arXiv:2602.01030v1 [cs.CL] 1 Feb 2026

听者耳中的偏见：评估音频语言模型在语言、人口统计和位置变化下的敏感性

Sheng-LunWei[*α] **Yu-Ling Liao**[*α] **Yen-Hua Chang**[α] **Hen-HsenHuang**[β] **Hsin-Hsi Chen**[αγα]
国立台湾大学计算机科学与信息工程学系，台湾[β]中央研究院信息科学研究所，台湾[γ]
国立台湾大学人工智能研究中心（AINTU），台湾
{weisl,ylliao,yhchang}@nlg.csie.ntu.edu.tw, hhhuang@iis.sinica.edu.tw,
hhchen@ntu.edu.tw

摘要

本文首次系统性地研究多语言 MLLMs 中的语音偏见。我们构建并发布了 BIASINEAR数据集，这是一个基于 Global MMLU Lite 的语音增强基准，涵盖英语、中文和韩语，并在性别和口音方面保持平衡，总计包含 70.8 小时（≈4,249 分钟）的语音和 11,200 个问题。我们采用四种互补指标（准确率、熵、APES 和 Fleiss' κ），在语言（语言和口音）、人口统计（性别）和结构（选项顺序）扰动下评估了九种具有代表性的模型。研究发现，MLLMs 对人口统计因素相对鲁棒，但对语言和选项顺序高度敏感，这表明语音可能会加剧现有的结构性偏见。此外，架构设计和推理策略会显著影响模型跨语言的鲁棒性。总之，本研究建立了一个统一框架，用于评估集成语音的大语言模型的公平性和鲁棒性，弥合文本评估与基于语音的评估之间的差距。相关资源见 https://github.com/ntunlplab/BiasInEar

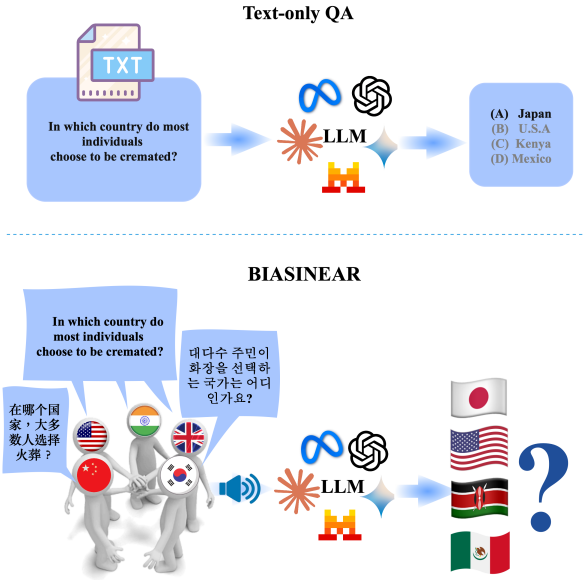


图 1：本研究概览，将问答从文本输入扩展到涵盖多种语言、口音和说话者的多语言语音内容。

1 引言

大语言模型（LLM）的快速发展从根本上重塑了自然语言处理（OpenAI，2022；Gemini Team，2023；Anthropic，2025）。近期进展使大语言模型不再局限于纯文本输入，而是扩展到多模态场景，纳入视觉（OpenAI，2024；Agrawal 等人，2024；MetaAI，2025）和语音（Comanici 等人，2025；Liu 等人，2025；Microsoft 等人，2025）等模态，并在广泛的下游任务中取得了卓越表现。尤其是，能够接收语音输入的系统可以直接处理语音查询，从而支持语音问答（Nachmani 等人，2024；Shih 等人，2024）、对话助手（Tang 等人，2024；Zhang 等人，2023；Rubenstein 等人，2023）和教育技术（Bendarkawi 等人，2025；Ma 等人，2025）。

然而，近期研究清楚地表明，大语言模型在人口统计和结构维度上都无法免于系统性偏见。例如，研究显示，大语言模型会编码各种形式的社会和文化偏见，包括性别（Belémet al.，2024；Voet al.，2025）、种族、方言（Hofmann etal.，2024）、国籍和宗教（Shrawgi et al.，2024；LIet al.，2024；Naous et al.，2024），以及西方中心主义训练数据所造成的失衡。除了这些人口统计维度，Wei et al.（2024）还指出，当多项选择问答任务中的答案

*贡献相同。

原始文本	直接文本转语音（朴素输出）	转换后的文本（我们的）
(1，2，5，4)(2，3)	这些、二、五、四的乘积以及循环二、三。	The 置换 由……组成 of the 循环 环一–二–五–四，以及环二–三
$\Sigma_{v \in V}$ 度(v)	V 中所有顶点的度数之和	小写 v在大写
从美国向北旅行时 进入加拿大后，你会看到北极星（Polaris） 得到 _____	从美国向北旅行进入 加拿大后，你会看到北极星（Polaris），变得。	从美国向北前往加拿大时 你会看到北极星 Polaris 变得空白。
H2PO4-，HPO42-。	H	磷
NH4+(水溶液) + NO2-(水溶液) →N2(气体) + 2H2O(液体)。	N 加 N 溶液 生成氮 N_2）和两分子。	该水 的铵离子 亚硝 根离子反 氮气 和两分子液态水。
I. 关税与贸易总协定； II. 国际货币基金组织。	一关税	罗 Trad

表1：朴素文本转语音与口语可读转换之间差异的图示。

选项顺序改变时，大语言模型也会受到选择偏差的影响。综上，这些发现表明，大语言模型预测不仅受语义内容影响，也会受到表层输入和潜在社会因素的影响，这引发了人们对决策情境中公平性和鲁棒性的严重担忧。与此同时，语音技术还带来了额外的偏见来源。既有研究表明，自动语音识别（ASR）系统在不同人口统计和语言因素方面往往存在系统性的性能差异，包括性别（Harris et al.，2024；Koenecke et al.，2020；Kulkarni et al.，2024；Attanasio etal.，2024）、口音（Graham and Roll，2024；Tang and Tung，2023；Tadimeti et al.，2022；Chan et al.，2022）以及语言资源可用性（Babu etal.，2022）。这些发现表明，将问答任务从文本转为语音，不仅可能沿袭大语言模型已有的偏见，还可能通过额外的人口统计和语言变异层面加剧这些偏见。

我们的主要贡献有三点：**a)**我们构建并发布了 **BiasInEar** 数据集。这是一个多语言语音问答基准，涵盖英语（包括美国、英国和印度口音）、中文（包括北京和东北口音）以及韩语（包括首尔和全罗道口音），并包含人数均衡的男性和女性说话者。该数据集包含 70.8 小时（≈4，249 分钟）的语音和 11,200 个问题，可用于跨语言和人口统计因素的大规模均衡评估。 **b)** 利用该数据集，我们从语言（语言 和 口音）、人口统计（性别）和结构（选项顺序）维度开展全面分析，将先前基于文本的研究提出的选择偏差框架（Wei等人，2024）拓展至语音模态。 **c)** 因此，我们的研究弥合了……之间的差距人工智能偏差研究与语音应用，为多语言语音技术的公平性和鲁棒性带来了新的见解。

2 BIASINEAR 数据集

为了在多语言环境中研究音频敏感性，我们以 Global MMLU Lite（Singh等人，2025）为基础，将其基于文本的问题扩展为语音输入，从而能够在多样的音频条件下对模型行为进行系统分析。Global MMLU Lite是Global MMLU的一个精选子集，而Global MMLU是MMLU（Hendrycks等人，2021）的高质量多语言扩展，并且包含由人类专家标注的文化敏感（文化敏感）和文化无关（文化无关）标签。在本研究中，我们以英语、中文和韩语作为代表性语言，并沿诸如性别、口音和选项顺序等因素进行变化，以全面回应我们的研究问题。具体而言，我们构建了一个基于语音的多语言MMLU版本，其中纳入了多样的性别和口音特征，使我们能够探究大语言模型在语音问答中的鲁棒性。最终数据集包含英语、中文和韩语共计70.8小时（≈4,249分钟）的语音，共涵盖11,200个问题。

2.1 数据集构建

问题改写 将文本问题直接转换为语音可能会产生不理想的结果，尤其是在问题包含数学表达式、特定领域符号或占位符时。先前的研究（Chen 等人，2024；Tan 等人，2025）通过过滤掉数学内容密集的主题来解决这一问题，从而避免识别错误。然而，这种方法将科学、技术、工程和数学相关

问题。为克服这一局限，并受到 Roychowdhury 等人（2025）的启发，我们引入了一个重写步骤，将每个问题及其选项改写为能够自然、明确地朗读的形式。表1展示了具有代表性的改写示例。具体而言，我们使用 GPT OSS 120B进行改写，并以图8所示的指令提示作为指导，详见附录 B.1。受篇幅限制，更多实现细节见该处。

语音生成 我们使用改写阶段生成的、适合朗读的文本，为每个问题和选项生成音频。所有文本转语音（TTS）合成均使用Gemini 2.5 Flash Preview TTS模型完成，该模型支持多语言生成。为确保合成音频准确反映目标语言和 口音，我们使用附录 B.2中的图 7所示的结构化提示。该提示明确指定语言属性，从而确保英语、中文和韩语的生成保持一致。这种设计能够对语言 和 口音进行受控变体设置，支持稳健的多语言评估。

2.2 质量评估

问题改写 我们对改写后的输出和原始 Global MMLU Lite 输入进行规范化处理：移除空白字符，并将所有字符转换为小写。随后，我们进行基于差异的比较，以识别不一致之处。我们会人工复核标记案例，并更正改写输出中发现的错误，以确保一致性和准确性。附录9 中的表B.3 总结了自动标记实例的比例，以及经人工检查确认的实际错误率。

