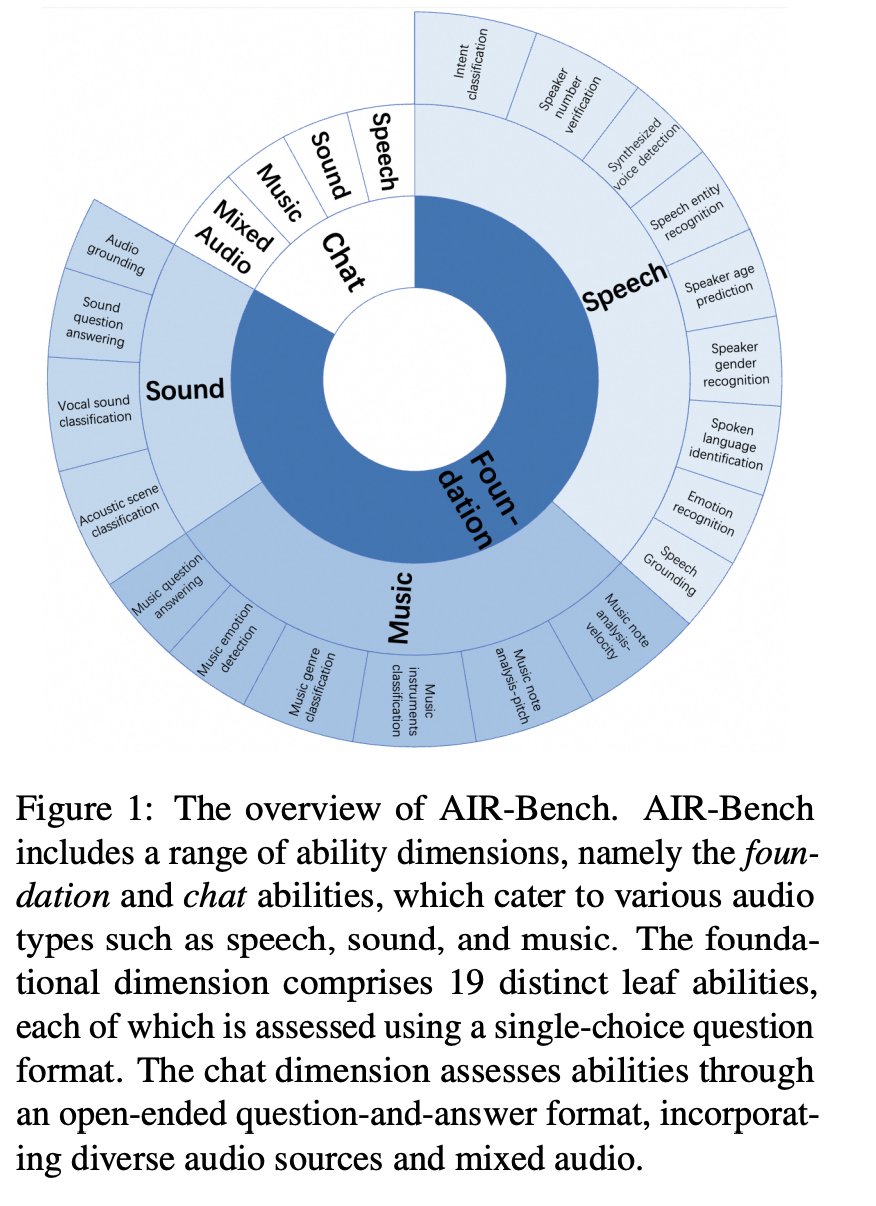
**音频分析 BenchMark 调研-AIR-Bench 篇**

|  |
| --- |
| * 速记： * Paper：https://arxiv.org/pdf/2402.07729 * Repo：https://github.com/OFA-Sys/AIR-Bench * LeaderBoard：https://github.com/OFA-Sys/AIR-Bench * 数据集地址： * Pfs：/mnt/pfs-mc0p4k/asr/team/group\_data/speaker/3oa\_data/AIR-Bench * Lpai： |

1. **简介**

* AIR-Bench是一个音频问答能力的基准测试，其包含了两部分不同的测试Foundation、Chat，分别使用客观指标（ACC）和主观指标（GPT评分），客观指标测试包含了 19 个不同的任务，主观指标测试由 2k 条开放性问答组成；BenchMark 能够从多个方面来考察 ALM 的感知和生成能力。



