EmoSURA：迈向对详细且长上下文情感语音描述文本的准确评估

Xin Jing [1,2], *Andreas Triantafyllopoulos* [1,2], *Jiadong Wang* [1,2], *ShahinAmiriparian*[1,2,3], *JunLuo*[3], *Bj¨ornSchuller* [1,2,4]

[1] CHI——健康信息学教席，慕尼黑工业大学医院，慕尼黑，德国，[2] 慕尼黑机器学习中心，慕尼黑，德国[3] 华为，荷兰，[4] GLAM，伦敦帝国学院，英国xin.jing@tum.de

arXiv:2603.09820v1 [cs.SD] 10 Mar 2026

摘要

近期语音描述生成模型的进展，使得人们能够为情感语音生成丰富、细致入微的描述文本。然而，这类描述文本的评估仍是一个关键瓶颈：传统的 N 元语法指标**无法**捕捉语义细微差别，而大语言模型评判器在处理长篇描述时，往往会出现推理不一致和上下文坍塌。在本研究中，我们提出了 EmoSURA，这是一种新颖的评估框架，将评估范式从整体评分转向原子级验证。EmoSURA 将复杂描述文本分解为原子感知单元，即针对声音或情感属性的独立陈述，并采用基于音频的验证机制，根据原始语音信号验证每个单元。此外，为解决标准化评估资源匮乏的问题，我们推出了**SURABench**，这是一个经过精心平衡和分层的基准。实验结果表明，与人类判断的相关性为正；与传统指标相比，EmoSURA 能更可靠地评估长篇描述。传统指标对描述文本长度较为敏感，因此呈现出负相关。

索引词：情感计算、语音理解、人机交互、计算副语言学

1. 引言

在大规模音频语言模型（ALM）近期发展的推动下，情感语音描述如今能够生成高度流畅、篇幅较长且细节丰富的自然语言描述，用于刻画说话人的声音特征、情绪状态和韵律风格 [1, 2, 3]。然而，我们仍然缺乏对什么样的生成式描述对于人类听众而言是恰当的深入理解——而这种理解对于指导未来的模型开发至关重要 [4, 5]。

由于任务本身的特性与现有评估指标之间存在不匹配，评估情感语音描述文本并非易事 [6]。传统的基于N元语法的指标关注表层词汇重叠，不适合评估自由形式的、基于感知的描述 [7]。语义相似度指标 [8, 9, 10]通过在学习得到的嵌入空间中进行运算，部分缓解了这一局限，但仍会受到文本长度的影响，且往往无法充分评估篇幅长、信息密集的描述文本。近来，大语言模型（LLM）凭借处理复杂自然语言的能力，被用作评审者 [11, 12]。然而，若直接输入包含丰富且细致情绪细节的长篇描述文本，LLM 评审方法容易出现信息丢失和推理不一致的问题，从而降低整体可靠性 [13, 14]。为降低评估此类长文本的复杂性，一些研究尝试在使用大语言模型评分前，将描述文本拆解为离散标签或关键词 [15, 16, 17]。然而，将文本与源音频脱离，根本上限制了模型的灵活性，使其无法直接依据实际声学信号来评估情绪描述。

为应对这些局限，我们提出**情**感**语**音**理**解**评**分（**EmoSURA**），这是一种针对情感语音描述文本的结构化评估框架。EmoSURA 的核心理念是将一段描述文本分解为一组最小原子感知单元（APU），其中每个单元都以一个完整且自包含的句子表达，用于描述单一的声音或情绪属性。EmoSURA 不对描述文本进行整体评估，而是通过查询音频语言模型，独立评估每个信息单元，以确定原始语音数据中是否存在相应的感知证据。为减轻幻觉和级联错误，单元级评估被设计为二元决策任务，模型仅需就原始音频中是否存在每个感知单元作出是/否判断。此前基于大语言模型的评估研究表明，这类受约束的决策方案能够带来更稳定、更一致的评估结果 [18]。最终得分同时反映各单元的感知有效性，以及候选描述文本与参考描述文本之间的一致性。这种“分解并评估”策略具有多项优势。通过围绕原子感知单元开展评估，EmoSURA 提升了可解释性，并支持细粒度的误差分析，因为每个评估步骤都会生成明确的中间决策，从而便于错误追踪和模型改进。该框架具有灵活性，无需重新设计指标即可引入新的评估维度。

本研究的贡献有三点：

- 我们提出了**EmoSURA**，这是一种新颖的细粒度评估框架，可将描述文本分解为原子感知单元，并利用基于音频的验证。
- 我们构建了**SURABench**，这是一个平衡且分层的基准数据集，为可复现的基于参考的评估提供标准化资源。
- 我们通过全面实验表明，与传统的 N 元语法指标和基于嵌入的指标相比，EmoSURA 与人类判断的相关性达到了**最先进水平**，证明了其在检测幻觉和评估细粒度副语言特征方面的鲁棒性，

