解码模糊情绪：音频语言模型中的测试时扩展

Hong Jia[1,2*], Weibin Li[3*], Jingyao Wu[4*], XiaofengYu[2], Yan Gao[5], Jintao Cheng[3], Xiaoyu Tang[3], FengXia[6], TingDang[1]

[1]墨尔本大学，[2]奥克兰大学，[3]华南师范大学[4]麻省理工学院，[5]剑桥大学 [6]皇家墨尔本理工大学

arXiv:2602.03873v1 [cs.SD] 1 Feb 2026

摘要

从人类语音中识别情绪，是实现具有社会意识的对话式人工智能的关键支撑。然而，尽管以往大多数工作将情绪识别框定为类别分类问题，现实世界中的情感状态往往模糊、重叠且依赖上下文，这给标注和自动建模都带来了重大挑战。近期的大规模音频语言模型（ALM）为在没有显式情绪监督的情况下进行细致的情感推理提供了新机会，但其处理模糊情感的能力仍未得到充分探索。与此同时，推理时技术（如测试时扩展（TTS））的进展已展现出在困难自然语言处理任务中改善泛化与适应性的前景，但其与情感计算的相关性在很大程度上仍不明确。在本工作中，我们提出了首个在测试时扩展条件下、面向音频语言模型的语音模糊情绪识别基准。我们的评估在三个著名的语音情绪数据集上，系统比较了八个最先进的音频语言模型和五种测试时扩展策略。我们进一步深入分析了模型容量、TTS与情感模糊性之间的交互，为模糊情绪理解所面临的计算与表征挑战提供了新见解。我们的基准为开发更稳健、具备上下文感知能力且富有情感智能的基于语音的人工智能系统奠定了基础，并突出了弥合模型假设与现实世界人类情绪复杂性之间差距的关键未来方向。

1 引言

理解人类对话中的情绪对于有效沟通和社会福祉至关重要，并且对于心理健康支持、教育和危机响应等现实世界应用日益重要（Elsayed 等人，2024年；Ramakrishnan 和 El Emary，2013年）。自动情绪识别现已部署于虚拟咨询、数字教育和行为健康评估等领域（Abd-Alrazaq 等人，2019年；Vistorte 等人，2024年；Flynn 等人，2020年），在这些领域中，系统必须准确且富有同理心地识别复杂情绪，以确保个人和社区获得积极结果。然而，当前大多数情绪识别方法依赖于简化假设，通常将情绪划分为快乐、悲伤或愤怒等离散类别（Niu 等人，2024年；Feng 和 Narayanan，2024年）。在实践中，情绪往往是模糊的、重叠的或微妙的，会将兴奋与紧张等状态结合起来，或呈现悲伤与沮丧等低强度混合。这些细微的情感表达对标注和建模都具有挑战性，即便对人类专家而言也是如此（Sethu 等人，2019年；Wu 等人，2024b）。因此，现有模型常常产生不可靠或具有误导性的输出，这在服务于代表性不足群体或弱势群体的应用中尤其令人关切。

近年来，大语言模型（LLM），尤其是Qwen2-Audio（Chu 等人，2024b）和Gemini（Google，2024年）等音频语言模型（ALM）的发展，为细致的情绪识别带来了新的机遇。得益于大规模多模态预训练，ALM 能够生成内容丰富、近似人类的情感状态描述，并展现出超越静态类别方法的潜力。然而，这些模型通常并未针对情绪识别进行专门训练。这引出了一个重要的研究问题：通用音频语言模型在没有显式情感监督的情况下，能够在多大程度上识别并推理人类语音中的模糊情感？

在此类情境下，提升音频语言模型情绪智能的一大障碍，是高质量模糊情绪

*同等贡献。

监督学习数据的稀缺性与主观性（Busso 等，2008；Sethu 等，2019；Dudzik 等，2024；Mower 等，2009；Dang 等，2017；Wu 等，2024a）。作为替代方案，测试时扩展（ TTS）（Snell 等，2024；Muennighoff 等，2025）应运而生，可在推理阶段扩展模型性能，而无需进行大量监督微调（Snell 等，2024；Muennighoff 等，2025）。这种适应性使测试时扩展尤其适用于模糊情绪识别等任务，在这类任务中，情绪线索的复杂性和细微差异可能存在显著变化。测试时扩展有望利用模型已有的知识，并在推理过程中适应感知不确定性，从而帮助模型更好地处理情绪模糊性，同时减少对昂贵且难以规模化的数据标注的依赖。鉴于情绪模糊性可能呈现出不同程度的强度和复杂性，另一个关键研究问题是：测试时扩展能够在多大程度上提升模糊情绪识别能力，更重要的是，能否动态适应不同程度的情绪模糊性？尽管现有研究大多聚焦于测试时扩展在自然语言处理任务中的认知智能应用（Bi 等，2024；Liu 等，2025a；Zhang 等，2025），但针对其在模糊情绪识别中的探索仍存在显著空白。

为弥补这一空白，我们提出了首个针对使用音频语言模型进行模糊情绪识别的测试时扩展策略基准。我们的贡献如下：

- 我们推出了一项全面的基准，评估八种最先进的音频语言模型在三个主流语音情绪数据集上的模糊情绪识别能力。
- 我们系统评估了面向音频语言模型的五种测试时扩展策略，从而能够在模糊情绪场景中进行具备上下文感知能力且灵活的推理。
- 我们深入分析了测试时扩展策略的有效性如何与不同程度的情绪歧义及模型规模相互作用，为开发具备情绪智能的音频语言模型提供了新见解。

我们的基准为评估和推进能够稳健应对现实世界中情绪模糊性的情绪识别系统奠定了基础，从而支持在高风险社会应用中打造更安全、更具同理心的人工智能。

2 相关工作

使用大语言模型进行语音情绪识别。基于语音的情绪识别（语音情绪识别）长期以来一直是情感计算的核心议题。传统方法主要侧重于将语音信号分类为离散情绪类别，或预测连续的唤醒度—效价值（Yang 和 Hirschberg，2018年）。大语言模型和音频语言模型的兴起为推进情绪识别开辟了新途径（Feng 和 Narayanan，2024年；Hu 等，2024年）。例如，Feng 等人（Feng 和 Narayanan，2024年）通过使用语音编码器对大语言模型进行微调，实现了基于语音的情绪识别。尽管取得了这些进展，目前的研究仍主要依赖明确无歧义的多数投票标签，往往难以捕捉情感线索中存在的细微差别和多样变化。

模糊情绪识别。 近期研究日益关注模糊情绪识别，并采用多种深度学习方法进行类别情绪识别和维度情绪识别（Wu 等人，2024a；Dang 等人，2017；Wu 等人，2024c），例如长短期记忆网络（Han 等人，2017）和序贯蒙特卡洛方法（Wu 等人，2022）。随着大语言模型的兴起，近期一项研究（Hong 等人，2024年）迈出了使用大语言模型进行模糊情绪识别的第一步。结果表明，这些模型在情绪歧义较低的情况下表现良好，但也凸显了其在高歧义场景中面临的重大挑战。不过，这项工作仅限于文本输入，未探索语音模型。另一项研究探讨了使用具有词元级表示的音频语言模型，并发现音频语言模型中的词元级表示能够解读模糊情绪（Halim 等人，2025）。然而，这项研究仅涉及单一音频语言模型，缺乏全面评估。

