（例如 Deepseek 模型）的输出标记为“偏好”，而性能较低的模型（例如 Yi-1.5-6B-Chat 模型）的输出则标记为“非偏好”。由此，每项任务生成了 1k 个样本，构成偏好数据集。在模型训练过程中，我们使用直接偏好优化（DPO）对模型进行微调，以提升其在各项任务上的性能。

数据集验证
对于大型数据集，人工标注并不现实。为此，我们采用随机抽样的方式验证数据集，选取具有代表性的样本进行人工核验。这些样本旨在体现数据集的整体特征，我们还进行了多轮验证，以确保其质量。此外，数据集的质量会直接影响模型训练的效果。实验结果表明，SongSage 在多项任务上的表现优于其他大语言模型，这也从侧面验证了数据集的质量。

SongSage 中潜在模型偏差分析
SongSage 的开发在文本音乐人工智能领域取得了显著进展，但仍需对模型训练数据和方法可能产生的潜在偏见进行全面审查。鉴于 SongSage 主要使用歌词和播放列表描述进行训练，偏见可能在多个维度上出现：

I. 文化与类型偏差：用于预训练的数据集可能无法同等充分地代表多元音乐文化和音乐类型。这可能导致模型在主流或占比过高的音乐类型上表现更好，而在小众或代表性不足的音乐风格上表现不佳。

II. 歌词内容偏差：歌词往往反映特定的主题、情绪或社会背景。如果训练数据不够均衡，模型可能会对某些情感基调或主题内容产生偏见，进而导致其在情感检测或歌词生成等任务中输出失衡。

III. 用户查询偏差：PlaylistSense 数据集用于评估模型理解用户查询的能力，其生成依据的是预先设定的视角。这可能会无意间引入偏差，涉及模型在训练期间接触到的查询类型或用户意图。

为应对这些潜在偏差，我们计划在未来工作中对训练数据进行详细分析，以识别其中的失衡或代表性不足的类别。此外，我们将实施平衡的数据增强策略，确保模型接触到多样化的音乐流派、文化和歌词主题。这将有助于减轻偏差，并提升模型在不同语境下的公平性和泛化能力。

未来工作与多模态集成
虽然当前版本的 SongSage 在基于文本的音乐任务方面表现出色，但我们也认识到，有必要拓展其能力，以应对更全面的音乐人工智能任务。音乐本质上具有多模态特性，涵盖音频、文本和符号记谱。未来的工作将着重把这些额外模态集成到 SongSage 的架构中，从而增强其对音乐的整体理解。

具体而言，我们计划将音频特征（如旋律、和声）和符号记谱（如乐谱）纳入模型的训练流程。这种多模态方法将使 SongSage 能够执行需要全面理解音乐的任务，例如基于音频特征推荐音乐、生成音乐

视频，以及创作兼顾歌词和旋律语境的新作品。通过利用多模态学习领域的最新进展，我们旨在开发更强大、用途更广泛的音乐人工智能模型，从而更好地满足音乐社区的多样化需求。

总而言之，通过多模态集成应对潜在偏见并拓展 SongSage 的能力，是未来工作的重点。我们致力于开发更加公平、全面的音乐人工智能模型，使其能够有效处理广泛的音乐相关任务。

超参数	设置
批次大小	8
GPU数量	16 (2×8)
训练轮数	1
精度	BF16
最大上下文长度	2048
权重衰减	0.1
预热比例	0.03
学习率	3e-5
梯度裁剪范围	1.0
梯度累积步数	8
优化器	AdamW
学习率调度器	余弦
总步数	7147

表7：预训练期间的超参数。

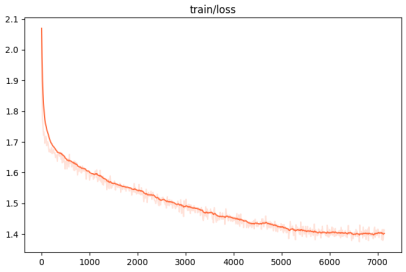

图8：预训练：训练损失曲线。

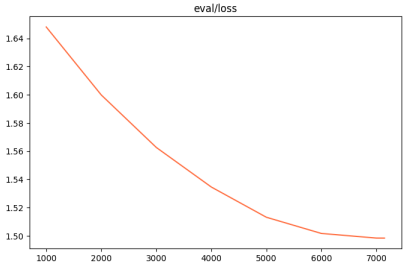

图9：预训练：评估损失曲线。

超参数	设置
批次大小	8
GPU数量	16 (2×8)
训练轮数	1
精度	BF16
最大上下文长度	2048
权重衰减	0.1
预热比例	0.03
学习率	1e-4
梯度裁剪范围	1.0
梯度累积步数	8
LoRA 秩	32
LoRA Alpha	64
LoRA Dropout	0.05
LoRA目标模块	ALL
学习率调度器	余弦

表 8：指令微调期间的超参数。

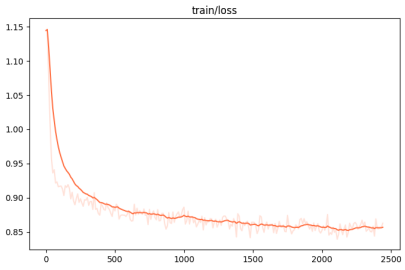

图 10：监督微调：训练损失曲线。

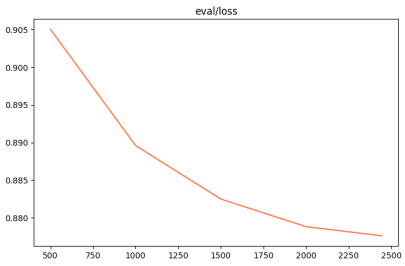

图 11：监督微调：评估损失曲线。

最佳描述评判提示

你是评估中文歌单描述的专家。你的任务是从
给出的八条描述中，根据以下标准选出最佳描述：

- 准确率：描述是否准确反映了歌单名称、标签和歌曲风格?
- 情感传达：描述是否传达了歌单的情感基调和氛围?
- 信息覆盖度：描述是否涵盖了主要风格和代表性歌曲?
- 场景适用性：描述是否展示了适用的使用场景?

