arXiv:2603.06991v1 [cs.SD] 7 Mar 2026

利用多模态大语言模型面向低资源分类的可解释音频属性自适应发现

Kosuke Yoshimura[**]、*Hisashi Kashima*

京都大学，京都，日本ykosuke@ml.ist.i.kyoto-u.ac.jp,
kashima@i.kyoto-u.ac.jp

摘要

在低资源音频分类的预测建模中，提取高准确率且可解释的属性至关重要。尤其在高可靠性应用中，可解释音频属性不可或缺。尽管由人类驱动的属性发现行之有效，但其低吞吐量成为瓶颈。我们提出一种利用多模态大语言模型（MLLMs）自适应发现可解释音频属性的方法。通过在AdaFlock框架中以多模态大语言模型替代人类，我们的方法显著加快了属性发现。该方法通过提示动态识别显著的声学特征，并构建基于属性的集成分类器。在多种音频任务上的实验结果表明，在大多数评估情形中，我们的方法优于直接的多模态大语言模型预测。整个训练在11分钟内完成，证明其是一种实用的自适应方案，超越了传统依赖人工的方法。**索引术语**：低资源学习、音频分类、属性发现、多模态大语言模型

1. 引言

从复杂的声学信号中提取人类可理解的见解，仍然是音频分析中的一项基本挑战。即使大规模生成模型不断兴起，在低资源场景中，实现高精度预测建模仍离不开可解释的属性工程。在此类场景下，训练参数量庞大的模型往往并不现实；相比之下，利用特定领域音频属性的轻量级模型在实用性和可解释性方面更具优势。此外，在高风险领域，人们对预测的可靠性和可解释性有着强烈需求。在这些领域中，基于明确定义的音频属性构建的透明模型优于黑箱推理。因此，在计算资源和可靠性均受到严格限制的实际环境中，以属性工程为核心的方法仍是一种极为有效的解决方案。

在本研究中，我们探讨一种高效定义并标注能够实现高预测性能的音频属性。传统上，表格数据的特征工程往往局限于现有变量的简单组合，难以提取高度创造性的特征 [1, 2]。相比之下，众包通过利用人类认知能力，能够为非结构化数据（如音频、图像和文本）实现创造性的属性发现 [3, 4]。然而，由人类完成从属性“定义”到“标注”全过程的传统方法面临实际

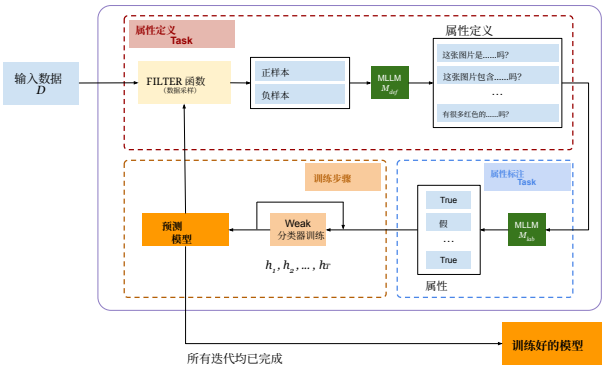

图1： 一个基于多模态大语言模型的属性发现与弱分类器训练迭代框架。根据预测结果自适应地更新采样权重，以发现属性并优化模型性能。

挑战，尤其需要过长的交付周期。为解决这一问题，我们提出一种新方法，利用多模态大语言模型（MLLMs）在短时间内实现可解释音频属性的自适应发现。我们的方法遵循“大语言模型在环”（LLM-in-the-loop）范式，其中多模态大语言模型在形式化算法框架中充当语义预言机，将高层语义判断委托给模型，同时保持符号控制结构。受AdaFlock [4], 启发，我们的方法用多模态大语言模型取代众包工作者，以自动完成音频数据的发现与标注。图1展示了我们方法的整体架构。为证明提出的方法的效用，我们使用四个不同模态的低资源数据集，对预测性能进行了比较实验。此外，为与众包中的交付周期进行比较，我们分析了训练处理时间。

本研究有三项贡献：（1）提出了一种利用多模态大语言模型发现可解释音频属性的自适应方法；（2）在四个音频数据集上的实验结果表明，在低资源场景下，基于属性的方法优于多模态大语言模型直接推理；（3）与众包方法相比，大幅缩短了属性发现和标注的交付周期。