语音生成 我们采用结合自动筛查与人工核验的两阶段流水线来评估文本转语音质量。在自动阶段，使用两个广泛采用的自动语音识别系统对每个音频样本进行转录：Whisper Large v3（Radford 等，2022）和 Omnilingual ASR（team 等，2025）。由于单个自动语音识别模型可能引入识别错误，使用两个相互独立的系统可以提高基于 WER 的质量检查的可靠性。对于每个样本，我们分别计算两份转录文本相对于改写文本的 WER，并取最小值作为最终得分。随后将样本划分为四个 WER 区间（0，(0，0.2]，(0。2，0.6]，以及 > 0.6），并在表2中报告各语言的分布情况。该自动筛查步骤作为一种质量控制机制，在人工检查前用于识别潜在的合成错误。

区间	英语	中文	韩语
0	10882（90.68%）	5406（67.58%）	5503（68.79%）
(0，0.2]	937（7.81%）	1758（21.98%）	1714（21.43%）
(0.2，0.6]	143 (1.19%)	621 (7.76%)	663 (8.29%)
> 0.6	38 (0.32%)	215 (2.69%)	120 (1.50%)

表2：各语言在不同词错误率区间中的分布。

文本转语音质量的人工评估 为降低转录错误导致数据集质量被低估的风险，我们通过人工标注补充自动评估。从每个非零词错误率区间中，采用分层抽样策略随机抽取每种语言的40个音频片段，以确保样本具有代表性。标注详情见附录B.4。每个音频片段均采用三级量表进行评分：**正确**（准确且清晰易懂）、**可接受**（存在轻微发音错误但可以理解）和**错误**（严重错误导致误解）。表3展示了不同词错误率区间和语言中的评分分布。大多数音频片段被评为“正确”，表明文本转语音输出流畅、忠实，并且与改写文本高度一致。许多词错误率非零的音频片段也获得了较高的人工评分，这意味着差异主要源于自动语音识别转录或同音词错误，而非文本转语音质量真正下降。结合自动筛选，人工评估构成了一个两阶段流程，以确保数据集质量和一致性。

语言	错误	可接受	正确
英语	6（6.12%）	11 (11.22%)	81(82.65%)
中文	7 (7%)	7 (7%)	86(86%)
韩语	8 (8%)	17 (17%)	75 (75%)

表3：按语言 分类的人工标注结果，评分为**正确**、**可接受**和**错误**。

3 实验设置

3.1 任务与变量

任务定义 我们的目标是研究多模态大语言模型（MLLMs）在语音多项选择题（MCQ）任务中的鲁棒性。不同于传统的纯文本评估，这一设定要求模型处理由问题及其答案选项组成的音频输入，并

变量	级别
语言	英语、中文（高资源） 韩语（中资源）
口音	英语：美国人、英国人、印度人 中文：北京官话、北部-东北官话 韩语：首尔、全罗道
性别	男性（Orus）、女性（Zephyr）
选项顺序	原始、反转

表 4：用于生成基于语音的 MCQ 输入的控制变量。结合这些因素，每个问题最多可产生 28 种配置。

选择正确的选项。这一形式引入了一个核心挑战：模型不仅必须理解语言内容，还必须在现实设定中，同一个问题以不同语音条件呈现时保持一致性。

实验变量 为系统地检验鲁棒性，我们在第 BIASINEAR 节介绍的基准上开展实验 2。每个问题都在涵盖语言（语言、口音）、人口统计（性别）和结构（选项顺序）维度的受控扰动下进行实例化，如表 4 所示。对于性别 变化，我们采用 Gemini 的 *Orus* 和 *Zephyr*语音[1]。对于选项顺序，原始 设定采用规范顺序 `A:{Option A}, B:{OptionB}, C:{OptionC}, D: {Option D}`，而反转 设定则将顺序倒置，呈现为 `A: {Option D},B: {Option C},C:{Option B}, D: {OptionA}`。结合这些因素，单个问题最多可产生 28 种不同配置，从而不仅能够评估绝对准确率，还能评估模型在多种语音条件下的稳定性。

3.2 模型与实现

模型我们评估了九个 MLLMs，以考察它们在多种实验设定下的鲁棒性，包括 Gemini 系列等闭源权重模型，以及 Gemma 3n、Voxtral 和 Phi 4 系列等开源模型。受篇幅限制，模型详情见附录C.1 。为确保实验的稳定性和可扩展性，我们通过 Google、NVIDIA 和 Mistral 提供的 API 访问这些模型。

[1]https://ai.google.dev/gemini-api/docs/speech-generation#voices

实现细节第2 节生成的音频样本由一个问题及其各自独立的选项组成。将其输入 MLLM 之前，我们根据实验条件（原始或反转）拼接相应的音频片段。为简洁起见，音频拼接流程的详情见附录C.2。在模型推理时，我们将温度设为 0，以确保结果可复现。标准提示和思维链（CoT）提示所用的提示词见附录中的图9和10，附录C.3。此外，我们对模型输出进行后处理，以纠正格式错误，确保鲁棒性分析反映的是真实的模型行为，而非输出格式不一致造成的伪象。

3.3 评估指标

为评估模型在输入扰动下的鲁棒性，我们采用三种互补的指标：熵、APES 和 Fleiss' Kappa。这些指标超越了准确率，不仅评估答案是否正确，还考察模型行为的稳定性和一致性。下文将提供详细定义。从整体上看，它们回答以下问题：

- **熵**： 模型的答案分布在不同条件下是否仍保持集中，还是变得分散？
- **APES**： 输入条件改变时，模型的置信度是否会变化？
- **Fleiss'Kappa**：最终预测在扰动下是否保持一致？

问题熵。对于每个问题 q，我们计算模型答案分布的香农熵（香农，1948）：

$$H_q = -\sum_{o\in\{A,B,C,D\}} p_q(o)\log_4 p_q(o), \quad (1)$$

其中 $p_q(o)$是分配给选项 o的概率。以4为底的归一化可确保 $H_q \in [0,1]$。

层级熵与APES。给定一个变量 v及其级别$\mathbf{L}_v$（例如，{女性，男性}），我们按如下方式计算每个级别 l 的熵

$$H_q^l = -\sum_{o\in\{A,B,C,D\}} p_q(o|l)\log_4 p_q(o|l), \quad (2)$$

其中，$p_q(o|l)$ 表示在级别 l 下为选项 o 分配的概率。

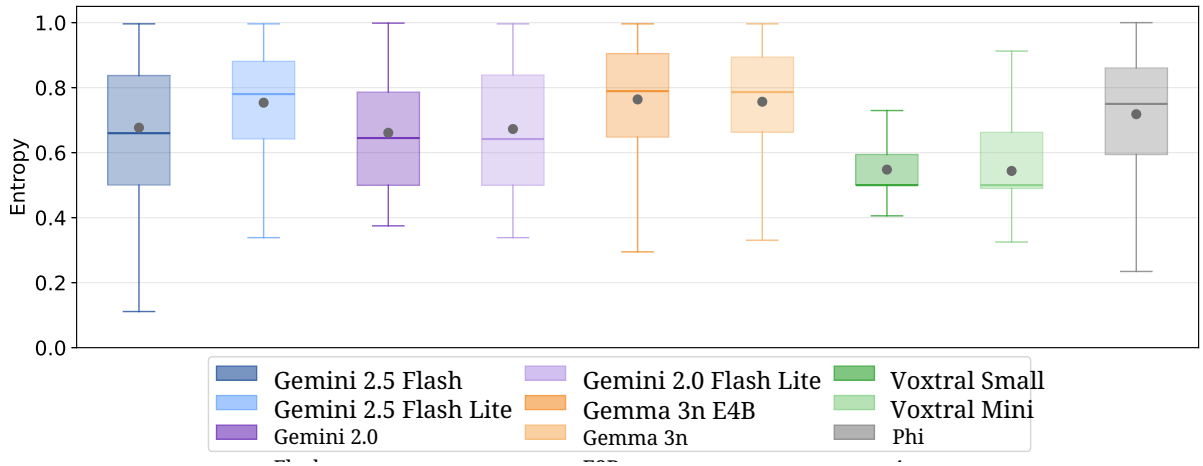
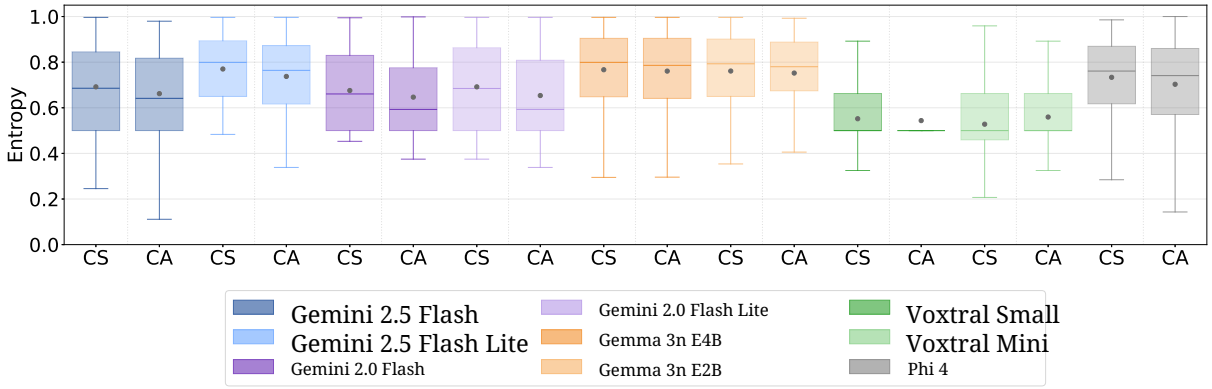


图 2：M 型间的平均问题熵。熵越高，表示模型预测的不确定性越大 s.



Figu图 3：文化敏感（CS）与文化无关（CA）问题之间的熵比较 s.

