RealTalk-CN: 一个具有跨模态交互分析的真实中文语音-文本对话基准

Enzhi Wang[1], **Qicheng Li**[1], **Shiwan Zhao**[1], **Aobo Kong**[1], **Jiaming Zhou**[1],
Xi Yang[2], Yequan Wang[2], Yonghua Lin[2], Yong Qin[1*]
[1]TMCC, College of Computer Science, Nankai University,
[2]Beijing Academy of Artificial Intelligence (BAAI), Beijing, China
Correspondence: liqicheng@mail.nankai.edu.cn, qinyong@nankai.edu.cn

Abstract

近年来，大型语言模型（LLMs）在多模态处理方面取得了显著进展，包括可以实现自然交互和执行特定任务的端到端语音语言模型，用于任务导向对话（TOD）系统。然而，现有的 TOD 数据集主要是基于文本的，缺乏评估基于语音的 LLMs 稳健性所必需的真实语音信号。此外，现有的语音 TOD 数据集主要是英文的，并且缺乏语音不流畅和说话者变化等关键方面。为了解决这些问题，我们引入了 RealTalk-CN，这是首个中文多轮次、多域语音-文本双模态 TOD 数据集，包括 5.4k 对话（60K 句子，150 小时）及配对的语音-文本注释。RealTalk-CN 捕获了多样的对话场景，并注释了自发语音不流畅现象，确保了对语音对话中真实世界复杂性的全面涵盖。此外，我们提出了一种新颖的跨模态聊天任务，真实模拟现实世界用户交互，允许在语音和文本模态之间动态切换。我们的评估涵盖了应对语音不流畅的稳健性、对说话者特性的敏感性以及跨域性能。广泛的实验验证了 RealTalk-CN 的有效性，为中文基于语音的 LLMs 研究建立了坚实的基础。数据和代码将会公开。

近些年，大型语言模型（LLMs）在多模态处理方面取得了显著突破。对于语音输入，这导致了端到端语音语言模型的出现（例如，GPT-4o、Qwen2-Audio 和 Baichuan-Omni）。这些模型可以直接与人类进行端到端语音对话并执行特定任务，提高了对话的效率、自然性和真实性。因此，传统基于文本的面向任务对话（TOD）系统也被转变为语音交互形式，以帮助用户实现特定目标。

然而，现有的面向任务的对话数据集主要是基于文本的，比如大规模的英文数据集 MultiWOZ [**?**] 或中文数据集如 CrossWOZ [**?**] 和 RiSAWOZ [**?**] 。这些数据集缺乏真实的语音信号，使得难以评估模型在自然言语不流利现象（例如填充词、犹豫、重复和自我纠正） [**?**] 或说话者变化（性别/年龄/地区口音） [**?**] 方面的健壮性。这严重限制了它们在评估基于语音的 LLM 时的适用性。

对于语音任务导向的对话数据，资源仍然稀缺。早期的英语数据集如 ATIS [**?**] 和 DSTC2/10 [**?** **?**] 是单轮语音理解数据集，评价场景有限。最近的 SpokenWOZ [**?**] 作为第一个大规模、多轮、多域的英语语音 TOD 数据集，推动了该领域的发展，提供了新的基准。然而，它没有标注不流利语音的语音信号，这对于评估语音驱动的 LLM 在 TOD 场景中的鲁棒性至关重要。此外，中文场景并没有类似的语音 TOD 数据集，而中文口语对话展示出独特的语言现象和社会文化特征 [**?**] ，且中国已经出现了强大的语音驱动甚至全模态 LLM 的快速发展（例如，Baichuan-Audio [**?**] ，GLM-4-Voice [**?**] ，Qwen-2.5-Omni [**?**] ）。这造成了中文语音驱动 LLM 的模型开发与评估之间不断扩大的差距。

此外，尽管之前的讨论集中于仅语音互动，现实世界的对话往往涉及语音和文本之间的模态切换。当前关于语音-文本多模态 TOD 系统的研究 [**?** **?**] 主要假设用户同时提供语音和文本输入，结合两种模态以改进响应。然而，这种方法与人们实际与语音助手互动的方式不符，他们自然地在不同回合中交替使用说话和打字——例如，在电子商务 AI 助理等语音

Preprint.

