# Bias in the Ear of the Listener: Assessing Sensitivity in Audio Language Models Across Linguistic, Demographic, and Positional Variations

**Sheng-Lun Wei**[*α] **Yu-Ling Liao**[*α] **Yen-Hua Chang**[α] **Hen-Hsen Huang**[β] **Hsin-Hsi Chen**[αγ]
[α]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
[β]Institute of Information Science, Academia Sinica, Taiwan
[γ]AI Research Center (AINTU), National Taiwan University, Taiwan
{weisl,ylliao,yhchang}@nlg.csie.ntu.edu.tw,
hhhuang@iis.sinica.edu.tw, hhchen@ntu.edu.tw

## Abstract

This work presents the first systematic investigation of speech bias in multilingual MLLMs. We construct and release the **BIASINEAR** dataset, a speech-augmented benchmark based on Global MMLU Lite, spanning English, Chinese, and Korean, balanced by *gender* and *accent*, and totaling 70.8 hours (≈4,249 minutes) of speech with 11,200 questions. Using four complementary metrics (accuracy, entropy, APES, and Fleiss' $\kappa$), we evaluate nine representative models under linguistic (*language* and *accent*), demographic (*gender*), and structural (*option order*) perturbations. Our findings reveal that MLLMs are relatively robust to demographic factors but highly sensitive to *language* and *option order*, suggesting that speech can amplify existing structural biases. Moreover, architectural design and reasoning strategy substantially affect robustness across languages. Overall, this study establishes a unified framework for assessing fairness and robustness in speech-integrated LLMs, bridging the gap between text- and speech-based evaluation. The resources can be found at https://github.com/ntunlplab/BiasInEar

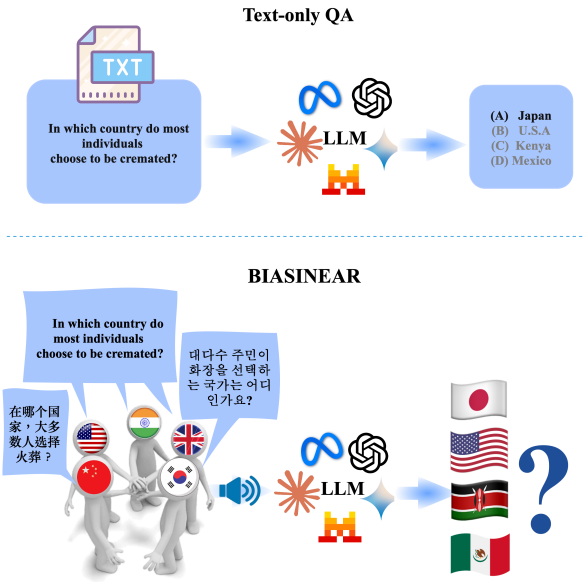


Figure 1: Overview of this work, which extends question answering from text inputs to multilingual spoken contents across languages, accents, and speakers.

## 1 Introduction

The rapid progress of large language models (LLMs) has fundamentally reshaped natural language processing (OpenAI, 2022; Gemini Team, 2023; Anthropic, 2025). Recent advances extend LLMs beyond text-only inputs to multimodal settings, incorporating modalities such as vision (OpenAI, 2024; Agrawal et al., 2024; Meta AI, 2025) and speech (Comanici et al., 2025; Liu et al., 2025; Microsoft et al., 2025), and achieving remarkable performance on a wide range of downstream tasks. In particular, systems that accept spoken inputs are capable of directly handle spoken queries, enabling applications in spoken question answering (Nachmani et al., 2024; Shih et al., 2024), conversational assistants (Tang et al., 2024; Zhang et al., 2023; Rubenstein et al., 2023), and educational technologies (Bendarkawi et al., 2025; Ma et al., 2025).

However, recent studies have made clear that LLMs are not free from systematic biases across both demographic and structural dimensions. For instance, they have demonstrated that large language models encode various forms of social and cultural bias, including gender (Belém et al., 2024; Vo et al., 2025), race, dialect (Hofmann et al., 2024), nationality, and religion (Shrawgi et al., 2024; LI et al., 2024; Naous et al., 2024), as well as imbalances arising from Western-centric training data. Beyond these demographic dimensions, Wei et al. (2024) demonstrate that LLMs also suffer from selection bias when the order of answer

[*]Equal contribution.



*Findings of the Association for Computational Linguistics: EACL 2026*, pages 1570–1589
March 24-29, 2026 

| Raw Text | Direct TTS (Naïve Output) | Converted Text (Ours) |
|---|---|---|
| (1, 2, 5, 4)(2, 3) | The product of the cycles one, two, five, four and the cycle two, three. | The permutation consisting of the cycle one–two–five–four, and the cycle two–three |
| $\Sigma_{v \in V}\ \deg(v)$ | the sum of the degrees of all vertices in V is v. | the sum over lowercase v in uppercase V of degree of v. |
| When traveling north from the United States into Canada you'll see the North Star (Polaris) getting _____ | When traveling north from the United States into Canada you'll see the North Star (Polaris) getting. | When traveling north from the United States into Canada you'll see the North Star Polaris getting blank. |
| H2PO4-, HPO42-. | H two P O four minus, H P O four two minus. | Dihydrogen phosphate, hydrogen phosphate. |
| NH4+(aq) + NO2-(aq) → N2(g) + 2H2O(l). | N H four plus aqueous plus N O two minus aqueous yields N two gas plus two H two O liquid. | The reaction between ammonium ion in aqueous solution and nitrite ion in aqueous solution yields nitrogen gas and two water molecules in liquid form. |
| I. GATT ; II. IMF. | One GATT two I M F. | Roman numeral one, General Agreement on Tariffs and Trad; Roman numeral two, International Monetary Fund |

Table 1: Illustration of the difference between naïve TTS and spoken-readable conversions.

options is altered in multiple-choice question answering tasks. Taken together, these findings show that LLM predictions are influenced not only by semantic content but by superficial inputs and latent social factors, raising serious concerns about fairness and robustness in decision-making contexts. At the same time, speech technologies introduce additional sources of bias. Prior research has shown that automatic speech recognition (ASR) systems often exhibit systematic performance disparities across demographic and linguistic factors, including gender (Harris et al., 2024; Koenecke et al., 2020; Kulkarni et al., 2024; Attanasio et al., 2024), accent (Graham and Roll, 2024; Tang and Tung, 2023; Tadimeti et al., 2022; Chan et al., 2022), and language resource availability (Babu et al., 2022). These findings suggest that transitioning QA tasks from text to speech may not only inherit existing LLM biases but amplify them through additional layers of demographic and linguistic variability.

Our main contributions are threefold: **a)** We construct and release the **BIASINEAR** dataset, a multilingual spoken QA benchmark covering English (with American, British, and Indian accents), Chinese (with Beijing and Northeastern accents), and Korean (with Seoul and Jeolla accents), with balanced male and female speakers. The dataset comprises 70.8 hours (≈4,249 minutes) of speech and 11,200 questions, enabling large-scale and balanced evaluation across languages and demographic factors. **b)** Leveraging this dataset, we perform comprehensive analyses across linguistic (*language* and *accent*), demographic (*gender*), and structural (*option order*) dimensions, extending the selection bias framework proposed in prior text-based studies (Wei et al., 2024) to the speech modality. **c)** Our study thus bridges the gap between LLM bias research and speech applications, offering new insights into fairness and robustness in multilingual speech technologies.

## 2 BIASINEAR Dataset

To investigate audio sensitivity in multilingual settings, we build upon the Global MMLU Lite (Singh et al., 2025) by extending its text-based questions into spoken inputs, enabling a systematic analysis of model behavior under diverse audio conditions. Global MMLU Lite is a curated subset of Global MMLU, a high-quality multilingual extension of MMLU (Hendrycks et al., 2021), and includes both culturally sensitive (CS) and culturally agnostic (CA) labels annotated by human experts. In this work, we focus on English, Chinese, and Korean as representative languages, each varying along factors such as *gender*, *accent*, and *option order*, to comprehensively address our research questions. Specifically, we construct a multilingual speech-based version of MMLU that incorporates diverse *gender* and *accent* features, allowing us to probe the robustness of LLMs in spoken question answering. The final dataset comprises 70.8 hours (≈ 4,249 minutes) of speech across English, Chinese, and Korean, covering 11,200 questions in total.

### 2.1 Dataset Construction

**Question Rewriting** A direct conversion of text-based questions into speech can yield undesirable outcomes, particularly when the questions contain mathematical expressions, domain-specific symbols, or placeholders. Prior work (Chen et al., 2024; Tan et al., 2025) has addressed this issue by filtering out math-intensive subjects, thereby avoiding recognition errors. However, this approach reduces the diversity of the dataset by excluding STEM-

related questions. To overcome this limitation, and inspired by Roychowdhury et al. (2025), we introduce a rewriting step in which each question and its options are reformulated into a format that can be naturally and unambiguously read aloud. Representative rewritten examples are presented in Table 1. Specifically, we employ the `GPT OSS 120B` to perform the rewriting, guided by the instruction prompt shown in Figure 8 of Appendix B.1. Additional implementation details are provided therein due to space constraints.