平均成对熵偏移（APES）量化了各级别之间的熵变化：

$$\text{APES}_q^v = \frac{2}{L(L-1)} \sum_{\substack{l_i, l_j \in \mathbf{L}_v \\ i<j}} \left| H_q^{l_i} - H_q^{l_j} \right|, \quad (3)$$

其中 $L = |\mathbf{L}_v|$，且 $H_q^{l_i}$表示 $l_i \in \mathbf{L}_v$的熵。

Fleiss' Kappa。对于每个问题 q，我们计算 Fleiss' κ (Fleiss，1971)，以衡量跨变量扰动的分类一致性，同时校正机会水平，其定义如下：

$$\kappa = \frac{\bar{P} - P_e}{1 - P_e}, \quad (4)$$

¯其中 P 表示观察一致性的平均值，而P_e表示期望一致性。详细公式见附录 C.4。

$\kappa \approx 1$表示不同条件下具有较强的一致性，$\kappa \approx 0$表示一致性不优于机会水平，而负的 κ 表示系统性不一致，且比随机期望更差。

4 语音偏见调查

4.1 总体观察

图 2 展示了九个评估模型的整体熵趋势。对于每个模型，我们计算所有设定下逐问题的熵，然后对每个设定的 400 个问题的结果取平均。结果表明，Gemini 和 Gemma 系列始终表现出较高的熵，说明其答案分布在多种条件下具有更大的不确定性。相比之下，Voxtral 系列表现出较低的熵和更窄的离散范围，反映出其预测更加集中且更具置信度；而 Phi 4 模型则呈现出更大的四分位距，表明其在不同问题上的预测置信度具有更大的变异性。在每个模型系列内部，较轻量的变体（例如 `Voxtral Mini` 与 `VoxtralSmall`）的平均熵略高于较大型号，这表明较小的参数规模总体上往往会导致行为稳定性较低以及更大的预测不确定性。

模型	口音			语言			选项顺序			性别		
	CS	CA	Δ	CS	CA	Δ	CS	CA	Δ	CS	CA	Δ
Gemini 2.5 Flash	**0.614**	0.593	0.021	**0.648**	0.623	0.025	**0.333**	0.279	0.054	**0.678**	0.650	0.028
Gemini 2.5 Flash Lite	**0.680**	0.647	0.033	**0.731**	0.692	0.039	**0.542**	0.459	0.083	**0.779**	0.740	0.039
Gemma 3n E2B	**0.468**	0.461	0.007	**0.579**	0.559	0.020	**0.604**	0.578	0.026	**0.731**	0.723	0.008
Gemma 3n E4B	0.490	**0.498**	-0.008	0.582	**0.587**	-0.005	**0.571**	0.540	0.031	**0.746**	0.738	0.008
Phi 4	0.461	**0.467**	-0.006	**0.566**	0.549	0.017	**0.499**	0.446	0.053	**0.697**	0.669	0.028

Table 5：不同变量下文化敏感（CS）与文化无关（CA）问题的平均熵 s.

	中文				英语				韩语			
性别	口音	原始	Rev.	Δ	口音	原始	Rev.	Δ	口音	原始	Rev.	Δ
♂	北京	**62.75**	61.50	1.25	美国人	**82.00**	79.75	2.25	全罗道	**61.50**	61.00	0.50
	东北	**64.75**	61.25	3.50	英国人	**81.00**	78.50	2.50	首尔	**63.00**	62.00	1.00
					印度人	**80.00**	79.25	0.75				
♀	北京	**66.75**	64.00	2.75	美国人	**80.00**	78.50	1.50	全罗道	**63.50**	58.50	5.00
	东北	**64.50**	57.75	6.75	英国人	**81.25**	76.75	4.50	首尔	**63.75**	62.00	1.75
					印度人	**80.50**	78.25	2.25				

表6：选项顺序条件下**Gemini 2.5 Flash**的准确率比较。每个单元格报告原始和反转选项顺序的平均准确率（%），其中 Δ 表示两者之间的差异。结果按语言（中文、英语、韩语）、口音和性别分组。

4.2 CS与CA的比较

Global MMLU Lite 基准将问题分为文化敏感（CS）和文化无关（CA）两类：前者需要依赖语境或特定文化知识，后者主要依赖领域知识。图 3展示了九个模型在 CS 和 CA 设定下的熵对比。总体而言，CA 问题始终表现出较低的熵，说明其答案分布更加集中且稳定，与基于事实的推理相一致；而 CS 问题则呈现出更宽的熵范围。如表5所示，这一 CS–CA 熵差距在各变量下均持续存在：对 口音 和 性别 的扰动仅引入了细微差异，而选项顺序 产生了最大的差距，表明对于具有文化背景的问题，模型对位置的敏感性更高。跨变量扰动下 CS/CA 稳健性的补充结果见附录D.1。

4.3 不同变量级别下的准确率

接下来，我们将按级别进行分析，考察不同变量级别如何影响模型行为，首先从准确率入手。表6 报告了`Gemini 2.5 Flash`在中文、英语和韩语中的表现。不同选项顺序配置之间的准确率差距为 0.5% 至 6.75%，且在所有语言-口音设定中，原始顺序的表现始终优于反转顺序。附录D.2中的表格11-17b展示了其他模型的结果，也呈现出相似的模式：大多数模型在原始配置下的准确率更高。这些发现表明，选项顺序会在模型预测中引入系统性偏差。其他变量的影响也在同一附录中进行了总结。

4.4 不同变量级别下的鲁棒性

分层分析 除正确性外，图 4还展示了不同模型系列的鲁棒性模式。 性别 和 口音 位于右下象限（高 κ、低 APES），表明其预测鲁棒，具有较强的层内一致性，且不确定性在不同层级间保持稳定。相比之下，语言位于中间区域（$\kappa \approx 0$、APES 较高），且不同模型系列之间存在显著差异，表明跨语言泛化仍是鲁棒性方面的一项关键挑战。请注意，Voxtral 系列不支持中文或韩语，因此未对该系列进行语言层面的分析。最后，选项顺序始终是所有模型中表现最弱的因素，

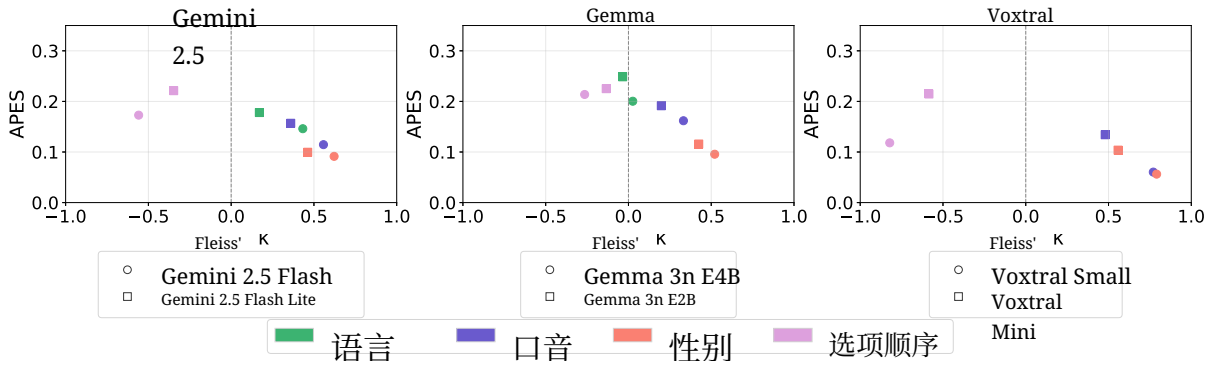


图 4：各模型系列中，Fleiss' κ 与 APES 的对比。语言、口音和性别 等变量的一致性和稳定性较高，而选项顺序 的 APES 较高、 κ较低，表明其敏感性较强。

模型	口音		选项顺序	
	直接克隆	直接克隆		
Gemini 2.5 Flash Lite	0.100	**0.096**	**0.155**	0.194
Gemini 2.5 Flash	0.060	**0.053**	**0.101**	0.132
Gemma 3n E2B	0.125	**0.027**	0.228	**0.215**
Gemma 3n E4B	**0.096**	0.098	0.179	**0.164**
Phi 4	**0.103**	0.112	0.208	**0.173**

表 7：APES 在口音 和选项顺序方面的比较。数值越低表示鲁棒性越高。

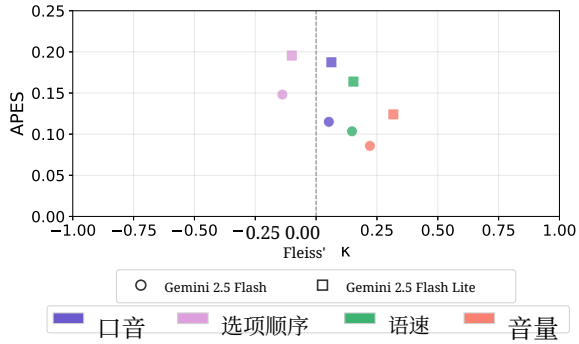


图 5：Fleiss' κ 与 APES 在口音、选项顺序、语速和音量变量方面的比较。

通常出现在左侧象限， κ 为负且 APES 相对较高，反映出模型对输入顺序具有明显的敏感性。总体而言，尽管模型对说话者相关因素（性别和口音）表现出相对更强的鲁棒性，但在这些扰动下的一致性仅达到“中等”至“较高”水平（$\kappa \approx 0.4$-0.8），而非鲁棒系统理想情况下应达到的“几乎完全一致”（$\kappa > 0.8$）。这一差距表明，提升语音鲁棒性仍有很大空间。

模型规模的影响 图4 还比较了三个代表性系列中不同规模下的模型鲁棒性。Gemini 2.0 系列的结果见附录 D.3 以供完整参考。

较大的模型在性别、口音和语言 变量上始终表现出更高的 κ和更低的 APES，表明其在输入扰动下的行为更稳定、更一致。对于选项顺序，较大的模型也能获得更低的 APES，尽管 κ 仍为负值，使得直接比较的意义较小。这些结果表明，参数削减会放大对输入扰动的脆弱性，使较小的模型或轻量变体不如其全规模对应模型鲁棒。

选项顺序变体 遵循 Wei等人(2024)，我们评估选项顺序 偏见是否能通过应用多种选项排列而超越单一反转实现泛化，包括原始、完全反转、词元反转和顺序反转。附录 D.4（表 21–22）的结果表明，选项重新排序对其他变量上的APES影响有限，并保留了因素排序（语言 > 口音 > 性别），且完全反转 顺序会引发最高的不确定性，与我们的主要发现一致。

