**音频理解能力汇总**

和@王欣@张奇@孙童沟通讨论，目前主流的比较全面的benchmark，代表的有MMAU、AIR-Bench。其中**MMAU**最具代表性。

**Benchmark**

**AIR-Bench**

* Paper：https://arxiv.org/pdf/2402.07729
* LeaderBoard：https://github.com/OFA-Sys/AIR-Bench
* 笔记：https://docs.feishu.cn/article/wiki/Anxgw21b6ig0yWk2UxDcOWPWnhc
* AIR-Bench（Audio Instruction Bench）是首个专为评估大型音频-语言模型（LALMs）设计的基准测试，由浙江大学 & 阿里巴巴联合开源，包含两个维度：基础基准和对话基准。基础基准由19个任务组成，包含约1.9万道单选题；对话基准则包含2000个开放式问答数据。

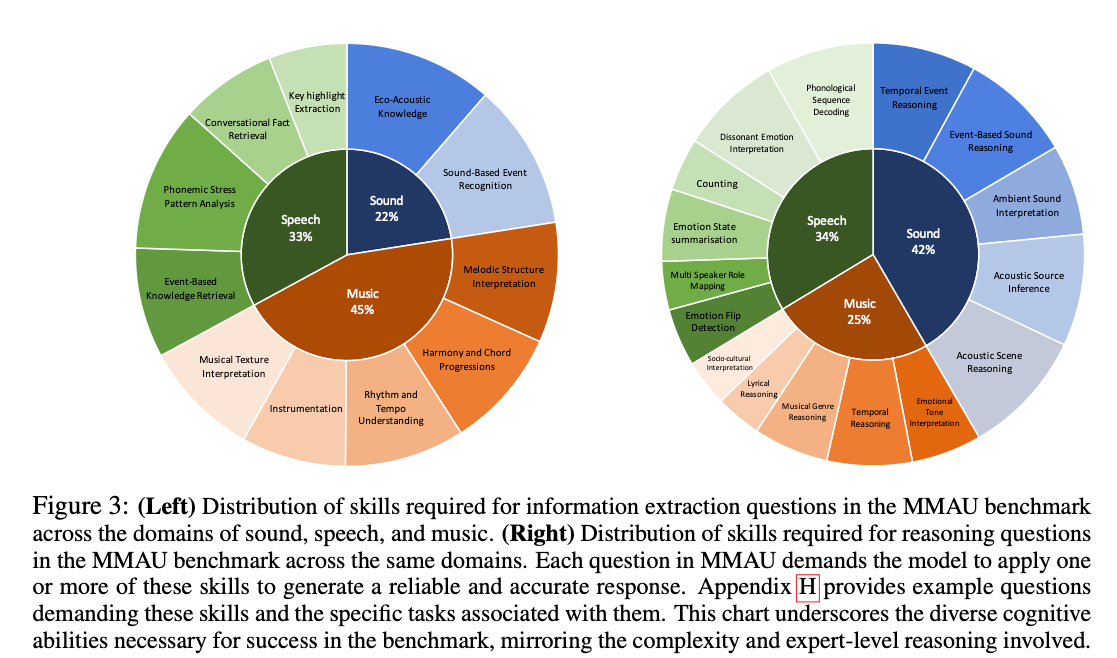
|  |  |  |  |
| --- | --- | --- | --- |
| 子基准 | 任务形态 | 数据量 | 设计目的 |
| **Foundation Benchmark（基础）** | 19 个单选题任务 | ≈19 k 题 | 快速、客观、可复现地检验模型在 **ASR、AST、SER、SEC、Music Tag** 等 19 项基础能力上的准确率。 |
| **Chat Benchmark（对话）** | **2 k 条开放式问答** | 2 k 条音频 | 旨在反映模型在真实应用场景中根据用户指令进行感知、推理和交互的能力。 |