**Voice Generation** We generate audio for each question and option using the spoken-readable text produced during the rewriting stage. All text-to-speech (TTS) synthesis is conducted with the `Gemini 2.5 Flash Preview TTS` model, which supports multilingual generation. To ensure that the synthesized audio accurately reflects the target *language* and *accent*, we use the structured prompt shown in Figure 7 of Appendix B.2. This prompt explicitly specifies linguistic attributes, enabling consistent generation across English, Chinese, and Korean. This design allows controlled variation in *language* and *accent*, supporting robust multilingual evaluation.

### 2.2 Quality Assessment

**Question Rewriting** We normalize the rewritten outputs and the original Global MMLU Lite inputs by removing whitespace and converting all characters to lowercase. We then conduct a diff-based comparison to identify discrepancies. Flagged cases are manually reviewed, and any detected errors in the rewritten outputs are corrected to maintain consistency and accuracy. Table 9 in Appendix B.3 summarizes the proportion of automatically flagged instances and the true error rate confirmed through manual inspection.

**Voice Generation** We assess TTS quality using a two stage pipeline that combines automatic screening with manual verification. In the automatic stage, each audio sample is transcribed using two widely adopted ASR systems, Whisper Large v3 (Radford et al., 2022)and Omnilingual ASR (team et al., 2025). Because a single ASR model may introduce recognition errors, using two independent systems improves the reliability of WER based quality checks. For each sample, we compute WER against the rewritten text for both transcripts and take the minimum value as the final score. Samples are then grouped into four WER ranges ($0$, $(0, 0.2]$, $(0.2, 0.6]$, and $> 0.6$), and per language distributions are reported in Table 2. This automatic screening step serves as a quality control mechanism to identify potential synthesis errors before human inspection.

| Interval | English | Chinese | Korean |
|---|---|---|---|
| 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%) |

Table 2: Distribution across WER intervals by language.

**Human Evaluation of TTS Quality** To mitigate the risk of transcription errors underestimating dataset quality, we complement automatic evaluation with manual annotation. From each nonzero WER bin, 40 clips per language are randomly sampled using a stratified strategy to ensure representativeness. Annotation details are in Appendix B.4. Each clip is rated on a three-level scale: **Correct** (accurate and intelligible), **Acceptable** (minor mispronunciations but understandable), and **Incorrect** (severe errors causing misunderstanding). Table 3 shows the distribution of ratings across WER bins and languages. Most clips are rated as "Correct", indicating that TTS outputs are fluent, faithful, and well-aligned with the rewritten text. Many clips with nonzero WER also receive high manual ratings, implying that discrepancies mainly stem from ASR transcription or homophone errors rather than genuine TTS degradation. Together with automatic filtering, human evaluation forms a two-stage process ensuring dataset quality and consistency.

| Language | Incorrect | Acceptable | Correct |
|---|---|---|---|
| English | 6 (6.12%) | 11 (11.22%) | 81(82.65%) |
| Chinese | 7 (7%) | 7 (7%) | 86(86%) |
| Korean | 8 (8%) | 17 (17%) | 75 (75%) |

Table 3: Manual annotation results by *language* with ratings of **Correct**, **Acceptable**, and **Incorrect**.

## 3 Experimental Setup

### 3.1 Task and Variables

**Task Definition** Our objective is to investigate the robustness of multimodal large language models (MLLMs) in spoken multiple-choice question (MCQ) tasks. Unlike conventional text-only evaluations, this setting requires models to process an

| Variable | Levels |
|---|---|
| Language | English, Chinese (high-resource)<br>Korean (medium-resource) |
| Accent | English: American, British, Indian<br>Chinese: Beijing Mandarin, Northeastern Mandarin<br>Korean: Seoul, Jeolla |
| Gender | Male (Orus), Female (Zephyr) |
| Option Order | Original, Reversed |

Table 4: Controlled variables used to generate speech-based MCQ inputs. Combining these factors yields up to 28 configurations per question.

audio input consisting of a question followed by answer options, and then select the correct choice. This formulation introduces a central challenge: models must not only comprehend the linguistic content but also maintain consistency when the same question is presented under varying speech conditions in realistic settings.

**Experimental Variables** To systematically examine robustness, we conduct our experiments on the **BIASINEAR** benchmark introduced in Section 2. Each question is instantiated under controlled perturbations spanning linguistic (*language*, *accent*), demographic (*gender*), and structural (*option order*) dimensions, as summarized in Table 4. For *gender* variation, we adopt the *Orus* and *Zephyr* voices from Gemini[1]. For *option order*, the *original* setting represents the canonical sequence `A: {Option A}, B: {Option B}, C: {Option C}, D: {Option D}`, while the *reversed* setting presents the sequence in reverse, `A: {Option D}, B: {Option C}, C: {Option B}, D: {Option A}`. By combining these factors, a single question can yield up to 28 distinct configurations, enabling evaluation not only of absolute accuracy but also of stability across diverse speech conditions.

### 3.2 Models and Implementation

**Models** We evaluate nine MLLMs to assess their robustness under diverse experimental settings, including closed-weight models such as the Gemini family and open-source models such as the Gemma 3n, Voxtral, and Phi 4 families. Model details are provided in Appendix C.1 due to space constraints. To ensure the stability and scalability of the experiments, we access the models through APIs provided by Google, NVIDIA, and Mistral.

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

**Implementation Details** The audio samples generated in Section 2 consist of a question followed by its separate answer options. Before inputting them into the MLLM, we concatenate the respective audio segments according to the experimental condition (*original* or *reversed*). Details of the audio concatenation pipeline are provided in Appendix C.2 for brevity. For model inference, we set the temperature to 0 to ensure reproducibility. The prompts used for standard and chain-of-thought (CoT) prompting are shown in Figures 9 and 10 in Appendix C.3. Additionally, we apply post-processing to the model outputs to correct formatting errors, ensuring that our robustness analysis reflects genuine model behavior rather than artifacts from output format inconsistencies.

### 3.3 Evaluation Metrics

To evaluate robustness under input perturbations, we employ three complementary metrics: entropy, APES, and Fleiss' Kappa. These measures go beyond accuracy by assessing not only correctness but also the stability and consistency of model behavior. Detailed definitions are provided below. At a high level, they address the following questions:

- **Entropy**: *Does the model's answer distribution remain concentrated or become scattered across conditions?*
- **APES**: *Does its confidence vary when input conditions change?*
- **Fleiss' Kappa**: *Does the final prediction stay consistent under perturbations?*

**Question Entropy.** For each question $q$, we compute the Shannon entropy (Shannon, 1948) of the model's answer distribution:

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

where $p_q(o)$ is the probability assigned to option $o$. Normalization with base 4 ensures $H_q \in [0, 1]$.

**Level Entropy and APES.** Given a variable $v$ with levels $\mathbf{L}_v$ (e.g., $\{\text{female}, \text{male}\}$), we compute entropy at each level $l$ as

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

where $p_q(o|l)$ denotes the probability assigned to option $o$ under level $l$.

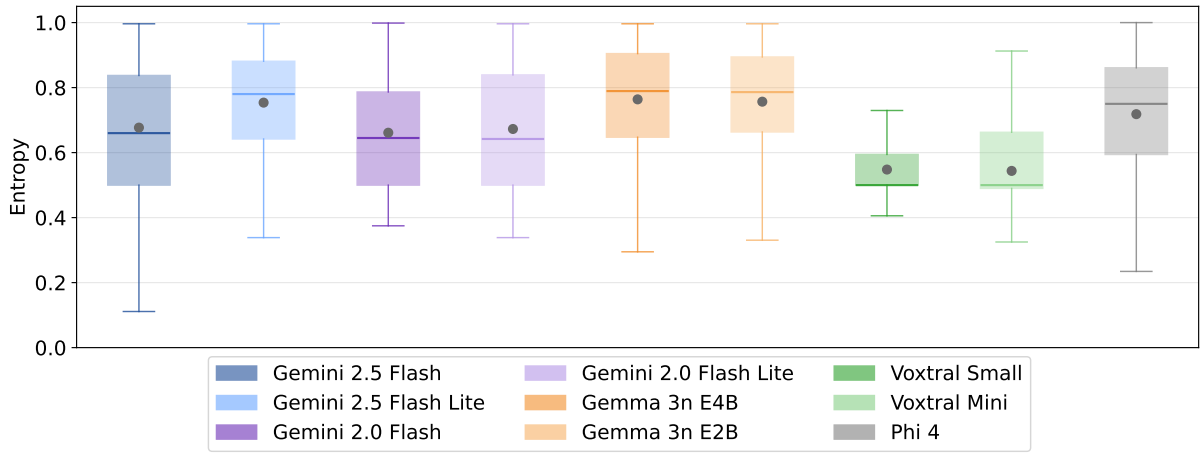


Figure 2: Mean question entropy across models. Higher entropy indicates greater uncertainty in model predictions.