2. 相关工作

本节回顾属性发现的发展历程，从结构化自动化方法逐步转向以人为中心的方法和基于大语言模型的方法，并讨论新兴的LLM在环范式，我们提出的方法正是基于这一范式。传统的自动化方法——

**表示通讯作者。

例如深度特征合成 [1, 2], ——主要面向表格数据，依赖关系结构中的数学变换，因此难以应用于非结构化格式。为处理更多样的数据，Flock [3] 和 AdaFlock [4]等众包框架利用人类认知能力，提取能够弥补模型不足的创新特征。尽管效果显著，高昂的人工成本仍是一大障碍。我们的方法利用多模态大语言模型自动化这一以人为中心的流程，在各种数据格式中高效实现属性发现，同时保持人类水平的可解释性。近年来，CAAFE [5] 等基于大语言模型的方法以及迭代优化脚本 [6]将语义推理引入特征工程。然而，这些最先进的技术仍主要关注表格元数据，缺乏直接分析原始多模态数据的能力。通过使用多模态大语言模型自动化众包中传统上由人工完成的自适应推理，我们提出的方法实现了不受传统数据格式限制的灵活属性发现。我们的方法与查询处理 [7, 8] 和聚类 [9, 10], 中采用的LLM在环范式一致：在这种范式中，大语言模型作为内部模块集成到明确的算法结构中，用于进行语义判断。

3. 问题设定

设 $\mathcal{D} = \{(x_i,\ y_i)\}_{i=1}^{N}$ 为带标签训练数据集，其中每个 $x_i \in \mathcal{X}^{\text{audio}}$ 都是变长音频信号（例如语音或其他声学事件），而 $y_i \in \{0, 1\}$ 是其对应的标签。我们考虑一种低资源音频分类设定，其中带标签训练样本的数量 N较少（例如仅有数百个）。我们的目标是学习一个分类器 H: $\mathcal{X}^{\text{audio}} \to \{0, 1\}$，使其能够泛化到未见过的音频输入。

在此类低资源场景中，直接训练或微调大型端到端模型通常计算成本高昂，而且由于数据稀缺，容易过拟合。除 $\mathcal{D}$外，我们假设可以使用能够处理音频信号的预训练基础模型。我们不对这些模型进行微调，而是利用它们定义一组语义明确的二元属性定义，并为每个音频信号分配属性标签。因此，学习问题是在带标签数据有限的情况下，基于这些由基础模型生成的属性构建一个数据高效分类器。

4. 提出的方法

我们提出一个使用两个多模态大语言模型 $\mathcal{M}_{def}$和 $\mathcal{M}_{lab}$的三阶段框架，以发现数据自适应属性：（1）属性定义：$\mathcal{M}_{def}$ 通过优先考虑当前模型失败的样本来定义新概念。（2）属性标注： $\mathcal{M}_{lab}$根据所定义的属性对每个数据点进行标注，以构建用于预测的数据集。（3）弱分类器训练：在这些发现的属性上训练一个提升模型。重复这一循环即可得到最终的属性集和模型，详见如图1和算法1所示。

4.1. 通过 $\mathcal{M}_{def}$进行采样与属性定义

$\mathcal{M}_{def}$ 充当自主发现引擎，利用开放词汇知识识别潜在判别模式。关键的是， $\mathcal{M}_{def}$ 仅接收分组后的样本（A组和B组），不提供任何明确的类别标签或特定领域的元数据。这确保属性发现过程完全由两组之间的声学对比驱动，

算法1基于多模态大语言模型的自适应属性发现。

要求： 多模态大语言模型 $\mathcal{M}_{def}$（用于定义）和 $\mathcal{M}_{lab}$（用于标注-），训练数据集 $D = \{(x_i,\ y_i)\}_{i=1}^{N}$，终止参-参数 δ_t 和 ϵ，每次迭代中每个标签的样本数 q，每次迭代的属性定义数量 k，迭代次数 T