**MMAU**

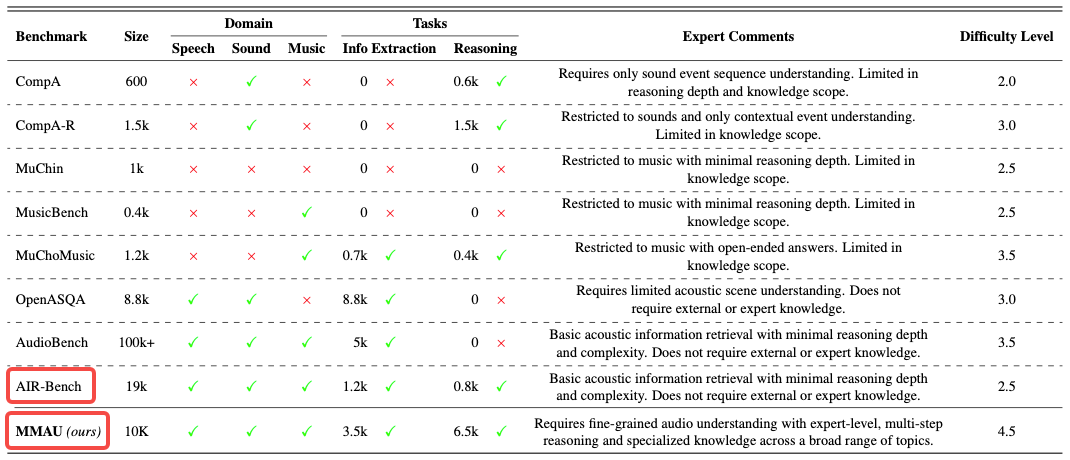
* Paper: https://arxiv.org/pdf/2410.19168
* Leaderboard: https://sakshi113.github.io/mmau\_homepage/
* MMAU（Massive Multi-Task Audio Understanding and Reasoning Benchmark）是目前音频领域难度最高、任务最全面的生成式评测基准之一，由美国马里兰大学、Adobe 等机构于 2024 年 10 月联合提出。**MMAU 在 27 项不同技能上对模型进行评估，其中问题分为两类：信息提取（3,499 题）和推理（6,501 题）。**

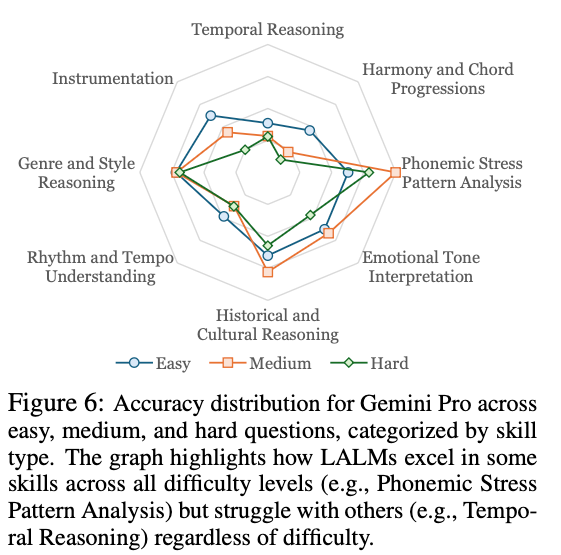
|  |  |  |
| --- | --- | --- |
| 子集 | 数量 | 作用 |
| **Foundation** | 3 499 题 | 传统信息检索（ASR、SEC、Music Tag…） |
| **Reasoning** | 6 501 题 | 复杂推理（因果、情感、场景、专业音乐知识…） |
| **test-mini** | 1 000 题 | 快速验证，27 任务均衡分布 |

* MMAU 分为 mini 集（1000 条）和全量集（10000 条），mini集合完全开源，全量测试集不完全开源，仅开源了问题+音频，答案闭源，如果想测试需要提交模型推理结果到[官方](https://eval.ai/web/challenges/challenge-page/2391/overview)来进行评测。



* MMAU 把音频大模型从“能听”推向“真能听懂”的极限考试——任务最难、领域最全、与人类专家差距最大，**是 2025 年音频理解研究必刷的“硬骨头”基准。**