2. **评测形式**

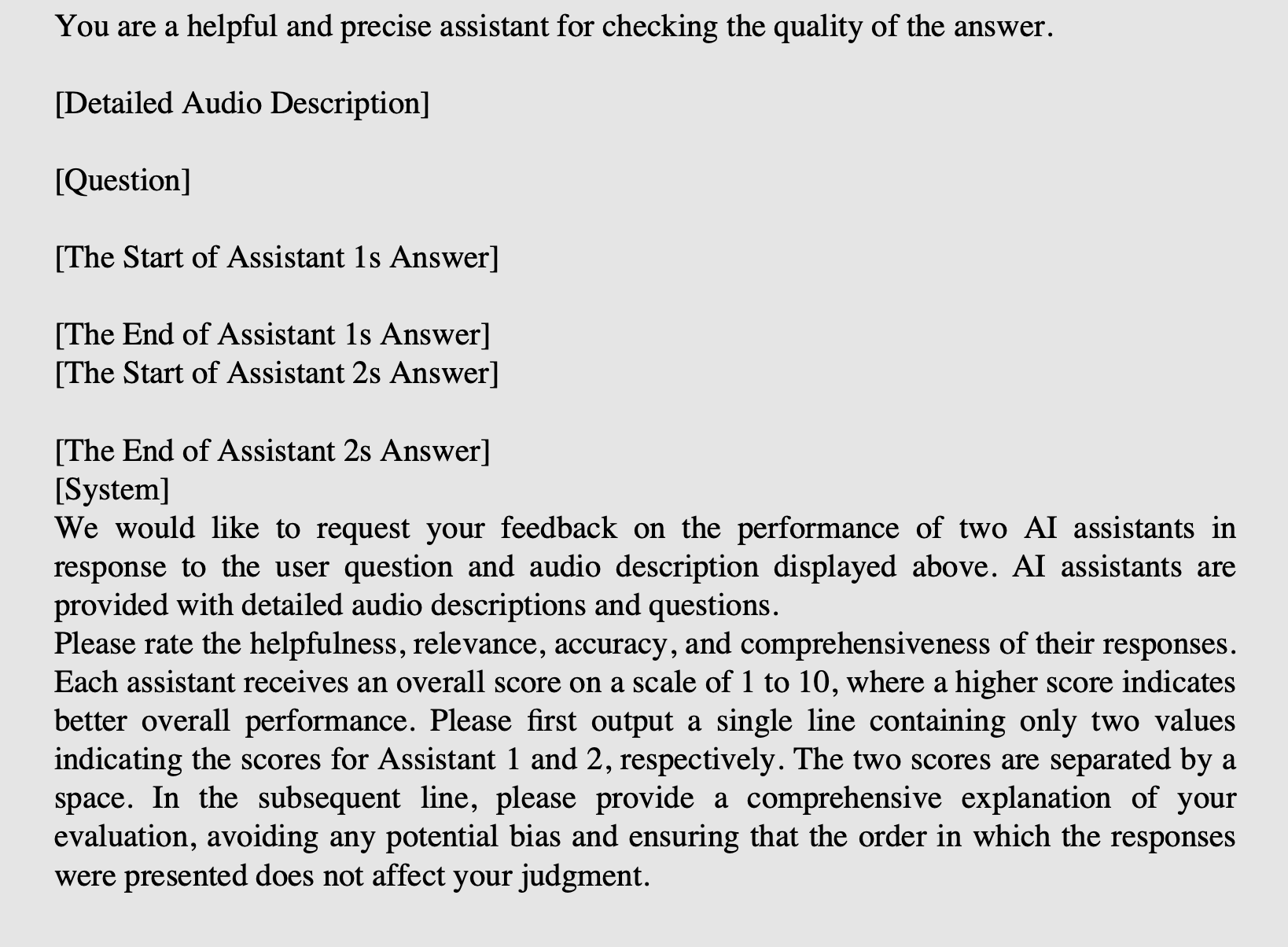
2.1 **Foundation / 基础能力**

* 使用选择题的形式，让 LLM 从给定的 A/B/C/D 选项中选出正确答案，最终计算 ACC 作为评价指标
* 数据样例：

|  |
| --- |
| JSON {  "path": "4515-11057-0078.flac",  "question": "What is the duration of 'him' being mentioned in the conversation?",  "choice\_a": "[1.77, 2.03]",  "choice\_b": "[0.34, 0.46]",  "choice\_c": "[2.54, 2.66]",  "choice\_d": "[2.34, 2.54]",  "answer\_gt": "[1.77, 2.03]",  "task\_name": "Speech\_Grounding",  "dataset\_name": "librispeech",  "uniq\_id": 1 } |