5 讨论

5.1 真实世界说话人变异性

本研究中的大多数实验依赖于文本转语音生成的语音，这可能会引发人们对所观察到的语音偏差是否能泛化到真实世界场景的担忧。为解决这一问题，我们进行了两项互补分析，以更好地近似真实世界说话人变异性和声学条件。

说话人身份真实性 为了超越纯合成语音，我们收集了三位真实说话者的短录音，他们分别代表美国、英国和印度英语口音，并使用 Chatterbox（Resemble AI，2025），这是一种神经语音克隆文本转语音模型，用于为每种口音生成完整的 400 道题目

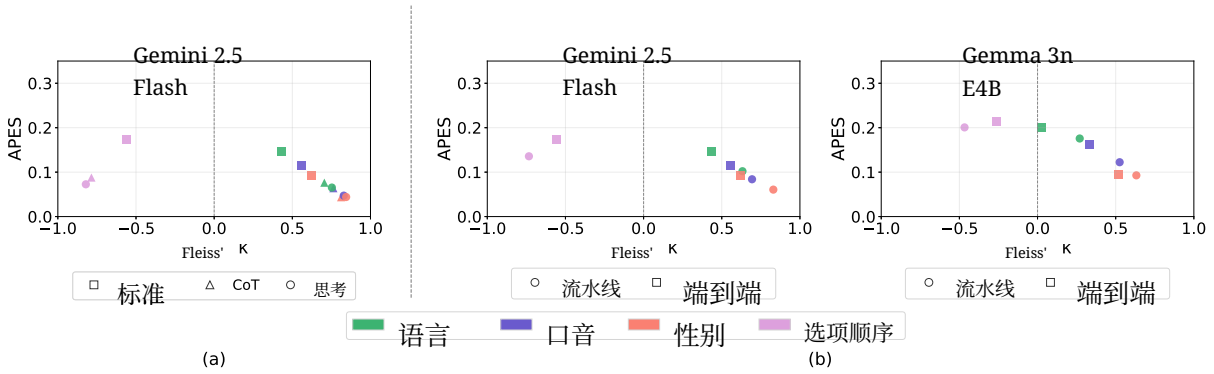


图6： (a) 推理复杂度和 (b) 架构范式对模型鲁棒性的影响。更高的推理复杂度和流水线设计带来更高的一致性（Fleiss' κ）和更低的不确定性（APES）。

的英语 Global MMLU Lite 数据集。如表7所示，在这些克隆语音中，核心趋势和模型相对排名保持一致。重要的是，偏差模式与直接 TTS 生成下观察到的情况高度吻合，这表明我们的发现并非特定合成语音造成的假象，而是在说话人特征更接近真实应用场景的条件下依然存在。

声学变异性为进一步考虑录音条件的变异性，我们以不同的语速（0.75×、1.0×和 1.25×）及响度水平（0.5×、1.0×和 1.5×）对克隆的英语数据进行扰动。图5 中的结果表明，语速变化引入的偏差明显大于音量变化，这一点体现在所有模型的 APES 值均较高。尽管存在这些扰动，我们研究的主要结论仍然稳定，进一步证实了在更真实的声学变异条件下，所观察到的语音偏差模式具有鲁棒性。

5.2 推理复杂度的影响

我们通过比较标准提示、思维链（CoT）提示（Wei 等人，2023）和显式推理（思考）模式，考察推理复杂度的提升如何影响模型鲁棒性。图6(a)报告了 Gemini 2.5 Flash 的结果。总体而言，思维链提示显著提高了一致性，Fleiss' κ 在性别、口音和语言方面的平均增幅分别为 19.01%、20.50% 和 27.20%。此外，它还提升了鲁棒性，性别、口音、语言和选项顺序方面的 APES 均值分别降低了 4.79% 、 5.07% 、 6.98 % 和 8.50%。在此基础上进一步启用显式推理，可在一致性和鲁棒性方面带来额外提升，超越仅使用思维链提示所取得的效果。这些结果表明，推理复杂度的增加能够减轻输入引发的变异，并使模型预测在多种扰动下更加稳定。

5.3 架构范式的影响

接下来，我们考察架构范式的作用，对比端到端多模态大语言模型与流水线设计。端到端模型直接处理音频，但可能依赖类似自动语音识别的层，从而过滤掉副语言线索（例如口音、性别）。为检验这一点，我们构建了一种流水线设置：模型先将音频转录为文本，再进行回答。此项比较旨在评估显式转录是否会去除影响模型在不同条件下鲁棒性的说话人相关线索。我们将此比较应用于两个具有代表性的模型：`Gemini 2.5 Flash`和`Gemma3n E4B`，它们因多模态输入能力强而入选。如图6 (b)所示，在流水线设置下，与各自的端到端对应模型相比，两个模型在κ 上的 Fleiss'一致性更高、APES 更低，且这一结果在不同语言、口音和性别条件下均成立。这一模式表明，显式转录能够抑制说话人相关变异，从而减轻口音和性别引发的偏差对最终预测的影响。综上，这些结果凸显了架构范式是提升鲁棒性的关键杠杆：流水线流程能够降低副语言敏感性，并促进模型在不同条件下表现得更加一致。

5.4 语音作为偏见放大器

在将观察到的鲁棒性差异归因于语音特有的属性之前，我们先考察这些模式是否反映了文本问答中已存在偏差的放大效应

模型	语言 文本	音频	选项顺序 文本	音频
Gemini 2.5 Flash Lite	**0.081**	0.178	**0.096**	0.221
Gemini 2.5 Flash	**0.080**	0.146	**0.085**	0.173
Gemma 3n E2B	**0.163**	0.249	**0.194**	0.225
Gemma 3n E4B	**0.148**	0.200	**0.167**	0.214
Phi 4 Multimodal	**0.208**	0.244	**0.192**	0.242

表 8：文本与音频输入在语言和选项顺序方面的 APES 对比。音频输入始终产生更高的 APES，表明既有的鲁棒性敏感性有所加剧。

过程。为此，我们比较文本和音频输入在两个已知会引发敏感性的变量上的表现，即 语言 和选项顺序。如表8所示，在音频条件下，所有模型的 APES 值都始终高于文本条件。这表明，语音中观察到的鲁棒性模式并非音频模态所独有，而是问题以口语形式呈现时，既有偏差被系统性放大的表现。这种跨模态比较是一项合理性检查，为我们的分析提供依据，并支持这样一种解读：语音主要放大了文本模型中已存在的敏感性，而非引入全新的偏差模式。

6 相关工作

自动语音识别中的语音偏见。 以往关于语音偏见的研究主要聚焦于自动语音识别（自动语音识别）系统，这些系统在性别、口音和语言方面持续表现出性能差异。性别相关偏见已有大量报道：YouTube 自动字幕中女性说话者的词错误率（词错误率）较高（Tatman，2017），Whisper Small 中男性语音的词错误率较高（ElGhazaly 等，2025），以及不同系统中因模型而异的偏差反转（Graham 和 Roll，2024）。口音偏见同样普遍，美国英语和加拿大英语的词错误率低于非母语口音（Graham 和 Roll，2024），而在 SQuAD SRC 等数据集中，不同地区口音之间也存在显著差异（Tang 和 Tung，2023）。在多语言环境中，高资源语言的表现持续优于低资源语言或声调语言，Meta 的 XLS-R 模型对中文和韩语的词错误率较高便是例证（Babu 等，2022）。总的来说，这些发现将自动语音识别偏见归因于性别、口音和数据不平衡的综合影响。我们的工作将这一研究方向拓展至研究范围从转录准确率扩展到**多语言多模态大语言模型中的语音偏见**，从而能够对语言和人口统计维度进行统一评估。

大语言模型鲁棒性。近期研究记录了大语言模型在人口统计和结构维度上存在的系统性偏见。即使没有明确标记，性别偏见依然存在（Belém等人，2024），模型表现出对女性的系统性不利（Vo等人，2025）；而种族和方言偏见则包括对非裔美国人英语的隐性负面刻板印象（Hofmann等人，2024），以及在复杂推理设定下更广泛的种族、国籍和宗教刻板印象（Shrawgi等人，2024）。文化偏见还源于西方中心主义训练数据占主导地位，导致多语言和多文化语境中的差异（LI等人，2024）。除人口统计因素外，Wei等人（2024）指出了选择偏差，即模型对多项选择题中选项顺序或符号格式等非语义线索存在结构性敏感。受这些发现启发，我们将选择偏差的研究从基于文本的评估拓展至语音模态，考察语音输入下的位置敏感性。

7 结论

本研究首次系统考察了多语言多模态大语言模型中的**语音偏见与鲁棒性**。我们推出了**BiasInEar**，这是一个基于 GlobalMMLU Lite 构建的语音增强基准，涵盖英语、中文和韩语，在性别和口音方面均衡分布，包含 11,200 个问题以及 70.8 小时（≈4，249 分钟）的语音。我们采用四项互补指标（准确率、熵、APES 和 Fleiss' κ），评估了来自四个模型系列的九个代表性模型，并分析其在语言、人口统计和结构因素方面的鲁棒性。结果表明，选项顺序导致的鲁棒性下降最为显著，而口音和性别则会造成幅度较小但持续存在的置信度变化。我们还发现，推理复杂度的提高和基于流水线的架构设计能够提升鲁棒性，且语音会系统性地放大文本环境中已有的偏见。总体而言，这些发现揭示了当前 MLLMs 中尚未充分探索的脆弱性，并为设计更公平、更稳定的语音集成式人工智能系统提供了实践参考。

局限性

语音生成尽管我们的数据集对语言、口音和性别进行了系统控制，但使用文本转语音（TTS）自动生成音频，在“口音”的定义和标准化方面带来了固有挑战。即便在同一种语言中，口音差异往往也呈连续变化，而非离散类别。不同地域或社会变体之间的界限模糊，有时甚至无法从语言学角度明确界定，因此难以确保当前设置能够充分涵盖人类语音中自然且连续的变化。此外，受计算资源限制，我们无法为每种条件合成更多语音变体。尽管如此，由于我们的数据集源自 Global-MMLU-Lite，覆盖了广泛的主题和语言，我们认为研究结果仍具有代表性，能够稳健地反映总体上的跨语言和副语言趋势。