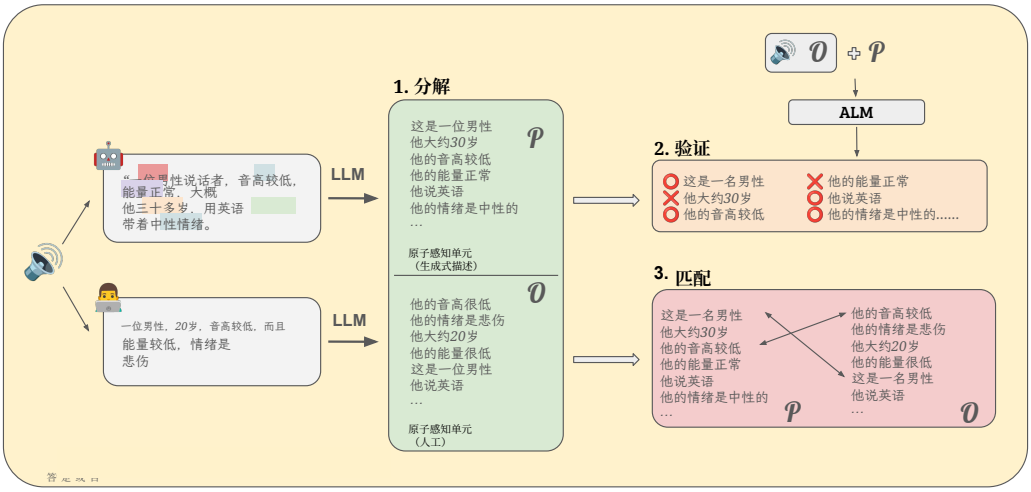


图 1：*EmoSURA* 框架。该框架包含三个步骤：*(1)* 使用大语言模型将描述文本分解为原子感知单元（*APU*）；*(2)* 使用音频语言模型根据原始音频验证生成的 *APU*；*(3)* 将生成的 *APU* 与基准进行匹配，以评估全面性。

语言细节。

本文其余部分安排如下。我们将在第 2 节和第 3 节分别介绍所提出的 EmoSURA 框架和 SURABench。在第 4 节中，我们将介绍实验设置、消融研究和实验结果，并进行详细分析。最后，在第 5 节中，我们总结研究发现并讨论未来方向。

2. EmoSURA

如图 1 所示，EmoSURA 是一个模块化评估框架，旨在评估情感语音与描述文本之间的一致性。EmoSURA 并非生成单一的综合评分，而是将评估过程分解为三个可解释的阶段：分解、验证和匹配。每个阶段都针对情感语音描述中的一种常见问题：句子级歧义、声学幻觉和内容表述不足。EmoSURA 以原子感知单元（APU）为粒度进行处理：这些单元是从描述文本中提取出的最小且自包含的语义断言。这种原子化方法能够进行细粒度分析，并明确推理出存在幻觉或缺失的内容。

2.1. 步骤 1：原子分解

EmoSURA 的第一步是将复杂描述文本分解为原子基本单元（APU）。关键的是，每个 APU 都被表述为一个**独立的陈述句**。这种格式至关重要，因为只有完整命题才具有明确的真值，从而使后续模型能够进行鲁棒的二元（是/否）验证。这种设计有效缓解了粗粒度句子级评估固有的歧义和语义纠缠 [19]。给定描述文本集合 $\mathcal{C}$，我们采用 Qwen2.5-7B-Instruct 作为分解引擎。令 $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$ 表示从生成式描述 $\mathcal{C}_{gen}$中提取的 APU 集合，令$\mathcal{O} = \{o_1, o_2, \ldots, o_M\}$表示从人工标注的参考描述 $\mathcal{C}_{ref}$中提取的对应集合。我们提示模型将每段描述文本解析为原子陈述，其中每个单元编码一个主谓宾关系或属性级事实。

2.2. 步骤 2：基于音频的验证

传统文本评估指标的一个关键局限在于，它们无法惩罚幻觉，即缺乏底层音频信号依据的情感或声学描述。为解决这一局限，EmoSURA 引入了音频依据验证阶段，明确评估生成的 APU 相对于声学证据的事实合理性。对于每个生成单元 $p_i \in \mathcal{P}$，我们采用 Qwen2-Audio-7B-Instruct 作为音频依据评判器。音频语言模型（ALM）会联合处理原始语音信号 $\mathcal{A}$ 和文本单元 p_i，并被明确提示执行二元 蕴含判断（是/否），以确定音频是否支持该陈述。这种设计遵循近年来评估文献中广泛采用的大语言模型评判范式，同时考虑到由于错误累积，较长的描述往往更容易出现幻觉 [7, 18, 20]。为确保鲁棒性，验证过程采用保守策略，优先拒绝可证伪的情感或声学幻觉，而非接受含糊不清或仅有部分依据的描述。形式化地，验证函数定义为：$V(p_i \mid \mathcal{A}) = \text{ALM}(\mathcal{A}, p_i) \in \{\text{是}, \text{否}\}$. 据此，音频支持的无幻觉单元集合为：$\mathcal{P}_{true} = \{p_i \in \mathcal{P} \mid V(p_i \mid \mathcal{A}) = \text{是}\}$.该验证阶段会得出一个精确率导向分数，反映相对于音频信号的事实正确性：$s_p = |\mathcal{P}_{true}|/|\mathcal{P}|$。