确保： 集成分类器 H_T

1: 对于每个 $i = 1,\ldots, N$，初始化权重 $w_i^1 \leftarrow 1/N$
2: 初始化集成分类器 $H_0(x) \leftarrow 0$
3: **对于** $t = 1$至 T **执行**
4: $S_F^+ \leftarrow \emptyset,\ S_F^- \leftarrow \emptyset,\ r \leftarrow 0$
5: **当** $|S^+{}_F| < q$ **或** $|S_F^-| < q$ **执行**
6: 样本 $(x,\ y,\ r)$，使用过滤 $(D,\ w^t,\ r,\ \delta_t,\ \epsilon)$
7: **如果** $x = \perp$**，则继续**
8: **如果** $y = +1$ **且** $|S_F^+| < q$，**则**
9: $S_F^+ \leftarrow S_F^+ \cup \{(x,\ y)\}$
10: **否则，如果** $y = -1$ **且** $|S_F^-| < q$，**则**
11: $S_F^- \leftarrow S_F^- \cup \{(x,\ y)\}$
12: $S_F \leftarrow S_F^+ \cup S_F^-$
13: 从 $F_t = \{f_t^1, \ldots, f_t^k\}$ 获取属性定义 S_F 使用 $\mathcal{M}_{def}$
14: 基于 F_t 的标签，使用 $\mathcal{M}_{lab}$
15: 基于 F_t训练弱学习器 h_t
16: $\gamma_t = \left(\sum_{i=1}^{N} \mathbf{1}[y_i = h_t(x_i)]\, w_i^t\right) / \left(\sum_{i=1}^{N} w_i^t\right) - 0.5$
17: $\alpha_t \leftarrow 0.5 \ln((0.5 + \gamma_t)/(0.5 - \gamma_t))$
18: $H_t(x) \leftarrow H_{t-1}(x) + \alpha_t h_t(x)$
19: 对于所有 $i, w_i^{(t+1)} \leftarrow (1 + \exp(y_i H_t(x_i)))^{-1}$
20: 归一化 $w^{t+1} \leftarrow w^{t+1} / \sum_i w_i^{t+1}$
21: **返回** $H_T(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$
22:
23: **函数** 过滤$(D, w^t, r, \delta_t, \epsilon)$
24: $r \leftarrow r + 1$ ▷ 第 t 步的调用次数
25: $\delta_t' \leftarrow \frac{\delta_t}{r\,(r+1)}$
26: $(D',\ w') \leftarrow (D,\ w^t)$ 的随机排列
27: **对于** $i = 0;\ i < \min(N,\ \frac{2}{\epsilon} \ln(1/\delta_t'))$; $i = i + 1$ **执行**
28: $(x,\ y)$ D' 的第 i 个元素
29: **返回** (x, y, r)，概率为 w_i'
30: **返回** $(\perp, \perp, r)$

使模型能够以自底向上、数据驱动的方式推导出显著属性。为启动这一过程， q 正例和 q 负例被分组并提供给 $\mathcal{M}_{def}$。模型的任务是生成 k 描述，以区分这两组，例如使用如下提示：“创建 k 属性来描述A组和B组之间的差异。”该提示确保每条描述都是一个是非问题，这种二元格式与决策树等弱分类器相契合，因为它们通过离散属性划分属性空间。此外，这种方法具有很高的可解释性，因为最终的集成决策可以追溯到一系列直观的语言查询，而这些查询是在没有任何先验标签偏差的情况下发现的。

为有效实施这一属性定义过程，呈现给 $\mathcal{M}_{def}$ 的样本集采用加权采样构建，采样权重依据当前模型的预测性能确定。此采样由“FILTER”函数管理，该函数实现了一种基于拒绝采样的策略，详见算法 1。具体而言，对于从数据集中随机抽取的样本 x_i ，会进行接受检验，以与其权重 $w_i' \in [0, 1]$成正比的概率接受 x_i。只有

被接受的样本才会纳入样本集。根据下文所述提升过程的学习结果，这些权重 w_i' 会被更新，使当前模型预测错误的样本获得更高权重。因此，困难样本会优先纳入所呈现的样本中。这种以错误为中心的采样确保 $\mathcal{M}_{def}$ 聚焦于当前模型的盲点，合成能够针对性地细化决策边界的属性，尤其是模型最容易出错的区域。针对每个属性定义，都会重复这一过程，直至样本集恰好包含 q 个正类样本和 q 个负类样本。