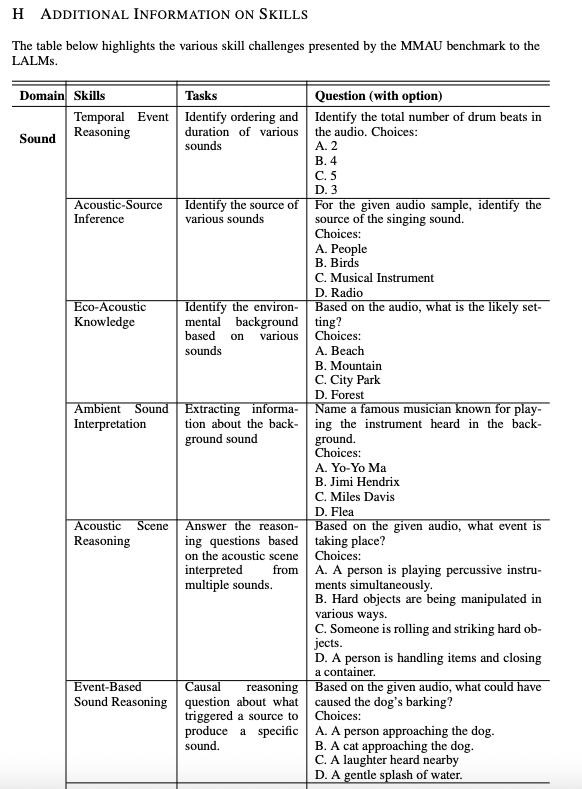


* [迁移DeepSeek-R1同款算法，小米让7B模型登顶音频理解推断MMAU榜单 - 智源社区](https://hub.baai.ac.cn/view/44197)

