面向描述性语音质量评估的校准-推理框架

Elizaveta Kostenok[1,2], *Mathieu Salzmann* [1], *Milos Cernak* [2]

[1] 洛桑联邦理工学院，洛桑，瑞士
[2] 罗技，洛桑，瑞士

lisakostenok@gmail.com, mathieu.salzmann@epfl.ch, mcernak@logitech.com

arXiv:2603.10175v1 [eess.AS] 10 Mar 2026

摘要

可解释的语音质量评估需要超越平均意见分（MOS），分析其 underlying 感知维度。为此，我们提出了一种新颖的后训练方法，使基础音频大语言模型能够进行多维推理、检测和分类音频瑕疵。首先，校准阶段使模型能够预测预定义的感知维度。其次，强化学习阶段采用结合维度特定奖励的组相对策略优化（GRPO），大幅提升对质量问题的描述准确性及其时间定位能力。通过这一方法，我们在多维 Qual-iSpeech 基准测试上取得了最先进的结果，平均 PCC 分数达到**0.71**，并在基于强化学习的推理驱动下，使 MOS预测提升**13%** 。此外，我们的细粒度 GRPO 奖励显著增强了模型在时间上精准定位和分类音频瑕疵的能力。

索引词：描述性语音质量评估、音频大语言模型、后训练

1. 引言

非侵入式语音质量评估方法主要针对平均意见分数，将其作为最接近人类感知的测量指标。尽管深度学习模型能够以较高的准确性对其进行近似 [1, 2, 3], ，但其“黑箱”式预测缺乏可解释性。为弥合这一差距，研究人员将质量分解为可懂度和自然度等感知维度 [4, 5, 6, 7]。尽管这些方法提供了更细致的视角，但它们局限于定量评估；既无法刻画具体的音频瑕疵类型，也无法在音频录音中定位这些瑕疵。

音频大语言模型（Audio LLMs）近期推动该领域转向可解释平均意见分数——生成描述性评估，以解释这些分数 [8, 9, 10, 11]。除了输出可验证的数值外，这些模型还能够利用自然语言推理对瑕疵进行分类，并定位其出现的时间。然而，当前的可解释平均意见分数系统更重视对话流畅性，而非诊断精度。由于描述质量评估任务并不存在于其预训练混合数据中，它们的推理往往缺乏依据；由于受到幻觉或错误的分维度预测的干扰，其MOS预测准确性通常不及传统的基于分数的方法。若要成为可靠的诊断工具，而非泛化的音频质量总结工具，这些模型需要经过定向对齐，严格确保维度和时间上的准确性。

为弥合这一差距，我们提出一种两阶段后训练方法，包括用于学习逐维度尺度的校准阶段，以及所实施的推理阶段通过带有细粒度、维度特定奖励的GRPO [12] 实现。应用于Audio Flamingo 3模型时， [13], 校准-推理框架[1] 在 QualiSpeech [8] 基准测试上创下新的最先进水平，在逐维度判断以及音频伪影的时间定位方面优于现有基线。

2. 相关工作

当前的描述性语音质量评估方法通常依赖于将用于特征提取的预训练音频编码器与用于文本生成的大语言模型相结合，随后在维度特定标注上进行微调。QualiSpeech [8] 的作者描述了一种两阶段流程（以下简称QualiSpeech-FT），两个阶段均采用监督微调（SFT），先学习各维度得分，再学习长篇描述。虽然第二阶段后音频瑕疵的分类与检测有所改善，但各维度得分预测略有下降。作者将Vicuna [14] LLM主干网络有限的推理能力视为其多模态系统中的潜在瓶颈，因为在根据各维度得分推导平均意见分方面，它的表现不如 GPT-4o-mini [15]。另一种方法SpeechQualityLLM [9], 在 NISQA [4] 数据集上联合微调音频编码器与大语言模型解码器，以自然语言解释依据预测平均意见分和各维度得分。该方法主要侧重于分数评估，将生成的描述视为补充，而未对文本推理的质量进行评估。