2.2 **Chat / 问答能力**

* 开放性问答，直接将 Query 作为输入，让 LLM 生成答案；最终使用 GPT-4 对生成的答案和原始答案分别打分，作为评价指标
* 打分 Prompt



* 数据样例：

|  |
| --- |
| JSON {  "meta\_info": "[{\"speaker\": \"speaker 1\", \"speaking time\": \"(0.00,13.39)\", \"transcription\": \"so it's a lot more than the like two thousand dollar computers that like you or i can afford you know [laughter] so ah they so they put a lot of effort and this is the kind of stuff that gets sold to the government\"}, {\"speaker\": \"speaker 2\", \"speaking time\": \"(5.33,6.88)\", \"transcription\": \"right right\"}, {\"speaker\": \"speaker 2\", \"speaking time\": \"(13.01,13.91)\", \"transcription\": \"mhm\"}, {\"speaker\": \"speaker 1\", \"speaking time\": \"(13.30,20.20)\", \"transcription\": \"and that's why and our government has a massive sum of money and that's why i\"}, {\"speaker\": \"speaker 2\", \"speaking time\": \"(19.29,21.60)\", \"transcription\": \"isn't it amazing how it has\"}, {\"speaker\": \"speaker 1\", \"speaking time\": \"(21.30,26.59)\", \"transcription\": \"i mean they they ah how what's the national debt six trillion plus\"}, {\"speaker\": \"speaker 2\", \"speaking time\": \"(25.88,28.08)\", \"transcription\": \"[laughter]\"}]",  "question": "What does the first speaker imply about the national debt?",  "answer\_gt": "That it's more than six trillion dollars.",  "path": "568.38\_596.46.wav",  "task\_name": "speech\_dialogue\_QA",  "dataset\_name": "fisher",  "uniq\_id": 0 } |

3. **数据样例**

3.1 **Foundation / 基础能力**

|  |
| --- |
| JSON  {  "path": "588703.mp3",  "question": "What emotive character does this tune portray?",  "choice\_a": "melodic",  "choice\_b": "ballad",  "choice\_c": "melancholic",  "choice\_d": "fast",  "answer\_gt": "melodic",  "task\_name": "Music\_Mood\_Recognition",  "dataset\_name": "MTJ-Jamendo",  "uniq\_id": 24681 } |

3.2 **Chat / 问答能力**

|  |
| --- |
| JSON {  "meta\_info": "[{\"speaker\": \"speaker 1\", \"speaking time\": \"(0.00,8.67)\", \"transcription\": \"like a license or something that just says they don't have to reveal why they do it but they do but then that would pre- then that would all ow them to you know do\"}, {\"speaker\": \"speaker 2\", \"speaking time\": \"(1.61,2.49)\", \"transcription\": \"yeah\"}, {\"speaker\": \"speaker 2\", \"speaking time\": \"(8.14,19.64)\", \"transcription\": \"yeah it cost it casts a lot of suspi cion on them because the employees been thinking oh this guy's going to cost me a lot in m- in medical insurance premiums he's gonna run up our costs for our healthcare [sigh] and um\"}, {\"speaker\": \"speaker 1\", \"speaking time\": \"(14.47,1 6.65)\", \"transcription\": \"oh right right\"}, {\"speaker\": \"speaker 2\", \"speaking time\": \"(19.44,29.63)\", \"transcription\": \"[mn] inadvertently they may want to you know not hire him because of that uh potential cost to the company o f whoa i bet you his health and premiums are really or his health costs is gonna really\"}, {\"speaker\": \"speaker 1\", \"speaking time\": \"(23.81,24.85)\", \"transcription\": \"yeah\"}]",  **"question": "What does 'like a license or something' refer to in the conversation?",**  **"answer\_gt": "It refers to an unnamed permit or authorization that excuses someone from having to disclose certain information.",**  "path": "494.54\_524.17.wav",  "task\_name": "speech\_dialogue\_QA",  "dataset\_name": "fisher",  "uniq\_id": 6 } |