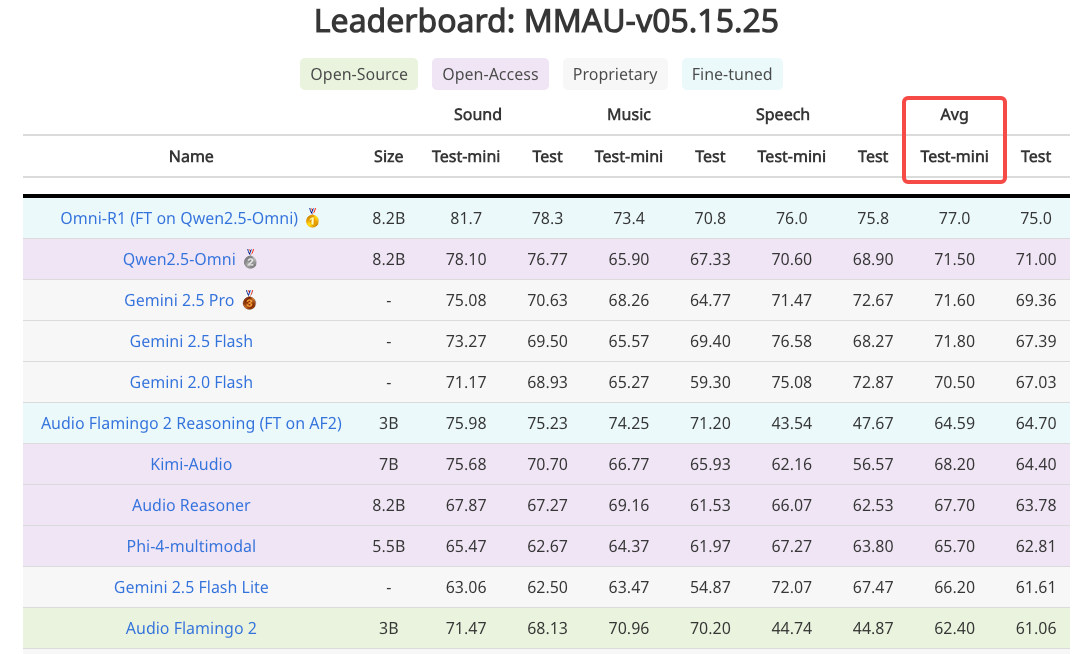
**音频理解能力汇总（MMAU）**

**MMAU 细分能力维度27项**

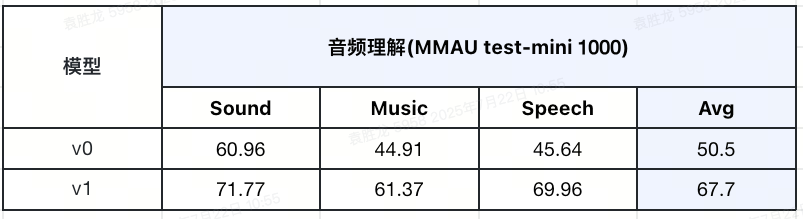
* 论文附录H：



**Leaderboard Top10:**



**OUR：**[**MMAU -test-mini 细分能力**](https://li.feishu.cn/sheets/EPrtsgKMUhdGW9tcXN4ckQSJnsd?sheet=vcPJbZ)



* @孙童翻译整理归类

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 类型 | 能力(带背景色为Reasoning) | 任务 | 任务归类 | 样例 | 现状（线上v1.5 oA模型）  [MMAU -test-mini 细分能力](https://li.feishu.cn/sheets/EPrtsgKMUhdGW9tcXN4ckQSJnsd?sheet=vcPJbZ)  **Avg: 50.50** | v1模型    **Avg: 67.70** | 数据情况：  A:数据比较完善  B:数据仍需补充  C:暂时没有该数据 | 数据备注(条) | 能力优先级：  p0：可落地的能力；  p1：刷榜需要覆盖的能力；  p2：小众能力、数据搜集困难 |
| Sound | 时序事件推理  (Temporal Event Reasoning) | 识别音频中的事件次数/时长 | 音频事件检测 |  | 47.91 | 54.16 | B | 只有音频事件检测，没有细节任务数据，数据量：44w | p1 |
| 声源识别  (Acoustic Source Inference) | 识别声音来源 | 音频事件检测 |  | 77.08 | 87.5 | A | 44w | p0 |
| 声学知识  (Eco-Acoustic Knowledge) | 音频场景检测 | 音频场景检测 |  | 70.21 | 82.97 | A | 12w | p0 |
| 环境音/音乐风格解析  (Ambient Sound Interpretation) | 背景音信息抽取 | 音频描述？ |  | 50 | 62.5 | C |  | p1 |
| 声学场景推理  (Acoustic Scene Reasoning) | 根据音频回答相关问题：音频描述 | 音频描述？ |  | 43.75 | 64.58 | A | 音频描述：500w  音频事件检测：44w | p1 |
| 音频事件推理  (Event-Based Sound Reasoning) | 音频事件因果推理 | 音频事件因果推理 |  | 72.91 | 79.16 | C |  | p1 |
| 音频事件检测  (Sound-Based Event Recognition) | 音频事件检测 | 音频事件检测 |  | 78.26 | 76.08 | A | 44w | p1 |
| Speech | 负面情感检测  (Dissonant Emotion Interpretation) | 多说话人场景的负面情感检测 | 情感识别 |  | 2.85 | 88.57 | C |  | p1 |
| 基于事件的知识抽取  (Event-Based Knowledge Retrieval) | 提取对话的核心知识 | 语音问答 |  | 75.75 | 81.81 | A |  | p1 |
| 计数问题  (Counting) | 说话人计数 | 语音问答 |  | 48.27 | 37.93 | C |  | p1 |
| 语音重音模式分析  (Phonemic Stress Pattern Analysis) | 识别语音中的重音模式 | 语音问答 |  | 47.16 | 43.39 | C |  | p1 |
| 情感状态总结  (Emotion State summarisation) | 多说话人情感识别 | 情感识别 |  | 29.54 | 72.72 | A | 情感识别：51w | p1 |
| 对话事实抽取  (Conversational Fact Retrieval) | 根据对话内容提取关键事实 | 语音问答 |  | 90.90 | 86.36 | A |  | p1 |
| 多说话人角色判断  (Multi Speaker Role Mapping) | 识别语音中每个说话人的角色 | 音频描述？  语音问答？ |  | 66.66 | 100.00 | A |  | p1 |
| 发音序列解码  (Phonological Sequence Decoding) | 识别单词的发音顺序 | 语音问答 |  | 51.02 | 77.55 | A |  | p1 |
| 情感转换检测  (Emotion Flip Detection) | 识别对话中的说话人情感是否有变化 | 情感识别 |  | 25.00 | 25.00 | A | 情感识别：51w | p1 |
| 对话意图检测  (Key highlight Extraction) | 识别对话意图 | 语音问答 |  | 85.71 | 95.23 | A |  | p1 |
| Music | 时序推理  (Temporal Reasoning) | 识别音乐/歌声中的时序信息 | 音乐描述 |  | 26.78 | 33.92 | C |  | p1 |
| 旋律推理  (Musical Genre Reasoning) | 旋律/曲风识别 | 曲风识别 |  | 38.23 | 64.7 | B | 5w | p1 |
| 歌词推理  (Lyrical Reasoning) | 歌词的情感、主题、含义识别 | 歌词理解 |  | 70 | 90 | C |  | p1 |
| 社会文化解读  (Socio-cultural Interpretation) | 识别音乐出现的社会背景 | 曲风识别 |  | 60 | 80 | B | 5w | p1 |
| 音乐旋律解读  (Melodic Structure Interpretation) | 推断旋律的组织结构和进展过程，以理解其模式、形式和情感表达。 | 音乐描述 |  | 51.51 | 63.63 | B |  | p1 |
| 和旋解读  (Harmony and Chord Progressions) | 和旋识别 | 和旋识别 |  | 27.27 | 54.54 | B |  | p1 |
| 节奏理解  (Rhythm and Tempo Understanding) | 节奏识别 | 节奏识别 |  | 30.43 | 58.69 | B | 5w | p1 |
| 歌声理解  (Musical Texture Interpretation) | 歌声理解 | 曲风识别 |  | 50 | 44.11 | B | 10w | p1 |
| 乐器识别  (Instrumentation) | 乐器识别 | 乐器识别 |  | 74.28 | 65.71 | B | 5w | p0 |
| 音乐情感识别  (Emotional Tone Interpretation) | 音乐情感识别 | 情感识别 |  | 60.6 | 81.81 | B | 5w | p1 |