请根据这些评估标准，选出最佳的播放列表描述，并
回复对应的字母。

- 以下是播放列表信息：
- 歌单名称是：{playlist_name}
- 歌单元描述是：{playlist_description}
- 歌单标签包括：{playlist_tag}
- 此歌单包含以下歌曲：{songlists}

选项如下：

A：{A}；B：{B}；C：{C}；D：{D}；E：{E}；F：{F}；G：{G}；H：{H}

根据这些评估标准，请选择最佳播放列表描述，并
只需回复对应的字母，无需任何解释。

图12：最佳描述评判提示。

模型	基于音乐的视角									
	文化		情感		Era		场景		Type	
	HR@1	HR@5	HR@1	HR@5	HR@1	HR@5	HR@1	HR@5	HR@1	HR@5
Llama-3-8B-Instruct	1.62	4.18	1.58	4.20	1.80	3.80	1.56	3.66	2.84	6.60
InternLM2.5-7B-Chat	0.02	0.12	0.04	0.12	0.04	0.14	0.02	0.10	0.02	0.20
Qwen2.5-7B-Instruct	3.48	8.48	3.08	7.32	2.90	6.74	2.70	6.64	6.20	13.54
GLM4-9B-Chat	2.82	6.40	2.90	6.48	2.38	5.50	3.18	6.60	6.02	12.98
Claude-3.5-Sonnet	**3.70**	**9.94**	**4.76**	**12.84**	**3.66**	**9.86**	**5.32**	**13.60**	**10.76**	**25.90**
GPT-4o	2.93	9.70	2.93	8.07	2.13	7.67	3.93	11.13	7.13	16.13

模型	基于用户的视角									
	特征		情感		习惯		社会角色		心理学	
	HR@1	HR@5	HR@1	HR@5	HR@1	HR@5	HR@1	HR@5	HR@1	HR@5
Llama-3-8B-Instruct	1.78	3.90	0.72	2.06	1.32	2.88	2.50	5.44	0.44	1.32
InternLM2.5-7B-Chat	0.02	0.08	0.02	0.08	0.00	0.06	0.06	0.14	0.00	0.12
Qwen2.5-7B-Instruct	2.94	7.18	1.18	3.60	2.24	5.32	5.08	10.84	1.06	3.32
GLM4-9B-Chat	3.12	6.10	1.26	3.26	2.06	4.80	4.76	10.00	1.36	3.70
Claude-3.5-Sonnet	**3.16**	**9.94**	**1.74**	**12.84**	**2.54**	**9.86**	**6.28**	**13.60**	1.54	4.34
GPT-4o	2.33	7.47	0.93	4.47	1.67	6.00	4.13	12.07	**2.33**	**6.00**

表 9：GPT-4o、Claude 3.5 Sonnet 和先进开源模型在英语 PlaylistSense 数据集的 10 个视角上的 HR@1(%) 和 HR@5(%)。粗体表示最佳性能，下划线标注的结果表示次优。

评分模板。

请从以下四个关键维度评估歌词续写：

创造力：

歌词在保持歌曲精髓的同时，展现出新颖独到的创意。

续写提出了新的视角，同时不失原有风格。

在文字游戏、隐喻和主题发展方面展现出独创性。

流畅度：

流畅性和节奏保持自然、顺畅的推进

续写部分始终保持一致的拍号与韵律

歌词与原作自然衔接，并展现出很强的可唱性

相关性：

续写部分与原歌词的主题和基调高度契合

意象和叙事方向始终保持一致

主题元素和结构方式展现出很强的连贯性

情感表达：

歌词有效地传达并展现了情感内容

情感历程保留或强化了原有情感

情感呈现展现出真实性和深度

关键词：[keywords]

前半部分歌词：[待续的前半部分]

歌词续写：[待评估的歌词]

图13：歌词评估提示。

数据集	来源	词元	样本数	类别	格式
pile	公开数据集	0.83B	18K	通用	文章
Falcon-RefinedWeb	公开数据集	0.80B	101K	通用	文章
维基百科	公开数据集	0.39B	588K	通用	文章
OpenChat	公开数据集	62.44M	43K	通用	chat
LinkSoul	公开数据集	0.6B	1.5M	通用	chat
GPT4-Alpaca	公开数据集	9.77M	49K	通用	chat
Dolly	公开数据集	3.12B	14K	通用	chat
IrishMAN	公开数据集	0.23B	868K	乐谱	chat
KernScores	公开数据集	276万	10K	乐谱	chat
巴赫四部合唱曲	公开数据集	44万	349	乐谱	chat
合成音乐聊天*	公开数据集	5.4亿	50K	乐谱	chat
音乐知识**	生成	0.22B	255K	音乐语言	chat
音乐摘要**	生成	0.21B	500K	音乐语言	chat
GSM8k	公开数据集	1.68M	7K	math	chat
math	公开数据集	7.03M	37K	math	chat
MathInstruct	公开数据集	55.50M	188K	math	chat
Camel-Math	公开数据集	27.76M	50K	math	chat
arxiv-math-instruct-50k	公开数据集	9.06M	50K	math	chat
Camel-Code	公开数据集	0.13B	366K	code	chat
OpenCoder	公开数据集	36.99M	28K	code	chat
歌词	网易云音乐	1.03B	1.8M	歌词	序列

表10：LyricBank 的分布。