为了相对于监督微调改进推理，较新的方法将强化学习（RL）整合到训练流程中。ALLD框架 [10] 采用直接偏好优化（DPO）式 [16] 奖励，同时通过令牌级蒸馏提高音频大语言模型与专家纯文本大语言模型输出之间的对齐。SQ-LLM [11]在对长篇描述进行监督微调之后应用强化学习，使用GRPO优化宽泛的响应级奖励（例如，有用性、准确性）。尽管这对文本推理而言是标准做法，但这种统一的奖励结构并未明确隔离针对各个语音质量维度的反馈。

为了解决这些局限性，我们提出的方法利用与各个质量维度直接关联的奖励信号来显式优化模型。此外，为了避免先前工作中观察到的推理瓶颈，我们在参数规模为70亿至80亿的先进音频大语言模型之上构建了我们的框架。

[1]音频样本和生成的评估可在 https://calibrationreasoningframework.github.io/audiollmdemo/ 获取；模型权重托管于https://huggingface.co/kostl/af3-sqa-grpo，源代码将在发表后发布。

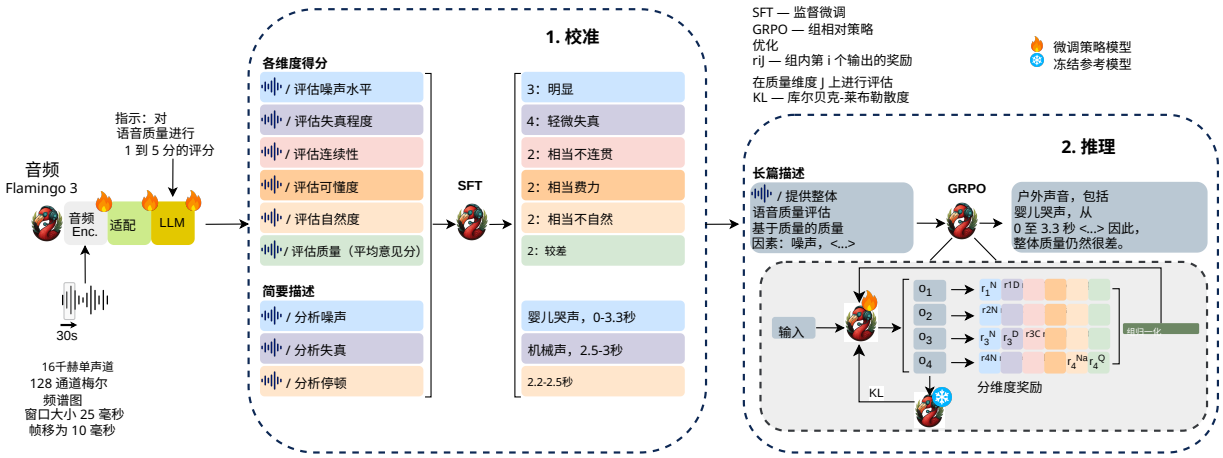


图1： 端到端微调流程概览。

3. 方法

我们的方法如图1所示。它包含两个阶段：校准和推理，我们将在下文中加以说明。

3.1. 校准

首先，我们使用监督微调对模型进行微调，使其能够进行多维度判断。对于每个方面，模型都会学习一个范围为 $[1,\ 5]$, 的质量分数，其中分数越高代表质量越高，并在提示中提供每个分数对应的评判标准。模型通过交叉熵损失优化真实答案的似然：

$$\mathcal{L}^{SFT}(\theta) = -\sum_{t=1}^{T} log p_\theta(y_t|y_{<t}, Emb[S, x_{prompt}])) \quad (1)$$

其中 $y_t, t \in [1, T]$ 为答案词元， $Emb[S, x_{prompt}]$ 为语音和指令的双模态嵌入。

这一阶段有两个目标：将能够区分质量的特征注入音频编码器的潜在空间，并使大语言模型能够高保真地解读这些信号。不同于以往工作 [8, 11]在后训练期间冻结音频编码器，我们校准阶段的核心是使音频编码器可训练，从而提升其对低层语音特征的敏感度。