测试时扩展。 测试时策略（TTS）无需额外训练即可增强模型推理能力，大致可分为思维链（CoT）提示（Wei 等人，2022），通过引导模型经历中间推理过程，以及带验证器的搜索（Wang 等人，2022）。思维链提示通过将复杂任务分解为连续的思考步骤（Fu 等人，2022），在提升大语言模型（LLMs）的推理能力方面展现出巨大潜力；近期研究还将其应用扩展至多模态

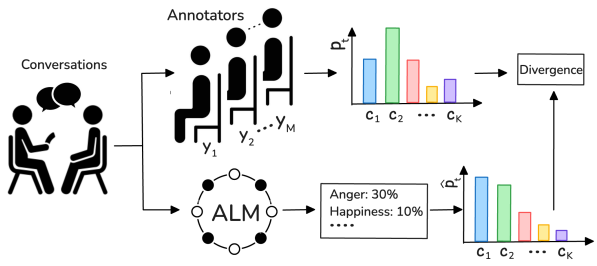


图 1：使用 ALM 进行模糊情绪识别的问题定义。

领域，例如音视频场景理解（Shu 等人，2023年）。验证器搜索方法通常包含两个主要阶段：生成多个候选输出，而不是仅生成单一预测；随后使用验证器或奖励模型选择最优输出。这些验证器可以是简单的启发式方法，也可以是复杂的、经过训练的奖励模型。通常，奖励模型本身是经过微调的大语言模型，用于评估、打分并对输出进行排序，从而识别最佳响应（Adler 等人，2024年）。尽管现有测试时计算方法效果显著，但其主要仍在自然语言处理、编程和数学任务上得到验证（Liu 等人，2025b）。由于缺乏明确的问题定义和合适的奖励模型，这些方法在音频任务，尤其是模糊情感方面的应用仍鲜有探索。

3 问题定义

如图1所示，对于给定的语音话语 $\boldsymbol{x}_t$，我们从 M 位独立人工评定者处获取情绪标注。每位评定者都会分配一个离散标签 $y_t^{(m)} \in \mathcal{C} = \{c_1, \ldots, c_K\}$，其中$\mathcal{C}$ 表示预定义的 K 情绪类别集合。单个话语的完整标注集记为 $\boldsymbol{y}_t = \{y_t^{(1)}, \ldots, y_t^{(M)}\}$。由于情绪感知本质上具有主观性和模糊性，标注者经常意见不一；我们将这种分歧视为情感歧义的指标，针对 $\boldsymbol{x}_t$。

根据这些标注，我们推导出经验情绪分布 $p_t \in \Delta^{K-1}$，作为情绪标签，其中

$$p_t = \text{SoftLabel}(\boldsymbol{y}_t) = \text{SoftLabel}(y_t^{(1)}, \ldots, y_t^{(M)}), \quad (1)$$

并且 Δ^{K-1} 表示（$K-1$）-单纯形：$\mathcal{C}$上的概率向量集合，即 $\sum_{k=1}^{K} p_{t,\,k} = 1$，其中 $p_{t,\,k} \geq 0$。标注者之间的分歧程度通过 p_t的熵来衡量，该熵反映了情绪模糊性的程度。一个 p_t 的熵值越高，表示歧义越大，反之亦然。由此得到一个数据集$\mathcal{D} = \{(\boldsymbol{x}_t,\ p_t)\}_{t=1}^{T}$，其中每个样本都包含语音话语及其对应的情绪概率分布。

本研究的目标有两个：首先，评估 ALM $f_\theta : \boldsymbol{x} \mapsto \hat{p} \in \Delta^{K-1}$ 从语音预测情绪分布的效果。其次，通过 TTS 在推理阶段增强情绪识别和推理能力：模型生成 N 个候选预测 $\{\hat{p}_t^{(1)}、\ldots\ldots、\hat{p}_t^{(N)}\}$，并应用评分函数 ϕ（例如奖励模型）来选择最终输出 $\hat{p}_t$。通过量化模型预测与真实值 $\ell(p_t,\ \hat{p}_t)$ 之间的分布差异，使用詹森–香农散度等指标对预测结果进行评估。散度越低，表示预测越准确。

4 方法

4.1 概述

我们的基准测试框架包含两个主要组成部分：(i) 对现有音频语言模型进行模糊情绪识别基准测试（ALM 基准测试）；(ii) 探究文本转语音在提升模糊情绪识别方面的作用（TTS 基准测试）。第一部分考察多种音频语言模型架构在多个数据集上的表现，重点关注其识别模糊情感表达的能力。第二部分着重评估文本转语音对提升情绪理解的影响，尤其是在不同程度的情绪模糊性下的影响。在音频语言模型基准测试中，我们分析了八种用于模糊情绪识别的最先进的音频语言模型。详细的提示设计见附录。本节其余部分将重点介绍文本转语音基准测试。

4.2 文本转语音方法

我们评估了五种不同的文本转语音方法：思维链提示（Wang 等人，2022）以及四种针对验证器的搜索，如图2所示。

4.2.1 思维链提示

我们引入思维链提示用于模糊情绪识别，利用关于语音情绪线索的先验知识，引导大语言模型进行逐步推理。思维链提示的具体细节详见附录。

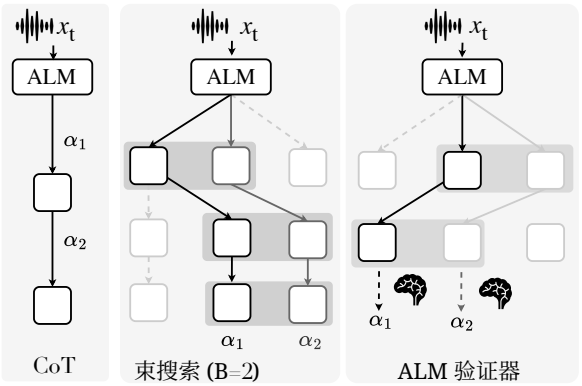


图 2：用于模糊情绪识别的测试时扩展策略。

4.2.2 使用 BestofN 的束搜索

Best-of-N 方法针对给定的音频输入 $\boldsymbol{x}_t$ 生成多个响应。该技术通常采用束搜索来探索多条解码路径。在每个解码步骤中，束搜索保留排名前 B 的候选序列，使模型能够考虑一系列高概率输出，从而同时提升多样性与准确率。

在每个解码步骤中，算法会扩展当前每个序列的所有可能后续词元，并仅保留累积对数概率最高的前 B 个序列。这一过程称为束搜索，会在每个时间步保留一组包含 B 个最可能候选序列的“束”，并迭代扩展这些序列，直至完成解码。解码后，通常会使用奖励模型评估最终的 B 个候选序列的质量。然而，由于目前仍缺少专为 ALM 和情绪任务设计的奖励模型，我们提出使用累积对数似然 α_b，$b \in [1, B]$ 作为评分指标来选择最终输出。

具体而言，我们考虑两种评分策略：i) 最大对数似然：选择对数似然最高的候选项作为 $\hat{p}_t^* = \hat{p}_t^{(\arg\max_b \alpha_b)}$；以及 ii) 加权组合：对束搜索生成的所有输出计算加权和，其中每个输出 $\hat{p}_t^{(b)}$ 均按其归一化对数似然 α_b加权。