LLM 应用中，询问关于产品详细信息的语音问题，然后在应用中输入文本的后续请求以获取折扣。这种常见的使用模式揭示了现有评估中的一个关键限制：当前的基准测试无法检验模型是否能够妥善处理这种动态模态切换。为了更好地评估语音基础 LLM 在 TOD 系统应用中的实际表现，我们需要设计反映对话中自然模态变化的任务。为了应对这些问题，我们提出：

- 一个新颖的数据集：第一个中文多轮、多领域的语音文本双模态任务导向对话数据集，RealTalk-CN，包括 5.4k 个对话（60K 个话语，150 小时）的配对语音文本注释。它标注了自发语音的不流利现象，涵盖了广泛的领域和多样的说话者，以反映真实世界对话的复杂性。所有数据都是真人与真人的语音对话。
- 创新任务: 一种新颖的跨模态聊天任务，用户和助手可以在对话中动态地在语音和文本模态之间交替（例如，语音查询之后通过短信或应用程序进行基于文本的反馈）。这种设计真实地模拟了人与助手之间的实际对话，能够全面评估模型的跨模态信息整合和上下文管理能力。
- 综合评估：一种系统的实验协议，涵盖四个数据集子集，结合多种开源和专有基线模型。评估跨越三个关键维度：(i) 在任务导向对话场景中对语音不流利现象（如语气词拖延、犹豫、重复、自我纠正）的鲁棒性，(ii) 对讲话者特征（性别/年龄/地区）的敏感性分析，(iii) 跨领域性能比较。我们采用基于 GPT-4 的自动化评估以确保一致和可扩展的度量标准。
- 严格的数据质量控制：数据收集和标注过程包括多项质量保证措施，包括标准化的脚本设计、受控的录音环境、均衡的说话者多样性表现、详细的标注指南以及严格的管理协议。这些程序集体地确保了为研究目的而设计的数据集的生态有效性和可靠性。

1 相关研究

表 **??** 总结了与其他相关数据集相比，我们数据集的各个方面。相关工作大致可以分为三类：

基于文本的任务导向对话数据集：该领域的英文资源包括 MultiWOZ [**?**]，这是一个广泛使用的数据集，涉及八个领域，包含超过一万个对话。对于中文方面，值得注意的数据集有 CrossWOZ [**?**]，其中包含六千个对话和十万两千个话语，以及 RiSAWOZ [**?**]，一个更为广泛的集合，包含 1.12 万个对话，十五万个话语，并覆盖十二个领域。这些数据集提供丰富的对话状态跟踪标注，但缺乏语音信号。

口语理解 (SLU) 数据集：大多数英语 SLU 数据集，例如 SNIPS [**?**]，依赖转录文本而没有考虑语音识别错误。现存最大的英语 SLU 资源是覆盖十八个领域的 SLURP [**?**]。相比之下，中国的研究在 CATSLU [**?**] 上有了初步进展，这是一个在 ICMI 2019 挑战中推出的多领域音频文本数据集。然而，这些数据集仅是单轮内容理解任务。

基于语音的任务导向数据集：现有基于语音的任务导向数据集仍然稀缺。早期的尝试如 DSTC2 [**?**] 和 DSTC10 [**?**] 仅提供了小规模的自动语音识别输出。SpokenWOZ [**?**] 作为首个大规模的英语语音文本基准但缺乏语音不流畅注释和说话者特征注释。此外，目前并不存在类似全面的中文数据集，形成了一个显著的缺口，阻碍了这一领域的研究进展。

在数据采集阶段，我们优先考虑语音质量和注释一致性。数据集是使用预先编写的脚本构建的，旨在反映自然语言的特征，包括随意的语法、口语词汇、短句结构和松散的句法组织 [**?**]。对话涵盖了多个领域，同时允许参与者在录音中即兴发挥以保持对话的真实性。关键是，收集的数据中有 10 % 特意保留了自发的语音不流畅，如重复、犹豫、自我修正和语气词拖延，以模拟真实世界的条件。