**MMAU未涵盖的细分能力7项**

* MMAU未涵盖，AIR-Bench中有的能力

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **类型** | **能力** | **任务归类** | **样例** | **数据情况：**  **A:数据比较完善**  **B:数据仍需补充**  **C:暂时没有该数据** | **能力优先级优先级：**  **p0：可落地的能力；**  **p1：刷榜需要覆盖的能力；**  **p2：小众能力、数据搜集困难** |
| Sound | 音频定位 | 音频事件检测 |  |  | p2 |
| Speech | 语音定位 | 语音问答 |  |  | p2 |
| 语种分类 | 语种分类 |  |  | p0 |
| 性别分类（生理性别） | 性别分类 |  |  | p0 |
| 年龄识别 | 年龄分类 |  |  | p0 |
| Music | MIDI音高分析 | 乐器（信号）识别 |  |  | p2 |
| MIDI力度分析 | 乐器（信号）识别 |  |  | p2 |

感谢@孙童提供的文档：

[音频分析 BenchMark 调研-AIR-Bench 篇](https://li.feishu.cn/docx/KBUPdZeT2ob7QVxeejzcAxkvnFd)

[音频分析BenchMark 调研-MMAU 篇](https://li.feishu.cn/docx/RKRmdmfxLoRdWPxXUCscya2MnwP)