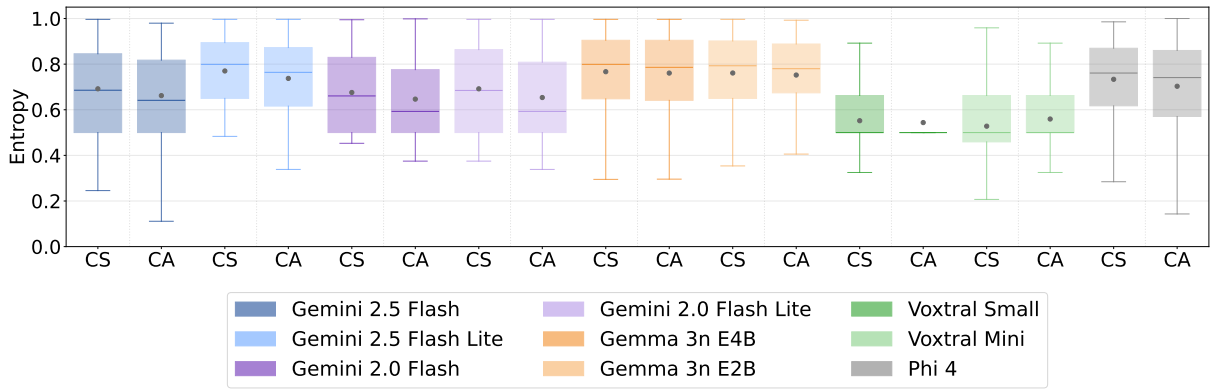


Figure 3: Entropy comparison between Culturally Sensitive (CS) and Culturally Agnostic (CA) questions.

The Average Pairwise Entropy Shift (APES) quantifies entropy variation across levels:

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

where $L = |\mathbf{L}_v|$, and $H_q^{l_i}$ be the entropy of $l_i \in \mathbf{L}_v$.

**Fleiss' Kappa.** For each question $q$, we compute Fleiss' $\kappa$ (Fleiss, 1971) to measure categorical agreement across variable perturbations while correcting for chance, defined as

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

where $\bar{P}$ is the average observed agreement and $P_e$ is the expected agreement. The detailed formulation is provided in Appendix C.4. $\kappa \approx 1$ indicates strong consistency across conditions, $\kappa \approx 0$ suggests agreement no better than chance, and a negative $\kappa$ reflects systematic disagreement worse than random expectation.

## 4 Investigation on Speech Bias

### 4.1 Overall Observation

Figure 2 illustrates the overall entropy trends across the nine evaluated models. For each model, we compute per-question entropy across all configurations and then average the results over 400 questions per setting. The results reveal that the Gemini and Gemma families exhibit consistently higher entropy, indicating greater uncertainty in their answer distributions under diverse conditions. In contrast, the Voxtral family show lower entropy with narrower dispersion, reflecting more concentrated and confident predictions, whereas the Phi 4 model displays a larger interquartile range, suggesting greater variability in prediction confidence across questions. Within each model family, the lighter variants (e.g., `Voxtral Mini` vs. `Voxtral Small`) exhibit slightly higher mean entropy than their larger counterparts, suggesting that smaller parameter scales tend to produce less stable behavior and greater prediction uncertainty overall.

| Model | Accent | | | Language | | | Option Order | | | Gender | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 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: Mean entropy of culturally sensitive (CS) vs. culturally agnostic (CA) questions across variables.

| | Chinese | | | | English | | | | Korean | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gender | Accent | Orig. | Rev. | Δ | Accent | Orig. | Rev. | Δ | Accent | Orig. | Rev. | Δ |
| ♂ | Beijing | **62.75** | 61.50 | 1.25 | American | **82.00** | 79.75 | 2.25 | Jeolla | **61.50** | 61.00 | 0.50 |
| | Northeastern | **64.75** | 61.25 | 3.50 | British | **81.00** | 78.50 | 2.50 | Seoul | **63.00** | 62.00 | 1.00 |
| | | | | | Indian | **80.00** | 79.25 | 0.75 | | | | |
| ♀ | Beijing | **66.75** | 64.00 | 2.75 | American | **80.00** | 78.50 | 1.50 | Jeolla | **63.50** | 58.50 | 5.00 |
| | Northeastern | **64.50** | 57.75 | 6.75 | British | **81.25** | 76.75 | 4.50 | Seoul | **63.75** | 62.00 | 1.75 |
| | | | | | Indian | **80.50** | 78.25 | 2.25 | | | | |

Table 6: Accuracy comparison across *option order* conditions for **Gemini 2.5 Flash**. Each cell reports the mean accuracy (%) for the *original* and *reversed* option orders, with Δ denoting the difference between them. Results are grouped by *language* (Chinese, English, Korean), *accent*, and *gender*.

### 4.2 Comparison Between CS and CA

The Global MMLU Lite benchmark categorizes questions as Culturally Sensitive (CS), which require contextual or culture-specific knowledge, and Culturally Agnostic (CA), which rely primarily on domain knowledge. Figure 3 presents entropy comparisons across nine models under CS and CA settings. Overall, CA questions consistently exhibit lower entropy, indicating more concentrated and stable answer distributions aligned with factual reasoning, whereas CS questions display broader entropy ranges. As shown in Table 5, this CS–CA entropy gap persists across variables: perturbations in *accent* and *gender* introduce only minor differences, while *option order* produces the largest gap, suggesting higher positional sensitivity for culturally grounded items. Complementary CS/CA robustness results under cross-variable perturbations are reported in Appendix D.1.

### 4.3 Accuracy across Variable Levels

We next perform a level-wise analysis to examine how different variable levels influence model behavior, beginning with accuracy. Table 6 reports the performance of `Gemini 2.5 Flash` across Chinese, English, and Korean. The accuracy gap between *option order* configurations ranges from 0.5% to 6.75%, with the *original* order consistently outperforming the *reversed* order across all language-accent settings. Results for other models, presented in Tables 11-17b in Appendix D.2, reveal a similar pattern: most models achieve higher accuracy under the *original* configuration. These findings indicate that *option order* introduces a systematic bias in model predictions. The effects of other variables are also summarized in same Appendix.

### 4.4 Robustness across Variable Levels

**Level-wise Analysis** Beyond correctness, Figure 4 visualizes robustness patterns across model families. The *gender* and *accent* lie in the lower-right quadrant (high $\kappa$, low APES), indicating robust predictions characterized by strong within-level agreement and stable uncertainty across levels. In contrast, *language* occupies intermediate regions ($\kappa \approx 0$, higher APES) with substantial variation across model families, suggesting that cross-lingual generalization remains a key robustness challenge. Note that Voxtral family do not support Chinese or Korean, and thus language-level analysis is omitted for this family. Finally, *option order* consistently emerges as the weakest factor across

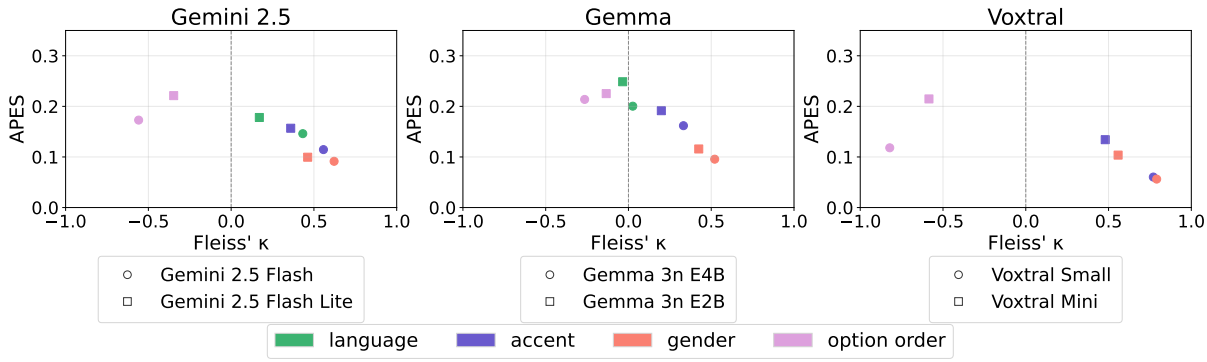


Figure 4: Fleiss' $\kappa$ versus APES across model families. Variables such as *language*, *accent*, and *gender* show higher agreement and stability, while *option order* yields higher APES and lower $\kappa$, indicating strong sensitivity.

| Model | Accent | | Option Order | |
|---|---|---|---|---|
| | Direct | Cloning | Direct | Cloning |
| 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** |

Table 7: APES comparison across *accent* and *option order*. Lower values indicate higher robustness.