在语音文本对齐方面，我们实施了一种严格的时间戳机制，用于标记话语边界和对话回合。录音会话在安静的室内环境中进行，使用专业麦克风和消费级智能手机麦克风，以确保设备多样性，代表真实使用场景。采用双重录音方法（专用录音机和智能手机）来捕捉真实的声学条件。

通过 300 名志愿者（性别比例为 1:1ś10 % ，年龄为 18-50 岁，符合正态分布，涵盖年轻人和中年人群）确保说话者的多样性。志愿者主要以普通话为主，同时允许轻微的地方口音。性别、年龄和地区分布如图 1 所示。每位参与者贡献了 50 次对话会话。

通过多轮验证并提供详细的指南来解决各种中文语音现象，以保持标注的一致性。转录内容必须严格匹配实际发音，同时容纳方言变化，例如将 “Liu nai” 转为标准的 “牛奶”。普通

Table 1: 四个子集的数据统计。Colloquial 表示文本内容包含上述不流畅的口语标记，而 System 则表示相反。Avg Intent Choices 表示每个话语作为答案的平均候选意图数量。MD 表示多域，SD 表示单域。M、R、S 和 H 分别代表语气词拖长、重复、自我纠正和犹豫。

Subsets	Samples	Avg Utterance Length	Avg Dialog Rounds	Avg Intent Choices	Avg Disfluency Markers
MD-Col	3,837	27.42	8.54	34.51	M: 0.12 R: 0.04 S: 0.11 H: 1.14
MD-Sys	3,837	19.27	7.73	34.77	-
SD-Col	892	25.61	8.14	25.90	M: 0.63 R: 0.07 S: 0.18 H: 0.52
SD-Sys	892	20.76	7.58	27.03	-

话语音特征，包括儿化音，在转录中被保留。标准参考文献用来验证专有名词，而数值表达则一致地用汉字呈现。为了保持韵律的真实性，填充词和话语标记被保留，并对刻意保留的包含不流畅的 10 % 的数据进行了特殊标注。在整个过程中实施了全面的质量控制措施。音

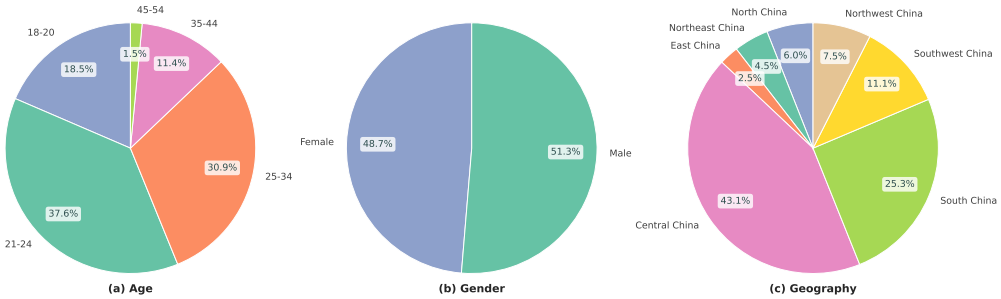

Figure 1: 说话者的分布。该数据集涵盖了大多数年龄组，同时性别比例几乎均等。它涵盖了中国所有主要地区，以探索不同地区口音对语音模型的影响。每个地区所包含的具体省份列在附录 A.1 中。

频片段保持了 0.2-0.3 秒的静音填充，时长优化在 5-6 秒（最长 12 秒）。5 % 的随机抽样协议确保槽值标注准确性超过 95 %。标注流程结合了迭代优化，首先通过三个样本批次的试点标注完善指南，然后再全面实施。数据集的伦理声明在附录 D.1 中进行了描述。