2.3. 步骤 3：语义匹配

虽然音频依据验证评估了事实正确性，但无法衡量生成式描述是否充分涵盖参考描述中的显著内容。为评估完整性并减少遗漏错误，Emo-SURA 在 APU 层级评估语义召回率。我们使用 Qwen2.5-7B-Instruct 对生成单元与参考单元进行语义对齐。对于每个参考单元 $o_j \in \mathcal{O}$，模型判断其是否在语义上蕴含于至少一个生成单元 $p_i \in \mathcal{P}$，或与之匹配。匹配的参考单元子集记为 $\mathcal{Q} \subseteq \mathcal{O}$。重要的是，除了与参考内容对齐的单元外，我们还明确

纳入经验证的生成单元，这些单元传达了正确但不属于参考内容的信息。这样可以避免将有效且有音频依据的描述误判为假阴性。因此，面向召回率的评分定义如下：

$$s_r = \frac{|\mathcal{Q}| + |\mathcal{P}_{true} \setminus \mathcal{Q}|}{|\mathcal{O}| + |\mathcal{P}_{true} \setminus \mathcal{Q}|}. \tag{1}$$

这一表述既奖励对参考内容的忠实覆盖，也奖励纳入更多可验证的描述性细节。

2.4. 最终评分

基于精确率导向分数 s_p 和面向召回率的得分 s_r，我们计算综合 F1得分，以平衡事实正确性与内容覆盖度：

$$s_f = 2 \cdot \frac{s_p \cdot s_r}{s_p + s_r}. \tag{2}$$

此外，为了明确评估描述丰富度，我们采用相同公式计算描述性F1得分 s'_f，但仅限于描述性APU。最终的 EmoSURA 得分定义如下：

$$\mathcal{F} = \frac{1}{2}\left(s_f + s'_f\right). \tag{3}$$

3. SURABench

为了建立高质量的评估基础，我们构建了**SURABench**，这是一个从 MSP-Podcast v1.11*Test1* 划分 [21]中整理而来的平衡评估基准。我们采用三阶段筛选流程，以确保声学适用性、标签可靠性和分布平衡。首先，我们剔除了过短（$< 3s$）而无法传达完整情绪语义，或过长（$> 8s$）而不利于稳定字幕生成的片段。随后，我们施加共识约束，仅保留效价和唤醒度评分标准差均为 ≤ 1.5 的话语。此步骤可剔除标注者间一致性较低的模糊样本。最后，我们采用分层网格采样策略，以缓解严重的类别不平衡问题。我们将效价-唤醒度空间（缩放至 1-7）离散为 10×10 网格。从每个网格单元中最多选取 15 个样本，并优先选择标注共识度最高（方差最低）的样本。如图 2 所示，该采样机制确保四个语义象限均得到均匀覆盖，同时减少中性语音的过度代表。最终数据集包含**1,018** 条话语，涵盖广泛的情绪类别。

为生成高保真描述文本，我们采用了一种广泛使用的混合标注流程，将声学特征提取与人工引导的大语言模型生成相结合 [22, 23, 24]。首先，我们遵循ParaCLAP [25, 26] 提取一组副语言特征作为客观证据，包括音高、音高变化、响度、抖动、微颤和语速。随后，专家标注员依据真实情感标签，为数据的一个代表性子集人工撰写了“黄金标准”描述。这些人工撰写的描述文本确立了期望的描述粒度与句法风格标准。最后，我们以“黄金标准”描述作为少样本示例提示GPT-4.1，为整个基准生成描述文本。