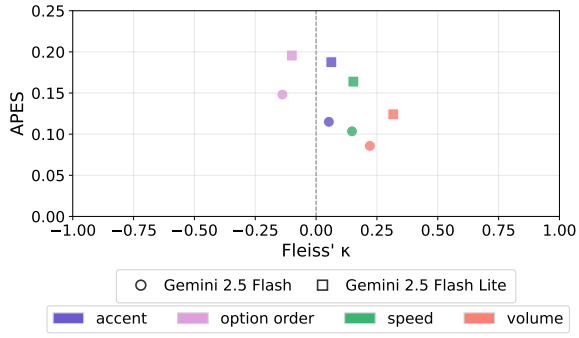


Figure 5: Fleiss' $\kappa$ versus APES across *accent*, *option order*, *speed*, and volume variables.

all models, typically appearing in the left quadrants with negative $\kappa$ and relatively high APES, reflecting pronounced sensitivity to input order. Overall, while models exhibit relatively greater robustness to speaker-related factors (*gender* and *accent*), agreement under these perturbations only reaches the "Moderate" to "Substantial" range ($\kappa \approx 0.4$-$0.8$), rather than "Almost perfect" ($\kappa > 0.8$) that robust system ideally require. This gap indicates substantial room for improving speech robustness.

**Impact of Model Scale** Figure 4 also compares model robustness across scales within three representative families. Results for Gemini 2.0 family are provided in Appendix D.3 for completeness. Larger models consistently demonstrate higher $\kappa$ and lower APES for the *gender*, *accent*, and *language* variables, indicating more stable and consistent behavior under input perturbations. For *option order*, larger models also achieve lower APES, although $\kappa$ remains negative, making direct comparison less meaningful. These results suggest that parameter reduction amplifies vulnerability to input perturbations, rendering smaller or lite variants less robust than their full-scale counterparts.

**Option Order Variants** Following Wei et al. (2024), we evaluate whether *option order* bias generalizes beyond a single reversal by applying multiple option permutations, including *original*, *fully reversed*, *token-backward*, and *order-backward*. Results in Appendix D.4 (Tables 21–22) show that option reordering has limited impact on APES across other variables, preserves the factor ranking (*language* > *accent* > *gender*), and that fully *reversed* orders induce the highest uncertainty, consistent with our main findings.

## 5 Discussion

### 5.1 Real World Speaker Variability

Most experiments in this work rely on TTS generated speech, which may raise concerns about whether the observed speech biases generalize to real world settings. To address this, we conduct two complementary analyses to better approximate real world speaker variability and acoustic conditions.

**Speaker Identity Realism** To move beyond purely synthetic voices, we collect short recordings from three real speakers representing American, British, and Indian English accents, and use Chatterbox (Resemble AI, 2025), a neural voice cloning TTS model, to generate the full 400 ques-

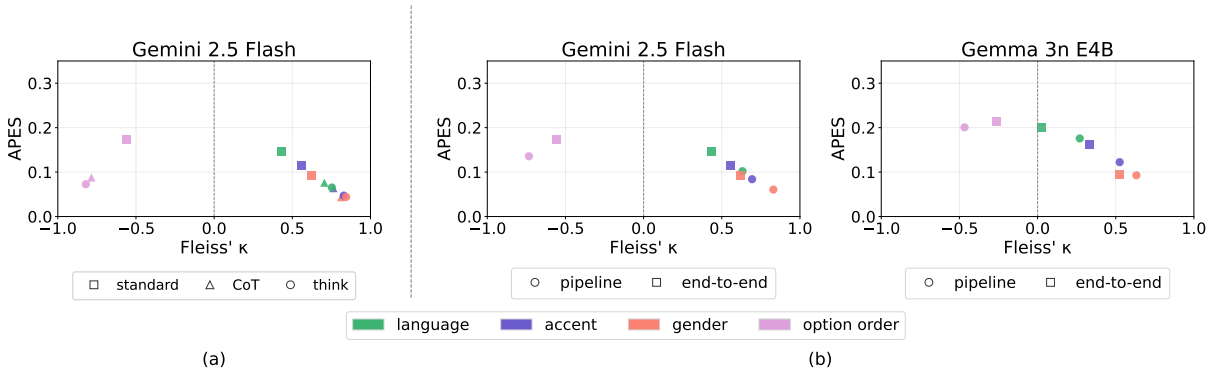


Figure 6: Effect of (a) reasoning complexity and (b) architectural paradigm on model robustness. Higher reasoning complexity and pipeline designs yield higher agreement (Fleiss' $\kappa$) and lower uncertainty (APES).

tion English Global MMLU Lite dataset for each accent. As shown in Table 7, the core trends and relative model rankings remain consistent across these cloned voices. Importantly, the bias patterns closely match those observed under direct TTS generation, suggesting that our findings are not artifacts of a specific synthetic voice, but persist under speaker characteristics closer to real world application conditions.

**Acoustic Variability** To further account for variability in recording conditions, we perturb the cloned English data with different speech rates (0.75×, 1.0×, and 1.25×) and loudness levels (0.5×, 1.0×, and 1.5×). Results in Figure 5 indicate that variations in speech rate introduce noticeably larger bias than changes in volume, as reflected by higher APES values across all models. Despite these perturbations, the main conclusions of our study remain stable, reinforcing the robustness of the observed speech bias patterns under more realistic acoustic variability.

### 5.2 Impact of Reasoning Complexity

We examine how increasing reasoning complexity affects model robustness by comparing standard prompting, chain-of-thought (CoT) prompting (Wei et al., 2023), and explicit reasoning (thinking) modes. Figure 6(a) reports results for `Gemini 2.5 Flash`. Overall, CoT prompting substantially improves agreement, with Fleiss' $\kappa$ increasing by an average of 19.01%, 20.50%, and 27.20% for *gender*, *accent*, and *language*, respectively. It also yields greater robustness, reflected in mean APES reductions of 4.79%, 5.07%, 6.98%, and 8.50% for *gender*, *accent*, *language*, and *option order*. Further enabling explicit reasoning leads to additional gains in both agreement and robustness beyond CoT prompting alone. These results suggest that increased reasoning complexity mitigates input induced variability and stabilizes model predictions under diverse perturbations.

### 5.3 Impact of Architectural Paradigm

We next examine the role of architectural paradigm, contrasting end-to-end multimodal LLMs with a pipeline design. While end-to-end models process audio directly, they may rely on ASR-like layers that filter out paralinguistic cues (e.g., *accent*, *gender*). To test this, we construct a pipeline setup where the model first transcribes the audio into text before answering. This comparison assesses whether explicit transcription removes speaker-dependent cues affecting robustness across conditions. We apply this comparison to two representative models, `Gemini 2.5 Flash` and `Gemma 3n E4B`, selected for their strong multimodal input capabilities. As shown in Figure 6 (b), under the pipeline setting, both models exhibit higher Fleiss' $\kappa$ and lower APES across *language*, *accent*, and *gender*, relative to their end-to-end counterparts. This pattern indicates that explicit transcription suppresses speaker-dependent variability, thereby mitigating accent- and gender-induced biases in the final predictions. Taken together, these results highlight architectural paradigm as a key lever for robustness, with the pipeline procedure reducing paralinguistic sensitivity and promoting more consistent behavior across conditions.

### 5.4 Speech as a Bias Amplifier

Before attributing the observed robustness differences to properties unique to speech, we examine whether these patterns reflect amplification of biases already present in text-based question answer-

| model | Language | | Option order | |
|---|---|---|---|---|
| | Text | Audio | Text | Audio |
| 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 |

Table 8: Comparison of APES between text and audio inputs across *language* and *option order*. Audio inputs consistently yield higher APES, indicating amplification of existing robustness sensitivities.

ing. We therefore compare text and audio inputs along two variables known to induce sensitivity, namely *language* and *option order*. As shown in Table 8, all models exhibit consistently higher APES values under the audio condition than under text. This indicates that the robustness patterns observed in speech are not unique to the audio modality, but correspond to systematic amplification of existing biases when questions are presented in spoken form. This cross-modal comparison serves as a sanity check that grounds our analysis, supporting the interpretation that speech primarily magnifies sensitivities already present in text-based models rather than introducing qualitatively new bias patterns.