2 RealTalk-CN 数据集概览

RealTalk-CN 是首个中文多轮多领域语音文本双模态任务导向对话数据集，包括 5.4k 个对话会话，其中 1.2k 为单领域，4.2k 为跨领域对话，总计超过 60k 个 utterances，由 113 位说话者贡献。数据集平均每个对话有 12.1 轮，包含 150 小时已验证的音频，覆盖几十个任务导向领域（例如，餐饮、交通、购物），通过真实的人与人互动。每个对话都附有全面的注释，包括对话状态（槽位）、意图、转录、utterance 级别的时间戳、说话者元数据以及自发语音不流利现象的标签（例如，填充停顿、重复、自我修正）。

作为一种面向口语语言的资源，RealTalk-CN 抓住了口头与书面交流风格之间的根本区别——即使在相同的语义环境下，口头对话也表现出随意的语法、口语词汇、片段结构和松散的句法组织。我们的脚本编写过程明确要求使用对话语言模式。关键是，口语的不流利现象为语言理解系统带来了额外挑战。尽管英语的 SpokenWOZ 数据集通过语音识别得出的噪音部分解决了这一问题，但缺乏对不流利类型的明确注释。RealTalk-CN 通过对自发语音不流利现象进行系统标注，推动了这方面的发展，使得可以创建现象特定的子集用于评估基于语音的大语言模型的稳健性。这一设计也促进了如语音不流利纠正之类的二次应用。如图 **??** 所示，我们定义了常见的不流利类别，指导发言者保持自然的对话流（包括自然而然地产生不流利现象），并实施严格的事后注释协议。

2.1 广泛的领域覆盖

RealTalk-CN 全面涵盖了 58 个任务导向对话 (TOD) 领域，包括天气、餐饮、旅行、新闻、购物、金融和医疗保健。它还包含 55 个意图和 115 个槽位类型，这些在之前的数据集中是不

可用的。详细的意图和槽位信息可以在附录 A.2 中找到。数据集被系统地组织为单领域和多领域对话，其中后者涉及 2-5 个交错的领域，以更好地模拟现实世界的场景。例如，与旅行相关的对话可以自然地包括天气查询和餐馆推荐。如图 2 所示，领域分布遵循长尾模式：高频领域（如旅行、天气）覆盖了常见的日常话题，而中频到低频领域确保了主题多样性。在多领域对话中，2,949 个会话涉及两个领域（占大多数），其次是有三个领域的 753 个会话。此外，数据集还包括跨越 4-5 个领域的复杂对话。

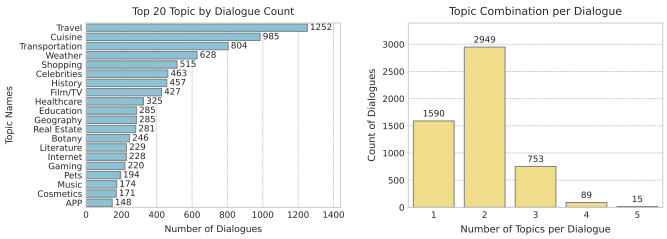

Figure 2: RealTalk-CN 的领域分布和多领域组合分布。

为了进行精细分析，我们根据两个标准将数据集划分为四个子集：(1) 单域与多域组成，以及 (2) 带有标注的语音不流利现象。测试集的详细统计数据在表 1 中提供。训练、验证和测试集划分的详细数据可在附录 A.3 中找到。多域对话相对于单域对话表现出显著更大的复杂性，这体现在更高的平均意图数量上，反映了跨域交互中用户需求的多样性。从语言学上看，标注不流利现象的子集由于自我修正、重复和语气助词拖延等现象，表现出更长的语句长度和更多的对话轮次。这些特征反映了真实的语音模式，并对于基于语音的 TOD 场景下的大语言模型（LLM）提出了更高的鲁棒性挑战。

当前对多模态对话系统的研究主要集中在用户和系统同时接收和处理多种模态（如语音和文本）的情境中。例如，之前的工作提出了多模态语音-文本对话数据集，其中评估任务涉及响应包含语音和文本模态的上下文，旨在通过聚合的语音嵌入增强文本表示，以改进响应。然而，这种同步多模态输入的模式在现实应用中很少出现。在实际的智能语音助手使用中，用户-系统交互通常跨多个回合，并伴随着动态的模态切换。例如，在餐厅预订场景中，用户可能最初通过语音询问，然后通过短信或移动应用继续对话，而不是同时通过语音和文本提供相同的内容。图 **??** 显示了这一模式的具体示例。