为聚合 B 预测的经验情绪分布 $\hat{p}_t^{(b)}$，我们使用狄利克雷混合模型（DMM）。每个预测 $\hat{p}_t^{(b)}$ 都作为狄利克雷分布的均值，其浓度参数为 $\boldsymbol{\beta}_b = \tau\, \hat{p}_t^{(b)}$，其中缩放因子$\tau$ 控制分布的尖锐程度。我们设置 $\tau = 1$，以确保所得分布是形式良好的概率分布（即其各分量之和为 1）。

为体现各项预测的相对重要性，我们使用上述归一化对数似然对狄利克雷分布的各分量进行加权：

$$w_b = \frac{\exp(\alpha_b)}{\sum_{j=1}^{B} \exp(\alpha_j)}. \tag{2}$$

DMM 下的期望聚合分布为：

$$\hat{p}_t^* = \sum_{b=1}^{B} w_b\, \mathbb{E}_{\mathrm{Dir}(\boldsymbol{\beta}_b)}[\boldsymbol{p}] = \sum_{b=1}^{B} w_b\, \frac{\boldsymbol{\beta}_b}{\sum_k \beta_{b,k}} = \sum_{b=1}^{B} w_b p_t^{(b)}, \tag{3}$$

其中 $\mathbb{E}_{\mathrm{Dir}(\boldsymbol{\beta}_b)}[\boldsymbol{p}]$ 表示狄利克雷分布的均值。这在数学上等价于加权求和，但 DMM 提供了一种有依据的方法来聚合预测，同时明确考虑各分布的相对重要性和不确定性。

加权组合使模型能够探索更多可能的输出，从而提高生成内容的多样性。这对于处理复杂或不确定的语音情绪尤其有益，因为此时涵盖广泛的可能输出更具优势。

4.2.3 ALM 验证器

另一种为排名前 B 个束打分的方法，是使用能力更强的 ALM 作为验证器。该验证器经过专门选择，具备更强的音频理解和情绪识别能力，因此能够更有效地评估主模型的响应。

验证器接收音频输入以及每个B 响应，并附有描述评估任务和评分标准的提示。随后，它会为每个响应打分。最终输出选择得分最高的响应；或者，也可以根据验证器的评分对所有响应进行加权组合，方法类似于公式（3）。

5 实验设置

数据集。我们在三个广泛使用的情绪识别数据集上评估所提出的方法：IEMO-CAP（Busso 等人，2008年）、MSP-Podcast（Lotfian 和 Busso，2017）和 CREMA-D（Cao 等人，2014年）。这些数据集涵盖了多种说话风格、情绪表达和标注规范，适合用于研究语音中的模糊情绪识别。

IEMOCAP（Busso 等人，2008年）是一个多模态数据集，包含约 12 小时的演员双人会话视听录音。每条话语均由多位评分者使用类别情感标签进行标注。在我们的实验中，我们聚焦于被标注为四种常见类别（开心、愤怒、悲伤和中性）的话语，共有 4,373 条话语用于分析。

MSP-Podcast（Lotfian 和 Busso，2017）由从播客录音中提取的自然语音片段组成，提供了生态效度更高的情感分布。评估选取了八种情绪，共包含 12,955 条话语。CREMA-D（Cao 等人，2014年）是一个受控数据集，由 91 名演员在不同强度下表达六种基本情绪的 7,400 多个视听片段组成。每个片段均由众多评分者进行标注。

实现细节。共评估了八种不同的音频语言模型，包括 Audio-Flamingo 2（Ghosh 等人，2025）、Qwen2.5-Omni（Xu 等人，2025）、Qwen2-Audio-Instruct（Chu 等人，2024a）等开源模型，以及 Fixie Ai（Fixie AI，2024a，b）推出的 Ultravox 系列的两个版本；此外还包括 Gemini2.5-pro 和 Gemini2.0-flash（Team 等人，2023年）以及 GPT-4o（Achiam 等人，2023年）等闭源模型。关于这些模型结构的详细信息请参见附录。

我们评估了五种文本转语音方法：思维链、*Best-of-N*（*BoN*），即从束搜索中选取得分最高的候选项；加权 *BoN*（*W-BoN*），定义见公式（3）；*ALM* 验证器（*ALM-v*），鉴于 GPT-4o 在通用自然语言处理任务中表现出色，我们使用它来选择最佳回答；加权 *ALM* 验证器（W-ALM-v），使用更强大的 ALM 的分数作为权重。我们使用 $B = 5$ 进行*BoN* 和*W-BoN*，并使用 $B = 3$ 进行*ALM-v* 和 *W-ALM-v*，束宽在$[2, 7]$[1]范围内优化。

评估。核心评估指标衡量预测的情绪分布与真实值情绪分布之间的相似度。我们使用 Jensen-Shannon 散度（JS）、巴塔恰里亚系数（BC）和R^2 作为主要评估指标。JS 散度量化预测分布与实际分布之间的差异（越低越好），BC 衡量分布相似度（越高越好），而 R^2 则评估拟合优度（越高越好）。所有指标的取值范围均为 0 到 1。尽管模型针对模糊情感的分布进行了优化，但良好的预测也应使主导情感与标注的多数投票结果一致。因此，我们还将预测的主导情感（例如，预测概率最高的情绪）与标注者的多数投票结果进行比较，使用准确率和 F1分数进行评估。

[1]代码将在论文录用后公开。

6 结果

6.1 模糊情绪识别

表1 展示了八种开源和闭源模型在模糊情绪识别任务上的表现。闭源模型的表现通常优于开源模型。Gemini 2.5 Pro 在几乎所有数据集和指标上都持续取得最佳或接近最佳的结果。值得注意的是，除 IEMOCAP 上的 BC 和 R^2 以及 CREMA-D 上的 JS 外，它在所有数据集和指标上均取得最佳结果；在这些例外情况下，它取得了第二好的分数。GPT-4o 和 Gemini 2.0 Flash 的表现也很出色；GPT-4o 在 CREMA-D 上创下了新的最低 JS（0.2935），表明其在该数据集上的类别分布匹配最为接近。

在开源模型中，Qwen2-Audio（7B）表现最佳，在 IEMOCAP、MSP-Podcast 和 CREMA-D 数据集上取得了最低的 JS 和最高的 BC。值得注意的是，它在 IEMOCAP 上取得了最高的 BC（0.6477），甚至超过了闭源模型。Qwen2.5-Omni（7B）在开源模型中，于 IEMOCAP、MSP-Podcast 和 CREMA-D 数据集上取得了最高的 R^2，这表明它更擅长捕捉这些数据集固有的方差。相比之下，Ultravox-v0.3（8B）和 Ultravox-v0.4（8B）在所有指标上的表现相近，但总体上都不及 Qwen2 系列模型。

尽管 Gemini 2.5 Pro 表现出色，Qwen2-Audio 仍在 IEMOCAP 和 MSP-Podcast 上缩小了差距，这表明高性能开源替代模型正变得越来越有竞争力。另一个值得注意的现象是，以十亿参数为单位衡量的模型规模，并不总能直接转化为更好的性能；7B 的 Qwen2 系列模型相较于规模更大的 8B Ultravox 系列模型，取得了具有竞争力的结果，便是一个例证。