## 6 Related Work

**Speech Bias in ASR.** Prior work on speech bias has primarily focused on automatic speech recognition (ASR) systems, which consistently exhibit performance disparities across gender, accent, and language. Gender related biases have been widely reported, with higher word error rates (WER) observed for female speakers in YouTube auto captions (Tatman, 2017), male speech in Whisper Small (ElGhazaly et al., 2025), and model dependent reversals across systems (Graham and Roll, 2024). Accent bias is similarly pervasive, with American and Canadian English yielding lower WERs than non native accents (Graham and Roll, 2024), and substantial variation across regional accents in datasets such as SQuAD SRC (Tang and Tung, 2023). In multilingual settings, high resource languages consistently outperform low resource or tonal languages, as shown by higher WERs for Chinese and Korean in Meta's XLS R model (Babu et al., 2022). Overall, these findings attribute ASR bias to the combined effects of gender, accent, and data imbalance. Our work extends this line of research beyond transcription accuracy to examine **speech bias in multilingual MLLMs**, enabling a unified evaluation across linguistic and demographic dimensions.

**LLM Robustness.** Recent studies document systematic biases in LLMs across demographic and structural dimensions. Gender bias persists even without explicit markers (Belém et al., 2024), with models exhibiting systematic disadvantages against women (Vo et al., 2025), while racial and dialectal biases include covert negative stereotypes toward African American English (Hofmann et al., 2024) and broader racial, national, and religious stereotypes under complex reasoning settings (Shrawgi et al., 2024). Cultural bias further arises from the dominance of Western centric training data, leading to disparities in multilingual and multicultural contexts (LI et al., 2024). Beyond demographic factors, Wei et al. (2024) identify selection bias, a structural sensitivity to non semantic cues such as option order or symbolic formatting in multiple choice questions. Motivated by these findings, our work extends the study of selection bias from text based evaluations to the speech modality, examining positional sensitivities under spoken inputs.

## 7 Conclusion

This work presents the first systematic study of **speech bias and robustness** in multilingual MLLMs. We introduce **BIASINEAR**, a speech-augmented benchmark built on Global MMLU Lite, covering English, Chinese, and Korean, balanced across *gender* and *accent*, and comprising 11,200 questions with 70.8 hours (≈4,249 minutes) of speech. Using four complementary metrics (accuracy, entropy, APES, and Fleiss' $\kappa$), we evaluate nine representative models from four model families and analyze robustness across linguistic, demographic, and structural factors. Our results show that *option order* induces the most pronounced robustness degradation, while *accent* and *gender* lead to smaller but consistent confidence shifts. We further demonstrate that increased reasoning complexity and pipeline-based architectural designs improve robustness, and that speech systematically amplifies biases already present in text-based settings. Together, these findings reveal underexplored vulnerabilities in current MLLMs and offer practical insights for designing fairer and more stable speech-integrated AI systems.

## Limitations

**Voice Generation** Although our dataset systematically controls for *language*, *accent*, and *gender*, the use of text-to-speech (TTS) for automated audio generation introduces inherent challenges in defining and standardizing "accent." Even within a single *language*, *accent* variation often exists on a continuous spectrum rather than as discrete categories. The boundaries between regional or social varieties are fuzzy and, in some cases, linguistically indeterminate, making it difficult to ensure that our current setup fully captures the natural and continuous variation present in human speech. Furthermore, due to computational constraints, we were unable to synthesize a larger number of voice variants for each condition. Nevertheless, because our dataset is derived from Global-MMLU-Lite, which covers a broad range of topics and languages, we believe that our results remain representative and robust in capturing overall cross-linguistic and paralinguistic trends.

**Evaluated Models** Due to computational and API interface constraints, this study evaluated only nine representative multimodal large language models (MLLMs) spanning both commercial and open-source categories. While these models capture diversity in architecture and reasoning pipelines, they do not fully cover the spectrum of existing systems. Some open-source models were excluded due to limited stability, insufficient scalability for large-scale inference, or the lack of a publicly available, stable, and efficient multimodal API. Future work could broaden the evaluation as standardized interfaces and reproducible deployment pipelines mature, enabling a more comprehensive assessment of cross-model consistency and generalization.

## Use of AI Assistants

We used ChatGPT as an assistant to refine the manuscript, improve clarity, and enhance the structure and readability. While the final content remains entirely our own, this assistance helped improve the overall presentation of our work.

## Acknowledgements

This work was supported by National Science and Technology Council, Taiwan, under grant NSTC 114-2221-E-002 -070 -MY3, NSTC 113-2634-F-002-003 -, and Ministry of Education (MOE) in Taiwan under grants NTU-114L900901.

## References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, 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, and 23 others. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Anthropic. 2025. Claude 3.7 sonnet and claude code. Accessed: 2025-09-29.

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21318–21340, Miami, Florida, USA. Association for Computational Linguistics.

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, and Michael Auli. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pages 2278–2282.

Catarina Belém, Preethi Seshadri, Yasaman Razeghi, and Sameer Singh. 2024. Are models biased on text without gender-related language? In *International Conference on Representation Learning*, volume 2024, pages 12876–12915.

Jad Bendarkawi, Ashley Ponce, Sean Chidozie Mata, Aminah Aliu, Yuhan Liu, Lei Zhang, Amna Liaqat, Varun Nagaraj Rao, and Andrés Monroy-Hernández. 2025. Conversar: Exploring embodied llm-powered group conversations in augmented reality for second language learners. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

May Pik Yu Chan, June Choe, Aini Li, Yiran Chen, Xin Gao, and Nicole Holliday. 2022. Training and typological bias in ASR performance for world Englishes. In *Interspeech 2022*, pages 1273–1277.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. 2024. Voicebench: Benchmarking llm-based voice assistants. *Preprint*, 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 Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Hend ElGhazaly, Bahman Mirheidari, Nafise Sadat Moosavi, and Heidi Christensen. 2025. Exploring gender disparities in automatic speech recognition technology. *Preprint*, arXiv:2502.18434.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Google Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *ArXiv*, abs/2312.11805.

Calbert Graham and Nathan Roll. 2024. Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2):025206.

Camille Harris, Chijioke Mgbahurike, Neha Kumar, and Diyi Yang. 2024. Modeling gender and dialect bias in automatic speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15166–15184, Miami, Florida, USA. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Dialect prejudice predicts ai decisions about people's character, employability, and criminality. *Preprint*, arXiv:2403.00742.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Ajinkya Kulkarni, Atharva Kulkarni, Miguel Couceiro, and Isabel Trancoso. 2024. Unveiling biases while embracing sustainability: Assessing the dual challenges of automatic speech recognition systems. In *Interspeech 2024*, interspeech 2024, pages 4628–4632. ISCA.

CHENG LI, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024. Culturepark: Boosting cross-cultural understanding in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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, and 87 others. 2025. Voxtral. *Preprint*, arXiv:2507.13264.

Rao Ma, Mengjie Qian, Siyuan Tang, Stefano Bannò, Kate M. Knill, and Mark J. F. Gales. 2025. Assessment of l2 oral proficiency using speech large language models. *Preprint*, arXiv:2505.21148.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal intelligence.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. 2024. Spoken question answering and speech continuation using spectrogram-powered LLM. In *The Twelfth International Conference on Learning Representations*.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2022. Introducing ChatGPT. `https://openai.com/blog/chatgpt`.

OpenAI. 2024. GPT-4o System Card. *arXiv preprint*. ArXiv:2410.21276.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Resemble AI. 2025. Chatterbox-TTS. `https://github.com/resemble-ai/chatterbox`. GitHub repository.

Sujoy Roychowdhury, Ranjani H.G., Sumit Soman, Nishtha Paul, Subhadip Bandyopadhyay, and Siddhanth Iyengar. 2025. Intelligibility of Text-to-Speech Systems for Mathematical Expressions. In *Interspeech 2025*, pages 2280–2284.

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, and 11 others. 2023. Audiopalm: A large language model that can speak and listen. *Preprint*, arXiv:2306.12925.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Min-Han Shih, Ho-Lam Chung, Yu-Chi Pai, Ming-Hao Hsu, Guan-Ting Lin, Shang-Wen Li, and Hung yi Lee. 2024. GSQA: An End-to-End Model for Generative Spoken Question Answering. In *Interspeech 2024*, pages 2970–2974.

Hari Shrawgi, Prasanjit Rath, Tushar Singhal, and Sandipan Dandapat. 2024. Uncovering stereotypes in large language models: A task complexity-based approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1841–1857, St. Julian's, Malta. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Divya Tadimeti, Kallirroi Georgila, and David Traum. 2022. Evaluation of off-the-shelf speech recognizers on different accents in a dialogue domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6001–6008, Marseille, France. European Language Resources Association.

Weiting Tan, Hirofumi Inaguma, Ning Dong, Paden D. Tomasello, and Xutai Ma. 2025. SSR: Alignment-aware modality connector for speech language models. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 56–75, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Yixuan Tang and Anthony K.H: Tung. 2023. Squad-src: A dataset for multi-accent spoken reading comprehension. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5206–5214. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Omnilingual ASR team, 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, and 14 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *Preprint*, arXiv:2511.09690.