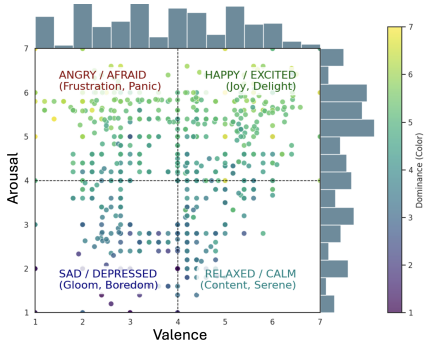


图 2：*SURABench* 在效价-唤醒度空间中的情绪分布。点的颜色表示支配度。边缘直方图（上方和右侧）展示了数据集的均匀性。

4. 实验与分析

4.1. 主观测试

我们对320个音频-描述对开展了主观平均意见分（平均意见分）评估，共有14名参与者（6名男性、8名女性，其中包括6名音频专家）。为确保代表性，这些样本对按效价–唤醒度环形模型与说话者性别进行了分层。评估者被要求先聆听音频，再以5分李克特量表对样本进行评分。样本被归入四个特定描述类别：（i）标准答案（标准答案）；（ii）篡改描述，其中特定事实细节被破坏，以检验错误检测能力；（iii）无约束 Qwen-Omni，代表长篇幅、自由风格生成；（iv）精炼 Qwen-Omni，代表简洁且遵循模式的输出。

如表 1 所示，所有基于规则的指标与人类判断之间的三种相关系数均呈负相关。如图 3 顶部一行所示，BLEU-4 和 CIDEr 的分布中零值占比很高，且偏向下界，这表明对于内容丰富、描述详尽的描述文本，传统指标严格的 n 元语法匹配要求无法为语义正确但词汇多样的描述赋予有意义的分数。与此同时，基于模型的评估指标表现不一。SPICE 仍呈负相关，而 MACE 则呈中等程度的正相关。相比之下，EmoSURA 与人工评分始终保持正向一致性。与 MACE 相比，EmoSURA 的秩相关更高，表明其根据人类偏好对样本进行排序时具有更好的一致性。

我们的统计分析还揭示了真实标注参考文本与模型生成描述文本之间存在根本性的分布偏移。目标参考文本的长度受到控制（$\mu = 459$ 个字符，$\sigma \approx 60$），而 Qwen-omni 等最先进模型则倾向于冗长，生成的描述文本平均长达 684 个字符，约为参考文本的 1.5 倍。长度差异对基于精确率的指标而言是致命的。由于 N 元语法指标会将不重叠的词元视为“插入错误”，模型额外生成的 200+ 个字符——即使其中包含有效且无幻觉的视觉细节——也会受到严厉惩罚。生成长度的高方差（$\sigma \approx 280$，极端情况下可达 1,318 个字符）进一步加剧了这一问题，导致图 3 中观察到的评分表现不稳定。与此同时，我们分析了音频大语言模型评审对指令的遵循情况。在整个测试集中，我们观察到格式失败率较低，仅为 5.61%；

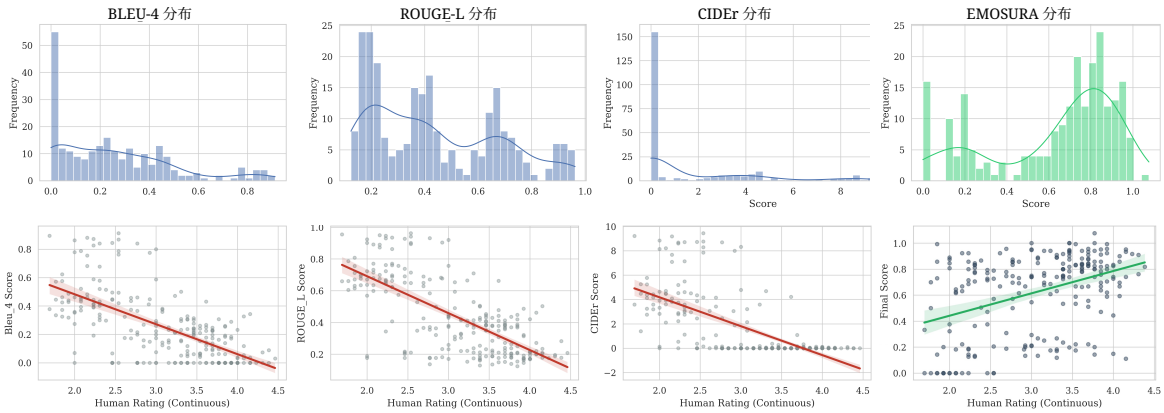


图 3： 基线指标与 *EMOSURA* 的分布图（上）及其相对于人工评分的散点图（下）。与呈现负相关（红线）的 *BLEU-4*、*ROUGE-L*和 *CIDEr* 不同，*EMOSURA* 与人工真实值呈现正向线性关系（绿线），表明其可靠性更高。

表 1： 自动评估指标与人工评分之间的相关性。性能采用皮尔逊相关系数（*PCC*）、肯德尔秩相关系数（$Kd\ \tau$）和逐样本 τ（$Sp\ \tau$）衡量。所有 p 值均 < 0.001。