评估模型 受计算资源和 API 接口限制，本研究仅评估了九个具有代表性的多模态大语言模型（MLLMs），涵盖商业和开源类别。尽管这些模型体现了架构和推理流水线方面的多样性，但并未完全覆盖现有系统的范围。部分开源模型因稳定性有限、缺乏大规模推理所需的可扩展性，或缺少公开、稳定且高效的多模态 API 而被排除在外。未来，随着标准化接口和可复现的部署流水线逐渐成熟，可以扩大评估范围，从而更全面地评估跨模型一致性和泛化能力。

人工智能助手的使用

我们使用 ChatGPT 作为助手来润色稿件、提升清晰度，并增强结构和可读性。尽管最终内容完全由我们自行完成，这一协助仍有助于改善整体呈现效果。

致谢

本研究得到台湾国家科学及技术委员会资助，项目编号为 NSTC 114-2221-E-002 -070 -MY3、NSTC 113-2634-F- 002-003 -；并获台湾教育部（MOE）项目 NTU-114L900901 资助。

参考文献

Pravesh Agrawal、Szymon Antoniak、Emma BouHanna、Baptiste Bout、Devendra Chaplot、Jessica Chud- novsky、Diogo Costa、Baudouin De Monicault、Saurabh Garg、Theophile Gervet、Soham Ghosh、Amélie Héliou、Paul Jacob、Albert Q. Jiang、Kartik Khandelwal、Timothée Lacroix、Guillaume Lample、Diego Las Casas、Thibaut Lavril，以及另外 23 位作者。2024。Pixtral 12b。 预印本，arXiv:2410.07073。

Anthropic. 2025. Claude 3.7 Sonnet 和 Claude Code。访问日期：2025-09-29。

Giuseppe Attanasio、Beatrice Savoldi、Dennis Fucci 和 Dirk Hovy。2024 年。曲折、隆起与卵石：多语言语音识别模型呈现性别表现差异。载于*2024 年自然语言处理实证方法会议论文集*，第 21318–21340 页，美国佛罗里达州迈阿密。计算语言学协会。

Arun Babu、Changhan Wang、Andros Tjandra、Kushal Lakhotia、Qiantong Xu、Naman Goyal、Kritika Singh、Patrick von Platen、Yatharth Saraf、Juan Pino、Alexei Baevski、Alexis Conneau 和 Michael Auli。2022 年。XLS-R：大规模自监督跨语言语音表征学习。载于*Interspeech2022*，第 2278–2282 页。

Catarina Belém、Preethi Seshadri、Yasaman Razeghi 和 Sameer Singh。2024。不含性别相关语言的文本是否存在模型偏见？载于国际表征学习会议，第 2024 卷，第 12876–12915 页。

Jad Bendarkawi、Ashley Ponce、Sean Chidozie Mata、Aminah Aliu、Yuhan Liu、Lei Zhang、Amna Liaqat、Varun Nagaraj Rao 和 Andrés Monroy-Hernández。2025。Conversar：为第二语言学习者探索增强现实中由具身大语言模型驱动的群体对话。载于计算机系统中人为因素会议扩展摘要论文集，CHI EA ’25，美国纽约。计算机协会。

May Pik Yu Chan、June Choe、Aini Li、Yiran Chen、Xin Gao 和 Nicole Holliday。2022。世界英语语音识别性能中的训练与拼写偏见。载于*Interspeech 2022*，第 1273–1277 页。

Yiming Chen、Xianghu Yue、Chen Zhang、XiaoxueGao、Robby T. Tan 和 Haizhou Li。2024 年。Voicebench：基于大语言模型的语音助手基准测试。预印本，arXiv:2410.17196。

Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Marcel Blistein、Ori Ram、Dan Zhang、Evan Rosen、Luke Marris、Sam Petulla、Colin Gaffney、Asaf Aharoni、Nathan Lintz、Tiago Cardal Pais、Henrik Jacobs- son、Idan Szpektor、Nan-Jiang Jiang 及其他 3416 位作者。2025 年。Gemini 2.5：凭借高级推理、多模态、长上下文和新一代智能体能力推动前沿发展。预印本，arXiv:2507.06261。

Hend ElGhazaly、Bahman Mirheidari、Nafise Sadat Moosavi 和 Heidi Christensen。2025 年。探索自动语音识别技术中的性别差异。预印本，arXiv:2502.18434。Joseph L.Fleiss。1971 年。衡量多位评定者之间名义量表的一致性。《心理学通报》，76(5):378–382。Google Gemini 团队。2023 年。Gemini：高度胜任的多模态模型系列。*ArXiv*，abs/2312.11805。

Calbert Graham 和 Nathan Roll。2024。评估 OpenAI 的 Whisper 自动语音识别：不同口音和说话者特征下的性能分析。《美国声学学会快报》，4(2):025206。

Camille Harris、Chijioke Mgbahurike、Neha Kumar 和 Diyi Yang。2024。自动语音识别中的性别与方言偏见建模。载于计算语言学协会论文集：*EMNLP 2024*，第 15166–15184 页，美国佛罗里达州迈阿密。计算语言学协会。

Dan Hendrycks、Collin Burns、Steven Basart、Andy Zou、Mantas Mazeika、Dawn Song 和 Jacob Steinhardt。2021。衡量大规模多任务语言理解能力。载于国际学习表征会议。

Valentin Hofmann、PratyushaRia Kalluri、DanJurafsky 和 Sharese King。2024。方言偏见可预测人工智能对人们品格、就业能力和犯罪倾向的判断。预印本，arXiv:2403.00742。

Allison Koenecke、Andrew Nam、Emily Lake、Joe Nudell、Minnie Quartey、Zion Mengesha、Connor Toups、JohnR. Rickford、Dan Jurafsky 和Sharad Goel。2020. 自动语音识别中的种族差异.《美国国家科学院院刊》，117(14):7684–7689。

Ajinkya Kulkarni、Atharva Kulkarni、Miguel Couceiro 和 Isabel Trancoso。2024。揭示偏见并拥抱可持续性：评估自动语音识别系统面临的双重挑战。载于 *Interspeech 2024*，Interspeech 2024，第 4628–4632 页。国际语音通信协会。CHENG LI、Damien Teney、Linyi Yang、Qingsong Wen、Xing Xie 和 Jindong Wang。2024。Culturepark：提升大型语言模型的跨文化理解能力。载于第三十八届神经信息处理系统年会。Alexander H. Liu、Andy Ehrenberg、Andy Lo、Clément Denoix、Corentin Barreau、Guillaume Lample、Jean-Malo Delignon、Khyathi Raghavi Chandu、Patrick von Platen、Pavankumar Reddy Muddireddy、Sanchit Gandhi、Soham Ghosh、Srijan Mishra、Thomas Foubert、Abhinav Rastogi、Adam Yang、Albert Q. Jiang、Alexandre Sablayrolles、Amélie Héliou 及其他 87 位作者。2025。Voxtral。预印本，arXiv:2507.13264。

Rao Ma、Mengjie Qian、Siyuan Tang、Stefano Bannò、Kate M. Knill 和 Mark J. F. Gales。2025。使用语音大型语言模型评估第二语言口语熟练度。预印本，arXiv:2505.21148。Meta AI。2025。Llama 4 阵容：原生多模态智能新时代的开端。

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dong- dong Chen, Junkun Chen,Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai及其他57人。2025.Phi-4-mini技术报告：通过LoRA混合实现的紧凑而强大的多模态语言模型.预印本，arXiv:2503.01743。

EliyaNachmani, Alon Levkovitch, RoyHirsch, Ju-lianSalazar, Chulayuth Asawaroengchai, SorooshMariooryad, Ehud Rivlin, RJSkerry-Ryan, andMichelleTadmorRamanovich.2024. 使用频谱图驱动的大语言模型进行语音问答和语音延续。载于第十二届国际学习表征会议。

Tarek Naous, Michael J Ryan, AlanRitter,and WeiXu.2024. 祈祷后喝啤酒？衡量大语言模型中的文化偏见。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第16366–16393页，泰国曼谷。计算语言学协会。

OpenAI.2022. 推出 ChatGPT。https://openai.com/blog/chatgpt。

OpenAI.2024. GPT-4o系统卡片。 *arXiv*预印本。ArXiv:2410.21276。

Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。2022。通过大规模弱监督实现稳健的语音识别。*arXiv* 预印本。

Resemble AI.2025. Chatterbox-TTS. https://github.com/resemble-ai/chatterbox. GitHub 仓库。Sujoy Roychowdhury、Ranjani H.G.、Sumit Soman、Nishtha Paul、Subhadip Bandyopadhyay 与 Siddhanth Iyengar。2025. 面向数学表达式的文本转语音系统的可懂度。载于*Interspeech 2025*，第2280–2284页。Paul K. Rubenstein、Chulayuth Asawaroengchai、Duc Dung Nguyen、Ankur Bapna、Zalán Borsos、Félix de Chaumont Quitry、Peter Chen、Dalia El Badawy、Wei Han、Eugene Kharitonov、Hannah Muckenhirn、Dirk Padfield、James Qin、Danny Rozenberg、Tara Sainath、Johan Schalkwyk、Matt Sharifi、Michelle Tadmor Ramanovich、Marco Tagliasacchi 以及其他11人。2023. AudioPaLM：一种能够说话和聆听的大语言模型。 预印本，arXiv:2306.12925。

C. E. Shannon. 1948. 通信的数学理论。 贝尔系统技术期刊，27(3):379–423。

Min-Han Shih、Ho-Lam Chung、Yu-Chi Pai、Ming-Hao Hsu、Guan-Ting Lin、Shang-Wen Li 和 Hung yi Lee。2024。GSQA：用于生成式口语问答的端到端模型。载于*Interspeech 2024*，第 2970–2974 页。