An Vo, Mohammad Reza Taesiri, Daeyoung Kim, and Anh Totti Nguyen. 2025. B-score: Detecting biases in large language models using response history. In *Forty-second International Conference on Machine Learning*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5598–5621, Bangkok, Thailand. Association for Computational Linguistics.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

## A Cost Analysis

All experiments involving TTS generation and model inference via the Gemini API incurred a total cost of under $550 USD. For the other models, the APIs provided by Mistral and NVIDIA were temporarily free during the experiment period.

## B Dataset Construction Details

### B.1 Question Rewriting

To perform the rewriting of questions and options, we employ the `GPT OSS 120B` model via the NVIDIA API, which ensures both stability and scalability. The model is prompted with task-specific instructions shown in Figure 8, which enforce eight conversion rules. These rules cover aspects such as reading mathematical expressions (e.g., "$x^2+y^2$"), disambiguating domain-specific terms using subject context (e.g., `"Na"` → `"sodium"` in chemistry), handling numbers and units (e.g., `"3kg"` → `"three kilograms"`), interpreting parentheses, and rendering placeholders like `"BLANK"` appropriately across different languages. This step ensures the generation of high-quality spoken-readable text prior to audio synthesis.

### B.2 TTS Generation Prompt

Figure 7 illustrates the prompt template used for TTS synthesis, showing how textual instructions, speaker characteristics, and prosodic cues are combined to guide the model toward more natural, expressive, and context-aware speech.

### B.3 Quality Assessment of Rewriting

As we mentioned in Section 2.2, to evaluate the reliability of the rewritten questions, we manually inspected a subset of automatically flagged cases. Table 9 summarizes the proportion of flagged instances and verified true errors across English, Chinese, and Korean.

### B.4 Quality Assessment of Voice Generation with Stratified Sampling

Sampling for quality assessment was stratified for representativeness and diversity. Within each WER interval, we pooled all accents and allocated per-accent quotas proportional to their sample counts using the largest-remainder method. For each question–answer pair, we selected one item (preferring the question-description row), then sampled in a subject-wise round-robin to balance subject coverage. When an interval lacked enough unique subjects or questions to meet the quota, we filled the remainder from the available pool.

```
Read this text in {language} with a {accent}
accent: {text}
```

Figure 7: Prompt template used for TTS synthesis. The placeholder `{text}` is substituted with the rewritten question or answer option.

| Language | Flagged Cases(%) | Verified Errors (%) |
|---|---|---|
| English | 477 (23.85%) | 20 (1.00%) |
| Chinese | 494 (24.70%) | 93 (4.65%) |
| Korean | 577 (28.85%) | 79 (3.95%) |

Table 9: Quality control statistics for rewritten questions. The table reports the number and percentage of automatically flagged cases and human-verified true errors.

## C Experimental Setup Details

### C.1 Models

As described in Section 3.2, we evaluate nine large multimodal models across both commercial and open-source settings. Table 10 summarizes these models, including their API endpoints provided by Google[2], NVIDIA[3], and Mistral[4], along with their corresponding providers.

### C.2 Audio Concatenation

For each question, we constructed an audio query by concatenating the question and four option descriptions with instruction tokens (*"question"*, *"A"* – *"D"*) pre-generateed by the TTS model. Instruction tokens were rendered in American, British, and Indian accents for English, and in American accent for Chinese and Korean.

The concatenation process followed the designated order: each instruction token placed before its corresponding content; four orderings (*original*, *reversed*, *order backward*, *token backward*) were implemented, with fixed pauses inserted between segments for clarity. To accommodate input length constraints (30s) in models such as `Gemma` and `Phi 4`, each final audio was further segmented into fixed-length chunks and exported as waveform files. This ensured consistent and compatible inputs across all experimental variables. This ensured consistent inputs across all experimental variables.

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

**[System]**
Convert the inputs into how a native {language} speaker would read them. Output only the results. Do not explain reasoning or revise the plain text content.
Inputs
Subject: <subject>
Question: <question>
Options: A) <option_a>, B) <option_b>, C) <option_c>, D) <option_d>

The conversion should follow the 8 rules:

1. Use the "Subject" to disambiguate domain terms or abbreviations (e.g., "ms" read as "millisecond" in Physics, "HPO42-" read as "hydrogen phosphate" in Chemistry, 'P' read as "phosphorus" in Chemistry).

2. Read math expressions naturally. Examples: "f(x)" $\rightarrow$ "f of x"; $Z_n \rightarrow$ "integers modulo n"; $(12)(123) \rightarrow$ "the permutation consisting of the cycle one-two, and the cycle one-two-three". If letter case matters, read it explicitly: "O(n) and o(n)" $\rightarrow$ "big-O of n and little-o of n".

3. Read numbers and units cleanly with correct plurals.

4. Read parentheses or brackets only if they affect meaning.

5. Read Roman-numeral labels (I, II, i, ii ...) as labels: "Roman numeral one", "Roman numeral two", etc.

6. Read "BLANK" in place of blank markers "_____"; number multiple blanks ("BLANK 1", "BLANK 2"); keep surrounding grammar; never guess the answer.

7. All conversions must be written directly in the {language} specified by the instructions. When the prompt contains "BLANK", render it in the language indicated (e.g., Chinese: 空格; Japanese: 空白; Korean: 공백). All other text must also be in the same language.

Output format (XML):

```
<output>
  <question>&lt;spoken rendering of the question&gt;</question>
  <option_a>&lt;spoken rendering of option A&gt;</option_a>
  <option_b>&lt;spoken rendering of option B&gt;</option_b>
  <option_c>&lt;spoken rendering of option C&gt;</option_c>
  <option_d>&lt;spoken rendering of option D&gt;</option_d>
</output>
```

Figure 8: Prompt for spoken-style rendering of MMLU-style items

| Model | API Endpoint | Provider |
|---|---|---|
| | *Commercial APIs* | |
| 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 |
| | *Open-source Models* | |
| 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 |

Table 10: Evaluated models.

### C.3 Model Inference and Post-processing

All models were queried through their official APIs with deterministic greedy decoding (temperature set to zero and candidate count set to one) to eliminate randomness. A unified text prompt was used to standardize outputs as shown in Figure 9 For chain-of-thought (CoT) inference, we used the prompt as shown in Figure 10 to standardize outputs. The maximum output length was capped at 4k tokens, sufficient to cover multiple-choice responses.

Model responses were post-processed to extract the final answer letter via pattern matching (e.g., `Answer:[[A]]`, `Answer: A`). The letter was mapped to an index per *option order*, and invalid outputs were marked as parsing failures.

### C.4 Derivation of Fleiss' $\kappa$

**Setup.** Fix a variable $v$, consider a set of items constructed by the other variables combination, indexed by $i = 1, \ldots, I$, where $I$ denotes the number of all variable combination except for $v$. Each item is assigned to one of $j$ categorical (A, B, C, D) by $n_i$ ratings. In our application, the $n_i$ ratings arise from the same model answering item $i$ under different levels of $v$. Let $n_{ij}$ denote the number of ratings that chose category $j$ for item $i$, so that $\sum_{j=1}^{J} n_{ij} = n_i$.

**Observed agreement per item.** For item $i$, the proportion of agreeing rater-pairs is

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

since each category-$j$ contributes $\binom{n_{ij}}{2}$ agreeing pairs and there are $\binom{n_i}{2}$ total unordered pairs.

**Expected agreement under chance.** Under the usual "random assignment with fixed marginals model", two independently drawn ratings match with probability

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

**Averaging observed agreement.** Aggregate the per-item agreements in (5) by the number of ratings:

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

This weighting treats each rating pair equally across items when $n_i$ varies.

```
Answer the following question. Output only
"Answer:[[LETTER]]", where LETTER is one of
A, B, C, D.
```

Figure 9: Standard Prompt template

```
Answer the following question. Let's
think step by step, and write concise
reasoning. The last answer line should be
"Answer:[[LETTER]]", where LETTER is one of
A, B, C, D.
```

Figure 10: CoT Prompt template

**Fleiss' Kappa.** Fleiss' $\kappa$ standardizes the excess agreement over chance:

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

with $\kappa = 1$ if $\bar{P} = 1$ (perfect agreement), $\kappa = 0$ if $\bar{P} = P_e$ (no better than chance), and $\kappa < 0$ when observed agreement falls below chance.