指标	皮尔逊相关系数 (ρ) ↑	Kd τ ↑	Sp τ ↑
基于规则的指标			
BLEU-4 [27]	-0.6419	-0.4494	-0.6916
ROUGE-L [28]	-0.7017	-0.4606	-0.6916
METEOR [29]	-0.5813	-0.4541	-0.6559
CIDEr [30]	-0.6640	-0.3732	-0.6175
SPIDER [31]	-0.6679	-0.4481	-0.7019
基于模型的评估指标			
SPICE [32]	-0.5728	-0.3874	-0.6240
MACE [8]	0.4283	0.2619	0.3709
所提方法			
EmoSURA（我们的方法）	**0.4391**	**0.3277**	**0.4480**

模型未能生成有效二元（“是”/“否”）token的情况。这表明 Qwen-Omni 能够有效遵循受约束的验证规范，且未出现显著的生成崩溃。然而，尽管表现优于基线模型，EmoSURA 与人类表现之间的皮尔逊相关系数（$\rho \approx 0.44$）表明，人类判断中的相当一部分差异仍未得到解释，这说明 EmoSURA 尚未完全捕捉人类认知评估的细微之处。

4.2. 扰动测试

我们基于 SURABench 构建了一个受控扰动评估集，系统地修改 GT字幕，使其与相应音频产生属性层面的不一致。具体而言，我们将声学幻觉定义为任何与标注的声学证据相矛盾的文本属性（例如说话者性别、情绪状态、发声事件或低层次声学特征）。我们以受控方式施加扰动，每个样本仅改变一个语义属性，同时保持声学语义一致性；例如，将人口统计属性从男性替换为女性时，会有意同步修改音高和嗓音质感描述 [33]。修改后的样本分布于三个预定义的幻觉类别中：情绪翻转、发声事件捏造和声学特征替换。完整定义请参阅补充材料第2节。

表2： *EmoSURA* 在不同幻觉类别下的检测性能。该系统对人口统计和声学事实错误表现出极高的敏感性。

扰动	注入	检测到	语速（%）
声学特征	**120**	**112**	**93.33**
– 性别	40	39	97.50
– 音高，节奏，音量等。	80	73	91.25
情绪	**40**	**33**	**82.50**
发声事件	**40**	**24**	**60.00**

如表2所示，EmoSURA的检测性能从低层声学特征到高层语义行为呈现出明显下降。该系统对物理声学扰动和情绪扰动表现出卓越的鲁棒性。这表明，对于性别、音高和节奏等基础帧级属性，音频与文本模态之间具有高度可靠的跨模态对齐。因此，EmoSURA能够利用其强大的声学敏感性来检测情感极性翻转。

相反，对伪造的发声事件（例如，在正常语音上幻觉出歌唱或啜泣）的检测构成了明显的性能瓶颈，降至60.00%。我们认为，发声事件需要复杂的长期时序建模和更高层次的语义抽象。这一性能差距可能表明，尽管EmoSURA是高度精确的“声学事实核查器”，但它难以捕捉复杂发声的时序动态。

5. 结论

本研究通过引入细粒度验证指标和标准化基准测试，解决了情感语音描述中的评估瓶颈。所提出的框架，

EmoSURA将评估方式从整体文本评分转向原子级验证方法。通过拆解描述文本，并对照原始音频验证各项主张，该方法明确针对长文本生成中常见的幻觉和句子级歧义问题。我们还构建了SURABench，这是一个带有长而详细描述文本的自发语音语料库。实验分析表明，与现有基于规则和基于嵌入的指标相比，EmoSURA与人类感知更为一致；这些现有指标往往会惩罚现代生成模型的冗长性。未来工作将聚焦于在强化学习中利用EmoSURA的反馈，直接优化描述生成模型的事实一致性。