3.2. 推理

其次，我们使用 GRPO算法微调校准检查点，以汇总分维度预测，并推理整体质量。其工作原理是收集数据、更新策略，然后立即使用更新后的策略来收集更好的数据。在文本生成的语境中，策略指模型在词汇标记上的输出分布 p_θ，数据指一批多模态输入——带有指令的语音录音。对于给定的输入 x，模型会采样一组 G 候选回答，并由定制的奖励机制为每个回答评分。组内奖励的比较将作为训练信号：

$$\hat{A}_i = \frac{R(y_i) - \bar{R}_{\text{group}}}{\sigma_{\text{group}}} \quad (2)$$

ˆ 优势 A_i 的符号和幅度用于衡量第 i 个回答的相对质量。随后更新模型权重，以提高高于平均水平的输出概率，并降低低于平均水平的输出概率。通过对策略模型 p_θ 与冻结参考模型 p_{ref} 的输出分布施加库尔贝克-莱布勒散度惩罚，对损失进行正则化，以防止奖励黑客：

$$\mathcal{L}_{GRPO}(\theta) = -\mathbb{E}_i \left[\log p_\theta(y_i \mid x)\hat{A}_i\right] + \lambda \mathrm{KL}(p_{\text{ref}} \parallel p_\theta) \quad (3)$$

3.2.1. 分维度奖励

GRPO 的有效性依赖于能够评估候选内容相关性的有界奖励函数。我们围绕从生成文本中提取的可验证锚点设计奖励：数值质量分数和简短的伪影描述。

第一种方法依赖结构化的质量描述：我们提取各维度得分，以估计**准确性** 奖励：

$$r_i^{\text{Score}} = \mathbb{1}[s_{\text{pred}} = s_{\text{ref}}] \quad (4)$$

对于附有简短描述的音频瑕疵，我们通过计算$\mathbf{e}$与参考 y_{ref}之间的嵌入余弦相似度，来计算**语义相似度** 。其中，嵌入由句子Transformer模型生成，并线性映射至 $[0, 1]$:

$$r_i^{\text{Semantic}} = \frac{1}{2}\left(\frac{\mathbf{e}_{\text{pred}} \cdot \mathbf{e}_{\text{ref}}}{\|\mathbf{e}_{\text{pred}}\|\|\mathbf{e}_{\text{ref}}\|} + 1\right) \quad (5)$$

通过将这些方法结合起来，我们兼顾严格的事实精确性与更广泛的上下文含义。总奖励 R 是所有维度上的准确性奖励与语义奖励之和。

第二种方法是使用独立的纯文本大语言模型进行多维评分。我们提示该模型按照图2所示评估每个生成结果。总奖励 R 估计为各维度奖励之和。

4. 实验设置

4.1. 数据

我们使用 QualiSpeech 语料库进行训练和评估，因为它包含核心感知维度的定量和定性特征

表1： 不同推理策略下，各维度语音质量评估的皮尔逊相关系数（*PCC*）。“自然度”至“平均意见分”列分别表示 *QualiSpeech* 数据集中的具体感知维度，“平均值”表示各维度皮尔逊相关系数值的平均值。

推理方法	学习编码器	校准	Nat.	噪声	失真	费力度	连续性	MOS	Avg.
基线									
QualiSpeech-FT-SALMONN	x	✓	0.56	0.65	0.57	0.61	0.45	0.63	0.57
QualiSpeech-FT	x	✓	0.61	0.68	0.58	0.60	0.49	0.64	0.60
SQ-LLM	✓	✓	0.71	0.72	0.60	0.65	0.44	0.68	0.63
Ours									
准确性+语义维度级	✓	✓	0.70	**0.78**	0.59	**0.69**	0.55	0.72	0.68
大语言模型评判维度级	✓	✓	**0.73**	0.77	0.68	0.67	**0.63**	**0.76**	**0.71**
消融实验									
QualiSpeech-FT	✓	✓	0.68	0.75	0.61	0.66	0.59	0.67	0.69
QualiSpeech-FT	✓	x	0.55	0.48	0.59	0.50	0.35	0.60	0.51
x（仅校准）	✓	✓	0.67	0.75	**0.69**	0.61	0.55	0.71	0.66