## D Detailed Experiment Results

### D.1 CS vs. CA under Variable Perturbations

We further stratify questions into Culturally Sensitive (CS) and Culturally Agnostic (CA) subsets to assess cross-variable robustness under *language*, *accent*, *gender*, and *option order* perturbations. Figures 11 and 12 show that across models and variables, CA items consistently exhibit lower uncertainty (APES) than CS items, indicating stronger robustness under cross-variable shifts

### D.2 Accuracy Analysis

Tables 11 to 17 provide accuracy comparison across *option order* grouped by *language*, *accent*, and *gender* for the other eight models: Gemini 2.5 Flash Lite, Gemini 2.0 Flash, Gemini 2.0 Flash Lite, Gemma 3n E4B, Gemma 3n E2B, Voxtral Small, Voxtral Mini, Phi 4 Multimodal. Tables 18 to 20 provide a factor-wise breakdown for Gemini 2.5 Flash, reporting accuracy comparisons grouped by *gender*, *accent*, and *language*, respectively. These stratified analyses clarify not only whether option reordering affects accuracy, but also how its impact interacts with speech-specific factors, yielding a more interpretable picture of model behavior under controlled variations.

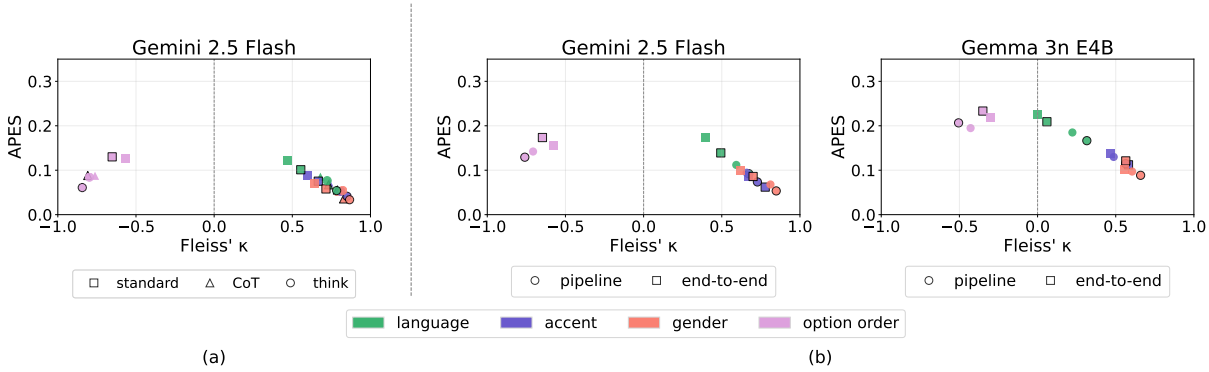


Figure 11: CS/CA stratified APES–$\kappa$ analysis under cross-variable perturbations (language, accent, gender, and option order). Unboxed markers denote CS items, while boxed markers denote CA items.

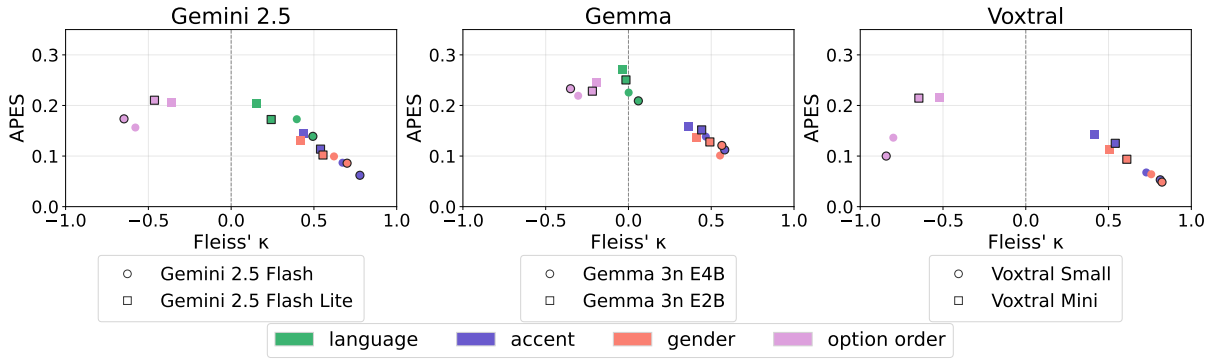


Figure 12: Grouped APES–$\kappa$ robustness plots across model families under cross-variable perturbations. Unboxed markers denote CS items, while boxed markers denote CA items.

| Chinese | | | | English | | | | Korean | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| input | original | reversed | Δ | input | original | reversed | Δ | input | original | reversed | Δ |
| Beijing ♀ | **46.75** | 45.00 | 1.75 | American ♀ | **62.25** | 61.00 | 1.25 | Jeolla ♀ | **43.75** | 42.50 | 1.25 |
| Beijing ♂ | **44.75** | 41.75 | 3.00 | American ♂ | **61.50** | 61.75 | -0.25 | Jeolla ♂ | **45.75** | 44.00 | 1.75 |
| Northeastern ♀ | **43.50** | 40.00 | 3.50 | British ♀ | **61.75** | 61.00 | 0.75 | Seoul ♀ | **44.50** | 43.50 | 1.00 |
| Northeastern ♂ | **41.60** | 40.35 | 1.25 | British ♂ | **60.75** | 58.50 | 2.25 | Seoul ♂ | **46.00** | 44.75 | 1.25 |
| | | | | Indian ♀ | **64.00** | 60.50 | 3.50 | | | | |
| | | | | Indian ♂ | **59.25** | 58.00 | 1.25 | | | | |

Table 11: Accuracy comparison across option-order settings for *Gemini 2.5 Flash Lite*, grouped by language, accent, and gender.

| Chinese | | | | English | | | | Korean | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| input | original | reversed | Δ | input | original | reversed | Δ | input | original | reversed | Δ |
| Beijing ♀ | **70.75** | 67.75 | 3.00 | American ♀ | **83.75** | 80.75 | 3.00 | Jeolla ♀ | **68.00** | 65.25 | 2.75 |
| Beijing ♂ | 69.00 | **69.50** | -0.50 | American ♂ | **83.25** | 80.00 | 3.25 | Jeolla ♂ | 66.00 | **67.00** | -1.00 |
| Northeastern ♀ | **68.25** | 66.00 | 2.25 | British ♀ | **81.75** | 80.50 | 1.25 | Seoul ♀ | **70.25** | 68.25 | 2.00 |
| Northeastern ♂ | 65.50 | **69.25** | -3.75 | British ♂ | **81.75** | 78.75 | 3.00 | Seoul ♂ | **67.25** | 65.00 | 2.25 |
| | | | | Indian ♀ | **81.50** | 80.00 | 1.50 | | | | |
| | | | | Indian ♂ | **83.00** | 80.00 | 3.00 | | | | |

Table 12: Accuracy comparison across option-order settings for *Gemini 2.0 Flash*, grouped by language, accent, and gender.

| Chinese | | | | English | | | | Korean | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| input | original | reversed | Δ | input | original | reversed | Δ | input | original | reversed | Δ |
| Beijing ♀ | **65.50** | 63.50 | 2.00 | American ♀ | **79.75** | 77.00 | 2.75 | Jeolla ♀ | 66.50 | **66.75** | -0.25 |
| Beijing ♂ | **66.25** | 64.75 | 1.50 | American ♂ | 77.75 | **78.25** | -0.50 | Jeolla ♂ | **62.75** | 61.25 | 1.50 |
| Northeastern ♀ | **66.75** | 61.75 | 5.00 | British ♀ | **78.00** | 74.25 | 3.75 | Seoul ♀ | **64.75** | 64.00 | 0.75 |
| Northeastern ♂ | 65.00 | **65.50** | -0.50 | British ♂ | **78.50** | 74.75 | 3.75 | Seoul ♂ | **66.25** | 62.50 | 3.75 |
| | | | | Indian ♀ | **79.00** | 77.25 | 1.75 | | | | |
| | | | | Indian ♂ | **77.75** | 77.25 | 0.50 | | | | |

Table 13: Accuracy comparison across option-order settings for *Gemini 2.0 Flash Lite*, grouped by language, accent, and gender.

| Chinese | | | | English | | | | Korean | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| input | original | reversed | Δ | input | original | reversed | Δ | input | original | reversed | Δ |
| Beijing ♀ | **37.00** | 34.75 | 2.25 | American ♀ | **59.75** | 53.25 | 6.50 | Jeolla ♀ | **36.25** | 32.25 | 4.00 |
| Beijing ♂ | 34.25 | **35.00** | -0.75 | American ♂ | **59.25** | 56.75 | 2.50 | Jeolla ♂ | **35.75** | 33.50 | 2.25 |
| Northeastern ♀ | 33.50 | **34.75** | -1.25 | British ♀ | **59.75** | 52.00 | 7.75 | Seoul ♀ | 33.25 | **34.50** | -1.25 |
| Northeastern ♂ | 37.00 | **37.50** | -0.50 | British ♂ | **61.25** | 51.00 | 10.25 | Seoul ♂ | **37.25** | 36.00 | 1.25 |
| | | | | Indian ♀ | **56.75** | 53.50 | 3.25 | | | | |
| | | | | Indian ♂ | **60.00** | 51.50 | 8.50 | | | | |