6. 参考文献

[1] J. Xu、Z. Guo、H. Hu、Y. Chu、X. Wang、J. He、Y. Wang、X. Shi、T. He、X. Zhu、Y. Lv、Y. Wang、D. Guo、H. Wang、L. Ma、P. Zhang、X. Zhang、H. Hao、Z. Guo、B. Yang、B. Zhang、Z. Ma、X. Wei、S. Bai、K. Chen、X. Liu、P. Wang、M. Yang、D. Liu、X. Ren、B. Zheng、R. Men、F. Zhou、B. Yu、J. Yang、L. Yu、J. Zhou和J. Lin，“Qwen3-omni技术报告”，*arXiv*预印本 *arXiv:2509.17765*，2025年。[2] C.-y. Huang、M.-H. Shih、K.-H. Lu、C.-Y. Hsiao和H.-y. Lee，“Speechcaps：利用多说话人说话风格描述推进基于指令的通用语音模型”，载于*2025*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。电气与电子工程师学会，2025年，第1–5页。[3] Y. Xu、H. Chen、J. Yu、Q. Huang、Z. Wu、S.-X. Zhang、G. Li、Y. Luo和R. Gu，“Secap：基于大语言模型的语音情感描述”，载于《*AAAI*人工智能会议论文集》，第38卷，第17期，2024年，第19323–19331页。[4] A. Triantafyllopoulos、I. Tsangko、A. Gebhard、A. Mesaros、T. Virtanen和B. W. Schuller，“计算机听觉：从任务特定机器学习到基础模型”，《*IEEE*会刊》，2025年。[5] B. Schuller、A. Mallol-Ragolta、A. P. Almansa、I. Tsangko、M. M. Amin、A. Semertzidou、L. Christ和S. Amiriparian，“情感计算已然改变：基础模型带来的颠覆”，*npj*《人工智能》，第2卷，第1期，第16页，2026。[6] S. Sarto、M. Cornia和R. Cucchiara，“多模态大语言模型时代的图像描述评估：挑战与未来展望”，*arXiv*预印本 *arXiv:2503.14604*，2025年。[7] Q. Ye、X. Zeng、F. Li、C. Li和H. Fan，“以文字作画：借助基准与对齐学习提升详细图像描述”，*arXiv*预印本 *arXiv:2503.07906*，2025年。[8] S. Dixit、S. Deshmukh和B. Raj，“MACE：利用音频评估音频描述系统”，载于*2025*年*IEEE*国际声学、语音与信号处理研讨会（*ICASSPW*）。电气与电子工程师学会，2025年，第1–5页。[9] T. Zhang、V. Kishore、F. Wu、K. Q. Weinberger和Y. Artzi，“BERTScore：使用BERT评估文本生成”，*arXiv*预印本 *arXiv:1904.09675*，2019年。[10] W. Zhao、M. Peyrard、F. Liu、Y. Gao、C. M. Meyer和S. Eger，“MoverScore：基于上下文嵌入与推土机距离的文本生成评估”，*arXiv*预印本 *arXiv:1909.02622*，2019年。[11] T. Guo、H. Chen、H. Liang、M. Qiang、B. Zeng、L. Sun、B. Cui和W. Zhang，“BRACE：稳健音频描述质量评估基准”，*arXiv*预印本 *arXiv:2512.10403*，2025年。[12] H. Wang、J. Zhao、Y. Yang、S. Liu、J. Chen、Y. Zhang、S. Zhao、J. Li、J. Zhou、H. Sun等，“SpeechLLM-as-Judges：迈向通用且可解释的语音质量评估”，*arXiv*预印本 *arXiv:2510.14664*，2025年。[13] M. A. Mohsin、M. Umer、A. Bilal、Z. Memon、M. I. Qadir、S. Bhattacharya、H. Rizwan、A. R. Gorle、M. Z. Kazmi、A. Mohsin等，“论大规模大语言模型的基本极限”，*arXiv*预印本 *arXiv:2511.12869*，2025年。[14] E. Barkett、O. Long 和 M. Thakur，《推理还不够：考察大语言模型中的真理偏见与谄媚》，*arXiv*预印本 *arXiv:2506.21561*，2025年。[15] A. Ando、T. Moriya、S. Horiguchi 和 R. Masumura，《因子条件化说话风格描述》，*arXiv*预印本 *arXiv:2406.18910*，2024年。[16] A. Spataru、E. Hambro、E. Voita 和 N. Cancedda，《知道何时停止：文本生成中的语义漂移研究》，*arXiv*预印本 *arXiv:2404.05411*，2024年。[17] H. Lightman、V. Kosaraju、Y. Burda、H. Edwards、B. Baker、T. Lee、J. Leike、J. Schulman、I. Sutskever 和 K. Cobbe，《让我们逐步验证》，载于第十二届国际学习表征会议，2023年。[18] N. Pham 和 M. Schott，《H-POPE：对大型视觉语言模型中幻觉的基于分层投票的探测评估》，*arXiv*预印本 *arXiv:2411.04077*，2024年。[19] L. Jing、R. Li、Y. Chen 和 X. Du，《FaithScore：对大型视觉语言模型中幻觉的细粒度评估》，载于计算语言学协会论文集：*2024*年自然语言处理实证方法会议，2024年，第5042–5063页。[20] M. Zhang、O. Press、W. Merrill、A. Liu 和 N. A. Smith，《语言模型幻觉如何滚雪球般加剧》，*arXiv*预印本 *arXiv:2305.13534*，2023年。[21] C. Busso、R. Lotfian、K. Sridhar、A. N. Salman、W.-C. Lin、L. Goncalves、S. Parthasarathy、A. R. Naini、S.-G. Leem、L. Martinez-Lucas 等，《MSP-Podcast语料库》，*arXiv*预印本 *arXiv:2509.09791*，2025年。[22] J. Bai、H. Liu、M. Wang、D. Shi、W. Wang、M. D. Plumbley、W.- S. Gan 和 J. Chen，《AudioSetCaps：利用大型音频与语言模型自动生成流程构建的增强型音频描述数据集》，*IEEE*音频、语音与语言处理汇刊，2025年。[23] X. Mei、C. Meng、H. Liu、Q. Kong、T. Ko、C. Zhao、M. D. Plumb-ley、Y. Zou 和 W. Wang，《WavCaps：由ChatGPT辅助标注、用于音频语言多模态研究的弱标注音频描述数据集》，*IEEE/ACM*音频、语音与语言处理汇刊，第32卷，第3339–3354页，2024年。[24] L. Sun、X. Xu、M. Wu 和 W. Xie，《Auto-ACD：用于音频语言表征学习的大规模数据集》，载于第*32*届*ACM*国际多媒体会议论文集，2024年，第5025–5034页。[25] X. Jing、A. Triantafyllopoulos 和 B. Schuller，《ParaCLAP——面向计算副语言任务的通用语言音频模型》，*arXiv*预印本 *arXiv:2406.07203*，2024年。[26] X. Jing、K. Zhou、A. Triantafyllopoulos 和 B. W. Schuller，《通过对比学习和扩散模型，利用自然语言指导增强情感文本转语音可控性》，载于*2025*年国际声学、语音与信号处理会议（*ICASSP*）。电气与电子工程师学会，2025年，第1–5页。[27] K. Papineni、S. Roukos、T. Ward 和 W.-J. Zhu，《BLEU：一种用于机器翻译自动评估的方法》，载于计算语言学协会第*40*届年会论文集，2002年，第311–318页。[28] J. Lin、H. Yin、W. Ping、P. Molchanov、M. Shoeybi 和 S. Han，《VILA：关于视觉语言模型的预训练》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第26 689–26 699页。[29] S. Banerjee 和 A. Lavie，《METEOR：一种与人类判断相关性更高的机器翻译评估自动评价指标》，载于计算语言学协会机器翻译和/或摘要内在与外在评估指标研讨会论文集，2005年，第65–72页。[30] R. Vedantam、C. Lawrence Zitnick 和 D. Parikh，《CIDEr：基于共识的图像描述评估》，载于*IEEE*计算机视觉与模式识别会议论文集，2015年，第4566–4575页。