总之，这些结果再次证实，当前模型在一定程度上能够识别模糊情感，闭源和开源

模型	IEMOCAP			MSP-Podcast			CREMA-D		
	JS ↓	BC ↑	R^2 ↑	JS ↓	BC ↑	R^2 ↑	JS ↓	BC ↑	R^2 ↑
开源模型									
Audio-Flamingo 2 (3B)	0.4920	0.5282	0.4151	0.6206	0.3908	0.2008	0.4539	0.5672	0.4756
Qwen2-Audio（7B）	0.3592	**0.6477**	0.3559	0.4128	0.6123	0.3601	0.3172	0.7051	0.3959
Qwen2.5-Omni（7B）	0.3855	0.5794	0.4664	0.4957	0.5039	0.3689	0.4159	0.6078	0.4875
Ultravox-v0.3（8B）	0.3694	0.5843	0.3973	0.4499	0.5725	0.3425	0.4438	0.5899	0.4296
Ultravox-v0.4（8B）	0.3717	0.5831	0.3956	0.4485	0.5758	0.3448	0.4400	0.5927	0.4336
闭源模型									
GPT-4o	0.3385	0.6217	0.4837	0.3737	0.6584	0.3964	**0.2935**	0.7434	0.4876
Gemini 2.0 Flash	0.3501	0.6163	**0.5148**	0.3980	0.6446	0.4057	0.3439	0.7082	0.4837
Gemini 2.5 Pro	**0.3378**	0.6332	0.5056	**0.3710**	**0.6716**	**0.4148**	0.2983	**0.7520**	**0.5296**

表1：三个基准数据集上用于模糊情绪识别的 ALM 对比。每列中的最佳和次佳结果分别以**粗体**和下划线标示。

方法	IEMOCAP JS ↓	IEMOCAP BC ↑	IEMOCAP R^2 ↑
不使用文本转语音	0.3592	0.6477	0.3559
CoT	0.3578（+0.4%）	0.6409(-1.0%)	0.3792（+6.5%）
BoN	0.3450 (+4.0)	0.6781 (+4.7)	**0.3815 (+7.2)**
W-BoN	**0.3364 (+6.3)**	**0.6915 (+6.8)**	0.3742 (+5.1)
ALM-v	0.3437 (+4.3)	0.6817 (+5.2)	0.3794 (+6.6)
W-ALM-v	0.3433 (+4.4)	0.6853 (+5.8)	0.3749 (+5.3)

方法	MSP-Podcast JS ↓	MSP-Podcast BC ↑	MSP-Podcast R^2 ↑
不使用文本转语音	0.4128	0.6123	0.3601
CoT	0.4393 (-6.4)	0.6138 (+0.2)	0.3179 (-11.7)
BoN	0.4600 (-11.4)	0.5636 (-8.0)	0.3410 (-5.3)
W-BoN	0.3960 (+4.1)	**0.6207 (+1.4)**	0.3642 (+1.1)
ALM-v	0.2949 (+28.6)	0.5945 (-2.9)	**0.3763 (+4.5)**
W-ALM-v	**0.2845 (+31.1)**	0.6110 (-0.2)	0.3669 (+1.9)

方法	CREMA-D JS ↓	CREMA-D BC ↑	CREMA-D R^2 ↑
无文本转语音	0.3172	0.7051	0.3959
CoT	0.3310 (-4.4)	0.7022 (-0.4)	0.3549 (-10.3)
BoN	0.3323 (-4.8)	0.6849 (-2.9)	0.4219 (+6.6)
W-BoN	**0.2908 (+8.3)**	**0.7286 (+3.3)**	0.4559 (+15.2)
ALM-v	0.3066 (+3.3)	0.7139 (+1.2)	**0.4645 (+17.3)**
W-ALM-v	0.2972 (+6.3)	0.7245 (+2.8)	0.4500 (+13.7)

表 2：使用 Qwen2-Audio 在三个基准数据集上比较 TTS方法。每个数据集的最佳和次佳结果以**粗体**和下划线标出。相对变化以百分比 (%) 的形式显示在括号中。

模型都展现出巨大的潜力。

6.2 文本转语音性能

表2 比较评估了 Qwen2-Audio 使用不同测试时扩展策略的表现；在开源模型中，Qwen2-Audio 的性能最佳。由于闭源模型无法提供文本转语音中束搜索所需的推理灵活性，因此未纳入考虑。总体而言，测试时扩展策略会显著影响模糊情绪识别，但其效果因数据集而异。在 IEMOCAP 上，BoN 在所有指标上均持续优于基线；而思维链的结果则不一，虽改善了 JS 和 R^2，但 BC 略有下降。

在所有方法中，加权 BoN 和加权 ALM 验证器在各项指标和数据集上均表现出色且稳定。尤其是，加权 BoN 在全部三个数据集上都取得了最高的 BC，相对提升最高达 6.8%。在 CREMA-D 上，它的 R^2 也排名第二。此外，它在 MSP-Podcast 和 IEMOCAP 上分别将 JS 提高了 4.1% 和 6.3%。

加权 ALM 验证器在 MSP-Podcast 上取得了最低的 JS 散度，相对降幅为 31.1%，并且在其他指标上也持续带来提升。在 IEMOCAP 上，它取得了第二好的成绩。同样，在 CREMA-D 上，它在两项指标中均排名第二。

有趣的是，未加权 BoN 在 MSP-Podcast 上表现出性能下降。这可能反映了该数据集情绪表达的自发性和细微性，因为选择单一“最佳”束无法捕捉完整的分布。相比之下，聚合策略（W-BoN 和 W-ALM-v）大幅扭转了这一趋势。例如，W-BoN 实现了最佳的总体平衡，在 CREMA-D 上取得最低的 JS （ 0.2908 ），并在 MSP-Podcast 上取得最高的 BC（0.6207）。这证实了，根据

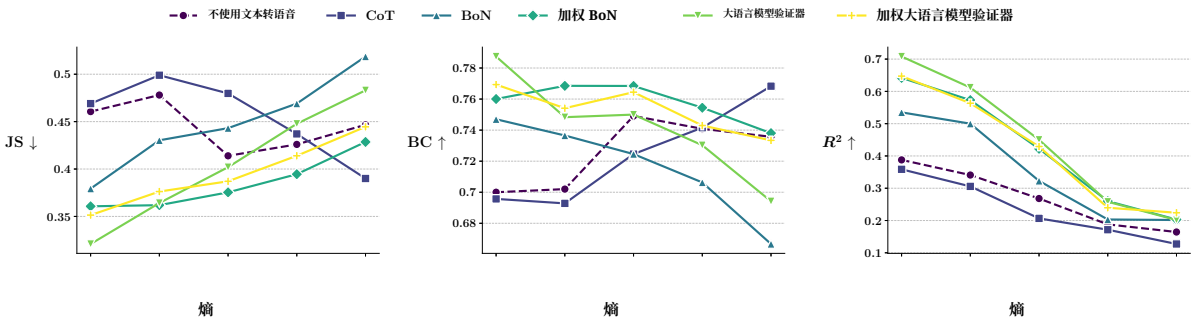


图3：T　S 在 CREMA- 上不同歧义程度（从低到高）下的表现，以熵量化　D.