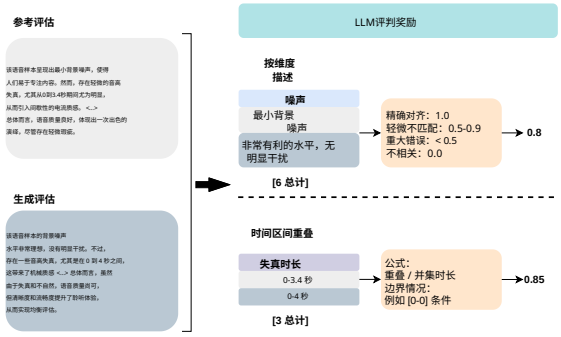


图 2：按维度划分的 *LLM*评判奖励。

。该语料库涵盖 12,450 段语音录音，并附有详细标注（训练/测试集按 85% / 15% 划分）。听众对七个感知维度的评分采用 [1, 5] 量表：语速[2], 噪声、失真、自然度、聆听费力程度、连续性、整体质量（平均意见分）。音频瑕疵的时间区间和简要描述包括：噪声、失真、不自然停顿。最后，所有质量因素都汇总在一段较长的自然语言描述中。

4.2. 模型

我们采用 Audio Flamingo 3 模型，因为在同等规模的开放权重模型中，它在先进的音频感知与推理基准测试 MMAU 上表现出色 [17]。我们还对 SALMONN-7B [18] 模型进行微调，以复现 QualiSpeech 论文中的结果。

4.3. 基线与对齐方法

我们将对平均意见分（MOS）的解释视为一项推理任务，并根据推理方法将对齐方法分为两类：监督微调或强化学习。我们选择 QualiSpeech-FT 和 SQ-LLM 作为基线，因为它们的微调课程与我们的校准-推理框架相当。

在基于监督微调的方法中，第二（推理）

[2]由于数据高度不平衡，故意不对该方面进行评估预测

表2： 不同微调阶段的超参数配置。*“Gr.s.”*、*“B.s.”*、*“L.r.”*分别指 *GRPO* 中的组大小、批量大小和学习率。

微调阶段ge	迭代次数	$G^{r.s.}$	B.s.	L.r.
校准	10k	–	8	1.5e-5
监督微调推理	10k	–	8	1.5e-5
群体相对策略优化推理	200	4	4	5e-6

阶段是针对长篇描述的思维链微调。我们使用AF3基础模型复现QualiSpeech-FT，以评估能力更强的LLM主干网络的影响。我们还比较了冻结与端到端学习的音频编码器以及单阶段流程，作为消融研究。

对于我们基于强化学习的方法，我们在针对长篇描述进行一轮监督微调（SFT）预热后应用GRPO，以稳定模型遵循预期响应结构的能力。我们通过将跨所有维度同时评估“有用性”“相关性”“细节”和“准确性”等广泛属性的**统**一偏好策略，转变为**按维度**策略，从而将我们的细粒度流程与SQ-LLM基线区分开来。在我们的方法中，奖励会被明确映射到Qual-iSpeech的各个质量维度。遵循SQ-LLM的设置，我们使用Qwen3 [19] 作为**LLM评审**。此外，为了研究特定维度优化的作用，我们实现了另一种由**准确性**和**语义相似度**奖励驱动的GRPO配置。这些奖励根据结构化响应中解析出的组件计算得出，为LLM评审提供了一种更易解释且计算效率更高的替代方案。语义相似度嵌入使用all-MiniLM-L6-v2模型 [20]生成。

我们在表2中总结了训练配置。所有模型均采用早停训练，并对所有线性层应用低秩适配 [21]，秩为 $r = 64$。

5. 结果

为评估解释的质量，我们采用 QualiSpeech 论文中的测试方法：提示 GPT-4o [22]从生成的文本中提取可验证的信号，例如分数、简要描述和时间区间。

表1展示了各维度的评估分数 res,

表3： 各种推理策略的描述准确性。我们报告伪影检测的 *F1*分数、由 *GPT-4o* 估算的伪影描述相关性分数（*Corr.*），以及检测到的时间区间与参考时间区间重叠部分的交并比（*IoU*）；长篇描述的 *ROUGE-L* 和相关性分数。