[31] S. Liu、Z. Zhu、N. Ye、S. Guadarrama 和 K. Murphy，《通过策略梯度优化SPIDER改进图像描述》，载于*IEEE*国际计算机视觉会议论文集，2017年，第873–881页。[32] P. Anderson、B. Fernando、M. Johnson 和 S. Gould，《SPICE：语义命题式图像描述评估》，载于欧洲计算机视觉会议。施普林格，2016年，第382–398页。[33] S. S. Alahmari、L. Hall、P. R. Mouton 和 D. Goldgof，《大语言模型抵御扰动的鲁棒性：S. Alahmari 等人》，《科学报告》，2025年。

生成式人工智能使用披露作者使用生成式人工智能工具协助进行语言编辑和稿件润色。该工具仅用于提升文本的清晰度、语法正确性和可读性。所有科学内容、分析、解读和结论均由作者完成并核实。

1 EmoSURA 提示词

提示词 1：分解

人设
你是一位资深语言学分析师，擅长将描述性文本拆解为可验证的原始信息单元，用以描述声音、语音和情绪模式。

任务
分析提供的输入文本。对于你识别出的每个声音、语音或情绪特征，将其提取为一个原始信息单元。输出必须是一个 JSON 列表，其中每个对象代表文本中一个可验证的属性。

规则。
1
仔细阅读输入文本，找出能够直接证明声音和情绪特征的关键词、短语和代词。2. 针对每条证据，在输出列表中创建一个对应的 JSON 对象。3. “fact”键必须包含一个完整、独立的句子，用于描述该特征。4. “evidence”键必须包含输入文本中支持该事实的原文引述。5. 如果某项属性没有证据，则不要为其创建对象。6. 最终输出只能是 JSON 列表，不得包含其他内容。### 输出模式与允许的术语 JSON 列表中的每个对象都必须符合以下结构：

- 事实：（字符串）一个用于描述该属性的完整句子。
- 属性：（字符串）必须是以下其中一项：“pitch”、“rate”、“volume”、“gender”、“emotion”。
- value：（字符串）该属性的具体术语。允许使用的术语如下：– 对于“pitch”：“low”、“normal”、“high”。– 对于“rate”：“slow”、“normal”、“fast”。
- 证据:（字符串）从输入文本中截取的确切子串。 ### 示例

示例 *1*输入：“他的声音低沉……” 输出：[{ "fact": "说话人的性别为男性。", "attribute": "性别", "value": "男性", "evidence": "他" }]### 输入文本