置信度分数对多个合理输出进行加权，而不是将其舍弃，对于建模情绪模糊性至关重要。

这些发现表明，仅依赖单一预测可能无法捕捉人类情绪细腻而多面的特性；然而，纳入多个不同的输出能够使模型更好地应对这种歧义。总体而言，性能的整体提升表明，通过探索更广阔的潜在空间来汇总多个合理回答，可以显著提升模型性能，因为这样能够在推理过程中充分利用预训练期间学到的基础知识。这种方法对于应对具有不同程度不确定性的模糊情绪识别尤为重要。

6.3 不同歧义程度下的文本转语音

为了更好地理解文本转语音如何提升不同歧义程度下的情绪识别效果，我们进一步评估了其在不同歧义程度下的表现。歧义程度根据标签分布的熵确定，熵越高，表明情绪越模糊，标注者分歧也越大。

根据熵值，歧义程度被划分为五个等分组。我们使用 CREMA-D 数据集比较了各歧义程度下的文本转语音性能，如图3所示。此分析报告各项指标的中位数，以减少偶发异常值对模型整体性能评估的影响。

JS 散度。不使用文本转语音的基础模型在低熵（模糊程度较低）的情绪上表现出乎意料地差，而在模糊程度较高的情绪上表现更好。这可能归因于 Qwen2-Audio 的训练方案：该方案通过在训练数据中纳入 30 多种不同类型的音频（包括非语音声音和音乐），着重于“通用音频理解”及相关任务。如此多样化的数据可能会带来更丰富的情绪表征学习。对于大多数文本转语音方法，JS 散度通常会随着熵的增加而上升，这表明随着情绪歧义的增加，模型预测与真实分布的偏差会进一步扩大，凸显了高熵数据的处理难度。

值得注意的是，尽管整体呈上升趋势，加权 BoN 和加权大语言模型验证器在所有熵值水平下始终优于基础模型。这种稳定的提升表明，利用多条束搜索路径能够提高推理质量。有趣的是，思维链呈现出不同于其他文本转语音方法的模式。随着熵值增加，其优势显著增强，这表明思维链中的逐步推理能够更有效地解读模糊情感。因此，思维链在高熵情境下表现最佳，但在低熵语境中的效果较差。

总体而言，结果表明，各种文本转语音方法都能提升对所有熵值区间内模糊情感的识别能力，但提升幅度因所用的文本转语音方法和熵值水平而异。

BC 与 R^2。对于 BC，我们观察到与 JS 散度相同的结论。关于 R^2，所有 TTS方法均呈下降趋势，反映出随着熵升高，准确估计分布的难度增加。尽管如此，加权 BoN、大语言模型验证器以及加权大语言模型验证器始终能相对基础模型带来提升。不同歧义水平下音频语言模型的详细性能见附录。

6.4 单类别情感识别

为进一步分析预测分布如何保留主导情感，我们将预测的

主导情感与标注中的多数投票结果进行了比较。预测的主导情感被确定为预测分布中概率最高的类别。

音频语言模型。 表 3 展示了 ALM 在不使用 TTS 时的表现。Gemini 系列和 GPT-4o 等闭源模型在各个数据集上的表现始终优于其他模型。Gemini 2.5 Pro 在 IEMOCAP 和 CREMA-D 数据集上的准确率最高，而 GPT-4o 在 MSP-Podcast 上表现出色。这一观察与总体分布预测一致，说明这些模型预测的情绪分布也能够捕捉主导情感。

在开源模型中，Qwen2.5 表现突出，而 Qwen2-Audio 在识别 IEMOCAP 和 MSP-Podcast 数据集中的主导情感时表现不佳。这表明不同版本的 Qwen 具备不同的能力：Qwen2.5 更擅长检测主导情感，而 Qwen2-Audio 对整体情绪分布的感知能力更强。

值得注意的是，在 CREMA-D 数据集上，若干模型取得了尚可的准确率，但宏平均F1分数却极低。例如，GPT-4o 的准确率达到 0.5418，但 F1 分数仅为 0.1231。我们的分析表明，这种差异是由于模型未能预测“厌恶”情绪，而该情绪占数据的 9.1%。这说明目前所有受评估的 ALM 都缺乏识别“厌恶”的能力，凸显了未来需要改进的方向。有趣的是，Qwen2-Audio 是唯一在预测“厌恶”方面取得部分成功的模型，因此在 CREMA-D 上获得了最高的 F1 分数。

带 TTS 的音频语言模型。表4显示，不使用 TTS 的基础模型在准确率和 F1 分数上始终表现不佳。大多数 TTS 方法都优于基础模型，其中 W-BoN 和 W-LLM 验证器这两种加权方法取得了最佳的整体结果。具体而言，W-BoN 在 IEMOCAP 上取得了最高的 F1 分数（0.3688），并在 MSP-Podcast 上同时取得了最高的准确率（0.4225）和 F1 分数（0.2373）；W-LLM 验证器则在 CREMA-D 上取得了最佳的 F1 分数（0.3630）。然而，CoT 和 BoN 等方法有时会逊于基线模型，尤其是在 MSP-Podcast 数据集上；该数据集包含更自然的情绪表达。这表明，当这些为探索情绪模糊性而优化的推理策略受到单一

	IEMOCAP		MSP-Podcast		CREMA-D	
模型	Acc	F1	Acc	F1	Acc	F1
开源模型						
Audio-Flamingo 2	0.3356	0.3071	0.2509	0.1217	0.5216	0.1715
Qwen2-Audio	0.2913	0.2061	0.3830	0.2142	0.3670	**0.3289**
Qwen2.5	0.4577	0.4246	0.4080	0.2643	0.5402	0.1341
Ultravox-v0.3	0.3894	0.3394	0.3902	0.2348	0.4676	0.1224
Ultravox-v0.4	0.3878	0.3422	0.3969	0.2379	0.4747	0.1238
闭源模型						
GPT-4o	0.4460	0.4227	**0.6236**	**0.3657**	0.5418	0.1231
Gemini 2.0 Flash	0.4580	0.4417	0.4337	0.2666	0.5357	0.1247
Gemini 2.5 Pro	**0.4697**	**0.4588**	0.4372	0.2885	**0.5927**	0.2555

表3：不同 ALM 在单类别情感分类上的性能比较。

	IEMOCAP		MSP-Podcast		CREMA-D	
方法	Acc	F1	Acc	F1	Acc	F1
不使用文本转语音	0.2913	0.2061	0.3830	0.2142	0.3670	0.3289
CoT	0.3310	0.3218	0.3357	0.1896	0.2999	0.3091
BoN	**0.3729**	0.3637	0.3461	0.2093	0.4391	0.3258
W-BoN	0.3669	**0.3688**	**0.4225**	0.2373	**0.5324**	0.3319
ALM-v	0.3386	0.3186	0.4171	**0.2533**	0.5126	0.3523
W-ALM-v	0.3431	0.3267	0.4131	0.2522	0.5230	**0.3630**

表4：使用 Qwen2-Audio 进行单类别情感分类时各种文本转语音方法的性能比较。

标签分类任务的限制时，其优势可能无法有效发挥，甚至可能造成干扰。总之，这些结果表明，尽管加权聚合策略能够持续提升不同数据集上的情感分类表现，但不同 TTS 方法的效果仍高度取决于具体任务和数据特征。

7 结论

在本研究中，我们系统评估了多种基准数据集上音频语言模型（ALM）及多种测试时扩展方法在模糊情绪识别方面的表现。结果清楚地表明，音频语言模型在一定程度上能够识别模糊情感，且闭源模型的表现通常更好。引入测试时扩展后，性能显著提升，尤其是采用 W-BoN 和 W-LLM 验证器等加权聚合策略时，表现大幅优于其他测试时扩展方法。这些发现凸显了采用多种方案应对情感歧义的重要性。此外，测试时扩展方法的有效性因数据集及其中存在的歧义程度而异。