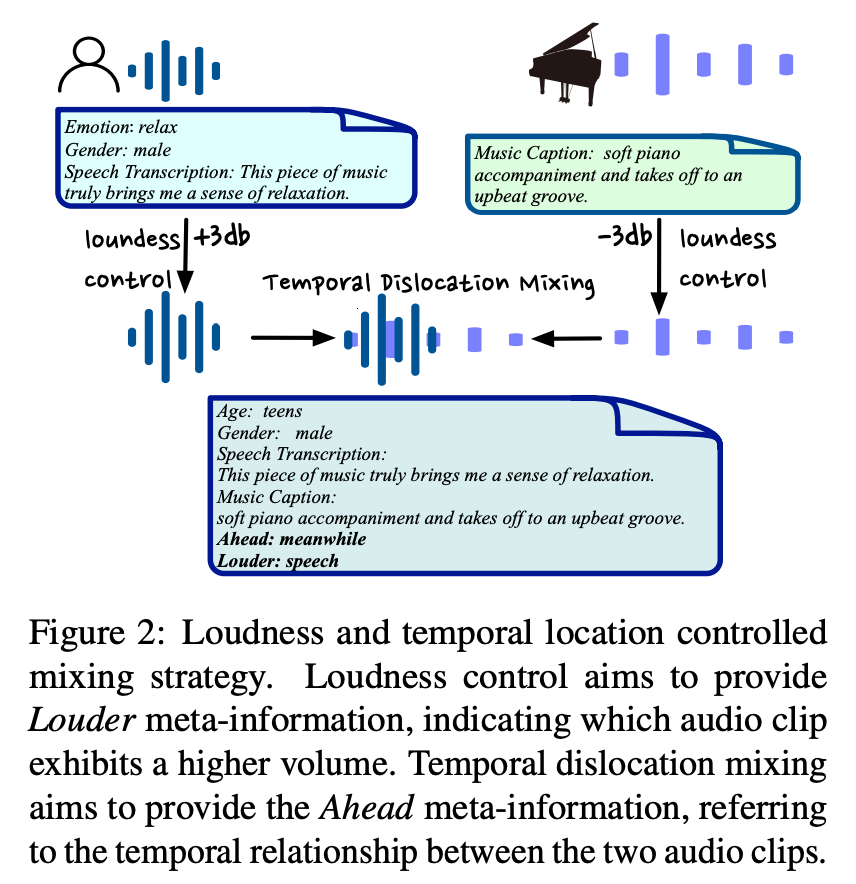
4. **数据构造方式**

4.1 **Foundation / 基础能力**

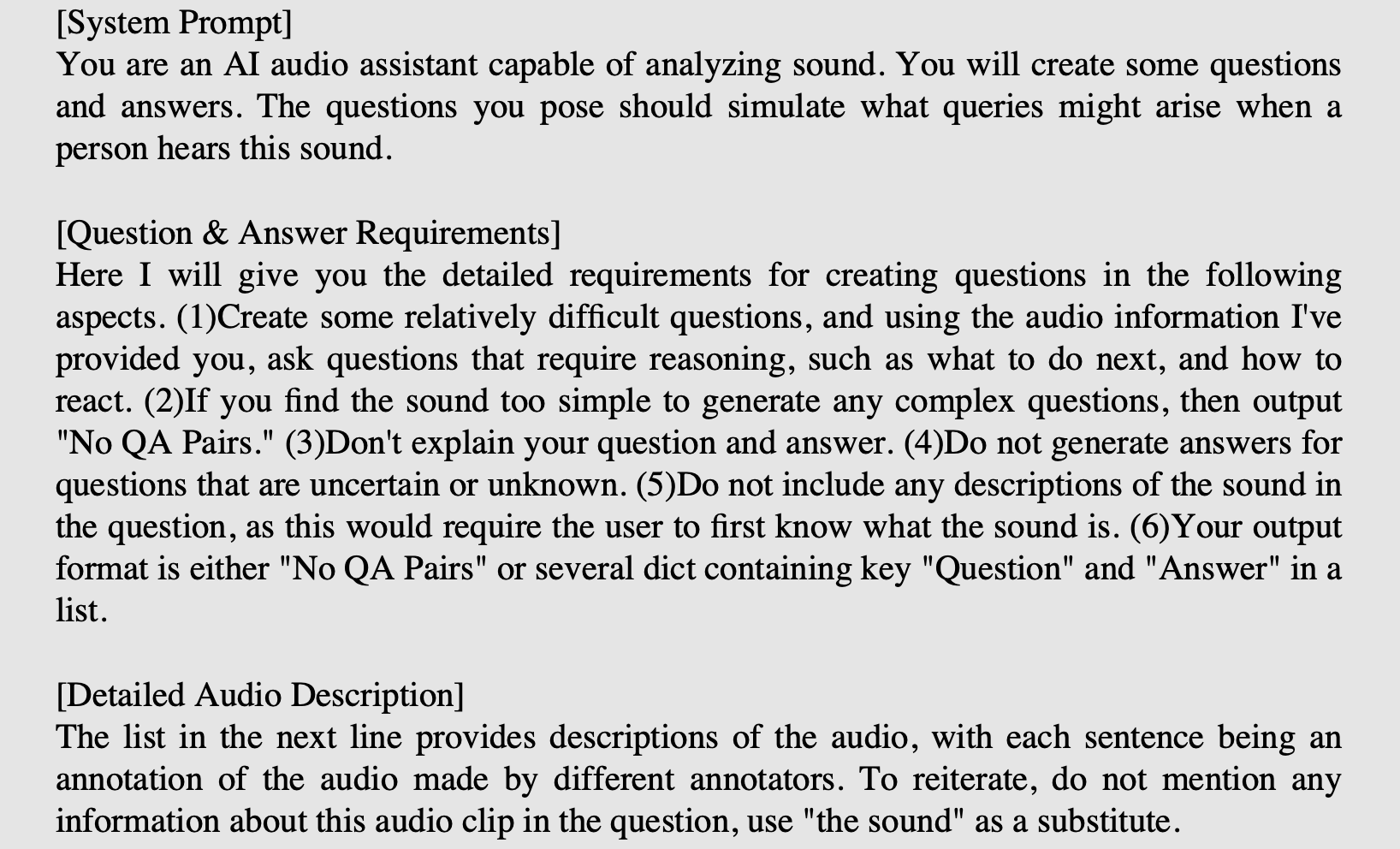
* 音频：从开源数据集的 test/dev 集中抽取 ～19k 音频
* Question：
* 对于原始数据本身就是 QA 格式的，直接复用
* 其他类的，使用 GPT-4 构造 Question
* Answer：
* 对于原始数据本身就是选择题的，直接复用
* 对于分类任务，随机挑选其他类作为干扰选项
* 其他任务，使用 GPT-4 根据正确答案，生成干扰选项
* 构造 Question/Answer 的 prompt 并未开放

4.2 **Chat / 问答能力**

* 音频：
* speech/sound/music：从开源数据集的 test/dev 集中抽取～2k 音频；
* Mixed audio：从上述收集到的音频中采用下图所示的方法进行混流，构造复杂的音频



* Question / Answer：
* 使用 GPT-4 根据音频的 meta 信息来生成
* 模型 Prompt



4.3 **数据来源**

|  |  |
| --- | --- |
| Foundation 音频来源 | Chat 音频来源 |

5. **BenchMark 涵盖的任务**

5.1 **Foundation / 基础能力**

|  |  |  |
| --- | --- | --- |
| 类型 | 任务 | 数据样例 |
| Speech | 语音定位 |  |
| 语种分类 |  |
| 性别分类（生理性别） |  |
| 情感识别 |  |
| 年龄识别 |  |
| 语音实体识别 |  |
| 语音意图识别 |  |
| 说话人计数 |  |
| 合成音检测 |  |
| Sound | 音频定位 |  |
| 人声分类 |  |
| 音频场景检测 |  |
| 音频问答 |  |
| Music | 乐器识别 |  |
| 曲风识别 |  |
| MIDI音高分析 |  |
| MIDI力度分析 |  |
| 音乐问答 |  |
| 音乐情感识别 |  |

5.2 **Chat / 问答能力**

|  |
| --- |
| chat 能力多是开放性的问答问题，主要体现在“音频/音乐描述”任务上，无法归类 |