Table 14: Accuracy comparison across option-order settings for *Gemma 3n E4B*, grouped by language, accent, and gender.

| Chinese | | | | English | | | | Korean | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| input | original | reversed | Δ | input | original | reversed | Δ | input | original | reversed | Δ |
| Beijing ♀ | **33.75** | 30.50 | 3.25 | American ♀ | **50.50** | 46.25 | 4.25 | Jeolla ♀ | **32.75** | 29.00 | 3.75 |
| Beijing ♂ | **37.25** | 28.00 | 9.25 | American ♂ | **52.00** | 48.50 | 3.50 | Jeolla ♂ | **31.50** | 30.25 | 1.25 |
| Northeastern ♀ | **32.50** | 29.50 | 3.00 | British ♀ | **51.00** | 47.75 | 3.25 | Seoul ♀ | **34.25** | 33.50 | 0.75 |
| Northeastern ♂ | **35.00** | 29.00 | 6.00 | British ♂ | **51.25** | 50.50 | 0.75 | Seoul ♂ | 28.75 | **33.50** | -4.75 |
| | | | | Indian ♀ | 48.25 | **48.75** | -0.50 | | | | |
| | | | | Indian ♂ | **49.00** | 47.50 | 1.50 | | | | |

Table 15: Accuracy comparison across option-order settings for *Gemma 3n E2B*, grouped by language, accent, and gender.

| Chinese | | | | English | | | |
|---|---|---|---|---|---|---|---|
| input | original | reversed | Δ | input | original | reversed | Δ |
| Beijing ♀ | **39.10** | 33.83 | 5.27 | American ♀ | **48.99** | 43.47 | 5.52 |
| Beijing ♂ | **33.67** | 31.16 | 2.51 | American ♂ | **46.12** | 42.86 | 3.26 |
| Northeastern ♀ | **34.84** | 30.08 | 4.76 | British ♀ | **46.25** | 45.25 | 1.00 |
| Northeastern ♂ | **33.75** | 33.50 | 0.25 | British ♂ | **45.75** | 43.75 | 2.00 |
| | | | | Indian ♀ | **47.75** | 44.75 | 3.00 |
| | | | | Indian ♂ | 42.46 | **43.97** | -1.51 |

Table 16: Accuracy comparison across option-order settings for *Phi 4 Multimodal*, grouped by language, accent, and gender.

| English | | | |
|---|---|---|---|
| input | original | reversed | Δ |
| American ♀ | **80.50** | 75.75 | 4.75 |
| American ♂ | **79.50** | 77.75 | 1.75 |
| British ♀ | **80.00** | 74.75 | 5.25 |
| British ♂ | **79.50** | 78.25 | 1.25 |
| Indian ♀ | **79.25** | 78.25 | 1.00 |
| Indian ♂ | **78.75** | 77.25 | 1.50 |

(a) Accuracy comparison for *Voxtral-Small-2507*.

| English | | | |
|---|---|---|---|
| input | original | reversed | Δ |
| American ♀ | **48.50** | 46.50 | 2.00 |
| American ♂ | **52.50** | 46.75 | 5.75 |
| British ♀ | **48.50** | 47.00 | 1.50 |
| British ♂ | **49.50** | 47.75 | 1.75 |
| Indian ♀ | **50.75** | 45.25 | 5.50 |
| Indian ♂ | **49.25** | 44.75 | 4.50 |

(b) Accuracy comparison for *Voxtral-Mini-2507*.

Table 17: Accuracy comparison of Voxtral models under different option-order settings, grouped by language, accent, and gender. Note that the language setting is limited to English, as Voxtral currently does not support Korean or Chinese.

| Chinese | | | | English | | | | Korean | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Setting | ♀ | ♂ | Δ | Setting | ♀ | ♂ | Δ | Setting | ♀ | ♂ | Δ |
| Beijing original | **66.75** | 62.75 | 4.00 | American original | 80.00 | **82.00** | -2.00 | Jeolla original | **63.50** | 61.50 | 2.00 |
| Beijing reversed | **64.00** | 61.50 | 2.50 | American reversed | 78.50 | **79.75** | -1.25 | Jeolla reversed | 58.50 | **61.00** | -2.50 |
| Northeastern original | 64.50 | **64.75** | -0.25 | British original | **81.25** | 81.00 | 0.25 | Seoul original | **63.75** | 63.00 | 0.75 |
| Northeastern reversed | 57.75 | **61.25** | -3.50 | British reversed | 76.75 | **78.50** | -1.75 | Seoul reversed | **62.00** | 62.00 | 0.00 |
| | | | | Indian original | **80.50** | 80.00 | 0.50 | | | | |
| | | | | Indian reversed | **78.25** | 79.25 | -1.00 | | | | |

Table 18: Accuracy comparison across *gender* conditions for *Gemini 2.5 Flash*. Each cell reports the mean accuracy (%) for *Female* and *Male*, with Δ denoting the difference (Female - Male). Results are grouped by *language* (Chinese, English, Korean), *accent*, and *option order*.

| Chinese | | | English | | | | Korean | | |
|---|---|---|---|---|---|---|---|---|---|
| Setting | Beijing | Northeastern | Setting | American | British | Indian | Setting | Seoul | Jeolla |
| original ♀ | **66.75** | 64.50 | original ♀ | 80.00 | **81.25** | 80.50 | original ♀ | 63.75 | **63.50** |
| original ♂ | 62.75 | **64.75** | original ♂ | **82.00** | 81.00 | 80.00 | original ♂ | **63.00** | 61.50 |
| reversed ♀ | **64.00** | 57.75 | reversed ♀ | **78.50** | 76.75 | 78.25 | reversed ♀ | **62.00** | 58.50 |
| reversed ♂ | **61.50** | 61.25 | reversed ♂ | **79.75** | 78.50 | 79.25 | reversed ♂ | **62.00** | 61.00 |

Table 19: Accuracy comparison across *accents* conditions for *Gemini 2.5 Flash*. Each cell reports the mean accuracy (%) for each accents, grouped by *language* (Chinese, English, Korean), *option order*, and *gender*.

| Setting | Chinese | English | Korean |
|---|---|---|---|
| original ♀ | 65.62 | **80.58** | 63.62 |
| original ♂ | 63.75 | **81.00** | 62.25 |
| reversed ♀ | 60.88 | **77.83** | 60.25 |
| reversed ♂ | 61.38 | **79.17** | 61.50 |

Table 20: Accuracy comparison across *language* conditions for *Gemini 2.5 Flash*. Each cell reports the mean accuracy (%) for each language, grouped by *option order* and *gender*.

| Model | Language | | | Accent | | | Gender | | |
|---|---|---|---|---|---|---|---|---|---|
| | 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 |

Table 21: APES under original and reversed option order for language, accent, and gender.

| Model | Original | Order Backward | Token Backward | Reversed |
|---|---|---|---|---|
| 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 |

Table 22: Mean entropy under different option-order perturbations. Lower values indicate more stable (less order-sensitive) behavior.

### D.3 Supplementary Results for Impact of Model Scale

Figure 13 presents the scaling results for the Gemini 2.0 series, showing the same trend as Figure 4: larger variants yield higher agreement (Fleiss' $\kappa$) and lower uncertainty (APES) across variables.

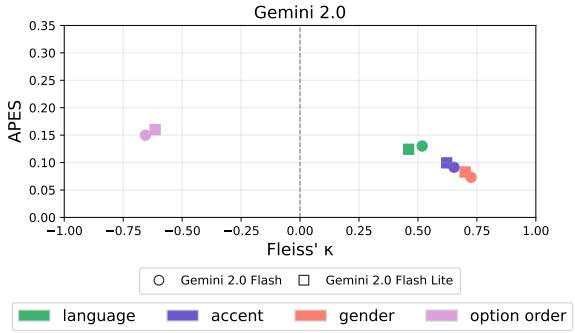


Figure 13: Fleiss' $\kappa$ versus APES by Gemini 2.0 family.

### D.4 Supplementary Results for Impact of Option Reordering

Table 21 reports APES under the original and fully reversed option orders, summarized by language, accent, and gender. For all models, the APES differences between the original and reversed settings are small across the three factors. In both settings, language yields larger APES values than accent and gender, and the ordering of factor magnitudes remains the same (*language* > *accent* > *gender*).

Table 22 reports mean entropy under four option-order configurations (original, order-backward, token-backward, and fully reversed). Across models, entropy values vary across configurations within a limited range. For each model, the table identifies which configuration attains the lowest mean entropy, with the lowest-entropy setting differing across models. Overall, these tables provide additional measurements of robustness under alternative option-order perturbations.