为了解决这一差距，我们提出了一项新的跨模态聊天任务，其中对话上下文包含混合的语音或文本话语，没有同时的模态呈现。关键挑战在于模型能够准确理解和跟踪分布在不同模态中的信息，同时有效整合这些异构输入以生成一致且连贯的回应。为了隔离模态转换对语音不流利影响的影响，我们特意采用语音模态来处理包含任何四种不流利标记的回合，而在其他情况下使用文本模态，从而创建一个动态切换的上下文。我们在 SpokenMMC 数据集上设计了多个任务以充分利用其潜力。我们使用与基于语音的 LLM 相同的零样本评估协议，并提供供研究人员使用的训练集。

标准任务导向对话任务，包括对话意图分类、槽填充和端到端聊天。按照 [**?**]，对于意图分类任务，我们计算准确率和与人工评估具有较强相关性的 PANDA 判别 [**?**] 估计方法。对于槽填充任务，我们使用经典的评估指标 F1 和联合目标准确率（JGA） [**?**]。任务的具体示例和我们的评估过程可以在附录 B.1 中找到。对于聊天任务，由于传统指标与人工判断的相关性较弱 [**?**]，我们按照 [**?**] [**?**] [**?**] 实施了基于 GPT-4 的自动评估。所有评估均使用 GPT-4o-mini [1] 进行，包括无参考评估和有参考评估的得分总和。评估提示可以在附录 C.1 中找到。

如前所述，跨模态聊天任务中，用户和助手在对话过程中动态地在语音和文本模态之间切换。

鲁棒性评估任务利用标注的语音不流利现象来检查模型对会话不连贯性的容忍度，使用在 Colloquial 子集上的性能差异作为评估指标。

我们评估了几种端到端的语音大型语言模型，包括 Qwen2-Audio-7B-Instruct [**?**]、Baichuan-Audio-Instruct [**?**]、GLM-4-Voice-9B [**?**]，以及最近的全模态基础模型（MiniCPM-o [**?**] 、Baichuan-Omni-1d5 [**?**] 、Qwen2.5-Omni-7B [**?**] ）。为了比较，我们还包括结合

[1] GPT-4o-mini-2024-07-18

Table 2: 对比模型在 RealTalk-CN 数据集上的意图分类（IC）和槽填充（SF）任务的表现。Acc 是意图分类的准确率，Pipeline 代表 Whisper-large-v3 + GPT-4o。GPT-4o-Audio 使用的是迷你版本。PAN. 代表 PANDA 评分。

Subsets	MD-Col				MD-Sys				SD-Col				SD-Sys				
Tasks	IC		SF		IC		SF		IC		SF		IC		SF		
Metrics	Acc	PAN.	F1	JGA	Acc	PAN.	F1	JGA	Acc	PAN.	F1	JGA	Acc	PAN.	F1	JGA	Average
Pipeline	53.56	53.56	45.90	26.09	54.83	54.83	48.81	31.99	59.75	59.75	38.55	20.68	62.44	62.44	45.17	28.52	46.68
Baichuan-Audio	30.70	30.70	48.60	30.46	28.20	28.20	54.11	40.66	27.47	27.47	39.96	23.94	30.49	30.49	47.15	33.80	34.53
GLM-4-Voice	26.40	26.40	10.48	19.41	19.49	19.49	9.31	39.59	32.51	32.51	9.58	15.31	28.36	28.36	10.64	19.19	21.69
Qwen2-Audio	24.76	24.78	47.67	30.67	18.14	18.26	52.76	25.48	27.47	27.50	38.58	23.78	23.09	23.15	45.69	32.92	30.29
Baichuan-Omni	36.17	36.19	48.06	28.88	34.53	34.54	52.99	39.81	38.68	38.79	39.99	24.42	34.53	34.53	46.34	31.34	37.49
MiniCPM-o	39.74	39.74	46.02	26.56	35.84	35.84	49.91	33.41	41.82	41.82	36.82	20.52	39.01	39.01	44.40	28.52	37.44
Qwen2.5-Omni	24.52	24.54	47.70	30.88	18.17	18.25	52.55	39.75	27.58	27.64	39.57	24.43	22.87	22.90	45.67	33.45	31.28
GPT-4o-Audio	46.31	46.31	51.53	31.93	45.04	45.04	53.65	38.39	48.21	48.21	43.16	24.27	49.10	49.10	48.45	33.10	43.86