4.2. 通过 $\mathcal{M}_{lab}$进行属性标注

在每次迭代中，我们将属性定义阶段发现的 k 是非问题呈现给所有训练实例，并让 $\mathcal{M}_{lab}$ 判定其真值。为降低计算成本，我们并不分别发起 $N \cdot k$ 次独立查询，而是同时呈现全部 k 问题。这将每次属性标注会话中的多模态大语言模型查询总数限制为 N。

4.3. 通过提升训练弱分类器

我们应用 AdaBoost 框架，利用多模态大语言模型依次发现的属性进行自适应学习。在每次迭代 $t = 1, \ldots, T$中，我们首先在当前权重分布 w^t 下，计算基于多模态大语言模型所定义属性 F_t 训练得到的弱分类器 h_t的边缘 γ_t，其计算方式为 $\gamma_t = (\sum_{i=1}^{N} \mathbf{1}[y_i = h_t(x_i)]w_i^t)/(\sum_{i=1}^{N} w_i^t) - 0.5$，其中 $\mathbf{1}[\cdot]$ 表示示性函数。利用该 γ_t，表示弱分类器 h_t 置信度的权重 α_t被确定为 $\alpha_t = 0.5\ln((0.5+\gamma_t)/(0.5-\gamma_t))$。所得 α_t用于更新临时集成分类器 $H_t($ x $=$ H_{t-1} x $+$ $\alpha_t h_t$ x $)$ $($ $)$ $($ $)$。相应地，每个实例 $w_i^{(t+1)}$ 的权重根据逻辑损失更新为 $w_i^{(t+1)} = (1 + \exp(y_i H_t(x_i)))^{-1}$。此处计算出的权重在下一步被归一化，使其总和为1，并通过 FILTER 函数充当概率分布，以优先采样当前模型误分类的“困难实例”。最后，执行最终预测的集成分类器 $H_T(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$通过对全部 T 次迭代所得弱分类器的加权和取符号来构建。

4.4. 推理

由于提出的方法定义了新的属性，因此必须在推理前为每个实例分配这些属性。因此，对于未见实例，首先使用 $\mathcal{M}_{lab}$ 进行属性标注，然后通过集成分类器 H_T进行推理。尽管这一过程会在推理期间引入多模态大语言模型的延迟，但计算开销高度可并行化，因为每个实例都可以独立标注。此外，可以通过批量处理多个实例来提高查询效率，这与训练期间采用的策略一致，从而确保实际可行性。

5. 实验

在本节中，我们将提出的方法与多模态大语言模型和基于特征的基线方法进行比较，以评估其性能。为确保结果可靠，我们报告所有随机过程在十次独立试验中的平均性能。

表 1：将提出的方法与两个基线（多模态大语言模型预测和 *LR*）进行比较。数值表示均值准确率 ± 标准差（%）；*LR*基线除外，其报告的是单一确定性分数。 Δ 表示我们的方法相较于多模态大语言模型预测基线的绝对提升。粗体 表示各数据集上的最佳性能。

	基线		提出的方法	
数据集	LR	多模态大语言模型预测	Ours	Δ
CREMA-D	70.00	69.00 ± 2.32	**72.45 ± 6.10**	+3.45
RAVDESS	54.00	66.60 ± 1.82	**68.55 ± 2.57**	+1.95
Coswara	**67.50**	48.10 ± 2.00	55.70 ± 2.20	+7.60
ESC-50	**94.00**	88.35 ± 0.88	87.15 ± 1.49	-1.20

我们的实现可在 [匿名链接][1]中获取。

5.1. 数据集

我们使用**CREMA-D** [11], **RAVDESS**[12], **Coswara** [13], 和 **ESC-50** [14], 评估了我们的方法，并将每个数据集调整为二分类任务，训练集和测试集均保持类别均衡（每类 $n = 100$）。对于 CREMA-D（91 名演员）和 RAVDESS（24 名演员），将快乐 指定为正类，而负类则由悲伤、愤怒、厌恶和恐惧（CREMA-D）或害怕 （RAVDESS）组成。对于 Coswara，我们选取质量评分为 1–2 的咳嗽较多.*wav* 样本，将正类 轻度/中度/无症状 标记为正类，将健康 标记为负类，并构建性别均衡的数据集（每类 50 名男性和 50 名女性）。对于 ESC-50，我们将十个类别归为两类：气态流体 （风、吸尘器、飞机、直升机、呼吸）和液态流体 （雨、海浪、倒水、洗衣机、饮水啜饮），并采用分层抽样以保持类别分布。