8 局限性

尽管我们的研究结果表明，测试时扩展（TTS）使音频语言模型能够超越僵化的分类方式，并在面对细微或混合情绪时展现出更强的不确定性感知能力，但仍存在若干局限。

首先，我们目前的评估受限于现有数据集和标注方案，而这些尚未体现不同文化、说话者和语境下自然情绪表达的全部多样性与复杂性。未来扩展基准时，若能纳入更多样的多语言数据，以及更丰富的标注格式（例如标注者分布、置信度分数或连续情绪表示），将在数据可用时带来更深入的洞见，并使评估更贴近现实。

此外，我们的工作采用了通用文本转语音方法。开发以情绪为重点的文本转语音策略具有令人振奋的潜力：这类策略可针对模糊情形动态调整推理、支持多假设输出，并提供校准效果更好的不确定性，所有这些都能进一步增强其在敏感场景中的可靠性。

参考文献

Alaa A Abd-Alrazaq、Mohannad Alajlani、Ali Abdallah Alalwan、Bridgette M Bewick、Peter Gardner 和 Mowafa Househ。2019年。心理健康领域聊天机器人的功能概述：一项范围综述。《国际医学信息学杂志》，132：103978。

Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat及另外1位作者。2023年。GPT-4技术报告。*arXiv* 预印本 *arXiv:2303.08774*。

Bo Adler、NiketAgarwal、Ashwath Aithal、Dong HAnh、Pallab Bhattacharya、Annika Brundyn、Jared Casper、Bryan Catanzaro、Sharon Clay、Jonathan Cohen及另外1位作者。2024年。Nemotron-4 340B技术报告。*arXiv* 预印本 *arXiv:2406.11704*。

Zhenni Bi、Kai Han、Chuanjian Liu、Yehui Tang和 Yunhe Wang。2024年。Forest-of-thought：通过扩展测试时计算增强大语言模型推理。*arXiv* 预印本 *arXiv:2412.09078*。

Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower、Samuel Kim、Jeannette N Chang、Sungbok Lee 和 Shrikanth S Narayanan。2008年。IEMOCAP：交互式情绪双人动作捕捉数据库。《语言资源与评估》，42:335–359。

Houwei Cao、David G Cooper、Michael K Keutmann、Ruben CGur、Ani Nenkova 和 Ragini Verma。2014年。Crema-d：众包情绪多模态演员数据集。*IEEE* 情感计算汇刊，5(4):377–390。Yunfei Chu、Jin Xu、Qian Yang、Haojie Wei、Xipin Wei、Zhifang Guo、Yichong Leng、Yuanjun Lv、Jinzheng He、Junyang Lin、Chang Zhou 和 Jingren Zhou。2024a。Qwen2-audio 技术报告。*arXiv* 预印本 *arXiv:2407.10759*。Yunfei Chu、Jin Xu、Qian Yang、Haojie Wei、Xipin Wei、Zhifang Guo、Yichong Leng、Yuanjun Lv、Jinzheng He、Junyang Lin 等。2024b。Qwen2-audio 技术报告。*arXiv* 预印本 *arXiv:2407.10759*。

Ting Dang、Vidhyasaharan Sethu、Julien Epps 和 Eliathamby Ambikairajah。2017。使用高斯混合回归对情绪预测不确定性的研究。载于 *INTERSPEECH*，第1248–1252页。

Bernd Dudzik、Tiffany Matej Hrkalovic、Chenxu Hao、Chirag Raman 和 Masha Tsfasman。2024年。情感计算中的不确定性：在数据收集实践中考量含义与上下文。载于 *2024*年第*12*届情感计算与智能交互研讨会及演示国际会议（*ACIIW*），第181–185页。IEEE。

Nelly Elsayed、Zag ElSayed、Navid Asadizanjani、Murat Ozer、Ahmed Abdelgawad 和 Magdy Bayoumi。2024年。使用监督式深度循环系统进行语音情绪识别，以实现心理健康监测。

Tiantian Feng 和 Shrikanth Narayanan。2024年。基础模型辅助的自动语音情绪识别：转录、标注和增强。载于 *ICASSP2024-2024 IEEE InternationalConfer-ence onAcoustics, Speech andSignal Processing(ICASSP)*，第12116–12120 页。IEEE。

Fixie AI。2024a。一款专为实时语音交互设计的快速多模态大语言模型。https://github.com/fixie-ai/ultravox/releases/tag/v0.3。访问日期：2025年7月31日。

Fixie AI。2024b。一款专为实时语音交互设计的快速多模态大语言模型。https://github.com/fixie-ai/ultravox/releases/tag/v0.4。访问日期：2025年7月31日。

Maria Flynn、Dimitris Effraimidis、Anastassia An- gelopoulou、Epaminondas Kapetanios、David Williams、Jude Hemanth 和 Tony Towell。2020年。评估面向临床研究的成年人和儿童自动情绪识别的有效性。《人类神经科学前沿》，14:70。Yao Fu、Hao Peng、Ashish Sabharwal、Peter Clark 和 Tushar Khot。2022。基于复杂度的提示方法，用于多步推理。*arXiv* 预印本 *arXiv:2210.00720*。

Sreyan Ghosh、Zhifeng Kong、Sonal Kumar、S Sak- shi、Jaehyeon Kim、Wei Ping、Rafael Valle、Di- nesh Manocha 和 Bryan Catanzaro。2025。Audio Flamingo 2：一款具备长音频理解和专家推理能力的音频-语言模型。*arXiv* 预印本 *arXiv:2503.03983*。

Google。2024年。Gemini 1.5：解锁跨越数百万词元上下文的多模态理解。*arXiv*预印本 *arXiv:2403.05530*。

Google。2025年。 Gemini 2.0 Flash——模型卡。技术报告，Google。

Jule Valendo Halim、Siyi Wang、Hong Jia 和 Ting Dang。2025年。词元级logits至关重要：深入探究用于模糊情绪识别的语音基础模型。*INTERSPEECH*。

Jing Han、Zixing Zhang、Maximilian Schmitt、Maja Pantic 和 Björn Schuller。2017年。从硬到软：通过建模感知不确定性，迈向更符合人类的情绪识别。载于第*25*届*ACM*国际多媒体会议论文集，第890–897页。

Xin Hong、Yuan Gong、Vidhyasaharan Sethu 和 Ting Dang。2024年。AER-LLM：利用大语言模型实现歧义感知情感识别。

Yongquan Hu、Shuning Zhang、Ting Dang、Hong Jia、Flora D Salim、Wen Hu 和 Aaron J Quigley。2024年。探索大规模语言模型，以评估用于心理健康的基于脑电的多模态数据。与 *UbiComp 2024* 联合举办的 *Well- Comp*。Runze Liu、Junqi Gao、Jian Zhao、Kaiyan Zhang、Xiu Li、Biqing Qi、Wanli Ouyang 和 Bowen Zhou。2025a。1B 大语言模型能否超越 405B 大语言模型？重新思考计算最优测试时扩展。*arXiv* 预印本 *arXiv:2502.06703*。Yexiang Liu、Zekun Li、Zhi Fang、Nan Xu、Ran He 和 Tieniu Tan。2025b。重新思考提示策略在大语言模型测试时扩展中的作用：概率论视角。*arXiv* 预印本 *arXiv:2505.10981*。Reza Lotfian 和 Carlos Busso。2017。通过从现有播客录音中检索情感语音，构建自然且情感平衡的语音语料库。*IEEE*情感计算汇刊，10(4)：471–483。Emily Mower、Athanasios Metallinou、Chul Min Lee、Ali Kazemzadeh、Carlos Busso、Sungbok Lee 和 Shrikanth Narayanan。2009。解读模糊情感表达。载于国际情感计算与智能交互会议论文集，第 1–8 页。Niklas Muennighoff、Zitong Yang、Weijia Shi、Xi- ang Lisa Li、李飞飞、Hannaneh Hajishirzi、Luke Zettlemoyer、Percy Liang、Emmanuel Candès 和 Tatsunori Hashimoto。2025。s1：简单的测试时扩展。*arXiv* 预印本 *arXiv:2501.19393*。