提示 2：音频验证

你是一名资深音频验证员。

任务
你将 获得一个“事实”、一个音频文件以及一组“标准答案”文本。你的任务是判断该“事实”是否 正确。

规则。
1
音频优先：你的判断必须仅基于音频中听到的内容。2. **标准答案为辅：**标准答案文本仅供参考。3. **冲突解决规则：**音频是最终的事实来源。

输出
- 1 → 事实正确
- 0 → 事实不正确> 标

准答案：{GT} 输入：> 事实：

提示 3：语义匹配

现在，你是一名音频语言学专家，负责匹配从两段描述文本中提取的两组原始信息单元。

原始信息单元集合以 JSON 格式的字典列表表示： [{"fact": [UNIT] , "identifier": [ID]}...] 。此外，标准答案集中的每个原始信息单元都附有唯一的 "id"。

匹配原始信息单元.....（为简洁起见，省略说明）.....

重要：请单独考虑集合中的每个原始信息单元。你只能输出格式如下的匹配结果： [{"fact": [UNIT] , "identifier": [ID] , "matched_oracle_id":[ORACLE ID]}...]其中，"identifier" 是可选项。对于名为 "matched_oracle_id" 的键，其值应为对应的 "id"。如果未匹配，则将 "matched_oracle_id" 设为 "None"。

> > > 标准答案集：{GT}> > > 原始信息单元集合：

2 案例研究：篡改描述文本的定性分析

为严格评估系统的鲁棒性，我们通过系统性扰动真值标注，手动构造了对抗性负样本（篡改后的描述文本）。如我们的分类体系所示，这些扰动涵盖四种复杂程度递增的不同类别。表 1 展示了每类扰动的定性示例，并突出了用于引入语义和声学幻觉的具体词汇替换。

- **类型 A：情感极性翻转**。针对情感维度，将情感效价和唤醒度反转，同时保留底层事实事件。

- **类型 B：属性交换**。通过改变说话人的生物身份（例如性别），并相应调整相关声学属性（例如音高），引入人口统计事实错误。
- **类型 C：事件捏造**。从根本上改变音频事件在时间和行为方面的性质，将正常语音替换为歌唱或啜泣等风格化发声。
- **类型 D：混合幻觉**。这是最严重的扰动，代表一种完整的语义崩溃，其中性别、情绪、发声行为和物理声学同时被反转。

如表 1 所示，生成的幻觉具有很强的内在一致性。例如，在构造 B 类幻觉（属性交换）时，扰动并非只是翻转人口统计标签（“男性”变为“女性”），还会同步更新相关的声学元数据（“低沉隆隆的音高”变为“高而细腻的音高”）。这种内在的逻辑一致性使这些负样本比随机打乱的文本更难被标准跨模态检索模型区分。

表 1：四种幻觉类别的案例。注入的幻觉和经过修改的声学描述以**红色粗体** 突出显示，以便与真实字幕（标准答案）进行对比。

Type	版本	描述文本
A	*GT*字幕	一位男性说话者以**饱满而富有活力的**声音念出这句台词，声音中洋溢着满怀**自信的热情**。他的语调**威严而充满活力——充满感染力**，展现出**权威感**，同时散发着**积极感**.....
	遭到破坏	一位男性说话者以**微弱、颤抖的**声音说道，声音中满是满是**焦虑与恐惧**。他的语调**充满恐惧与犹豫**，展现出**脆弱感**，同时散发着**消极感**.....
B	*GT*字幕	**男性说话人介绍自己**.....以一种**粗哑而威严的**男子气概——的方式。尽管**音高低沉隆隆**.....他的声音仍带有一种**强烈的强势**锋芒.....音质**粗糙而沙哑**.....
	遭到破坏	一位**女性**说话人介绍**自己**.....以一种**柔和温婉的**方式。尽管**音高高而细腻**..... 她的声音带有**轻盈、空灵的**质感.....纹理**平滑清晰**.....
C	*GT*字幕	一位女说话人**以**“.....”这句台词，带着**活泼欢快的活力感**和**自信的气场**。 她的声音**清晰流畅**，高声且**快速地**说着，使她的话语带有一种**生动活泼的自信**.....
	遭到破坏	一位女说话人 **唱着** 这句“.....”时，带着一种 **旋律优美、如歌剧般的能量**。 她的声音**富有音乐性和节奏感**，高声地 **以以歌曲般的方式**，让她的话语带有一种**艺术表演感**.....
D	*GT*字幕	一位**女性**说话人**以沉着、欢快的**语调说出这句台词“.....” 她的声音虽然**轻柔而低沉**，音量不大，却带有一种**自信、笃定的**品质。她说话时节奏**平稳、不急不缓**.....
	遭到破坏	一位**男性**说话人**大喊**着台词“.....”并带着**愤怒、攻击性的**语调。他的声音**洪亮而响亮**，音量很大，带着**敌意，威胁性的** 品质。他说话的**节奏慌乱急促**.....