5.2. 实验设置

模型： 我们采用三种原生音频模型作为提出的方法和多模态大语言模型预测基线的骨干模型：来自 OpenAI 的 *GPT-audio-mini* [15] (gpt-audio-m) 和*GPT-audio* [16] (gpt-audio)，以及来自 Google 的 *Gemini-3-flash- preview* [17] (gemini-3-f-p)。**基线：** 我们将提出的方法与两个基线进行比较：(1) **LR**：使用 CLAP [18] 特征并通过 L-BFGS 求解器（借助 scikit-learn [19]）实现的逻辑回归；(2) **多模态大语言模型预测**：由多模态大语言模型直接进行零样本预测。**评估方案：** 为确保结果不受多模态大语言模型生成随机性的影响，我们将采样温度设为 1。0，并报告十次试验的平均准确率。对于 LR基线，由于 CLAP 特征提取和 L-BFGS 求解器（使用固定初始化）均为确定性过程，因此报告单次代表性得分。**超参数：** 对于我们的方法，所有数据集均设置 $q = 10$ （正类 / 负类样例）、 $T = 10$ （总迭代次数）和 $k = 10$ （每次迭代的属性数）。我们采用决策树桩 [20] 作为弱分类器。

5.3. 性能比较

表 1 展示了我们的评估对比结果。为确保比较公平并隔离框架的影响，我们在多模态大语言模型预测基线和提出的方法中均采用 gemini-3-f-p 作为核心模型。提出的方法在四个数据集中的三个上始终优于多模态大语言模型

[1]论文被接收后将发布代码。

表2： *gemini-3-f-p* 为每个数据集发现的重要性排名前三的属性。

数据集	属性
CREMA-D	说话者的语气通常是正面或积极的吗？；是否说话者听起来较为放松，而非紧张或神经质？；说话者的声音听起来愉快吗？；
RAVDESS	说话人的语气是否更偏向正类-中性情绪，而非负类-中性情绪？；说话人听起来是否兴奋？；是否说话人的声音表现出痛苦或激动的迹象？；
Coswara	咳嗽后是否能听到明显的吸气声——吗？；咳嗽听起来是否持续较长时间或在开始后仍久久不止？；在样本持续期间，咳嗽的频率是否有所增加在样本持续期间？；
ESC-50	环境听起来是否有风或微风？；是否有背景中是否存在流动水的声音？；是否声音中是否包含飞溅声或滴水声？；

预测，在四个数据集中的三个上取得了准确率提升，分别是在Coswara（+7.60%）和 CREMA-D（+3.45%）上的提升。这些结果凸显了我们的方法在增强多模态大语言模型推理能力方面的有效性。

然而，与LR基线的比较揭示了更为细致的性能特征。我们的方法在情感识别任务中明显优于LR，尤其是在CREMA-D（72.45% 对 70.00%）和RAVDESS（68.55% 对 54.00%）上。相比之下，LR基线在其他领域仍表现更优：其在环境声音任务（ESC-50）上的准确率达到94.00%，在医疗音频数据集（Coswara）上达到67.50%，两者均超过了提出的方法的表现。总体而言，在判别信息主要源自语义和概念表征的低资源语音分类任务中，提出的方法表现出显著的有效性。相反，在由低层声学统计特征主导的任务中，连续声学嵌入似乎具有更强的判别能力。

5.4. 定性分析：发现的属性

表2总结了 gemini-3-f-p 提取的属性，其中重要性由可靠性评分 α_t 定义。尽管无法访问标签，多模态大语言模型仍识别出了与真实概念高度语义对齐的属性。对于情感语音（CREMA-D、RAVDESS），它自主捕捉了效价（如“欢快”“愉悦”）和唤醒度（如“兴奋”“痛苦或焦躁”）指标。在 Coswara 中，模型提取了具有临床相关性的标记，例如“可听见的吸气声”和咳嗽频率的时间变化。对于 ESC-50，属性描述了“有风”或“流动的水”等物理质感。这些结果验证了我们的方法能够将原始音频映射到结构化、可解释的语言空间，从而为集成决策提供依据。