Hari Shrawgi、Prasanjit Rath、Tushar Singhal 和 Sandi- pan Dandapat。2024。揭示大语言模型中的刻板印象：一种基于任务复杂度的方法。载于 欧洲计算语言学协会第*18*届会议论文集（第*1*卷：长篇论文），第 1841–1857 页，马耳他圣朱利安。计算语言学协会。

Shivalika Singh、Angelika Romanou、Clémentine Four- rier、David Ifeoluwa Adelani、Jian Gang Ngui、Daniel Vila-Suero、Peerat Limkonchotiwat、Kelly Marchi- sio、Wei Qi Leong、Yosephine Susanto、Raymond Ng、Shayne Longpre、Sebastian Ruder、Wei-Yin Ko、Antoine Bosselut、Alice Oh、Andre Martins、Leshem Choshen、Daphne Ippolito 及另外4位作者。2025。Global MMLU：理解并解决多语言评估中的文化和语言偏见。载于 计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文），第 18761–18799 页，奥地利维也纳。计算语言学协会。

Divya Tadimeti、Kallirroi Georgila和David Traum。2022. 对话领域中不同口音上现成语音识别器的评估。载于《第十三届语言资源与评估会议论文集》，第6001–6008页，法国马赛。欧洲语言资源协会。

Weiting Tan、Hirofumi Inaguma、Ning Dong、Paden D. Tomasello和Xutai Ma。2025. SSR：面向语音语言模型的对齐感知模态连接器。载于《第*22*届国际口语翻译会议（*IWSLT 2025*）论文集》，第56–75页，奥地利维也纳（现场与线上）。计算语言学协会。

Changli Tang、Wenyi Yu、Guangzhi Sun、Xianzhao Chen、Tian Tan、Wei Li、Lu Lu、Zejun MA和Chao Zhang。2024. SALMONN：面向大语言模型的通用听觉能力。载于第十二届国际学习表征会议。

Yixuan Tang和Anthony K.H: Tung。2023。Squad- src：一个用于多口音口语阅读理解的数据集。载于《第三十二届国际人工智能联合会议论文集，*IJCAI-23*》，第5206–5214页。国际人工智能联合会议组织。主会场。

Rachael Tatman. 2017. YouTube 自动字幕中的性别偏见与方言偏见。载于 以下论文集：首届 *ACL* 自然语言处理伦理研讨会，第53–59页，瓦伦西亚，西班牙。计算语言学协会。

Omnilingual ASR 团队，Gil Keren、Artyom Kozhevnikov、Yen Meng、Christophe Ropers、Matthew Setzler、Skyler Wang、Ife Adebara、Michael Auli、Can Balioglu、Kevin Chan、Chierh Cheng、Joe Chuang、Caley Droof、Mark Duppenthaler、Paul-Ambroise Duquenne、Alexander Erben、Cynthia Gao、Gabriel Mejia Gonzalez以及其他14人。2025。Omnilingual ASR：开源多语言语音识别，面向 1600+种语言。预印本，arXiv:2511.09690。

An Vo、Mohammad Reza Taesiri、Daeyoung Kim 和 Anh Totti Nguyen。2025。B-score：利用响应历史检测大语言模型中的偏见。载于第四十二届国际机器学习会议。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed Chi、Quoc Le 和 DennyZhou。2023。思维链提示可引发大语言模型进行推理。预印本，arXiv:2201.11903。

Sheng-Lun Wei、Cheng-Kuang Wu、Hen-Hsen Huang 和 Hsin-Hsi Chen。2024。揭示选择偏见：探索大语言模型中的顺序和词元敏感性。载于计算语言学协会论文集：*ACL 2024*，第5598–5621页，泰国曼谷。计算语言学协会。

Dong Zhang、Shimin Li、Xin Zhang、Jun Zhan、Pengyu Wang、YaqianZhou 和 Xipeng Qiu。2023。SpeechGPT：赋能大语言模型，使其具备内在的跨模态对话能力。载于计算语言学协会论文集：*EMNLP 2023*，第 15757–15773 页，新加坡。计算语言学协会。

A 成本分析

所有涉及文本转语音生成以及通过 Gemini API 进行模型推理的实验，总费用不到 550 美元。对于其他模型，Mistral 和 NVIDIA 提供的 API 在实验期间暂时免费。

B 数据集构建详情

B.1 问题改写

为改写问题和选项，我们通过 NVIDIA API 使用`GPT OSS 120B` 模型，以确保稳定性和可扩展性。我们向模型提供图 8所示的任务专属指令，其中规定了八条转换规则。这些规则涵盖了多个方面，例如读取数学表达式（如 “x^2+y^2”）、结合学科背景消除领域特定术语的歧义（例如，在化学中将 “Na” $\rightarrow$ “sodium” ）、处理数字和单位（例如，将 “3kg” $\rightarrow$ “three kilograms”）、解读圆括号，以及在不同语言中恰当地呈现如“空白” 这样的占位符。此步骤可确保在音频合成之前生成高质量、适合朗读的文本。

B.2 文本转语音生成提示词

图 7 展示了文本转语音合成所用的提示词模板，说明了文本指令、说话人特征和韵律线索如何结合起来，引导模型生成更加自然、富有表现力且能理解语境的语音。

B.3 改写质量评估

正如我们在第 2.2 节所述，为了评估改写后问题的可靠性，我们手动检查了自动标记案例中的一部分。表9 总结了英语、中文和韩语中被标记实例及经核实的真实错误所占比例。

B.4 分层抽样的语音生成质量评估

质量评估的抽样采用分层方式，以确保代表性和多样性。在每个词错误率区间内，我们汇总所有口音，并使用最大余数法，按照各口音的样本数量比例分配口音配额。对于每个问题–答案对，我们选取一个项目（优先选择问题描述行），随后按学科轮转进行抽样，以平衡学科覆盖范围。当某个区间中没有足够的不同主题或问题来满足配额时，我们从可用池中填补剩余部分。

```
请用{language}和{accent}口音朗读这段文本
口音：{text}
```

图7：用于文本转语音合成的提示词模板。占位符{text}将替换为改写后的问题或答案选项。

语言	标记案例（%）	已核实错误（%）
英语	477 (23.85%)	20 (1.00%)
中文	494 (24.70%)	93（4.65%）
韩语	577（28.85%）	79（3.95%）

表 9：改写后问题的质量控制统计数据。该表报告了自动标记案例以及经人工核实的真实错误的数量和百分比。

C 实验设置详情

C.1 模型

如第 3.2 节所述，我们在商业和开源两种设定下评估了九种大型多模态模型。表 10总结了这些模型，包括由 Google 提供的 API 端点[2]，NVIDIA[3], 和 Mistral[4], 以及各自对应的提供商。

C.2 音频拼接

对于每个问题，我们将问题和四个选项的描述与由文本转语音模型预先生成的指令标记（“问题”、“*A*”–“*D*”）拼接，构成音频查询。英语的指令标记采用美国、英国和印度口音录制，中文和韩语则采用美国口音录制。

拼接过程遵循指定顺序：每个指令词元都置于对应内容之前；采用了四种顺序（原始、反转、选项顺序反转、词元顺序反转），并在各片段之间插入固定停顿，以确保清晰度。为了满足 Gemma 和 Phi 4 等模型的输入长度限制（30 秒），每段最终音频还被分割成固定长度的分块，并导出为波形文件。这确保了所有实验变量下的输入一致且兼容。这确保了所有实验变量下的输入一致。

[2]http://aistudio.google.com[3]https://build.nvidia.com[4]https://console.mistral.ai

[系统]
将输入转换为母语为{language}的人会如何朗读的形式。只输出结果。不要解释推理过程，也不要修改纯文本内容。输入主题： <subject>问题： <question>选项：A) <option_a>，B) <option_b>，C) <option_c>，D) <option_d>

转换应遵循8条规则：

1. 使用“主题”来消除领域术语或缩写的歧义（例如，在物理学中将“ms”读作“毫秒”，在化学中将“HPO42-”读作“磷酸氢根”，将“P”读作“磷”。）
2. 自然地读出数学表达式。示例：“f(x)” → “f of x”； Z_n → “模 n 整数”；(12)(123) →“由置换构成，包含循环一-二和循环一-二-三”。如果大小写有区别，请明确读出：“O(n) 和 o(n)” → “n 的大 O 和 n 的小 o”。
3. 清晰读出数字和单位，并使用正确的复数形式。
4. 只有在圆括号或方括号影响含义时才读出来。
5. 将罗马数字标签（I、II、i、ii.....）作为标签读出：“罗马数字一”“罗马数字二”等。
6. 用“空白”代替空白标记“_____”；对多个空白进行编号（“空白 1”“空白 2”）；保留周围的语法结构；绝不猜答案。
7. 所有转换都必须直接使用指令指定的{language}来书写。当提示词包含"空白"，请使用指定的语言呈现（例如，中文：空格；日语：空白；韩语：공백）。所有其他文本也必须使用相同的语言。

输出格式（可扩展标记语言）：

```
<输出>
  <问题>&lt;问题的口语呈现&gt;</问题>
  <选项_a>&lt;选项 A 的口语呈现&gt;</选项_a>
  <option_b>&lt;选项 B 的口语化呈现&gt;</option_b>
  <option_c>&lt;选项 C 的口语化呈现&gt;</option_c>
  <option_d>&lt;选项 D 的口语化呈现&gt;</option_d>
</output>
```

图8：用于以口语风格呈现 MMLU 类题目的提示词

模型	API 端点	提供商
	商业 *API*	
Gemini 2.5 Flash	gemini-2.5-flash	Google
Gemini 2.5 Flash Lite	gemini-2.5-flash-lite	Google
Gemini 2.0 Flash	gemini-2.0-flash	Google
Gemini 2.0 Flash Lite	Gemini 2.0 Flash Lite	Google
	开源模型	
Gemma 3n E4B	google/gemma-3n-e4b-it	Nvidia NIM
Gemma 3n E2B	google/gemma-3n-e2b-it	Nvidia NIM
Voxtral Small	voxtral-small-2507	Mistral
Voxtral Mini	voxtral-mini-2507	Mistral
Phi 4 Multimodal	microsoft/phi-4-multimodal-instruct	Nvidia NIM