Minxue Niu、Mimansa Jaiswal 和 Emily Mower Provost。2024年。从文本到情绪：揭示大语言模型的情绪标注能力。预印本，arXiv:2408.17026。

SrinivasanRamakrishnan 和 Ibrahiem MMElEmary。2013年。人机交互中的语音情绪识别方法。《电信系统》，52：1467–1478。

Vidhyasaharan Sethu、Emily Mower Provost、Julien Epps、Carlos Busso、Nicholas Cummins 和 Shrikanth Narayanan。2019年。情绪表征的模糊世界。*arXiv* 预印本 *arXiv:1909.00360*。

Fangxun Shu、Lei Zhang、Hao Jiang 和 Cihang Xie。2023年。用于视频理解的音视频大语言模型。*arXiv* 预印本 *arXiv:2312.06720*。

Charlie Snell、Jaehoon Lee、Kelvin Xu 和 Aviral Kumar。2024年。最优扩展大语言模型的测试时计算，比扩展模型参数更有效。*arXiv* 预印本 *arXiv:2408.03314*。

Gemini 团队、Rohan Anil、Sebastian Borgeaud、Jean-Baptiste Alayrac、Jiahui Yu、Radu Soricut、Johan Schalkwyk、Andrew M Dai、Anja Hauth、Katie Millican 和另外 1 位作者。2023年。Gemini：一个功能强大的多模态模型系列。*arXiv* 预印本 *arXiv:2312.11805*。

Angel Olider Rojas Vistorte、Angel Deroncele-Acosta、Juan Luis Martín Ayala、Angel Barrasa、Caridad López-Granero 和 Mariacarla Martí-González。2024年。在学习环境中整合人工智能以评估情绪：系统性文献综述。《心理学前沿》，15:1387089。Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。2022。自洽性提升语言模型中的思维链推理。*arXiv* 预印本 *arXiv:2203.11171*。Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 及另外 1 人。2022。思维链提示引发大语言模型中的推理。《神经信息处理系统进展》，35:24824–24837。Jingyao Wu、Ting Dang、Vidhyasaharan Sethu 和 Eliathamby Ambikairajah。2022。一种用于预测模糊情绪状态的新型序列蒙特卡洛框架。载于 *ICASSP 2022—2022 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 8567–8571 页。IEEE。Jingyao Wu、Ting Dang、Vidhyasaharan Sethu 和 Eliathamby Ambikairajah。2024a。用于歧义感知连续情绪预测的双重约束动力神经常微分方程。载于 *Interspeech2024*，第 3185–3189 页。

Jingyao Wu、Ting Dang、Vidhyasaharan Set hu 和 Eliathamby Ambikairajah。2024b。情绪识别系统必须接纳歧义。载于 *2024*年第*12*届情感计算与智能交互国际会议研讨会及演示（*ACIIW*），第 166–170 页。IEEE。

Wen Wu、Bo Li、Chao Zhang、Chung-Cheng Chiu、Qi- ujia Li、Junwen Bai、Tara N Sainath 和 Philip C Woodland。2024c。《处理情绪歧义：从域外检测到分布估计》。*arXiv* 预印本 *arXiv:2402.12862*。

Jin Xu、Zhifang Guo、Jinzheng He、Hangrui Hu、Ting He、Shuai Bai、Keqin Chen、Jialin Wang、Yang Fan、Kai Dang 等人。2025。《Qwen2. 5-omni 技术报告》。*arXiv* 预印本 *arXiv:2503.20215*。

Zixiaofan Yang 和 Julia Hirschberg。2018年。《使用深度神经网络从波形和频谱图预测唤醒度和效价》。载于 *INTERSPEECH*，第 3092–3096 页。

Qiyuan Zhang、Fuyuan Lyu、Zexu Sun、Lei Wang、WeixuZhang、Zhihan Guo、Yufei Wang、Irwin King、Xue Liu 和 Chen Ma。2025。《什么、如何、何处，以及效果如何？大语言模型测试时扩展综述》。*arXiv* 预印本 *arXiv:2503.24235*。

A 技术附录和补充材料

A.1 提示

表格1展示了未使用思维链的音频语言模型（ALM）所用的示例提示。背景部分为模型提供基本情境。任务部分明确要求模型从预定义类别列表中预测目标话语的情绪概率分布。最后，输出约束部分将模型的输出限制为特定格式，以便后续进行自动评估。

在 CREMA-D 数据集的上下文中，情绪是通过演员的表达方式传达的，而非通过 12 个固定目标话语的语义内容。因此，为了向 ALM 提供更明确的指示并改善结果，我们修改了提示。具体而言，我们将“预测目标话语的情绪概率分布”改为“预测音频的情绪概率分布”。

表1：IEMOCAP数据集离散情绪分类的提示模板。

背景 两位说话者正在交谈。	
目标话语	是的。 我想，我一直都是这样。 但它正在离我而去。
Task	预测以下内容的情绪概率分布 以下选项中的目标话语： ['中性状态', '快乐', '愤怒', '悲伤']。 同时考虑上下文和声学特征。
输出约束	1. 严格生成以下JSON结构： {"中性状态":浮点数， "快乐":浮点数, "愤怒":浮点数, "悲伤":浮点数} 输出前，请检查你的输出格式是否符合 输出符合要求 我已提供。 2. 概率之和必须等于 1.0。 3. 除字典外，不要包含任何解释或文本。

思维链提示还加入了如表 2所示的逐步指令。ALM-Verifier 的提示设计如表 3所示。

A.2 实现细节

模型结构本研究使用的模型包括七个音频大语言模型。

表 2：带思维链的情绪识别提示模板

背景	两位说话者正在进行对话。
目标话语	是的，我想我一直都是。 但它正在离我而去。
Task	预测以下选项中 目标话语的情绪概率分布： ['中性状态', '快乐', '愤怒', '悲伤']。 同时考虑上下文和声学特征。
CoT 指令	• **第**一，仔细聆听语气、语调，停顿、声音能量及其他声学特征——征。 • **第**二，检查文本内容，考虑用词、情感倾向和语境线索。 • 识别所有情绪线索，即使是细微的线索。 • 评估每种情绪是否存在。如果存在多种情绪，估计它们相对的强度和存在程度。
输出约束	1. 完全按照以下JSON结构生成： {"中性状态":float, ...} 输出之前，请检查你的 输出是否符合 我提供的要求。 2. 概率之和必须等于 1.0。 3. 不要包含任何解释或文本，包字典之外。

表3：带有 ALM-Verifier 的情绪识别提示模板

背景	你正在评估模型对一项情绪-分布任务的预测。
Role	User
User 提示	预测以下选项中 目标话语的情绪概率分布： ['中性状态', '快乐', '愤怒', '悲伤']。 同时考虑上下文和声学特-征。
模型输-put	{"Anger": 0.1, {"Happy": 0.5, . . . }
Task	仅基于情感内容进行评分： 分布在多大程度上反映了 话语和上下文的情感内容?
评分指南	• **0.0–0.3**：明显错误或无关 • **0.4–0.7**：部分合理或有些不匹配 • **0.8–1.0**：与预期的情绪分布相符 • 仅回复一个介于 0.0 和 1.0.