5.5. 属性定义中的多模态大语言模型差异分析 tion

表3展示了用于属性定义的不同多模态大语言模型（$\mathcal{M}_{def}$）的影响，其中 $\mathcal{M}_{lab}$ 固定为 gemini-3-f-p。没有单一模型能在所有数据集上始终占据主导地位，不过 gpt-audio 在情绪识别（CREMA-D 上为 75.30%）和环境声音分类（ESC-50 上为 88.10%）方面取得了最高准确率。gemini-3-f-p 和 gpt-audio-m 也展现出

表3： 不同属性定义模型（$\mathcal{M}_{def}$）对准确率的影响。数值表示均值准确率 ± 标准差（%）。 Δ_{max} 表示 $\mathcal{M}_{def}$ 个模型的最高与最低均值准确率之差。粗体 数值表示各数据集上的最佳表现。

	属性定义模型（$\mathcal{M}_{def}$）			
数据集	**gpt-audio-m**	**gpt-audio**	**gemini-3-f-p**	Δ_{max}
CREMA-D	73.55 ± 4.05	**75.30 ± 3.94**	72.45 ± 6.10	2.88
RAVDESS	**68.65 ± 3.81**	66.90 ± 1.63	68.55 ± 2.57	1.75
Coswara	54.20 ± 1.84	53.35 ± 3.13	**55.70 ± 2.20**	2.35
ESC-50	87.25 ± 1.99	**88.10 ± 2.07**	87.15 ± 1.49	0.95

表 4： 不同 $\mathcal{M}_{def}$ 下每个数据集的训练时间为 *[分钟]*（均值 ± 标准差）。 $\mathcal{M}_{lab}$ 固定为 *gemini-3-f-p*。

	属性定义模型（$\mathcal{M}_{def}$）		
数据集	**gpt-audio**	**gpt-audio-m**	**Gemini-3-f-p**
CREMA-D	7.89 ± 1.00	7.72 ± 1.11	8.57 ± 0.81
RAVDESS	9.87 ± 1.56	9.74 ± 1.81	8.84 ± 1.35
Coswara	10.53 ± 1.54	10.30 ± 1.78	8.43 ± 0.84
ESC-50	9.63 ± 1.02	9.50 ± 0.68	9.11 ± 0.92

具有竞争力的表现，分别在 Coswara 和 RAVDESS 上取得最佳结果。这些结果表明，尽管模型能力会影响属性质量，但最佳 $\mathcal{M}_{def}$ 取决于任务的声学特征。值得注意的是，所有数据集上的性能差距都很小，最大差异（Δ_{max}）仅为 2.88%。这种一致性表明，只要定义模型具备足够的多模态推理能力，能够将音频属性转化为描述性文本，我们的框架就能很好地适应不同定义模型的选择。

5.6. 训练时间分析

与传统的人在回路方法相比，提出的方法的主要优势在于能够大幅缩短完成训练所需的时间。如表 4 所示，使用 gpt-audio 作为 $\mathcal{M}_{def}$ 时，Coswara数据集的平均处理时间最长，约为 10.53 分钟；其他配置所需时间更短（7.72–10.53 分钟）。与受招募和响应延迟拖慢的人在回路研究不同，我们纯粹的计算方法消除了这类开销。这种高效性使我们能够在实际可行的时间范围内快速开展试错。

6. 结论

本文提出了一种利用多模态大语言模型自适应发现可解释音频属性的方法。通过将多模态大语言模型融入定义与标注流程，我们的方法能够自动发现适用于低资源任务的判别性属性。在四个数据集上的评估结果表明，我们的方法在三种情况下优于多模态大语言模型预测。值得注意的是，在情感识别任务（CREMA-D 和 RAVDESS）中，该方法还优于使用 CLAP嵌入的 LR基线。对多模态大语言模型变体的分析表明，该框架具有稳健性：不同属性定义模型的最终准确率差异极小。最后，整个流程耗时不到 11 分钟，为依赖人工的属性工程提供了一种实用且可解释的替代方案。