Whisper-Large-V3 [**?**] 与纯文本大型语言模型（GPT-4o）和 GPT-4o-Audio-mini [2] 的管道方法，旨在衡量当前开源/闭源语音大型语言模型与传统管道方法之间的性能差距。我们的模型评估基于 [**?**] 中的代码。

2.2 结果 & 讨论

Table 3: 模型在 RealTalk-CN 数据集的聊天任务中的表现。流水线表示 Whisper-large-v3 + GPT-4o，分数是 GPT-4o-mini-Audio 的得分，满分为 5。* 注意在输入时，GPT-4o-Audio-mini 不支持助手端的语音模式。

Models	MD-Col	MD-Sys	SD-Col	SD-Sys	Avg
Pipline	8.92	9.12	8.84	9.12	9.00
Baichuan-Audio	7.44	7.80	7.79	7.68	7.67
GLM-4-Voice	8.30	8.54	8.24	8.39	8.37
Qwen2-Audio	7.82	8.11	7.85	8.06	7.96
Baichuan-Omni	7.32	7.51	7.34	7.72	7.47
MiniCPM-o	8.22	8.41	8.19	8.33	8.29
Qwen2.5-Omni	7.83	8.14	7.78	8.04	7.95
Gpt-4o-Audio-mini	8.66 *	8.79 *	8.71 *	8.77 *	8.73 *

语音不流利影响槽位填充和聊天任务。在表 2 的口语子集上，大多数模型在槽位填充任务中的性能显著下降，例如 Pipeline 方法的 JGA 值从 31.99 下降到 26.09，而 Baichuan-Audio 在 MD-Col 子集上从 40.66 下降到 30.46，而意图分类任务并没有受到显著影响。这种差异源于模型在处理语音特征和语义信息方面的性能差异。意图分类任务主要依赖于捕捉用户句子中的核心语义，即使在不流利的语音表达情况下，模型也可以通过上下文语义保持高准确率。然而，在槽位填充任务中，模型需要在连续语音流中准确识别和提取特定槽位信息，这依赖于模型捕捉关键信息的能力。当语音中存在自我重复、语法错误或非标准表达时，模型在槽位边界和内容提取中容易混淆，这对模型提出了更高的要求。聊天任务（如表 3 所示）需要获取核心语义，提取一些关键信息，并在整合后作出响应。在口语子集上，性能也有所下降。

流水线方法和端到端模型各有优缺点：在表格 2 中展示的意图分类任务以及表格 3 中的聊天任务中，流水线方法表现优于端到端模型。该结果可归因于在流水线方法中，Whisper-large-v3 将语音转换为文本后，GPT-4o 负责文本语义理解和意图分类，充分利用了 GPT-4o 在文本理解方面的强大能力，而这两个任务更依赖语义。相比之下，虽然端到端模型可以直接从语音输入中捕捉语音和文本信息，但它在语义理解能力上往往下降。然而，在槽填充任务中，端到端模型表现优于流水线方法。在表格 2 中，许多模型的 F1 和 JGA 值高于流水线方法。槽填充等任务更依赖语音中的细节，而端到端模型能够更好地适应语音质量和表达的清晰度（如不流畅性），更有利于捕获语音的详细信息。此外，计算所有任务性能的平均值后，流水线方法保持了领先地位，而 GPT-4o-Audio-mini 排在第二位，并在多个任务中总体上领先于其他端到端模型，表明其在语音理解和多模态特征融合方面具有更强的能力。