•**Qwen2.5-Omni** （Xu 等，2025）：阿里巴巴推出的一款先进开源多模态模型，旨在无缝处理

音频、视觉和文本输入。它支持音频理解、语音交互和跨模态推理。本研究采用其 7B 参数版本，该版本在各种音频语言集成任务中表现出色。

• **Qwen2-Audio-Instruct**(Chu 等人，2024b）：阿里巴巴开发的 Qwen2-Audio 是一款先进的音频文本模型，旨在理解并响应口语指令。它能够处理多种音频输入，以进行分析或生成文本。我们的实验使用了该模型的 70 亿参数版本。

•**Audio-Flamingo 2**（Ghosh 等人，2025）：英伟达的 Audio-Flamingo 2 基于原始 Flamingo 架构构建，提升了对较长音频片段进行推理的能力。尽管其语言核心相对紧凑，仅有 30 亿参数，但在音频理解方面，它在 20 多项标准基准测试中的表现超越了许多同类模型。

•**Ultravox-v0.3**（Fixie AI，2024a）：Ultravox-v0.3 由 Fixie AI 开发，是一款大型语音转文本模型，专为低延迟实时转录和语音理解而设计。它基于改进的 Llama-3 8B 架构构建，擅长处理长篇音频，并能在噪声环境中准确转录语音，因此适用于先进的对话式人工智能应用。

• **Ultravox-v0.4**（Fixie AI，2024b）：作为前代产品的升级版本，Fixie AI 的 Ultravox-v0.4 进一步提升了转录准确率和鲁棒性。该版本通过架构优化，显著降低了词错误率（WER），并增强了对不同口音和声学环境的适应能力。它保留了实时处理的核心优势，同时在复杂语音理解任务中表现更出色。

•**Gemini-2.0-Flash**（Google，2025年）：作为 Google 于 2025 年推出的 Gemini-2.0 系列的精简版本，Gemini-2.0-Flash 专为低延迟音频处理而设计。它能够高效执行语音相关功能，例如识别和合成，同时降低计算开销。

表4：解析策略

策略	描述
JSON 词典-nary	尝试将字符串解析为类似以下形式的 JSON 对象 {"愤怒": 0.1, "开心": 0.5, ...}。如果成-功，则将字典转换为概率向量并对其进行归一化。
情绪 List	处理类似 ['厌恶' 的原始字符串, '中性'] 或 ["开心"]。匹配的情绪会被映射到对应的索引，并在这些索引上创建均匀分布。无法识别的条目将被忽略。
浮点型 List	检测方括号内的数值序列（例如，[0.3、0.2、0.5]）。如果至少有两个值被找到，则将其解释为概率向量。该向量随后会补齐或截短，以适应预期的长度，并进行归一化。
关键词匹配	搜索已知情绪关键-词（例如，悲伤、愤怒）在原始字符串中的直接提及。如果找到，则返回独热向量，并将匹配的情绪设为 1.0。
单一浮点型	如果原始输出仅包含一个数字或一个仅含单个元素的列表（例如，0.54 或 [0.54]），则不会将其视为有效的情绪分布，并会将其丢弃。

• **Gemini-2.5-Pro**(Google，2024年）：该模型是 Gemini-1.5 系列中的高容量迭代版本，专为涉及音频的多模态任务而设计。凭借扩展的训练数据和架构改进，它在情绪检测、多语言音频理解和语音转文本转换等领域取得了出色的成果。

• **GPT-4o**（Achiam 等人，2023年）：OpenAI 的 GPT-4o 引入了一个统一框架，用于处理音频和文本输入。其能力涵盖多种任务，从自然流畅的语音合成到实时转录和对话推理，使其成为适用于跨模态交互的多功能模型。

输出处理 值得注意的是，Audio-Flamingo 2 的输出并未始终遵循我们预先设定的格式，我们认为这是由其模型规模较小所致。因此，我们实施了表4中所示的一些策略，最终 IEMOCAP、MSP-Podcast 和 CREMA-D 数据集中的有效数据比例分别仅为 57.9%、66.5% 和 87.2%。相比之下，所有其他模型的输出均完全符合要求，因此可以使用其完整数据集。

A.3 进一步分析

不同难度级别下 ALM 的表现 模型的性能层级总体上与表 1中的分析一致。总体而言，闭源模型优于

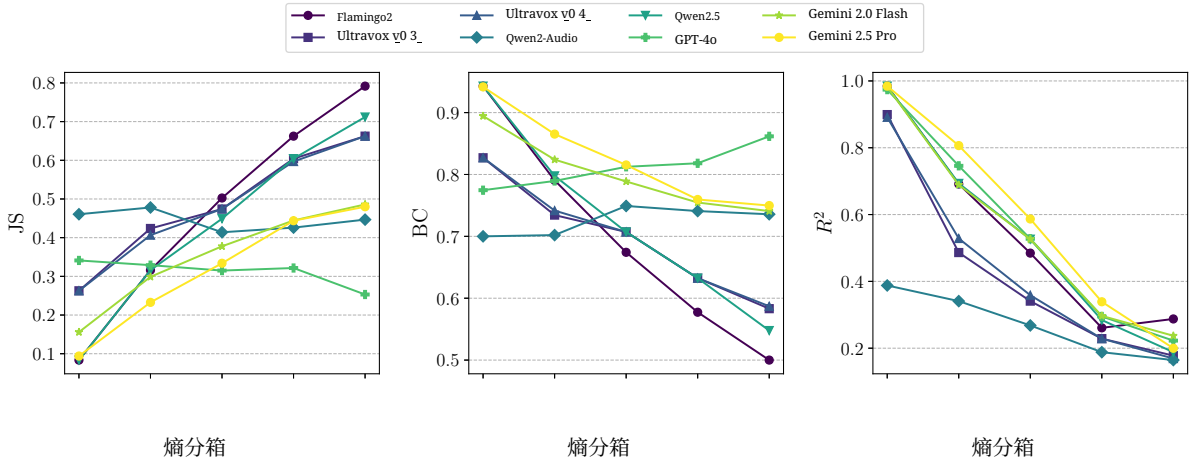


图1：音频语言模型在 CREMA-D 数据集上不同熵分箱中的性能。本分析呈现中位数，以减轻偶发异常值对整体模型性能评估的影响。

开源模型，其中 Gemini 2.5 Pro 的整体表现最佳。尤其值得关注的是，Qwen2-Audio 和 GPT-4o 在 JS 散度和 Bhattacharyya 系数指标上展现出显著的稳健性。随着数据熵增加，其他模型的表现有所下降，而 Qwen2-Audio 的得分几乎保持不变，GPT-4o 的得分甚至出现了反直觉的提升。这表明这两种模型在解读复杂情绪状态方面具备更强的能力。