7. 生成式人工智能使用披露

我们仅使用生成式人工智能工具（ChatGPT 和 Gemini）来润色稿件的语法并提升其清晰度。

8. 参考文献

[1] J. M. Kanter 和 K. Veeramachaneni，“深度特征合成：迈向数据科学工作的自动化”，载于 DSAA 2015，2015 年。[2] F. Horn、R. Pack 和 M. Rieger，“用于自动特征工程与选择的 autofeat Python 库”，载于 ECML PKDD 2019，2020 年。[3] J. Cheng 和 M. S. Bernstein，“Flock：混合式众包机器学习分类器”，载于 CSCW 2015，2015 年。[4] R. Takahama、Y. Baba、N. Shimizu、S. Fujita 和 H. Kashima，“AdaFlock：面向人在环预测建模的自适应特征发现”，载于 AAAI 2018，2018 年。[5] N. Hollmann、S. M¨uller 和 F. Hutter，“用于自动化数据科学的大语言模型：推出 CAAFE，实现上下文感知自动特征工程”，载于 NeurIPS 2023，2023 年。[6] J. Nam、K. Kim、S. Oh、J. Tack、J. Kim 和 J. Shin，“利用具备决策树推理能力的大语言模型优化表格数据的特征生成”，载于 NeurIPS 2024，2024 年。[7] Z. Cheng、T. Xie、P. Shi、C. Li、R. Nadkarni、Y. Hu、C. Xiong、D. Radev、M. Ostendorf 和 L. Zettlemoyer，“在符号语言中约束语言模型”，载于 ICLR 2023，2023 年。[8] P. Glenn、P. Dakle、L. Wang 和 P. Raghavan，“BlendSQL：一种可扩展方言，用于在关系代数中统一混合式问答”，载于 ACL 2024，2024 年。[9] Y. Zhang、Z. Wang 和 J. Shang，“Clusterllm：作为文本聚类向导的大语言模型”，载于 EMNLP 2023，2023 年。[10] J. Diaz-Rodriguez，“以摘要作为质心，实现可解释且可扩展的文本聚类”，载于 ICLR 2026，2026 年。[11] H. Cao、D. G. Cooper、M. K. Keutmann、R. C. Gur、A. Nenkova 和 R. Verma，“CREMA-D：众包情感多模态演员数据集”，IEEE 情感计算汇刊，2014 年。[12] S. R. Livingstone 和 F. A. Russo，“瑞尔森情感语音与歌曲音视频数据库（RAVDESS）”，2018 年。[在线]。获取地址：https://doi.org/10.5281/zenodo.1188976[13] D. Bhattacharya、N. K. Sharma、D. Dutta、S. R. Chetupalli、P. Mote、S. Ganapathy、C. C、S. Nori、S. K. K、S. Gonuguntla 和 M. Alagesan，“Coswara：用于 SARS-CoV-2 感染远程筛查的呼吸音与症状数据集”，2022 年。[在线]。获取地址：https://doi.org/10.5281/zenodo.7188627[14] K. J. Piczak，“ESC：环境声音分类数据集”，载于 MM 2015，2015 年。[15] OpenAI，“GPT-audio-mini”，https://developers.openai.com/api/docs/models/gpt-audio-mini，2024 年，访问日期：2026-03-04。[16] ——，“GPT-audio”，https://developers.openai.com/api/docs/models/gpt-audio，2024 年，访问日期：2026-03-04。[17] Google，“Gemini 3”，https://ai.google.dev/gemini-api/docs/models，2024 年，访问日期：2026-03-04。[18] Y. Wu、K. Chen、T. Zhang、Y. Hui、T. Berg-Kirkpatrick 和 S. Dubnov，“通过特征融合与关键词到描述增强实现大规模对比式语言-音频预训练”，载于 ICASSP 2023，2023 年。[19] F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cournapeau、M. Brucher、M. Per-rot 和 ´Edouard Duchesnay，“Scikit-learn：Python 中的机器学习”，机器学习研究杂志，2011 年。[20] Y. Freund 和 R. E. Schapire，“在线学习的决策论推广及其在提升中的应用”，《计算机与系统科学杂志》，1997 年。