表10：评估模型。

C.3 模型推理与后处理

所有模型均通过其官方 API 进行查询，并采用确定性贪心解码（将温度设为零、候选数量设为一）以消除随机性。我们使用统一的文本提示词来规范输出，如图9 对于思维链（CoT）推理，我们使用图10所示的提示词来规范输出。最大输出长度限制为4k词元，足以涵盖多项选择题回答。

通过模式匹配对模型响应进行后处理，以提取最终答案字母（例如，Answer:[[A]]，Answer: A）。根据选项顺序将该字母映射为索引，并将无效输出标记为解析失败。

C.4 Fleiss' κ 的推导

设置。 固定一个变量 v，考虑由其他变量组合构成的一组条目，并将其索引为 $i = 1,\ldots, I$，其中 I 表示除 v 之外所有变量的组合数。每个条目由 n_i 个评分分配到 j 个分类（A、B、C、D）之一。在我们的应用中，n_i 个评分来自同一模型在不同 v 级别下回答条目 i。设 n_{ij} 表示条目 i 中选择类别 j 的评分数量，因此 $\sum_{j=1}^{J} n_{ij} = n_{i\circ}$

每个条目的观察一致性。 对于条目 i，评分者对之间的一致比例为

$$P_i = \frac{1}{n_i(n_i-1)} \sum_{j=1}^{4} n_{ij}(n_{ij}-1), \qquad (5)$$

因为每个类别-j 贡献 $\binom{n_{ij}}{2}$ 个一致的评分对，而总共有 $\binom{n_i}{2}$ 个无序评分对。

机会水平下的期望一致性。 在通常的"固定边际的随机分配模型"下，两个独立抽取的评分相同的概率为

$$P_e \;=\; \sum_{j=1}^{J} p_j^2, \text{ where } p_j \;=\; \frac{\sum_{i=1}^{I} n_{ij}}{\sum_{i=1}^{I} n_i}. \qquad (6)$$

对观察一致性求平均。 根据评分数量对（5）中的各项一致性进行汇总：

$$\bar{P} \;=\; \frac{\sum_{i=1}^{I} n_i\, P_i}{\sum_{i=1}^{I} n_i}. \qquad (7)$$

这种加权方式在 n_i 变化时，对各个项目中的每个评分对一视同仁。

> 请回答以下问题。仅输出"Answer:[[LETTER]]"，其中 LETTER 为 A、B、C、D 之一。

图 9：标准提示词模板

> 请回答以下问题。让我们一步一步地思考，并写出简洁的推理。最后一行答案应为"Answer:[[LETTER]]"，其中 LETTER 为 A、B、C、D 之一。

图 10：思维链提示模板

Fleiss' Kappa。 Fleiss' κ 将超出机会水平的额外一致性标准化：

$$\kappa \;=\; \frac{\bar{P} - P_e}{1 - P_e}, \qquad (8)$$

¯，若 $\kappa = 1$ 则为 $P = 1$（完全一致），若 $\kappa = 0$ 则为$P = P_e$（不优于机会水平），而当观察一致性低于机会水平时为 $\kappa < 0$。

D 详细实验结果

D.1 变量扰动下的CS与CA

我们进一步将问题划分为文化敏感（CS）与文化无关（CA）子集，以评估跨变量鲁棒性，在语言、口音、性别以及选项顺序扰动下。图11 与 12 表明，在各模型与各变量中，CA题目始终比CS题目表现出更低的不确定性（APES），说明其在跨变量变化下具有更强的鲁棒性

D.2 准确率分析

表格11 至 17 提供了以下模型在不同选项顺序下，按语言、口音和性别 分组的准确率比较：Gemini 2.5 Flash Lite、Gemini 2.0 Flash、Gemini 2.0 Flash Lite、Gemma 3n E4B、Gemma 3n E2B、Voxtral Small、Voxtral Mini 和 Phi 4 Multimodal。表格 18至 20 则对 Gemini 2.5 Flash 进行逐因素分析，分别按性别、口音和语言报告准确率比较。这些分层分析不仅阐明了选项重新排序是否会影响准确率，还揭示了其影响如何与语音特定因素相互作用，从而更清晰地呈现模型在受控变化下的模型行为。

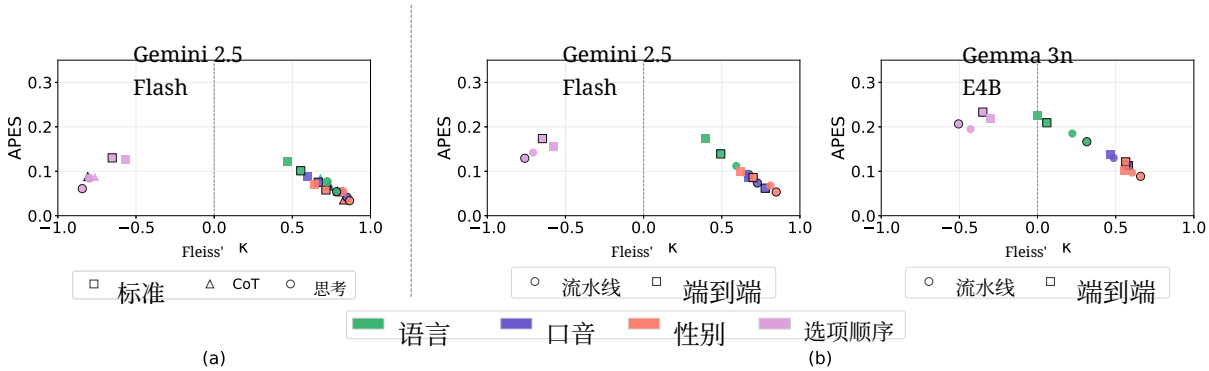


图 11：　跨变量扰动（语言、口音、性别、a）下按 CS/CA 分层的 APES–κ 分析　　nd 选项顺　　序）。无框标记表示文化敏感项目，方框标记表示文化无关项目。

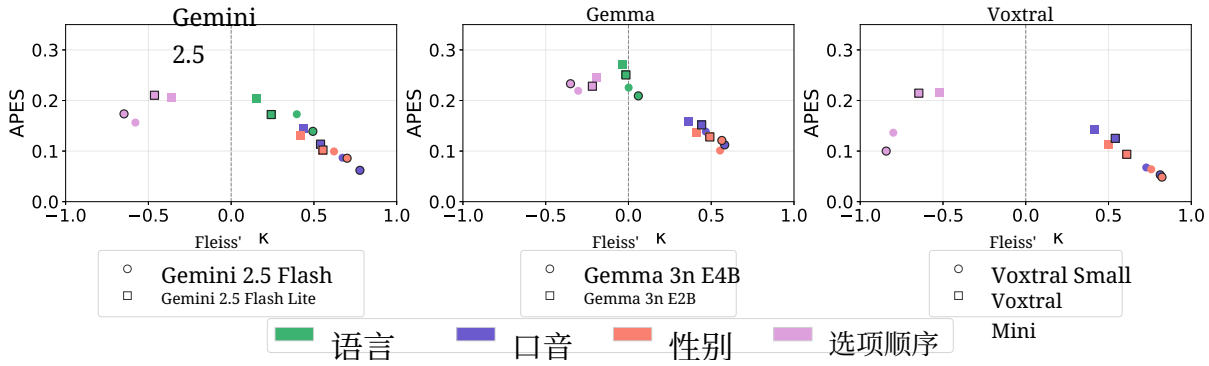


图 12：跨变量扰动下各模型系列的分组 APES–κ 鲁棒性图。无框标记表示文化敏感项目，方框标记表示文化无关项目。

中文				英语				韩语			
输入	原始	倒序	Δ	输入	原始	倒序	Δ	输入	原始	倒序	Δ
北京 ♀	**46.75**	45.00	1.75	美国人 ♀	**62.25**	61.00	1.25	全罗道 ♀	**43.75**	42.50	1.25
北京 ♂	**44.75**	41.75	3.00	美国人 ♂	**61.50**	61.75	-0.25	全罗道 ♂	**45.75**	44.00	1.75
东北 ♀	**43.50**	40.00	3.50	英国人 ♀	**61.75**	61.00	0.75	首尔 ♀	**44.50**	43.50	1.00
东北 ♂	**41.60**	40.35	1.25	英国人 ♂	**60.75**	58.50	2.25	首尔 ♂	**46.00**	44.75	1.25
				印度人 ♀	**64.00**	60.50	3.50				
				印度人 ♂	**59.25**	58.00	1.25				

表11：*Gemini2.5 Flash Lite*在不同选项顺序设置下的准确率比较，按语言、口音和性别分组。

中文				英语				韩语			
输入	原始	倒序	Δ	输入	原始	倒序	Δ	输入	原始	倒序	Δ
北京 ♀	**70.75**	67.75	3.00	美国人 ♀	**83.75**	80.75	3.00	全罗道 ♀	**68.00**	65.25	2.75
北京 ♂	69.00	**69.50**	-0.50	美国人 ♂	**83.25**	80.00	3.25	全罗道 ♂	66.00	**67.00**	-1.00
东北 ♀	**68.25**	66.00	2.25	英国人 ♀	**81.75**	80.50	1.25	首尔 ♀	**70.25**	68.25	2.00
东北 ♂	65.50	**69.25**	-3.75	英国人 ♂	**81.75**	78.75	3.00	首尔 ♂	**67.25**	65.00	2.25
				印度人 ♀	**81.50**	80.00	1.50				
				印度人 ♂	**83.00**	80.00	3.00				