多域复杂性主要影响意图分类能力。在多域数据集上，意图分类任务的性能显著低于单域数据集。例如，在表 2 中，与 SD-Col 相比，MD-Col 的意图分类的 PANDA 通常减少了 2-5 个百分点，而槽填充任务的性能没有受到显著影响。这个差异反映了模型在处理语义多样性和上下文切换方面的局限性。意图分类任务本质上依赖于模型在语义空间中正确分类用户意图，而多域场景涉及多个任务和上下文，因此模型需要具有更强的跨域语义泛化能力。

[2]GPT-4o-mini-音频预览

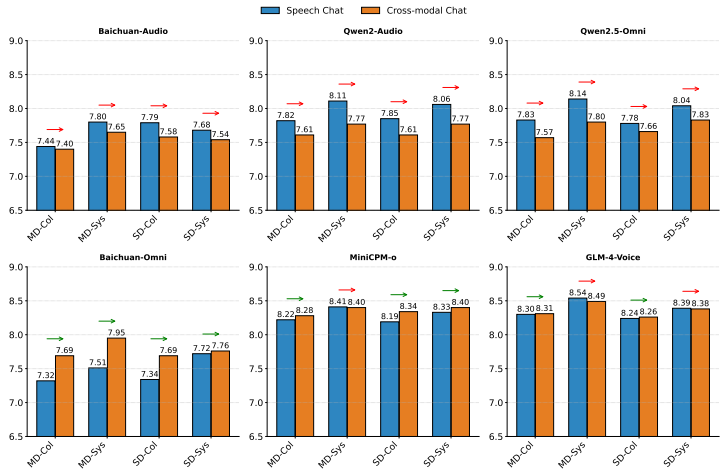

Figure 3: 纯语音聊天任务和跨模态聊天任务之间的性能对比。结果使用配对 t 检验 [?] ($p<0.05$) 进行了分析，差异显著的任务是 Qwen2-Audio、Baichuan-Audio、Baichuan-Omni 和 Qwen2.5-Omni，而差异不显著的任务是 MiniCPM-o 和 GLM-4-Voice。详细过程可见附录 C.3。

然而，当前的端到端语音模型在面对域切换时往往缺乏语义表示和上下文适应性，可能错误地混淆不同域的语义，从而影响意图分类的准确性。相比之下，槽填充任务在多域场景中表现得更为稳定，因为它依赖于对特定槽的识别。模型只需要识别预定义的槽信息，而域的变化对这些槽的定义影响不大。

跨模态聊天任务中的性能差异。图 3 展示了语音基础模型在纯语音聊天任务和跨模态聊天任务中的多样化表现，揭示了不同模型行为。第一类包括表现退化的模型，如 Baichuan-Audio、Qwen2-Audio 和 Qwen2.5-Omni。这些模型在跨模态场景中显示出一致的指标下降，例如，Qwen2-Audio 的 MD-Col 分数从 7.82 降至 7.61。通过详细的案例研究，我们发现模型在整合和响应异构模态信息时存在一些问题，包括遗忘不同模态的前一个上下文的关键信息和语义，响应丰富度质量降低，以及响应文本退化。详细案例可以在附录 E.1 中找到。第二类是表现稳定的模型，包括 GLM-4-Voice 和 MiniCPM-o。特别是，第三类包含性能提升的模型。Baichuan-Omni 的 MD-Col 分数显著增加，从 7.32 提升到 7.69，这表明该模型在对话历史中受益于文本模态替换。

2.3 说话人验证

说话者相关属性，如年龄 [?] [?] 和区域口音 [?] 显著影响语音模型的性能。现有研究 [?] [?] 表明，当前语音模型在处理老年用户的语音时，由于与年龄相关的声带退化而表现出显著的性能下降 [?] 。虽然 [?] 调查了全球英语口音变化对语音 LLM 的影响，但对中文语音 LLM 的类似研究仍然稀缺，特别是在面向任务的对话领域。鉴于中国广泛的地理分布及多样的普通话口音和广泛的年龄分布，我们系统地分析了在面向任务的对话场景中，年龄和地区口音对基于语音的 LLM 的影响，结果如图所示 4 。