推理方法	简要描述									长篇描述	
	噪声			失真			不自然停顿			ROUGE-L	相关系数
	F1	相关系数	IoU	F1	相关系数	IoU	F1	相关系数	IoU		
基线											
QualiSpeech-FT-SALMONN	0.44	0.49	0.49	0.49	0.67	0.78	0.53	0.40	0.34	0.38	0.75
QualiSpeech-FT	0.56	0.52	0.44	0.73	0.71	0.81	0.56	0.40	0.25	0.40	0.78
SQ-LLM	0.72	0.57	0.56	0.82	0.68	0.74	0.61	0.39	0.33	0.45	0.82
Ours											
准确性+语义维度级	**0.77**	0.68	0.61	0.63	0.70	0.80	0.54	0.42	0.46	0.47	0.80
大语言模型评判维度级	**0.77**	0.75	**0.72**	0.84	0.85	**0.84**	**0.67**	**0.44**	**0.47**	**0.51**	**0.83**
消融实验											
x（仅校准）	0.75	**0.85**	0.67	**0.88**	**0.90**	0.81	0.56	**0.44**	0.45	0.12	0.19

并特别着重于 MOS 预测，因为它是根据各质量维度推导得出的。采用维度级奖励的 LLM 评审方法整体表现最为出色，超越了此前最先进的方法，平均皮尔逊相关系数（PCC）达到最高的 0.71，MOS 皮尔逊相关系数达到 0.76。在基于强化学习的方法中，我们提出的两种维度级奖励策略在 MOS 和平均分数方面均严格优于统一奖励基线（SQ-LLM）。为评估各个方法组件的影响，我们开展了一系列消融研究。首先，考察 SFT 基线发现，将 LLM 主干网络从 SALMONN 升级为 AF3（QualiSpeech-FT），平均 PCC 仅提升了 0.03；而解冻音频编码器则带来了显著的 0.12 提升。这表明，与单纯扩大语言模型规模相比，针对任务对声学特征进行对齐，对于准确推导 MOS 更有帮助。最后，单阶段消融显示，仅推理模型的分维度预测 PCC 显著下降，最多达 0.20。虽然仅校准模型的平均 PCC 达到较高的 0.66，但它在结构上无法生成自然语言解释，正如我们在后文所示。

表3展示了用于表征音频瑕疵的描述性准确度指标。简要描述通过F1分数、GPT-4o估算的相关性以及噪声、失真和不自然停顿的时间定位交并比（IoU）进行评估；长篇描述则使用ROUGE-L [23] 和GPT-4o相关性进行评估。与现有基线相比，我们的维度级奖励方法表现出明显优势。LLM评审维度级模型的整体表现最强，在长篇描述的ROUGE-L和相关性分数上均取得最高值，并且对噪声和不自然停顿的IoU定位最精准。此外，我们的准确性和语义相似度维度级奖励方法表现也很突出，位居第二，优于基于监督微调的方法以及SQ-LLM中的统一奖励。消融研究凸显了两阶段流程的必要性：仅校准模型虽然在简要描述上取得了较高的数值准确性，但由于其结构受到简短回答的限制，在长篇生成任务中表现崩溃。最重要的是，我们的维度级方法保留了校准带来的数值精度提升，成功避免了QualiSpeech和SQ-LLM中出现的性能退化。总体而言，这些发现表明，维度特定奖励可防止模型混淆不同的音频瑕疵，从而提升描述质量和时间定位准确性。

尽管所提出的校准-推理框架显著提升了诊断精度，但它仍存在一定局限。首先，在校准阶段解冻音频编码器会引入相当大的计算开销，相比使用冻结编码器的方法更为明显。其次，我们的细粒度奖励设计本质上依赖于QualiSpeech基准中预定义的瑕疵分类体系；因此，模型可能难以准确推理训练分布中不存在的新型分布外退化（例如超低比特率编解码器瑕疵）。

6. 结论

总之，本研究强调了可解释语音质量评估所需的关键架构与训练组件。我们证明，采用端到端可训练音频大语言模型的两阶段校准-推理流程，能够有效平衡数值精度与丰富的描述能力。使组相对策略优化奖励具备维度感知被证明至关重要：这一细粒度方法在QualiSpeech基准上取得了0.76的MOS皮尔逊相关系数，相较此前的监督微调当前最优提升了13%，同时在分维度预测准确性与长文本描述一致性上达到最高。此外，尽管LLM评审为GRPO提供了最可靠的训练信号，更简单的准确率与语义相似度奖励同样极具竞争力。在保持校准精度的同时释放结构化推理能力，我们的分维度奖励框架对局部诊断被证明高度有效。