表12：*Gemini 2.0 Flash*在不同选项顺序设置下的准确率比较，按语言、口音和性别分组。

中文				英语				韩语			
输入	原始倒序		Δ	输入	原始倒序		Δ	输入	原始倒序		Δ
北京 ♀	**65.50**	63.50	2.00	美国人 ♀	**79.75**	77.00	2.75	全罗道 ♀	66.50	**66.75**	-0.25
北京 ♂	**66.25**	64.75	1.50	美国人 ♂	77.75	**78.25**	-0.50	全罗道 ♂	**62.75**	61.25	1.50
东北 ♀	**66.75**	61.75	5.00	英国人 ♀	**78.00**	74.25	3.75	首尔 ♀	**64.75**	64.00	0.75
东北 ♂	65.00	**65.50**	-0.50	英国人 ♂	**78.50**	74.75	3.75	首尔 ♂	**66.25**	62.50	3.75
				印度人 ♀	**79.00**	77.25	1.75				
				印度人 ♂	**77.75**	77.25	0.50				

表13：按语言、口音和性别分组的*Gemini2.0 Flash Lite*在不同选项顺序设置下的准确率比较。

中文				英语				韩语			
输入	原始倒序		Δ	输入	原始倒序		Δ	输入	原始倒序		Δ
北京 ♀	**37.00**	34.75	2.25	美国人 ♀	**59.75**	53.25	6.50	全罗道 ♀	**36.25**	32.25	4.00
北京 ♂	34.25	**35.00**	-0.75	美国人 ♂	**59.25**	56.75	2.50	全罗道 ♂	**35.75**	33.50	2.25
东北 ♀	33.50	**34.75**	-1.25	英国人 ♀	**59.75**	52.00	7.75	首尔 ♀	33.25	**34.50**	-1.25
东北 ♂	37.00	**37.50**	-0.50	英国人 ♂	**61.25**	51.00	10.25	首尔 ♂	**37.25**	36.00	1.25
				印度人 ♀	**56.75**	53.50	3.25				
				印度人 ♂	**60.00**	51.50	8.50				

表14：按语言、口音和性别分组的*Gemma 3n E4B*在不同选项顺序设置下的准确率比较。

中文				英语				韩语			
输入	原始倒序		Δ	输入	原始倒序		Δ	输入	原始倒序		Δ
北京 ♀	**33.75**	30.50	3.25	美国人 ♀	**50.50**	46.25	4.25	全罗道 ♀	**32.75**	29.00	3.75
北京 ♂	**37.25**	28.00	9.25	美国人 ♂	**52.00**	48.50	3.50	全罗道 ♂	**31.50**	30.25	1.25
东北 ♀	**32.50**	29.50	3.00	英国人 ♀	**51.00**	47.75	3.25	首尔 ♀	**34.25**	33.50	0.75
东北 ♂	**35.00**	29.00	6.00	英国人 ♂	**51.25**	50.50	0.75	首尔 ♂	28.75	**33.50**	-4.75
				印度人 ♀	48.25	**48.75**	-0.50				
				印度人 ♂	**49.00**	47.50	1.50				

表15：*Gemma 3n E2B*在不同选项顺序设置下的准确率比较，按语言、口音和性别分组。

中文				英语			
输入	原始倒序		Δ	输入	原始倒序		Δ
北京 ♀	**39.10**	33.83	5.27	美国人 ♀	**48.99**	43.47	5.52
北京 ♂	**33.67**	31.16	2.51	美国人 ♂	**46.12**	42.86	3.26
东北 ♀	**34.84**	30.08	4.76	英国人 ♀	**46.25**	45.25	1.00
东北 ♂	**33.75**	33.50	0.25	英国人 ♂	**45.75**	43.75	2.00
				印度人 ♀	**47.75**	44.75	3.00
				印度人 ♂	42.46	**43.97**	-1.51

表16：按语言、口音和性别分组的 *Phi4Multimodal* 在不同选项顺序设置下的准确率比较。

英语			
输入	原始倒序		Δ
美国人 ♀	**80.50**	75.75	4.75
美国人 ♂	**79.50**	77.75	1.75
英国人 ♀	**80.00**	74.75	5.25
英国人 ♂	**79.50**	78.25	1.25
印度人 ♀	**79.25**	78.25	1.00
印度人 ♂	**78.75**	77.25	1.50

(a) *Voxtral-Small-2507* 的准确率比较。

英语			
输入	原始倒序		Δ
美国人 ♀	**48.50**	46.50	2.00
美国人 ♂	**52.50**	46.75	5.75
英国人 ♀	**48.50**	47.00	1.50
英国人 ♂	**49.50**	47.75	1.75
印度人 ♀	**50.75**	45.25	5.50
印度人 ♂	**49.25**	44.75	4.50

(b) *Voxtral-Mini-2507* 的准确率比较。

表17：按语言、口音和性别分组，比较 Voxtral 模型在不同选项顺序设置下的准确率。请注意，语言设定仅限英语，因为 Voxtral 目前不支持韩语或中文。

中文				英语				韩语			
设定	♀	♂	Δ	设定	♀	♂	Δ	设定	♀	♂	Δ
北京原始	**66.75**	62.75	4.00	美国人原始	80.00	**82.00**	-2.00	全罗道 原始	**63.50**	61.50	2.00
北京 反转	**64.00**	61.50	2.50	美国人反转	78.50	**79.75**	-1.25	反转	58.50	**61.00**	-2.50
东北 原始	64.50	**64.75**	-0.25	英国人 原始	**81.25**	81.00	0.25	首尔 原始	**63.75**	63.00	0.75
东北 反转	57.75	**61.25**	-3.50	英国人 反转	76.75	**78.50**	-	首尔 反转	**62.00**	62.00	0.00
				印度人原始	**80.50**	80.00	0.50				
				印度人反转	**78.25**	79.25	-1.00				

表18：*Gemini 2.5 Flash*在不同性别条件下的准确率比较。每个单元格报告女性和男性的平均准确率（%），其中 Δ 表示差值（女性 - 男性）。结果按语言（中文、英语、韩语）、口音和选项顺序分组。

中文			英语				韩语		
设定	北京	东北	设定	美国人	英属印度人		设定	首尔	全罗道
原始 ♀	**66.75**	64.50	原始 ♀	80.00	**81.25**	80.50	原始 ♀	63.75	**63.50**
原始 ♂	62.75	**64.75**	原始 ♂	**82.00**	81.00	80.00	原始 ♂	**63.00**	61.50
反转 ♀	**64.00**	57.75	反转 ♀	**78.50**	76.75	78.25	反转 ♀	**62.00**	58.50
反转 ♂	**61.50**	61.25	反转 ♂	**79.75**	78.50	79.25	反转 ♂	**62.00**	61.00

表19：准 确率对比，按口音 条件划分，模型为*Gemini 2.5 Flash*。每个单元格报告平均准确率 cy（%） 各口音，按语言 （中文、英语、韩语）、选项顺序和性别分组。

设定	中文	英语	韩语
原始 ♀	65.62	**80.58**	63.62
原始 ♂	63.75	**81.00**	62.25
反转 ♀	60.88	**77.83**	60.25
反转 ♂	61.38	**79.17**	61.50

表20：语言 条件下*Gemini2.5 Flash*的准确率比较。每个单元格报告各语言的平均准确率（%），并按选项顺序 和性别分组。

模型	语言			口音			性别		
	Ori	Rev	Δ	Ori	Rev	Δ	Ori	Rev	Δ
Gemini 2.5 Flash	0.182	0.190	0.008	0.156	0.168	0.012	0.103	0.111	0.008
Gemini 2.5 Flash Lite	0.238	0.252	0.014	0.218	0.226	0.008	0.123	0.129	0.006
Gemma 3n E2B	0.288	0.289	0.001	0.213	0.211	-0.002	0.166	0.163	-0.003
Gemma 3n E4B	0.271	0.265	-0.006	0.194	0.193	-0.001	0.158	0.142	-0.016
Phi 4 Multimodal	0.290	0.285	-0.005	0.172	0.188	0.016	0.156	0.167	0.011

表21：语言、口音和性别在原始和反转选项顺序下的 APES。

模型	原始	选项顺序反转	词元顺序反转	反转
Gemini 2.5 Flash	**0.267**	0.328	0.309	0.297
Gemini 2.5 Flash Lite	**0.440**	0.503	0.499	0.445
Gemma 3n E2B	0.517	0.492	**0.489**	0.524
Gemma 3n E4B	0.472	0.494	**0.465**	0.493
Phi 4 Multimodal	0.388	**0.353**	0.366	0.415

表22：不同选项顺序扰动下的平均熵。数值越低表示行为越稳定（对顺序越不敏感）。

D.3 模型规模影响的补充结果

图13 展示了 Gemini 2.0 系列的缩放结果，其趋势与图4相同：较大的变体在各变量上具有更高的一致性（Fleiss' κ）和更低的不确定性（APES）。

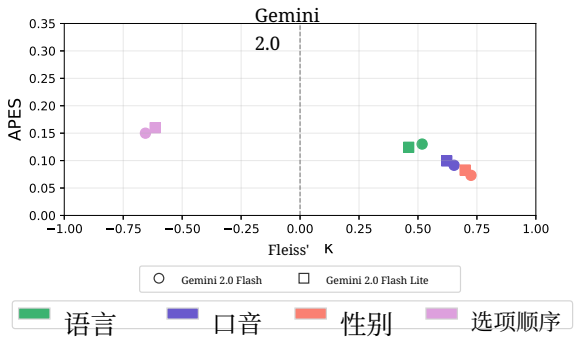


图 13：Gemini 2.0 系列的 Fleiss' κ 与 APES。

D.4 选项重新排序影响的补充结果

表 21 汇报了原始和完全反转的选项顺序下的 APES，并按语言、口音和性别进行了汇总。对于所有模型，原始设定与反转设定之间的 APES 差异在这三个因素上都很小。在两种设定下，语言对应的 APES 值均高于口音和性别，且因素大小的排序保持不变（语言 > 口音 > 性别）。

表 22 汇报了四种选项顺序配置（原始、顺序反转、词元反转和完全反转）下的平均熵。对于各模型而言，不同配置下的熵值变化幅度有限。该表列出了每个模型平均熵最低的配置，而最低熵设定因模型而异。总体而言，这些表格提供了在其他选项顺序扰动下鲁棒性的补充测量结果。