从任务的角度来看，模型在意图分类和槽值填充时对说话者特征表现出不同的敏感性。少数模型在意图分类上显示出统计学上的显著差异，特别是在年龄组之间，除了在中国西北和东北地区的地理分布上有显著变化。相反，槽值填充对年龄和地区因素都表现出明显的易感性（$p<0.05$），这加强了我们在 5.3 节的发现，即细粒度的语义解析任务更容易受到语音变异性的影响。

人口统计分析显示，45-54 岁用户的槽填充性能显著下降，伴随着强烈的负向 Cohen's d 值，这证实了声带老化和训练数据分布不匹配的复合挑战。地区比较进一步表明，华南和华北用户的表现优于西南或西北地区，这表明训练数据中普通话标准的接近性更好地服务于语言中心区域，而外围地区更强的方言特征降低了模型的泛化能力。

模型的适应能力差异很大。在槽填充任务中，Baichuan-Audio 和 GPT-4-Audio-mini 在主流用户中表现优异，但对老年用户的表现降低了超过 20 %，这表明对声乐老化的鲁棒性有限。相反，Qwen 模型在各年龄组中保持了一致的表现。在区域适应方面，MiniCPM-o、GPT-4-

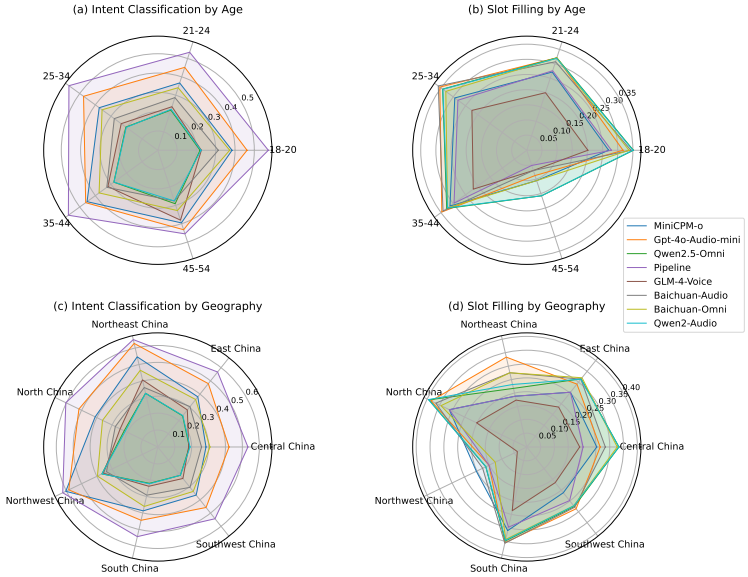

Figure 4: 雷达图展示了说话者的年龄和地理位置对对话意图分类和插槽填充任务的影响。结果是模型在整个 RealTalk-CN 数据集上的平均表现，按年龄和地理位置（指祖籍）进行分组。我们进行了整体的 Kruskal-Wallis H 检验 [**?**] ($p<0.05$)，并进行了比较的 Mann-Whitney U 检验 [**?**] ($p<0.05$) 来判定显著性，并计算了 Cohen's d 效应量 [**?**] 。详情见附录 C.2。

Audio-mini 和 Qwen 展现了优越的跨区域泛化。这些发现突出了增强模型公平性的关键方向——提高在不同用户群体之间的稳定性。

3 结论

在本文中，我们介绍了 RealTalk-CN，这是首个大规模的中文语音文本双模态对话基准，全面捕捉了语音中的不流畅表达、多样的说话者特征以及跨模态交互。我们的评估展示了此数据集在语音鲁棒性、说话者适应性和跨模态一致性方面对模型进行基准测试的有效性。我们提出的跨模态聊天任务进一步揭示了模型在处理动态模态切换方面的局限性。RealTalk-CN 为中文多模态对话研究设立了新的标准，提供了推进基于语音的语言模型的关键资源。