未来的工作中，我们计划将这一框架扩展至更广泛的声学领域，包括音乐和空间音频。此外，我们希望通过加入程序化的、基于信号处理的奖励（例如算法化的削波或dropout检测器），增强 GRPO 推理阶段，从而进一步使大语言模型的时序预测以客观的声学现实为依据，减少对计算成本高昂的 LLM评判奖励流水线的依赖。

7. 生成式人工智能使用披露

作者感谢使用 Google Gemini 模型润色文稿（例如查找语法问题），并协助完成表格和图形的 LaTeX 排版。

8. 参考文献

[1] C. K. A. Reddy、V. Gopal和R. Cutler，“DNSMOS：用于评估噪声抑制器的非侵入式感知客观语音质量指标”，载于*ICASSP 2021 - 2021 IEEE*声学、语音与信号处理国际会议（*ICASSP*），2021，第6493–6497页。[2] T. Saeki、D. Xin、W. Nakata、T. Koriyama、S. Takamichi和H. Saruwatari，“UTMOS：面向VoiceMOS挑战赛2022的UTokyo-SaruLab系统”，载于*Interspeech 2022*会议论文集，2022，第4521–4525页。[3] A. Ragano、J. Skoglund和A. Hines，“SCOREQ：基于对比回归的语音质量评估”，神经信息处理系统进展，第37卷，105 729，2024 105 702页 – 。[4] G. Mittag、B. Naderi、A. Chehadi和S. M¨oller，“NISQA：一种利用众包数据集进行多维语音质量预测的深度卷积神经网络自注意力模型”，载于*Interspeech 2021*会议论文集，2021，第2127–2131页。[5] C. K. A. Reddy、V. Gopal和R. Cutler，“DNSMOS P.835：用于评估噪声抑制器的非侵入式感知客观语音质量指标”，载于*ICASSP 2022 - 2022 IEEE*声学、语音与信号处理国际会议（*ICASSP*），2022，第886–890页。[6] A. Kumar、K. Tan、Z. Ni、P. Manocha、X. Zhang、E. Henderson和B. Xu，“Torchaudio-SQUIM：torchaudio中的无参考语音质量和可懂度测量”，载于*ICASSP 2023 - 2023 IEEE*声学、语音与信号处理国际会议（*ICASSP*），2023，第1–5页。[7] W. Wardah、R. P. Spang、V. Barriac、J. Reimes、A. Llagostera、J. Berger和S. M¨oller，“SQ-AST：一种用于语音质量预测的基于Transformer的模型”，载于*Interspeech 2025*会议论文集，2025，第2335–2339页。[8] S. Wang、W. Yu、X. Chen、X. Tian、J. Zhang、L. Lu、Y. Tsao、J. Yamagishi、Y. Wang和C. Zhang，“QualiSpeech：包含自然语言推理与描述的语音质量评估数据集”，载于计算语言学协会第*63*届年会论文集（第*1*卷：长文）。奥地利维也纳：计算语言学协会，2025年7月，第23 588–23 609页。 [在线]。可获取：https://aclanthology.org/2025.acl-long.1150[9] M. Monjur和S. Nirjon，“SpeechQualityLLM：基于大语言模型的语音质量多模态评估”，2025。 [在线]。可获取：https://arxiv.org/abs/2512.08238[10] C. Chen、Y. Hu、S. Wang、H. Wang、Z. Chen、C. Zhang、C.-H. H. Yang和E. S. Chng，“音频大语言模型可以作为描述性语音质量评估者”，载于国际学习表征会议，2025。 [在线]。可获取：https://openreview.net/forum?id=U42TkrEDzb[11] H. Wang、J. Zhao、Y. Yang、S. Liu、J. Chen、Y. Zhang、S. Zhao、J. Li、J. Zhou、H. Sun、Y. Lu和Y. Qin，“Speechllm-as-judges：迈向通用且可解释的语音质量评估”，2025。 [在线]。可获取：https://arxiv.org/abs/2510.14664[12] Z. Shao、P. Wang、Q. Zhu、R. Xu、J. Song、M. Zhang、Y. K. Li、Y. Wu和D. Guo，“Deepseekmath：推动开放语言模型数学推理的极限”，2024。[在线]。可获取：https://arxiv.org/abs/2402.03300[13] S. Ghosh、A. Goel、J. Kim、S. Kumar、Z. Kong、S. gil Lee、C.-H. H. Yang、R. Duraiswami、D. Manocha、R. Valle和B. Catanzaro，“Audio Flamingo 3：借助完全开放的大型音频语言模型推进音频智能”，载于第三十九届神经信息处理系统年会，2025。 [在线]。可获取：https://openreview.net/forum?id=FjByDpDVIO[14] L. Zheng、W.-L. Chiang、S. Ying、Z. Hao、Z. Wu、S. Hwang、H. Xue、Z. Wu、E. P. Xing、J. E. Gonzalez和I. Stoica，“用MT-bench和聊天机器人竞技场评判以大语言模型为评判者”，载于神经信息处理系统进展，2023。[15] OpenAI，“GPT-4o mini：推进高性价比智能”，https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/，2024。[16] R. Rafailov、A. Sharma、E. Mitchell、C. D. Manning、S. Ermon和C. Finn，“直接偏好优化：你的语言模型其实是一个奖励模型”，载于第三十七届神经信息处理系统大会，2023。 [在线]。可获取：https://arxiv.org/abs/2305.18290[17] S. Sakshi、U. Tyagi、S. Kumar、A. Seth、R. Selvakumar、O. Nieto、R. Duraiswami、S. Ghosh和D. Manocha，“MMAU：大规模多任务音频理解与推理基准”，2024。[在线]。可获取：https://arxiv.org/abs/2410.19168[18] C. Tang、W. Yu、G. Sun、X. Chen、T. Tan、W. Li、L. Lu、Z. MA和C. Zhang，“SALMONN：面向大语言模型的通用听觉能力”，载于第十二届国际学习表征会议，2024。 [在线]。可获取：https://openreview.net/forum?id=14rn7HpKVk[19] A. Yang、A. Li、B. Yang、B. Zhang、B. Hui、B. Zheng、B. Yu、C. Gao、C. Huang、C. Lv、C. Zheng、D. Liu、F. Zhou、F. Huang、F. Hu、H. Ge、H. Wei、H. Lin、J. Tang、J. Yang、J. Tu、J. Zhang、J. Yang、J. Yang、J. Zhou、J. Zhou、J. Lin、K. Dang、K. Bao、K. Yang、L. Yu、L. Deng、M. Li、M. Xue、M. Li、P. Zhang、P. Wang、Q. Zhu、R. Men、R. Gao、S. Liu、S. Luo、T. Li、T. Tang、W. Yin、X. Ren、X. Wang、X. Zhang、X. Ren、Y. Fan、Y. Su、Y. Zhang、Y. Zhang、Y. Wan、Y. Liu、Z. Wang、Z. Cui、Z. Zhang、Z. Zhou和Z. Qiu，“Qwen3技术报告”，2025。 [在线]。可获取：https://arxiv.org/abs/2505.09388[20] N. Reimers和I. Gurevych，“Sentence-BERT：使用孪生BERT网络的句子嵌入”，载于*2019*年自然语言处理经验方法会议论文集。计算语言学协会，2019 [ ] 1908 10084 。在线。可获取：https://arxiv.org/abs/ 。[21] E. J. Hu、Y. Shen、P. Wallis、Z. Allen-Zhu、Y. Li、S. Wang、L. Wang和W. Chen，“LoRA：大语言模型的低秩适配”，载于国际学习表征会议，2022。 [在线]。可获取：https://openreview. net/forum?id=nZeVKeeFYf9[22] A. Hurst、A. Lerer、A. P. Goucher、A. Perelman、A. Ramesh、A. Clark和OpenAI，“GPT-4o系统卡”，2024。 [在线]。可获取：https://openai.com/index/gpt-4o-system-card/[23] C.-Y. Lin，“ROUGE：用于摘要自动评估的软件包”，载于文本摘要分支。巴塞罗那，西班牙：计算语言学协会，2004年7月，第74–81页。 [在线]。可获取：https://aclanthology.org/W04-1013/